Univerzita Pardubice

Fakulta ekonomicko-správní

Analýza témat patentů pro účely konkurenčního zpravodajství

Diplomová práce

Václav Jaroš, 2025

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2024/2025

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Václav Jaroš**
Osobní číslo:	**E23043**
Studijní program:	**N0613A140041 Aplikovaná informatika – Data Science pro business**
Téma práce:	**Analýza témat patentů pro účely konkurenčního zpravodajství**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování

Cílem práce je charakterizovat analytické metody používané v konkurenčním zpravodajství, shrnout současné přístupy k analýze témat v textu, provést sběr textů patentových žádostí pro vybrané odvětví, analyzovat jejich témata na základě velkých jazykových modelů a identifikovat technologické trendy.
Osnova:

- Analytické metody používané v konkurenčním zpravodajství
- Textová analýza témat
- Sběr dat
- Analýza témat v patentových datech
- Identifikace technologických trendů

Rozsah pracovní zprávy: **cca 50 stran**
Rozsah grafických prací:
Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam doporučené literatury:

JEON, Eunji; YOON, Naeun; SOHN, So Young. Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa. *Technological Forecasting and Social Change*, 2023, 186: 122130.
KIM, Mujin; PARK, Youngjin; YOON, Janghyeok. Generating patent development maps for technology monitoring using semantic patent-topic analysis. *Computers & Industrial Engineering*, 2016, 98: 289--299.
TIAN, Chen; ZHANG, Junyan; LIU, Dayong; WANG, Qing; LIN, Shen. Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model. *Technology Analysis & Strategic Management*, 2024, 36.9: 2084-2099.
TSENG, Yuen-Hsien; LIN, Chi-Jen; LIN, Yu-I. Text mining techniques for patent analysis. *Information Processing & Management*, 2007, 43.5: 1216-1247.
VENUGOPALAN, Subhashini; RAI, Varun. Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change*, 2015, 94: 236-250.

Vedoucí diplomové práce: **prof. Ing. Petr Hájek, Ph.D.**
Centrum pro vědu a výzkum

Datum zadání diplomové práce: **1. září 2024**
Termín odevzdání diplomové práce: **30. dubna 2025**

L.S.

prof. Ing. Jan Stejskal, Ph.D. v.r.
děkan

prof. Ing. Petr Hájek, Ph.D. v.r.
garant studijního programu

V Pardubicích dne 1. září 2024

Prohlašuji:

Práci s názvem Analýza témat patentů pro účely konkurenčního zpravodajství jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 7 /2019 Pravidla pro odevzdávání, zveřejňování a formální úpravu závěrečných prací, ve znění pozdějších dodatků, bude práce zveřejněna prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 27. 6. 2025

Václav Jaroš v. r.

PODĚKOVÁNÍ

Na tomto místě bych rád poděkoval vedoucímu diplomové práce prof. Ing. Petru Hájkovi, Ph.D., za odborné vedení, cenné rady, podnětné připomínky a vstřícný přístup. Poděkování patří také mé rodině za podporu během celého studia.

ANOTACE

Cílem diplomové práce je představit analytické metody používané v konkurenčním zpravodajství, analyzovat témata patentových dat a identifikovat technologické trendy. Úvodní část se věnuje oblasti konkurenčního zpravodajství. Práce se následně zaměřuje na představení problematiky patentové analýzy a analýzy témat v textových datech. Praktická část demonstruje využití neuronového modelu témat BERTopic k analýze témat patentů v oblasti fotografické techniky a identifikaci technologických trendů.

KLÍČOVÁ SLOVA

Konkurenční zpravodajství, patentová analýza, textová analýza, analýza témat, BERTopic

TITLE

Topic Modeling of Patents for Competitive Intelligence

ANNOTATION

The aim of this thesis is to present analytical methods used in competitive intelligence, analyse topics in patent data, and identify technological trends. The introductory part focuses on the field of competitive intelligence. The thesis then presents the issues of patent analysis and topic modelling in textual data. The practical part demonstrates the use of the BERTopic neural topic model to analyse patent topics in the field of photographic technology and to identify technological trends.

KEYWORDS

Competitive Intelligence, patent analysis, text analysis, topic modeling, BERTopic

OBSAH

SEZNAM ILUSTRACÍ A TABULEK

Ilustrace

Tabulky

SEZNAM ZKRATEK A ZNAČEK

BERT	Bidirectional Encoder Representations from Transformers
BI	Business Intelligence
BoW	Bag of Words
CI	Competive Intelligence
CPC	Cooperative Patent Classification
CRISP-DM	Cross-industry standard process for data mining
c-TF-IDF	class-based Term Frequency-Inverse Document Frequency
DTM	Dokument-term matice
IPC	International Patent Classification
KINs	Key Intelligence Needs
KIQs	Key Intelligence Questions
KITs	Key Intelligence Topics
LDA	Latentní Dirichletova alokace
LSA	Latentní sémantická analýza
NMF	Nezáporná maticová faktorizace
NTMs	Neural Topic Models
OHE	One-hot encoding
PLSA	Pravděpodobnosntí latentní sémantická analýzy
SVD	Singular Value Decomposition
TDM	Term-dokument matice
TF-IDF	Term Frequency-Inverse Document Frequency
WIPO	World Intellectual Property Organization

ÚVOD

Společnosti působící na volném trhu čelí neustálému konkurenčnímu tlaku. Vedení podniků i další kompetentní osoby proto musí neustále podnikat správná rozhodnutí, aby se společnost s tímto tlakem vypořádala. Rozhodovací proces tak představuje každodenní, zásadní a velmi náročnou činnost.

V minulosti proto byla vyvinuta řada nástrojů, které mají za cíl tento proces usnadnit. Jedním z těchto nástrojů je konkurenčního zpravodajství, kterým se tato práce zabývá. Jak bude podrobněji rozvedeno v úvodní kapitole práce, sběr a zpracování relevantních informací z externího prostředí a jejich distribuce kompetentním osobám má potenciál společnosti poskytnout významnou konkurenční výhodu.

V současné době přitahují pozornost odborníků i širší veřejnosti nástroje z oblasti umělé inteligence, především velké jazykové modely (Large Language Models – LLMs). Tyto modely jsou schopny na základě trénování na velkém objemu dat nalézat optimální odpovědi, sumarizovat velké množství informací nebo generovat nový obsah. Velké jazykové modely tak mohou představovat významný nástroj pro zpracování velkého objemu dat, se kterým se mohou potýkat právě pracovníci v oblasti konkurenčního zpravodajství.

Tato diplomová práce si proto klade za cíl charakterizovat základní analytické metody konkurenčního zpravodajství. Práce je poté podrobněji věnována metodám analýzy témat v textu ve spojení s cenným zdrojem informací, kterým jsou patentová data. Tato data představují významný zdroj technologických informací. V kombinaci s vhodnými analytickými metodami, jako jsou výše uvedené metody pro analýzu témat v textu, mohou nabídnout cenné informace o současných technologických trendech nebo předpovědi budoucího směřování konkrétní technologie nebo celého odvětví.

Praktická část této práce ilustruje propojení patentových dat, moderních metod analýzy témat v textu, které jsou založeny na LLMs, a konkurenčního zpravodajství. V této části práce bude využit model BERTopic pro identifikaci technologických trendů v odvětví fotografických zařízení, které bylo zvoleno s ohledem na jeho dynamičnost a aktuálnost. bylo zvoleno s ohledem na jeho dynamičnost a aktuálnost.

1 KONKURENČNÍ ZPRAVODAJSTVÍ

Úvodní kapitola si klade za cíl představit problematiku konkurenčního zpravodajství (Competitive Intelligence – CI). V jednotlivých podkapitolách se postupně zaměří na vymezení pojmu CI, představení jednotlivých fází procesu CI a také na informační zdroje, jež jsou pracovníkům CI k dispozici. Závěr kapitoly přináší základní přehled analytických metod a nástrojů, které tvoří nezbytnou součást CI.

1.1 Vymezení pojmu konkurenční zpravodajství

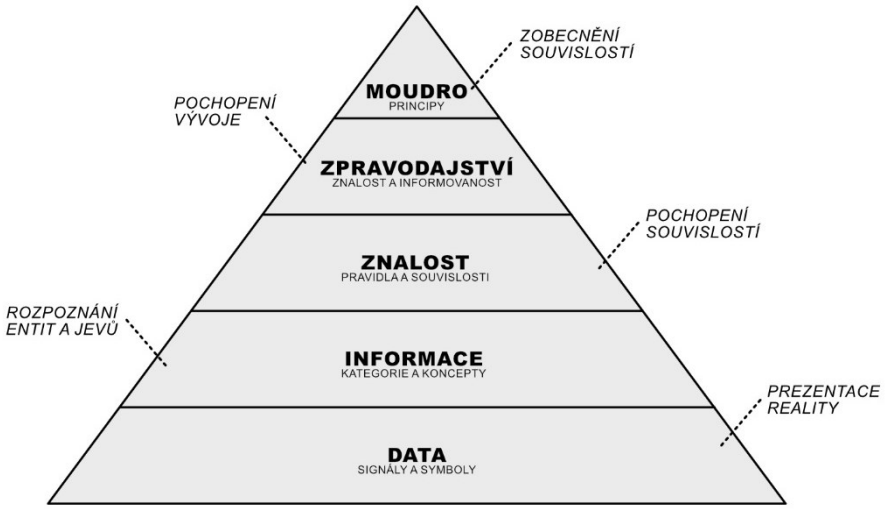
Zpravodajství v podnikatelském prostředí, které vychází z prostředí vojenského, se začalo etablovat přibližně na přelomu 60. a 70. let 20. století [1]. Nově vzniklá oblast podnikatelské činnosti zaměřená na využívání informací z externího prostředí byla označována různými názvy, jako je například „commercial intelligence“, „competitive information“, „corporate intelligence“ nebo právě „competitive intelligence“, což je termín, který se nakonec ustálil [2].

V českém jazyce se pojem competitive intelligence překládá jako konkurenční zpravodajství. Odborná literatura [3; 4] se však shoduje na tom, že tento překlad plně nevystihuje význam původních anglických slov a může být v některých ohledech zavádějící. Z tohoto důvodů bude i v této práci nadále upřednostňován pojem competitive intelligence, respektive jeho zkrácená podoba – CI.

Samotný pojem zpravodajství, který je odvozen od anglického výrazu intelligence, lze vnímat jako další stupeň tzv. znalostní pyramidy, jak ilustrují například Bartes [4] nebo Rodenberg [1]. Je ovšem vhodné zmínit, že CI představuje pouze jednu z dílčích oblastí intelligence v rámci firmy. Pojem CI je často dáván do souvislosti například s pojmem Business Intelligence (BI) [1; 4; 3]. Vztah CI a BI lze podle Bartese [4] vnímat různým způsobem. Na základě současných poznatků je podle Bartese je na CI nahlíženo nejčastěji jako na samostatný informační systém, jehož smyslem je zejména získávání informací o konkurenci z externího prostředí. Odlišný pohled nabízí Rodenberg [1], který CI společně s BI a organisational intelligence vnímá jako součást tzv. enterprise intelligence.

Bartes [4] a Rodenberg [1] sice vnímají výše zmiňovanou pyramidu odlišně, první čtyři stupně jsou ovšem vnímány jednotně. Na nejnižší úrovni pyramidy se nachází data, nad nimi stojí informace. Třetí stupeň tvoří znalost a nad těmito třemi stupni poté stojí zpravodajství.

Podle Bartese [4] je nutné vnímat zpravodajství jako nadstavbu znalostí, tedy *„nejen pochopení současného stavu, ale i vývoje zkoumaného problému umožňující vytvořit kvalitní podklad pro strategické rozhodování vrcholového managementu firmy.“* Jak ukazuje obrázek č. 1, nad zpravodajství Bartes umisťuje ještě moudro, jež chápe jako zobecnění souvislostí a z nich odvozené principy.



Obr. 1 Zpravodajská pyramida
Zdroj: vlastní zpracování podle [4]

Rodenberg [1] ve svém pojetí pyramidy ukazuje souvislost všech tří výše zmiňovaných složek enterprise intelligence. Na prvních čtyřech stupních rozlišuje, jestli data poskytují informace o vnitřním nebo o vnějším prostředí společnosti. Na základě toho poté rozlišuje CI a BI. Nad zpravodajství Rodenberg dále staví ještě dvě úrovně, které představují proces rozhodování společně se strategickým dopadem a akci jako implementaci všech nižších úrovní pyramidy. Výše postavené stupně pyramidy pak Rodenberg vnímá jako součást zmiňované organisational intelligence.

Na příkladu pyramidy je tedy možné vidět, že zpravodajství lze chápat jako obohacenou informaci. Bose [5] ve své studii uvádí, že CI je „*proces a produkt zároveň*,“ přičemž produktem myslí právě zpravodajství. Vymezení pojmu CI poté nabízí ve své studii Pellissier a Nenzhelele [6]. Z této studie vyplývá, že neexistuje jednotná definice pojmu competitive intelligence, přičemž různé definice se liší zejména v tom, které stránky CI vystihují a které naopak opomíjí. Pellissier a Nenezhelele v této studii porovnávali 50 různých definic CI, současně tak vypozorovali 12 charakteristických znaků CI, které shrnují v navrhované univerzální definici. Tato definice vymezuje CI jako *„proces nebo postup, který vytváří*

a distribuuje využitelné zpravodajské informace plánováním, etickým a legálním sběrem, zpracováním a analýzou informací z interního a externího nebo konkurenčního prostředí, přičemž záměrem je usnadnit rozhodování a poskytnout podniku konkurenční výhodu.“

Výše uvedené poznatky lze tedy shrnout tak, že CI je proces využívaný v podnikatelské praxi, jehož smyslem je získat konkurenční výhodu prostřednictvím sběru informací a jejich následnou analýzou, přičemž výsledným produktem tohoto procesu je zpravodajství.

Vyzdvihovanými aspekty CI jsou legalita a etika celého procesu. Tento aspekt procesu CI je jednak obsažen v uvedené definici, ale zvlášť na něj upozorňují jednotliví autoři odborné literatury. Například Bartes [4] uvádí, že „*pracovník firmy podílející se na jakékoliv činnosti spadající do procesu competitive intelligence se nesmí dopustit neetického jednání a v žádném případě se nesmí dopustit jednání nezákonného.“* Etický přístup společně s expertními znalostmi pak umožňuje vykonávat práci účinně a efektivně [7]. Na tomto místě je taky možné se vrátit k problematice českého překladu výrazu CI. Molnár [3] i Bartes [4] ukazují, že výraz konkurenční zpravodajství může být spojován s výrazem průmyslová špionáž, což označuje nelegální postup sběru informací.

V neposlední řadě lze upozornit na další významný aspekt CI, kterým je orientace do budoucna. Silná společnost je schopna ovlivnit strategická rozhodnutí konkurence, vedení společnosti proto musí mít přehled o schopnostech a úmyslech konkurence [4]. Informace, které jsou v rámci procesu CI zpracovávány, umožňují firmě právě předvídat vývoj v konkurenčním prostředí [5].

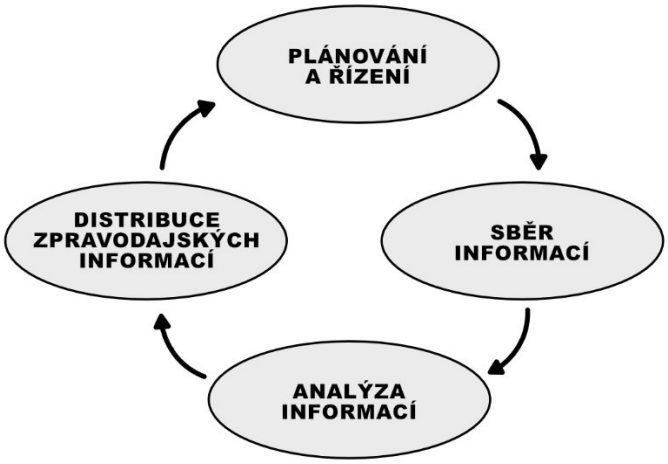
Pro ilustraci konkrétních přínosů CI lze uvést seznam, který sestavil Kahaner [7]. Podle tohoto seznamu CI může přinést společnosti konkurenční výhodu například díky tomu, že dokáže předpovědět změny na trhu nebo chování konkurence, odhalit nové konkurenty, identifikovat nové technologie a produkty nebo rozpoznat nové akviziční příležitosti.

1.2 Proces Competitive Intelligence

McGonagle a Vella [2] rozlišují dva hlavní aspekty CI. Prvním aspektem je sběr surových dat o konkurenci a tržním prostředí z veřejných zdrojů. Druhým je transformace těchto dat na informace, jež je možné využít během rozhodovacího procesu. Tyto dva základní aspekty jsou vnímány jako součást celého procesu CI. Proces CI může mít například podobu lineárního procesu, pyramidy nebo právě cyklu, přičemž v jednotlivých přístupech se otiskují zkušenosti různých manažerů a pracovníků CI [4].

Přestože existují různé způsoby nahlížení na proces CI, řada autorů [3; 4; 8] se shoduje na tom, že nejpoužívanějším modelem procesu CI je tzv. zpravodajský cyklus, jenž vychází ze zpravodajského cyklu Ústřední zpravodajské služby (CIA). Například Molnár [3] jej považuje za základní stavební kámen celého CI. Pellissier a Nenzhelele [8] uvádí, že model zpravodajského cyklu je rozšířen díky svým vlastnostem, jako je nepřetržitost cyklu nebo návaznost jednotlivých kroků.

Zpravodajský cyklus je rozdělován do čtyř fází [3; 7], nicméně někteří autoři nabízejí další pohledy na tento proces. Například Bose [5] prezentuje pětifázový model zpravodajského cyklu, který v porovnání s níže uvedeným čtyřfázovým modelem zahrnuje navíc fázi, jejímž smyslem je poskytnout pracovníkům CI zpětnou vazbu a optimalizovat proces do budoucna.



Obr. 2 Zpravodajský cyklus
Zdroj: vlastní zpracování

Jak je možné vidět na obrázku č. 2, jednotlivými fázemi zpravodajského cyklu jsou [3]:

- plánování a řízení zpravodajské činnosti;
- sběr informací;
- analýza informací;
- distribuce zpravodajských informací.

1.2.1 Plánování a řízení zpravodajské činnosti

Plánování a řízení je výchozí fází zpravodajského cyklu, a z tohoto důvodu také v některých ohledech nejdůležitější fází [7]. Odborná literatura [7; 5] zdůrazňuje, že jde o fázi, která vyžaduje zapojení jednak pracovníků CI, ale také vedení společnosti vzhledem k tomu, že

v rámci této fáze je nezbytné vymezit konkrétní požadavky na výsledný produkt. Podle Kahanera [7] tato fáze vyžaduje 3stranný přístup:

- vymezení potřeb uživatele CI / vedení společnosti;
- sestavení plánu sběru dat a analýzy;
- informování uživatele o sestaveném plánu.

Potřeby společnosti převedené na požadavky CI bývají označovány jako Key Intelligence Needs (KINs) nebo jako Key Intelligence Topics (KITs). Definováních těchto potřeb bývá považováno za jednu z nejtěžších [9] nebo nejdůležitějších [10] aktivit. Muller [10] uvádí, že KINs představují základní rámec pro veškeré další činnosti spojené s tvorbou zpravodajství.

S KINs, popřípadě KITs se pojí další výraz, kterým je poté Key Intelligence Questions (KIQs). Molnár [3] spatřuje podstatu fáze plánování a řízení právě ve vymezení KITs a stanovení konkrétních otázek (KIQs), jenž je potřeba zodpovědět. V neposlední řadě je podle Molnára také nutné vymezit si priority, jako je například požadovaný datum předání výstupů.

KITs tedy představují klíčová témata, která vedení firmy vnímá jako důležitá pro své rozhodování [1]. KITs mohou představovat oblasti týkající se aktuálního rozhodovacího problému, konkrétního subjektu, jako jsou například konkurenti nebo partneři, nebo varování před budoucími hrozbami [3]. Pro ilustraci, jak může vypadat vymezení konkrétních klíčových otázek KIQs, lze poukázat na studii [11] věnující se využití otevřených dat v procesu CI. Autoři této studie pracují specificky s daty z medicínské oblasti. Jedna z klíčových otázek poté zní: *„Je užívání antidepresiv v České republice na vzestupu?“* V tomto případě tak autoři studie chtějí pomocí analýzy dat prokázat, zdali je užívání antidepresiv v ČR na vzestupu, nebo toto tvrzení neplatí.

1.2.2 Sběr informací

Druhá fáze cyklu spočívá ve sběru informací, které jsou relevantní k vymezeným potřebám společnosti, respektive ke klíčovým otázkám. Podle Boseho [5] tato fáze zahrnuje *„identifikaci všech potenciálních zdrojů informací, průzkum identifikovaných zdrojů a shromáždění všech potřebných dat legální a etickou cestou.“* Shromážděné informace mohou být průběžně ukládány do tzv. poznatkové databáze tak, aby s nimi bylo možné pracovat [3].

Pracovníci CI mají k dispozici celou řadu informačních zdrojů, které mohou v rámci analýzy využít. Jednotlivé zdroje lze kategorizovat například z hlediska legality nebo původnosti, jak bude podrobněji rozebráno v kapitole 1.3.

1.2.3 Analýza informací

Informace jsou pouze výchozím bodem procesu, důležité je zejména to, jak budou tyto informace zpracovány a vyhodnoceny [7]. Podle Bartese [12] lze na fázi analýzy nahlížet jako na královskou disciplínu CI, přičemž Bartes tuto fázi popisuje jako „*náročný a specifický proces analýzy a syntézy informací, jehož smyslem je vytvořit přidanou hodnotu, kterou lze využít při strategickém rozhodování vrcholového managementu*.“ Na základě uvedených poznatků lze říci, že analýza získaných informací je klíčovou součástí celého procesu CI, a to z toho důvodu, že právě analýzou informací a jejím vyhodnocením vzniká zpravodajství.

V rámci analýzy získaných informací lze využít různorodé nástroje. Stejně tak, jako mají pracovníci CI k dispozici celou řadu zdrojů informací, mají i k dispozici širokou paletu nástrojů. Tyto nástroje budou ve stručnosti představeny v kapitole 1.4.

1.2.4 Distribuce zpravodajských informací.

Poslední fází zpravodajského cyklu je distribuce získaných zpravodajských informací – produktu celého procesu. Tuto fázi je možné popsat také jako fázi procesu, kdy manažeři obdrží odpovědi na své otázky [7].

Kahaner [7] upozorňuje na to, že tato fáze je také fází, kdy nejvíce projektů CI selže, proto je nezbytné, aby výsledný produkt splňoval určitá kritéria, jako jsou reflexe požadavků managementu, konkrétnost, včasnost a důvěryhodnost. Požadavky na produkt zpravodajství mohou být charakterizovány také pomocí tzv. 4R [3]:

- right time (ve správný čas);
- right quality (v požadované kvalitě);
- right place (na správném místě);
- right product (ve správné podobě).

Výsledný produkt zpravodajství může mít mnoho podob. Může se jednat například o osobní předání informacím nebo uveřejnění výsledků v informačním systému společnosti [3]. Někteří lidé vnímají informace lépe vizuálně, někteří pak auditivně nebo kinesteticky (motoricky), forma zpravodajství proto musí reagovat na preference příjemce zprávy [7].

1.3 Informace v konkurenčním zpravodajství

Tato podkapitola přináší přehled informačních zdrojů, jež je možné v rámci procesu CI využít. Současně představuje i možné způsoby klasifikace informačních zdrojů, které mohou pomoci

pracovníkům při výběru vhodného zdroje. Záměrem této kapitoly je také poukázat na možné způsoby vyhodnocení kvality informací.

1.3.1 Kvalita informací

Jak ukazují následující příklady, k zhodnocení informací lze přistupovat různými způsoby. Například Molnár [3] říká, že „*nikdy nelze předpokládat, že všechny informace jsou přesné a spolehlivé,"* přičemž bez patřičného zhodnocení informací může dojít k tomu, že v důsledku práce se špatnými informacemi budou vyvozeny klamné závěry. Prvním kritériem je podle autora spolehlivost informace, přičemž ta je dána „důvěryhodností zdroje". Druhým kritériem je přesnost informace. Při hodnocení tohoto aspektu informace lze podle Molnára vycházet z vlastních zkušeností nebo z porovnání různých zdrojů.

Podle Bartese [4] lze hodnotit inherentní kvalitu informace a pragmatickou kvalitu informace neboli hodnotu informace. Při vyhodnocování inherentní kvality informace je možné sledovat několik kritérií, které Bartes nazývá „základní znaky kvality". Těmito znaky mohou býti například srozumitelnost informace, relevantnost, úplnost nebo pravdivost. Hodnotou informace pak Bartes popisuje jako poměr mezi užitkem informace a celkovými náklady, které byly na získání informace vynaloženy.

1.3.2 Klasifikace informačních zdrojů

Pracovníci CI mají k dispozici, jak již bylo zmíněno, velké množství různorodých zdrojů. Právě i z toho důvodu je nezbytné vybrat vhodné zdroje, které mohou poskytnout informace odpovídající na vymezené otázky.

Vzhledem k množství různorodých informačních zdrojů, je vhodné informační zdroje rozdělit do skupin na základě společných vlastností. Kategorizace může usnadnit výběr vhodného zdroje informací a zároveň odpovědět na některé otázky týkající se kvality informace.

Zdroje bývají rozdělovány například na interní a externí zdroje. Data mohou být také v různých formátech, kdy lze rozpoznat, jestli se jedná o data strukturovaná nebo nestrukturovaná [3]. Někteří autoři [3; 4; 7] pak zdroje člení také podle dostupnosti. Mezi klasifikační kritéria patří také původnost informací, které zdroje poskytují [1; 3; 4; 7].

Dostupnost zdrojů

Molnár [3] i Bartes [4] shodně rozlišují tři kategorie zdrojů, a to bílé zdroje, šedé zdroje a černé zdroje. Bílé zdroje jsou zdroje, které jsou publikované, resp. veřejně přístupné. Do této

kategorie lze zařadit například výroční zprávy, zpravodajské články, patenty nebo odborné časopisy a konferenční materiály [3].

Molnár [3] nazývá šedé zdroje polopublikovanými zdroji, Bartes [4] pak zdroji nepublikovanými. Molnár tyto zdroje popisuje jako zdroje, jejichž získání je složitější než u bílých zdrojů, což je ale vyváženo tím, že mohou mít značnou informační hodnotou.

Poslední skupinu tvoří černé zdroje, což jsou zdroje uzavřené neboli utajované. Oba autoři [3; 4] shodně uvádí, že využití těchto zdrojů je nepřípustné, jelikož se nejedná o legální způsob získávání informací, a jak již bylo zmíněno, legalita je jedním ze základních znaků procesů CI. Bartes [4] ovšem doplňuje, že informace uložené v černých zdrojích lze získat i legální cestou, a to například zakoupením licence.

Původnost zdrojů

Zdroje lze klasifikovat také na základě původnosti informací, které poskytují. Autoři zdroje nejčastěji dělí na primární a sekundární, ale lze rozlišit i terciární zdroje [4].

Primární zdroje mnohdy představují významný zdroj informací pro zpravodajskou analýzu [4]. Tyto zdroje poskytují informace přímo z pramenů, představují proto neupravené informace, které lze považovat za přesné [7]. Naopak závěry odvozené z informací obsažených v sekundárních zdrojích nemusí být plně kvalitní vzhledem k tomu, že tyto informace byly už jednou zpracovány [4]. Na druhou stranu tyto zdroje mohou poskytovat i určité výhody. Jak bude rozvedeno v druhé části práce, například sekundární patentové databáze poskytují přístup k patentům z několika patentových úřadů, čímž mimo jiné usnadňují sběr dat pro plánovanou analýzu. Terciární zdroje, jako jsou například encyklopedie, pak shrnují informace ze sekundárních zdrojů [4].

1.3.3 Přehled informačních zdrojů

Pro sestavení přehledu již konkrétních zdrojů, lze využít například výše zmíněné rozdělení na základě dostupnosti, jak jej uvádí Kahaner [7]. Skupinami jsou v tomto případě veřejné zdroje a neveřejné zdroje. Informační zdroje by však bylo možné rozdělit i z dalších hledisek, jako je rozdělení na klasické zdroje dat a elektronické zdroje dat [4].

Veřejné (publikované) zdroje dat

Do skupiny veřejných zdrojů lze zařadit [7]:

- vládní zdroje informací;
- média (noviny, televize apod.);
- patenty;
- databáze;
- profesní organizace;
- internet.

Výše uvedený seznam představuje pouze základní přehled veřejných informačních zdrojů. Do každé z těchto skupin by bylo možné zařadit například konkrétní webové stránky, případně by bylo možné seznam rozdělit do dalších podskupin. Seznam naráží také na limity, jako je například překrývání jednotlivých podskupin. To lze ilustrovat na skupině patentů a databází. Přístup k patentům poskytují patentové databáze, které by však mohly být vnímány jako podskupina databází.

Neveřejné (nepublikované) zdroje dat

Jak již bylo uvedeno, opatření informací z neveřejných zdroj vyžaduje větší úsilí než opatření veřejně dostupných informacích. Nepublikované informace lze získat například dotazováním konkrétních osob, přímým pozorováním nebo sběrem informací během veletrhů [7].

1.4 Analytické metody používané v konkurenčním zpravodajství

Cílem této podkapitoly je charakterizovat analytické metody používané v CI. Problém však může představovat množství nástrojů, jež existuje, jak ukazuje například Bartes [4]. Tento autor uvádí, že CI lze popsat jako metodologii skládající se z dílčích metod. Popsání všech exitujících metod je proto podle autora úlohou pro samostatnou publikaci.

Jak ale ukazují následující studie a publikace, vypořádání se s výše uvedeným problémem nabízí rozdělení jednotlivých metod do kategorií. Jeden z pohledů nabízejí například Bartes [4] nebo Bose [5], kteří metody dělí z hlediska fáze procesu CI, ve které nacházejí své uplatnění. Bartes [4] ve své publikaci popisuje vybrané metody, které se využívají v jednotlivých fázích procesu. Bose [5] se pak ve své studii věnuje nástrojům pro sběr dat a pro analýzu dat. Tato kapitola se zaměřuje na fázi analýzy, nadále proto bude popisován především přístup Boseho.

Bose [5] rozpoznává čtyři základní formy analýzy v CI: dedukci, indukci, rozpoznání vzorců a analýzu trendů. Nástroje a metody, které analytici využívají, proto musí umožňovat induktivní a deduktivní uvažování a také rozpoznávání vzorců, přičemž k zpracování velkého objemu napomáhají počítačové nástroje [5]. Detailní přehled nástrojů výpočetní techniky využívající se v procesu CI nabízí ve své studii Olszak [13]. Autorka charakterizuje 10 typů nástrojů, které dělí do dvou skupin (viz kapitola 1.4.1).

Bose [5] zároveň upozorňuje, že získané informace je potřeba zasadit do strategického kontextu k čemuž slouží druhá skupina nástrojů. Analytické přístupy v rámci CI lze tedy rozdělit do dvou hlavních skupin. První skupina budou tvořit nástroje z oblasti datové analytiky, druhá skupina se pak bude věnovat strategické analýze.

1.4.1 Datová analytika

Olszak [13] technologie a nástroje pro práci s informacemi dělí do dvou skupin: na metody pro exploraci dat (data exploration) a na metody pro využití dat (data exploitation). Jak ukazuje níže uvedená tabulka č. 1, v první skupině se nachází například nástroje pro prediktivní modelování a data mining, web mining nebo agentové modelování. Do druhé skupiny poté autorka řadí nástroje jako jsou dashboardy, nástroje pro interaktivní vizualizaci nebo balanced scorecard.

Tab. 1 Nástroje pro práci s informacemi

Explorace dat	**Využití**
Prediktivní modelování a data-mining	Dashboardy
Text mining	Nástroje pro interaktivní vizualizaci
Web mining	Balanced scorecard
Agentové orientované modelování	Architektura orientovaná na služby
Exponenciální modely náhodných grafů	
Vyhledávací aplikace	

Zdroj: [13]

Zatímco první skupina nástrojů se soustředí především na odhalování nových znalostí, druhá skupina napomáhá porozumění stávajícím znalostním bázím [13]. Jednotlivé nástroje jsou však často na sobě závislé. Například vizualizační nástroje, jako jsou výše zmíněně dashboardy a nástroje pro interaktivní vizualizaci, mívají v sobě integrované i analytické nástroje [5].

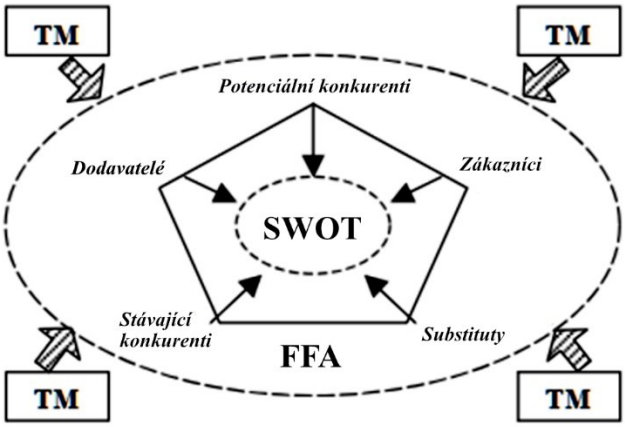
Data mining lze popsat jako iterativní proces získávání znalostí z dat, který využívá metody, jako jsou například klasifikace, hledání asociací a korelací, shlukování a regrese [14]. Text mining neboli textová analytika pak představuje rozšíření metod data miningu pro analýzu a zpracování nestrukturovaných textových informací, přičemž využívá nástroje a metody, jako jsou například extrakce informací, sumarizace, kategorizace, hluboké učení apod. [15]. Web mining ve spojení s text miningem společně tvoří skupiny nástrojů pro analýzu nestrukturovaných webových dat, jako jsou například e-maily, webové stránky nebo sociální sítě [13].

CI pak pracuje s různými konkrétními metodami, jako jsou například textová a obsahová analýza, analýza citlivosti, analýza konkurenta nebo analýza možného vývoje odvětví [4]. Tyto metody však obvykle pracují s informacemi, které byly získány za pomoci výše uvedených technologií, popřípadě se jedná právě o nepočítačové metody spadající do kategorie strategické analýzy. Jako příklad lze použít analýzu patentů, které se věnuje navazující část této práce. Tento konkrétní typ analýzy zaměřený na práci s patentovými informacemi využívá mimo jiné právě metody text miningu nebo různé vizualizační přístupy (viz podkapitola 2.4).

Bose [5] poté vyzdvihuje, že právě nástroje, které jsou založené na metodách text miningu a web miningu nebo také nástroje pro vizualizaci, usnadňují sběr dat a zároveň urychlují a zefektivňují fázi zpravodajské analýzy.

1.4.2 Strategická analýza

Přestože počítačové nástroj plní klíčovou úlohu při zpracování velkého objemu dat, závěrečná část analytické fáze procesu CI je stále závislá na neautomatizovaných metodách [5]. Jako názorný příklad propojení analytických nástrojů se strategickými nástroji lze zmínit studii [16] ukazující, jak mohou být zkombinovány metody text miningu a nástroje strategické analýzy, jako jsou Porterova analýza pěti sil a SWOT analýza. Autoři této studie pomocí kombinace uvedených nástrojů vytvořili nový model MiEDec, který je vyobrazen na obrázku č. 3. Tento koncept využívá informace získané pomocí metod TM jako vstup pro analýzu pěti sil, která je následně zkombinována se SWOT analýzou.



Obr. 3 Metoda MiEDec

Zdroj: [16]

SWOT analýza přestavuje v odborné literatuře [1; 3; 7] pravděpodobně nejrozšířenější nástroj strategické analýzy. SWOT lze popsat jako základní a efektivní způsob, jak vystihnout vnitřní charakteristiky (silné a slabé stránky) a externí charakteristiky (příležitosti a hrozby) dané společnosti [7]. Například Rosenberg pak ukazuje, jak využít celou řadu nástrojů jako je metoda scénářů, KITs, matice GE, BCG matice, SPACE analýza, kritické faktory úspěchu, The Competitive Assessment Model apod. [1]

2 PATENTOVÁ ANALÝZA

Patentová analýza představuje kombinaci několika samostatných činností, jako je práce s patentovými databázemi nebo extrakce informací z patentových dokumentů a jejich následná analýza [17]. Analýza patentů je z hlediska CI schopna poskytnout například předpověď vývoje konkrétního odvětví nebo informace o vývojovém potenciálů konkurence [4]. Tato metoda je proto velmi cenným nástrojem, který může pracovníkům CI poskytnout žádané odpovědi.

Úvodní část této kapitoly vyzdvihuje význam informací ukrývajících se v patentových dokumentech. Navazující podkapitoly se pak již podrobněji věnují vymezení základních pojmů, patentovým databázím a práci s nimi, nakonec také konkrétním metodám a technikám, jenž se v analýze patentů uplatňují.

2.1 Patentové informace

Pod pojmem patentové informace lze chápat veškeré informace, které se nacházejí v patentových dokumentech nebo také informace získané statistickou analýzou patentových přihlášek [18]. Tyto informace jsou obecně vnímány jako velmi cenné. Kahaner [7] uvádí, že *„více než 75 % informací, které skýtají patenty registrované ve Spojených státech amerických, nejsou zveřejněny nikde jinde.“*

Konkrétní využití patentových informací lze ilustrovat na řadě existujících odborných studií. Jeong a Yoon [19] pomocí analýzy témat v patentech definovali pět hlavních technologických témat v oblasti rozšířené reality. Tato témata byla následně zkoumána v kontextu vývojových a patentových aktiv významných společností působících v daném odvětví. Další studie [20] pak ilustruje využití patentových informací pro identifikaci technologií s potenciálem pro další výzkum v oblasti digitální terapeutiky, čímž usnadňuje plánování vývoje produktů. Analýzou patentových informací je tedy možné získat informace o budoucím vývoji v konkrétním odvětví, na základě čehož lze učinit rozhodnutí v oblasti vývoje.

2.2 Patent

Světová organizace duševního vlastnictví (World Intellectual Property Organization – WIPO) [21] popisuje patent jako „*výlučné právo na vynález, které přináší prospěch vynálezcům tím, že jim poskytuje právní ochranu jejich vynálezů a je zároveň prospěšné pro společnost tím, že poskytuje veřejnosti přístup k technickým informacím, a tak urychluje inovace.“*

V České republice toto právo upravuje Zákon č. 527/1990 Sb., o vynálezech, průmyslových vzorech a zlepšovacích návrzích [22]. Dle uvedeného zákona mohou být patentovány vynálezy, které jsou *„nové, jsou výsledkem vynálezecké činnosti a jsou průmyslově využitelné.“* Zákon říká, že vynález lze považovat za *„nový, není-li součástí stavu techniky.“* Současným stavem techniky je pak vše, co bylo zveřejněno dříve, než přede dnem, kdy přihlašovatel nabývá práva přednosti.

Zákon [22] chrání osoby, jimž náleží právo na patent tak, že všem ostatním zakazuje bez souhlasu majitele předmět patentu využívat (vyrábět, nabízet, uvádět na trh apod.) přímým i nepřímým způsobem. Zákon dále upřesňuje, kdo je majitel patentu a komu tak právo na patent přísluší, vyčerpání práv nebo omezení účinků patentu. Právo na patent po určité době zaniká, přičemž tato doba je stanovena na 20 let od podání přihlášky. Patent zaniká také v případech, kdy majitel patentu nezaplatí příslušný poplatek nebo v případě, že se majitel patentu vzdá. Patent může být také úřadem dodatečně zrušen, jestliže nesplňuje nějakou ze zákonem stanovených podmínek.

Jedním z přínosů patentů je tedy to, že umožňují přístup k technologickým informacím. Analýza těchto informací může společnosti poskytnout cenné informace týkající se budoucího vývoje v daném odvětí. Jestliže pak firma tyto informace využije a podaří se ji díky tomu vyvinout vlastní technologii, může si svůj vynález patentovat a využít tak i druhého benefitu, kterým je ochrana, kterou patent poskytuje.

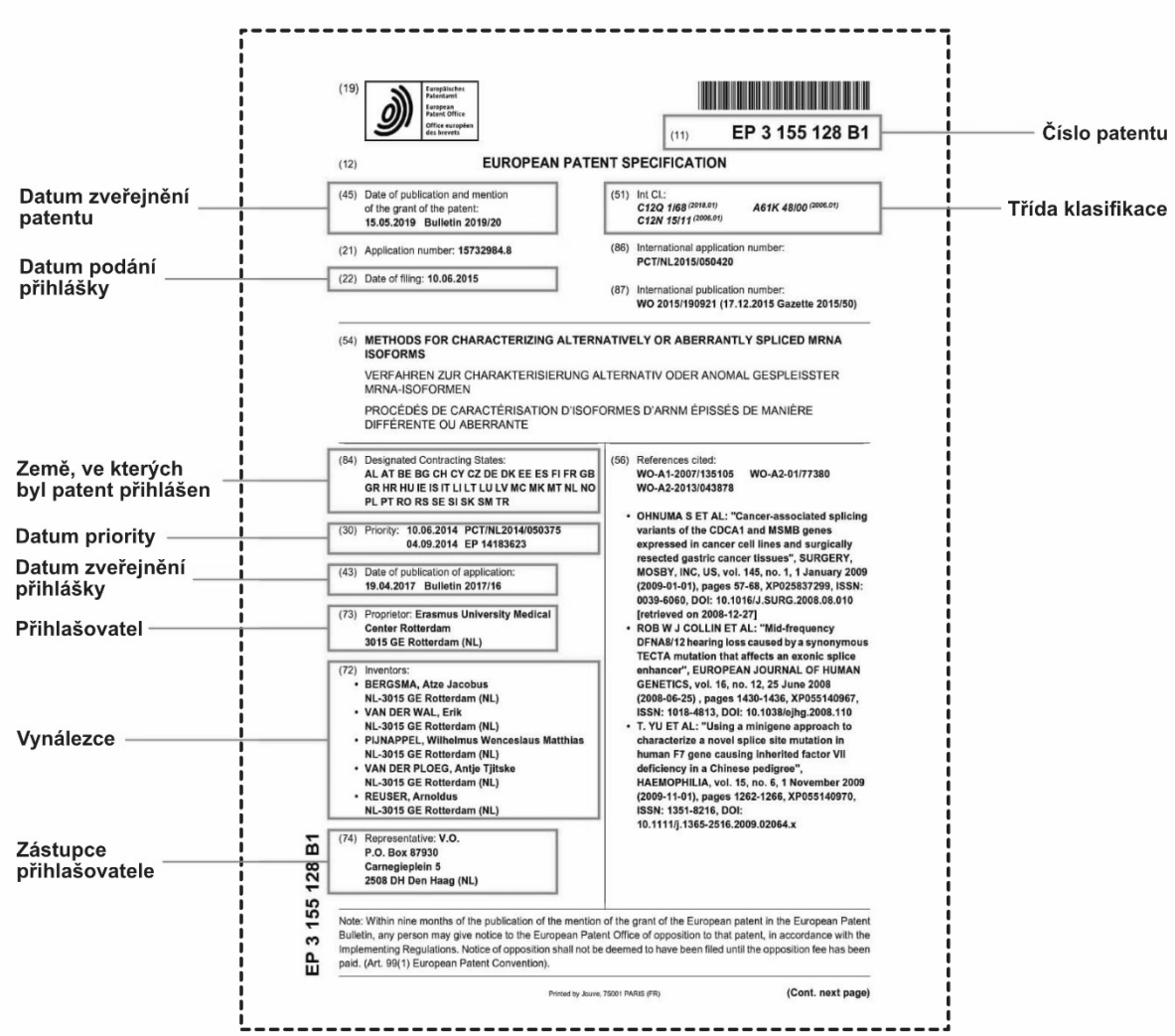
2.2.1 Struktura patentového dokumentu

Podle Oldhama [23] se lze na patenty dívat dvěma způsoby: z právního pohledu a jako na specifický typ dokumentu. Pro analýzu patentů je pak důležité pochopení obou pohledů. Právní pohled a z něho vyplývající přínosnost informací uložených v patentových dokumentech byl nastíněn v předchozích odstavcích, tato podkapitola si klade za cíl představit druhý pohled.

Jestliže se na patenty díváme druhým způsobem, tedy jako na typ dokumentu, lze říci, že to jsou textové dokumenty, které lze rozdělit na dílčí části, jež jsou společné všem patentovým dokumentům, a proto je lze považovat za semistrukturovaná data [24]. Toto uspořádání přináší určité výhody, jako je například usnadnění vyhledávání patentů [25].

Patentové dokumenty je obvykle možné rozdělit na 4 hlavní části [26]:

- bibliografické údaje;
- popis patentu;
- patentové nároky;
- abstrakt.



Obr. 4 Evropský patent
Zdroj: [27], upraveno

Na obrázku č. 4 je možné vidět strukturu evropského patentu, konkrétně pak úvodní stránku zahrnující bibliografické údaje. Bibliografické údaje zahrnují například číslo patentu, jména a adresy, datumy, jako je datum podání přihlášky nebo datum zveřejnění apod. [26] Důležitými datumy jsou z hlediska analýzy patentů zejména datum priority, který vymezuje právní ochranu před patenty, které byly podány později, a datum zveřejnění patentu, který je společně s číslem patentu jedním z nejdostupnějších údajů v databázích [23].

Další části patentu tvoří abstrakt shrnující základní vymezení vynálezu, detailní popis samotného vynálezu a nároky definující konkrétní části vynálezu, které si patent právně nárokuje [26].

2.2.2 Patentové rodiny

První patentová přihláška (prioritní přihláška) se stává rodičem všech patentových dokumentů (patenty, přihlášky), které popisují stejný vynález a byly podány na jednom nebo více patentových úřadech [23].

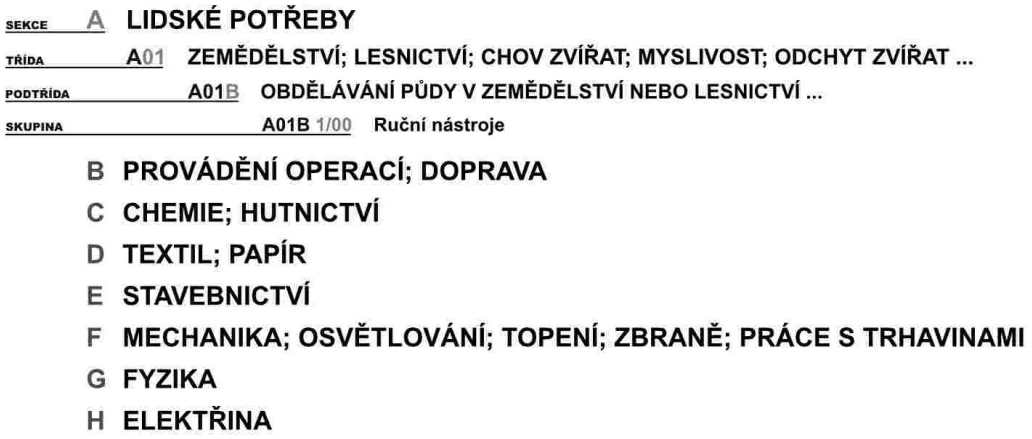
2.2.3 Klasifikace patentů

Klasifikační systém třídí patenty do jednotlivých kategorií a podkategorií, což umožňuje snadnější vyhledávání patentů. Z pohledu patentové analýzy jsou klasifikační systémy důležité například při sběru dat, kdy usnadňují vyhledávání v databázích [28]. Patentový dokument může být označen i více klasifikačními kódy, přičemž jeden bývá považován za hlavní a zbývající za vedlejší [24].

Klasifikačních systémů existuje hned několik. Za dva nejdůležitější systémy jsou považovány [24; 28] Mezinárodní patentové třídění (International Patent Classification – IPC) a Kooperativní patentové třídění (Cooperative Patent Classification – CPC). Odborná literatura [23; 24; 25; 28] pak mezi významné klasifikační systémy řadí i některé další třídění. Uváděnými systémy jsou například japonské klasifikační systémy (File forming term a File Index), klasifikační systém Spojených států amerických (United States Patent Classification) nebo komerční klasifikační systém Derwent. Zmiňované dva nejvýznamnější systémy jsou podrobněji popsány níže.

International Patent Classification (IPC)

IPC neboli Mezinárodní patentové třídění je spravováno WIPO. Třídění vzniklo na základě Štrasburské dohody z roku 1971. Technologie jsou zde rozděleny hierarchicky do jednotlivých skupin a podskupin, přičemž každá skupina i podskupina je označena určitým symbolem (viz obr. 5) Na nejvyšší úrovni se nachází 8 hlavních sekcí, na nejnižší úrovni pak lze nalézt přibližně 70 000 skupin [29].



Obr. 5 Struktura mezinárodního patentového třídění
Zdroj: vlastní zpracování podle [30]

Třídník je pravidelně aktualizován a WIPO jej publikuje ve dvou jazycích – v angličtině a ve francouzštině. Český překlad Mezinárodního patentového třídění [30] je publikován na webových stránkách Úřadu průmyslového vlastnictví. Společně s třídníkem úřad nabízí také překlad návodu k mezinárodnímu patentovému třídění [31], který si klade za cíl objasnit, jak a pro jaké účely třídník využívat.

Cooperative Patent Classification (CPC)

Klasifikační systém CPC vychází z IPC a klasifikačních systémů Evropského patentového úřadu (EPO) a patentového úřadu Spojených států amerických (USPTO). CPC tvoří veškeré kódy IPC a také nové kódy CPC. CPC k původním 8 třídám IPC přidává třídu Y, která je věnována novým technologiím. [32]

2.3 Práce s patentovými databázemi

Jedním ze základních kroků každé analýzy je sběr potřebných dat, přičemž je nezbytné, aby tato data byla kvalitní a vztahovala se ke vymezenému předmětu analýzy. Přístup k patentovým dokumentům poskytují patentové databáze.

2.3.1 Patentové databáze

Patentové databáze jsou *„repositáře dat uchovávající údaje týkajících se vydaných patentů a zveřejněných přihlášek.*“ [33]. V současné době existuje široká nabídka databází odkud lze získat patentová data. Jednotlivé databáze se ovšem odlišují nabízenými funkcemi nebo tím, jestli jsou zpoplatněné. Analytici, kteří plánují pracovat s patentovými daty by měli podle Trippeho [24] porovnat na jedné straně náklady na pořízení a časovou náročnost zpracování

získaných dat na druhé straně. Veřejně přístupné databáze mohou přinést restrikce například v počtu záznamů, které si může uživatel stáhnout, v nabídce dostupných polí nebo v tom, jaké formáty dat nabízejí ke stažení [23].

Jedním z nejdůležitějších parametrů, který by měl být brán na zřetel při výběru databáze pro vyhledávání patentů, je podle Jürgensa a Herrero-Solana [34] to, jaké je územní pokrytí jednotlivých databází. Dalšími parametry, které výše uvedení autoři [34] sledovali v rámci porovnání databází Patentscope, Espacenet a Depatisnet, jsou nabízené vyhledávací funkce, výsledky vyhledávání, dostupné bibliografické údaje, automatický překlad dokumentů do dalších jazyků nebo možnosti exportu dat. Jednotlivé databáze se od sebe lišili například tím, jaké klasifikační systémy lze využít při vyhledávání, nebo tím, která pole bylo možné exportovat. Například databáze Patentscope jako jediná umožňovala i stažení obrázků.

Jak bylo rozvedeno výše, jednotlivé patentové databáze se od sebe liší, můžeme je proto na základě vybraných parametrů rozdělit do určitých skupin. Na databáze lze nahlížet například z hlediska toho, jestli je přístup k databázi zpoplatněn nebo jestli se jedná o národní nebo mezinárodní databáze [33]. Možnou kategorizaci patentových databází nabízí Trippe [24], který patentové databáze dělí obdobně jako lze rozdělit veškeré informační zdroje v CI, a to na primární zdroje a sekundární zdroje, jež dále člení na volně-přístupné sekundární zdroje a zpoplatněné sekundární zdroje.

Primární zdroje

Primární zdroje jsou takové zdroje, které jsou spravovány příslušnými orgány – patentovými úřady. Typickými vlastnostmi těchto databází je podle Trippeho [24] například to, že vyhledávání a stažení bibliografických údajů nebývá zpoplatněno, databáze umožňují vyhledávání v anglickém uživatelském prostředí nebo to, že obvykle umožňují pouze stažení celých dokumentů a neumožňují stažení jednotlivých polí.

90 % všech patentových přihlášek mají přijmout 3 největší patentové úřady — patentový úřad Spojených států amerických, Evropský patentový úřad a Japonský patentový úřad [33]. Databázi patentů spravuje v České republice již zmiňovaný Úřad průmyslového vlastnictví, přičemž v této v této databázi jsou dostupné *„přihlášky vynálezů zveřejněné od roku 1991, udělené patenty od č. 1, evropské patenty platné na území ČR a zapsané užitné vzory.“* [35].

Sekundární zdroje

Sekundární zdroje jsou veškeré zdroje, které vychází z primárních zdrojů, což je jejich hlavní výhodou, jelikož často kombinují data z několika patentových úřadů a rozšiřují tak možnosti vyhledávání [24]. Komerční zdroje poté přináší některé výhodné funkce, jako jsou například překlady informací v patentech, základní grafické a statistické analýzy, práce s patentovými rodinami nebo klasifikaci patentů podle chemické struktury, biologických sekvencí apod. [33].

2.3.2 Přehled patentových databází

Základní přehled patentových databází nabízí Oldham [23], který ve své publikaci vyzdvihuje především volně dostupné databáze poskytující mezinárodní data, jako jsou Patentscope, Espacenet, LATIPAT, USPTO, Google Patents, DEPATISnet, databáze patentů OECD nebo The Lens. Ze zpoplatněných databází a nástrojů pro vyhledávání patentů Oldham zmiňuje databáze Derwent, PatSnap. Dimensions nebo Questel Orbit.

Přehled 50 patentových databází poskytuje WIPO na webové stránce *inspire.wipo.int/wipo-inspire*. Na této stránce lze databáze filtrovat na základě různých vlastností a poskytovaných funkcí. Uvedená stránka zároveň nabízí nástroje pro porovnání zvolených databází.

2.3.3 Vyhledávání v patentových databázích

Podle WIPO [18] je možné přistupovat k vyhledávání patentů pomocí těchto kritérií:

- klíčová slova;
- patentové klasifikace;
- datumy;
- identifikační čísla (číslo přihlášky, číslo publikace apod.);
- jména (majitel patentu, podavatel).

Klíčová slova jsou slova nebo slovní spojení, které určitým způsob vystihují podstatu patentované technologie. Vyhledávání pomocí klíčových slov může být upřesněno booleovskými operátory (AN, OR, NOT), díky kterým lze výrazy kombinovat, případně určité výrazy vyloučit. Některé databáze pak umožňují stematizaci, kdy vyhledávače místo celého původního výrazu vyhledávají kmen slova, což urychluje a rozšiřuje hledání. [18]

Podle Clarkeho [36] je také nezbytné vzít v potaz, jestli omezit vyhledávání pouze na název patentu nebo na celý text. Clarke argumentuje, že názvy mohou být nejasné nebo neinformativní, vyhledávání v celém textu zvyšuje pravděpodobnost nalezení všech

relevantních dokumentů, ale zároveň mohou být vyhledány nerelevantní dokumenty. Optimální variantou tak může podle Clarkeho být i omezení vyhledávání na abstrakt dokumentu.

Použití klíčových slov jako primární způsob vyhledávání v patentové databázi ovšem není považováno za vhodné [33; 36], a to hned z několika důvodů. Problematickými aspekty vyhledávání pomocí klíčových slov jsou podle Clarkeho [36] například mnohoznačnost některých slov, užití synonym pro označení stejného předmětu nebo specifický jazyk patentů, kdy může být i obyčejný předmět označen nejasným názvem. Ku příkladu Clarke uvádí označení svinovacího metru souslovím *„přímočaré srovnávací zařízení“* (linear comparison device). Použití klíčových slov může být problematické také z pohledu jazykových rozdílů mezi jednotlivými dokumenty [18].

Namísto vyhledávání pomocí klíčových slov odborná literatura [33; 36] navrhuje využití jednoho z klasifikačních systémů (viz kapitola 2.2.3), přičemž klíčová slova je možné využít pro upřesnění vyhledávání. Clarke [36] za optimální považuje postup v následujících krocích:

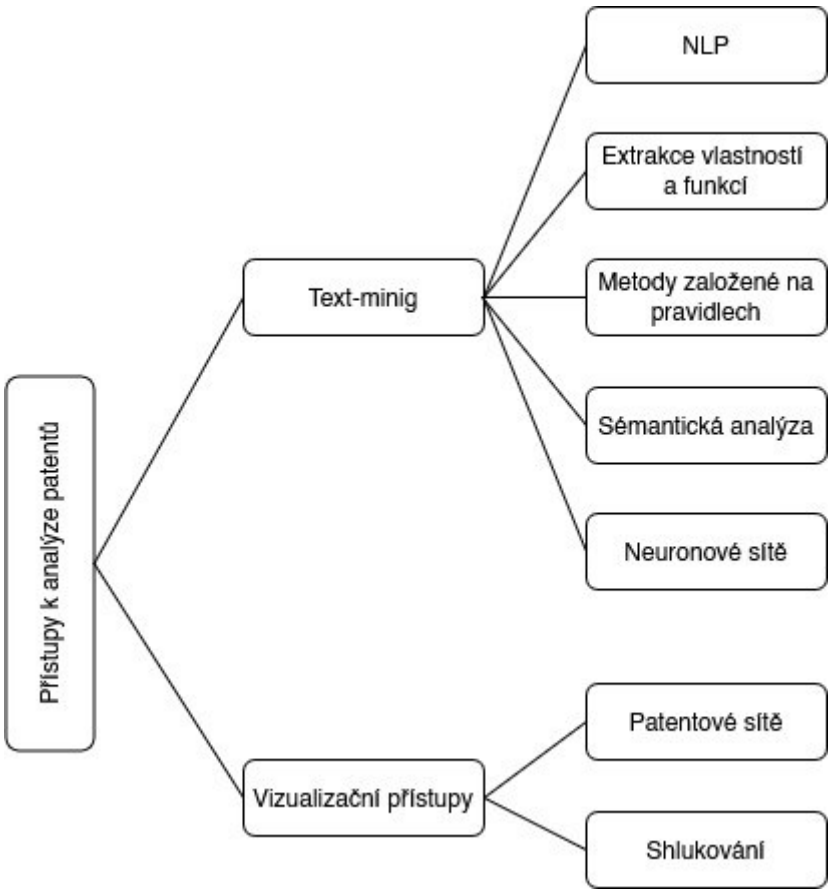
1. výběr vhodného klasifikačního systému;
2. vyhledání odpovídajících patentů;
3. dodatečná úprava vyhledávání za pomocí klíčových slov.

V souvislosti s výše zmíněným postupem se nabízí otázka, jak určit správnou třídu klasifikace. Jedním ze způsobů, jak k tomuto problému přistupovat je vyhledávání klíčových slov v třídníku WIPO. Dalším přístup vyhledávání popisuje Trippe [24], který tento problém řeší z pohledu „reverzního inženýrství“. Postup je poté následující: vyhledávání v názvech patentů; výběr několika relevantních patentových dokumentů; identifikace klasifikačních symbolů, nových termínů a dalších poznatků; opakování postupu, dokud nejsou sesbírány všechna potřebná data.

2.4 Nástroje a metody analýzy patentů

Patentové dokumenty obsahují nespočet různých údajů. Jak již bylo zmíněno, patenty lze z tohoto důvodu považovat za semistrukturovaná data. Analýza patentů tak propojuje mnoho odlišných disciplín, jako jsou například analýza patentových dat, analýza odborné literatury, čištění dat, zpracování textových dat nebo také strojové učení, mapování a vizualizace dat [28].

Výsledky analýzy patentové analýzy je možné prezentovat v podobě patentových grafů nebo patentových map, přičemž patentové grafy jsou vizualizací strukturovaných dat, patentové mapy jsou poté vizualizací nestrukturovaných dat [37].



Obr. 6 Přístupy k analýze patentů
Zdroj: [17], upraveno

Možný způsob klasifikace jednotlivých metod poskytují například Abbas a kol. [17]. Tito autoři identifikují dvě hlavní větve metod a přístupů využívané při analýze patentových dokumentů. Na obrázku č. 6 lze vidět první dvě úrovně tohoto rozdělení.

První z větví zahrnuje techniky zpracování textových dat, které se využívají pro získávání informací ze strukturovaných a nestrukturovaných textových. Do této větve patří metody založené na principech zpracování přirozeného jazyka, na principech sémantické analýzy, na pravidlech, na neuronových sítí nebo na extrakci vlastností a funkcí.

Druhou větev tvoří metody založené na vizualizaci, které pak podle autorů [17] usnadňují manažerům a expertním pracovníkům analýzu informací právě díky tomu, že informace prezentují ve vizuální podobě. Do této větve patří patentové sítě a metody založené na shlukování dat.

Na klasifikaci metod patentové analýzy lze nahlížet také z hlediska typu úloh, jak ukazují Jiang a Goetz [38]. Základními dvěma typy úloh jsou analýza a generování. Analýzu patentových dat lze dále rozdělit na klasifikační úlohy, úlohy získávání informací, hodnocení kvality a úlohy umožňující zaměřené na porozumění technologického vývoje, jako jsou například předpovědi vývoje.

3 ANALÝZA TÉMAT V TEXTU

Jednou z metod, které lze využít pro analýzu patentů, je textová analýza témat. Cílem této kapitoly je proto objasnit základní pojmy z této oblasti. První podkapitola se věnuje vymezení základních pojmů. V navazujících podkapitolách jsou představeny možnosti kategorizace modelů témat, principy předzpracování textových dat a způsoby reprezentace textových dat, které modely témat využívají. V závěru kapitoly jsou představeny principy nejvyužívanějších modelů témat.

3.1 Vymezení základních pojmů

Téma vyjadřuje něco, co lze chápat spíše intuitivně, ve stručnosti jej však můžeme popsat jako vyjádření základní myšlenky textu pomocí konkrétního termínu nebo jako pravděpodobnostní rozložení slov ze všech slov uvažovaného korpusu [14]. Druhý zmiňovaný způsob vyjádření tématu je výstižnější, jelikož umožňuje popsat i složitější témata a zároveň řeší problém mnohoznačných termínů [14]. Pravděpodobností rozložení slov však lze nahradit významností jednotlivých slov. Tento způsob reprezentace využívají například neuronové modely, jako je BERTopic (viz 3.5.5).

Pojem modelování témat pak lze popsat různými způsoby. Kherwa a Bansal [39] jej charakterizují jako *„statistický přístup pro odhalení skrytých významových vazeb ve velkém souboru dokumentů.“* Alghamdi a Alfalqi spatřují podstatou modelování témat v identifikaci dokumentů, jež sdílí stejné vzorce výskytu slov [40]. Další studie [41] popisuje modelování témat jako modelování tří entit: konstruktů (slov), kolekcí (dokumentů) a témat. Tato studie poté popisuje téma jako shluk konstruktů, který je idealizovaným popisem kolekce. Uvedené definice lze shrnout tak, že témata jsou v definicích popisované skryté struktury nebo vazby slov, jež lze vyjádřit pomocí pravděpodobnostního rozložení slov. Modelování témat je poté proces vedoucí k nalezení zmiňovaných skrytých vazeb nebo vzorců.

Modelování témat umožňuje automaticky organizovat, sumarizovat a zpracovávat velké množství textových dat [41]. Jak ukazují studie, tyto vlastnosti lze využít různými způsoby. Modelování témat lze uplatnit v oblastech jako je vývoj software, bioinformatika nebo analýza sociálních sítí [39]. Alghamdi a Alfalqi [40] poukazují na vědecký výzkum jako na typický příklad aplikace modelů pro analýzu vývoje témat v čase. Tato metoda dle autorů může sloužit jako nástroj pro poskytnutí přehledu nad vývojem přemýšlení o stanoveném předmětu zkoumání nebo pro identifikaci relevantních dokumentů. Modelování témat ve

spojení s patentovými dokumenty je pak možné využít například pro identifikaci technologických trendů [42] nebo pro automatizovanou klasifikaci patentů [43].

Modelování témat lze mimo jiné využít i při práci s netextovými daty, kdy lze koncept slov nahradit jinými entitami s podobnou strukturou, jako jsou například segmenty obrazových dat nebo geny v genových sadách [41].

3.2 Klasifikace modelů témat

Výrazy analýza témat a modelování témat neodkazují na jednu konkrétní metodu, ale zahrnují v sobě celou řadu odlišných přístupů. Jednotlivé modely se vyvíjely postupně, liší se v předpokladech o vstupních datech, reprezentaci dokumentů i tématech, a lze je využít pro různorodé účely [41].

Kherwa a Bansal [39] rozdělují modely do dvou základních kategorií. První z kategorií je tvořena nepravděpodobnostními modely neboli algebraickými modely, jako je Latentní sémantická analýza (LSA) nebo Nezáporná maticová faktorizace (NMF). Do druhé kategorie se poté řadí pravděpodobnostní modely jako je Latentní Dirichletova alokace (LDA) nebo Pravděpodobnostní LSA (PLSA). Novější studie [41] ke dvěma zmíněným kategoriím připojuje další dvě. Celkem tak lze rozlišit čtyři kategorie:

- algebraické modely;
- fuzzy modely;
- pravděpodobnostní modely;
- neuronové modely.

Základním principem algebraických modelů témat, jako je LSA nebo NMF, je rozklad matic, přičemž tyto modely reprezentují data v podobě sady slov, na jejímž základě je následně sestavena dokument-term matice [39]. Algebraické modely témat jsou jednoduché a efektivní, problematickým aspektem však může být řídkost matic [41].

Fuzzy modely fungují na principu měkkého shlukování, což znamená, že každý objekt může částečně patřit do každého shluku [44]. Jestliže shluk představuje určité téma, dokument tak může být přiřazen k více tématům. Fuzzy modely se uplatňují v oblasti text miningu nebo také v medicínských a vzdělávacích oblastech [41].

Pravděpodobnostní modely byly hlavním předmětem výzkumu od roku 2003, kdy byla představena LDA, do roku 2015, kdy se výzkum začal přesouvat směrem k vývoji modelů na

bázi neuronových sítí [41]. Pravděpodobnostní model, jako je LDA, vyžaduje v porovnání s LSA vyšší výpočetní výkon [45], výhoda pravděpodobnostních modelů ale spočívá v tom, že jsou jednoduché, intuitivní, škálovatelné a interpretovatelné [41]

Nejnovější přístup k modelování témat představují neuronové modely témat, souhrnně označované jako Neural Topic Models (NTMs). Neuronové modelování témat, jak ukazují například Wu a kol. [46], zahrnuje poměrně širokou škálu modelů, které jsou jednak vystavěny na různých principech, jednak umožňují různé způsoby modelování témat (hierarchické modelování, modelování vývoje témat v čase apod.).

NTMs bývají vyzdvihovány z několika důvodů. Zengul a kol. [45] ve své studii na základě porovnání modelů LSA, LDA a Top2Vec, konstatují, že *„algoritmus Top2Vec lze považovat za nadřazený ostatních algoritmů díky tomu, že nevyžaduje tak velkou závislost na lidském vstupu a tak velké úsilí během předzpracování dat.“* Problematickou stránku NTMs může na druhou stranu představovat složitá interpretace nastavení jednotlivých parametrů [41]. NTMs se potýkají také s neexistencí ustálených přístupů pro vyhodnocení modelu a citlivostí na nastavení hyperparametrů [46].

Modely lze kategorizovat také z jiného pohledu, a to z pohledu, který byl naznačen již u neuronových modelů. Například Alghamdi a Alfalqi [40] rozlišují dvě kategorie modelů. První skupinu tvoří modely zaměřené čistě na nalezení témat v textu, jako jsou již zmiňované modely LSA, LDA, PLSA apod. Druhé skupině dali autoři studie název „topic evolution models“ (modely vývoje témat), což jsou modely, jejichž cílem je popsat vývoj témat v čase. Autoři v rámci této studie upozorňují, že modelování témat bez zohlednění vývoje témat v čase může vést k matoucím výsledkům.

Modely pro modelování vývoje tématu v čase lze vnímat také jako rozšíření LDA, jak ve své studii ukazují Kherwa a Bansal [39]. Autoři v této studii mimo jiné prezentují další modely, jež vycházejí z LDA, mezi které se řadí například supervizované modely, hierarchické modely, vícejazyčné modely nebo korelované modely.

3.3 Předzpracování dat

Předzpracování dat je nezbytnou součástí každého procesu vytěžování informací z dat. To lze ilustrovat například na metodice CRISP-DM (Cross-Industry Standard Process for Data Mining), která představuje rámec pro získávání informací z dat. CRISP-DM je flexibilní 6fázový model, přičemž jednou z fází toho cyklu je právě předzpracování neboli příprava dat

[47]. Tato podkapitola se proto ve stručnosti zaměřuje na předzpracování textových dat, a to zejména v kontextu modelování témat.

Proces předzpracování dat může společně s nastavením hyperparametrů značně ovlivnit kvalitu výsledných témat [41]. Pro ilustraci toho, jaký má předzpracování dat vliv na kvalitu výsledného modelu, je možné zmínit studii Dennyho a Spirlinga nesoucí název *Text Preprocessing For Unsupervised Learning* [48], ve které autoři hodnotí kvalitu modelů LDA v souvislosti s podniknutými kroky předzpracování dat.

Proces předzpracování textových dat je možné rozdělit na několik dílčích kroků. Vijayarani a kol. [49] považují za základní kroky předzpracování textových dat extrakci, odstranění stop slov a stematizaci. Extrakcí slov je v tomto případě rozuměna tokenizace neboli rozdělení obsahu dokumentu na jednotlivá slova nebo přesněji tokeny. Anadakumar a Padmavathy [50] považují za základní shodné tři kroky, v rozšířeném přehledu však uvádějí i další kroky jako je například označení slovních druhů (Part-of-Speech Tagging). Uvedené tři základní kroky jsou popsány níže, přičemž jsou nejdříve uvedeny příklady poukazující na to, jak se může předzpracování dat lišit na základě použitých analytických nástrojů.

Karl, Wisnowski a Rushing [51] v souvislosti s LSA a LDA popisují dílčí kroky předzpracování dat jako sjednocení znakové sady, odstranění nadbytečných informací a následné zpracování jednotlivých slov. Tím autoři míní odstranění interpunkce, převedení všech znaků na malá písmena (minuskule), odstranění čísel z textu, odstranění příliš dlouhých slov a odstranění slov, které se v celém korpusu vyskytují buď velmi často, nebo naopak výjimečně. Některé kroky jsou ovšem podle autorů volitelné. Například to, jestli budou z korpusu odstraněna čísla, bude záviset na tom, jaké cíle si prováděná analýza klade.

Jak již bylo zmíněno, výhodou některých neuronových modelů (Top2Vec, BERTopic) oproti klasickým metodám modelování témat je to, že nevyžadují tak náročný proces předzpracování dat. Dle studie [52] zaměřené na porovnání metod předzpracování textových dat v souvislosti s nejmodernější přístupy k NLP, se však dá říci, že předzpracování dat bývá v kontextu nových modelů, jako jsou transformer modely podceňováno. Volba správného přístupu může naopak podle autorů studie zvýšit efektivitu a výkonost nejnovějších modelů.

3.3.1 Tokenizace

Tokenizaci lze považovat za prvotní úlohu procesu získávání informací z textových dat (text miningu) [28]. Princip tokenizace spočívá v rozdělení textu na jednotlivá slova, fráze, symboly nebo věty, kterým se říká tokeny, přičemž se odstraňují některé znaky, jako je

například interpunkce [50]. Například větu „Patentová analýza je důležitým nástrojem CI.“ lze rozdělit na tokeny: patentová; analýza; je; důležitým; nástrojem; CI.

3.3.2 Odstranění stop slov

Odstranění stop slov znamená odstranění slov, která se v daném jazyce vyskytují s vysokou frekvencí. Na tyto slova lze nahlížet také jako na slova bez informační hodnoty [50]. Obvykle se může jednat například o spojky nebo předložky, ale je možné vytvořit vlastní seznam stop slov pro konkrétní typ dokumentů. Konkrétně v patentových dokumentech, jak ukazuje Oldham [28], se často vyskytují slova jako metoda (method), zařízení (device) nebo přístroj (apparatus).

3.3.3 Lematizace a stematizace

Pojmy lematizace a stematizace lze z pohledu zpracování přirozeného jazyka popsat jako proces, kdy se „*libovolné slovo daného přirozeného jazyka převádí buď na jeho základní, slovníkový gramatický tvar, nebo na holý kmen, případně ještě na menší část společnou celé skupině slov…“* [53]. Zatímco v anglickém jazyce se pojmy stematizace a lemmatizace odlišují, jak upozorňuje Oldham [28], v českém jazyce se využívá především výraz lematizace – výraz stematizace odkazuje specificky na převod na kmen slova, který se anglicky nazývá stem [53].

3.4 Reprezentace textu

Patil a kol. [54] uvádí, že transformace textu do numerické podoby ve formě vektoru nebo matice a následné modelování představují dva základní kroky každé úlohy zpracování přirozeného jazyka. První ze těchto kroků lze popsat pojmy, jako je reprezentace textu nebo vnoření slov.

Výše zmiňovaní autoři [54] dělí způsoby reprezentace textu do několika kategorií. První kategorii tvoří statistické metody založené na frekvenci výskytu jednotlivých slov. Autoři tuto kategorii dále dělí na podskupiny metod, přičemž mezi konkrétními metodami autoři uvádí metody, jako jsou One Hot Encoding (OHE), Bag-of-Words (BoW), Term-Frequency (TF), Inverse-Document-Frequency (IDF) apod. Dalšími skupinami jsou poté podle autorů metody založené na pravidlech a metody reprezentace textu využívající neuronové sítě.

Abdelrazek a kol. [41] poukazují na to, že modely pro analýzu témat v textu využívají různé typy reprezentace textu. Některé modely jsou podle autorů založeny na konceptu sady slov (BoW), kdy není bráno v potaz pořadí jednotlivých slov neboli konstruktů. Další modely

naopak pracují s reprezentací, jež pořadí konstruktů zohledňuje. Nejnovější modely témat podle autorů v neposlední řadě využívají kontextuální reprezentaci textu, která zachycuje významové vztahy mezi slovy v dokumentu a vychází z tzv. transformer modelů jako je model BERT.

Reprezentaci textových dat pomocí sady slov využívají nejenom algebraické modely témat, jak již bylo uvedeno, ale také pravděpodobnostní modely, jako je například LDA. Algebraické a pravděpodobnostní modely témat jsou tedy založeny na statistických přístupech k reprezentaci textu. Neuronové modely témat poté vychází právě z přístupů k reprezentaci textu využívajících neuronové sítě.

Přístupy k reprezentaci textu založené na neuronových sítí automaticky extrahují syntaktické a významové vlastnosti textu, přičemž vytváří vektorové reprezentace, které zachycují význam jednotlivých slov [54]. Princip těchto přístupů reprezentace vyjadřuje pojem distribuční sémantika, která vychází z myšlenky, že význam každého slova je vymezen významem slov v jejichž blízkosti se dané slovo často vyskytuje [55].

Přístupy založené na neuronových sítích mohou vytvářet statické nebo dynamické reprezentace, přičemž dynamické reprezentace jsou schopny postihnout i mnohoznačnost některých slov, jelikož berou v úvahu kontext, ve kterém se dané slovo vyskytuje [54].

Mezi rozšířené neuronové modely témat patří modely LDA2Vec [41; 56], Top2Vec [45; 57] nebo BERTopic [20; 41; 57]. Modely LDA2Vec a Top2Vec jsou založeny na statické reprezentaci textu. Oba modely vychází z algoritmu Word2Vec. Zatímco LDA2Vec využívá pro vytvoření textové reprezentace přímo algoritmus Word2Vec, Top2Vec využívá odvozený algoritmus Doc2Vec [58]. Dynamický neboli kontextuální způsob reprezentace poté využívá výše zmíněný model BERTopic [59].

3.4.1 Sada slov

Sada slov neboli BoW představuje základní princip reprezentace textu u algebraických i pravděpodobnostních modelů témat [41]. Tento způsob reprezentace využívá „*vektor, jenž má tolik složek, kolik je slov (termínů) ve slovníku nebo v souboru dokumentů.*“ [60].

Podstata reprezentace v případě modelování témat spočívá poté v tzv. dokument-term matici (DTM) [51], kterou lze vnímat jako rozšíření sady slov. DTM je matice jejíž řádky reprezentují jednotlivé dokumenty a sloupce jednotlivá slova obsažená v korpusu. Transpozicí DTM vznikne term-dokument matice (TDM), která sice není ideálním způsobem

reprezentace slov v případě modelování témat, ale nachází uplatnění v řadě jiných úloh zpracování přirozeného jazyka [51].

Jednotlivé prvky sady slov mohou být reprezentovány třemi způsoby, a to binárně, počtem výskytů slova v dokumentu (TF) nebo pomocí TF-IDF [60]. V praxi to znamená, že se jednotlivé prvky v DTM vyjadřující výskyt příslušného slova v daném dokumentu nahradí příslušnými hodnotami TF-IDF.

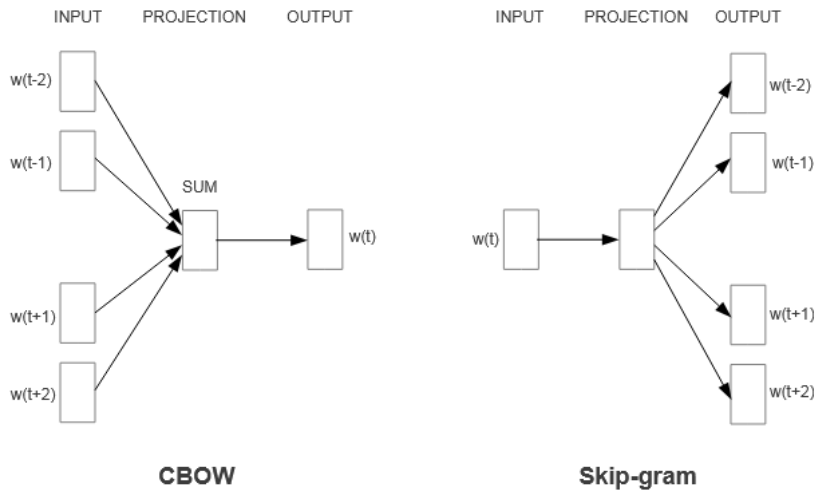
TF-IDF

Metodu TF-IDF (3.1) lze popsat jako kombinaci dvou samostatných metod, a to TF (Term Frequency) a IDF (Inverse Document Frequency) [61]. TF vyjadřuje váhu daného termínu v dokumentu, nebere ovšem v potaz ostatní dokumenty, což řeší právě metoda IDF, která porovnává celkový počet dokumentů s počtem dokumentů, kde se daný termín vyskytuje [54]. Výsledná váha *W* vypočítaná za pomoci metody TF-IDF, je tedy součinem TF a IDF. TF vyjadřuje podíl výskytu termínu *t* v dokumentu *d* vůči celkovému počtu termínů v dokumentu *d.* IDF je poté logaritmem podílu celkového počtu dokumentů *N* vůči počtu dokumentů, ve kterých se daný termínu *t* vyskytuje:

$$W_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \tag{3.1}$$

3.4.2 Word2Vec

Word2Vec je metoda vyvinutá vědci ve společnosti Google, která pomocí neuronových sítí automaticky vytváří vektorovou reprezentaci slov v uvažovaném korpusu [55]. Mikolov a kol. [62] ve své studii představují dvě architektury pro vytvoření vektorové reprezentace textu (obr. 7). První z nich se nazývá Continuous Bag-of-Words (CBOW), přičemž tato metoda je založena na predikci cílového slova na základě okolních slov tvořící kontext daného slova. Druhá architektura se pak nazývá Continuous Skip-gram. Model založený na této metodě se naopak učí predikovat okolní slova na základě vstupního slova tak, aby ve výstupní vrstvě měly vysokou pravděpodobnost slova právě slova, která se nacházela v blízkosti uvažovaného slova.



Obr. 7 Architektury Continuous Bag-of-Words a Skip-gram
Zdroj: [62]

Obě architektury (obr. 7) zjednodušují strukturu dopředných a rekurentních neuronových sítí, jejichž výpočetní náročnost je dána především nelineárními skrytými vrstvami, a tak jsou oproti těmto modelům výpočetně méně náročné [54]. Výsledný vektor každého slova lze získat poté, co proběhne trénování modelu, přičemž tento vektor je tvořen váhami spojení vstupního slova se skrytou vrstvou [55].

3.4.3 BERT

Model BERT (Bidirectional Encoder Representations from Transformers) lze spolu s modely, jako jsou Context2Vec, CoVe, GPT nebo ELMo, zařadit do skupiny modelů, které vytváří dynamické reprezentace, jež jsou proměnné v návaznosti na aktuální kontext [54].

BERT [63] je založen na architektuře transformer modelů. Tato architektura byla představena v článku *Attention Is All You Need* [64], který publikovala společnost Google v roce 2017. V článku navrhovaný model se skládá ze dvou částí, kterými jsou kodér a dekóder. Pomocí kodéru jsou vstupní symboly převáděny do kontextuálních vektorových reprezentací. Na základě těchto reprezentací je generována výstupní sekvence, přičemž symboly jsou generovány postupně, kdy každý vygenerovaný symbol představuje dodatečný vstup.

Kodér obsahuje několik vrstev, které se dále skládají z dopředné neuronové sítě a tzv. self-attention vrstvy, díky které je model schopen vytvořit reprezentaci na základě pozic všech symbolů v předchozí vrstvě [64].

BERT vychází z výše popsané kodérové části modelu [59]. Oproti jiným kontextuálním modelům tento model pracuje s oboustranným kontextem [54]. To znamená, že model vytváří reprezentaci daného na základě kontextu z levé i z pravé strany [54].

3.5 Modely témat

Konkrétních modelů témat existuje celá řada, jak již bylo prezentováno v kapitole zabývající se klasifikací těchto modelů (viz kapitola 3.2). Popsat všechny modely proto není v možnostech této práce. Tato kapitola si tak klade za cíl podrobněji představit principy nejpoužívanějších modelů témat.

Využití modelů témat pro analýzu patentů lze ilustrovat na některých studiích. Kim, Park, Yoon [65] ve své studii prezentují využití LDA pro vytvoření vývojových map patentů. Využití této metody pro analýzu patentů pak ilustrují i další studie [66; 67].

Problematickým aspektem LDA metody může být, jak již bylo zmíněno, její výpočetní náročnost. V některých případech tak může být vhodné upřednostnit například Latentní sémantickou analýzu [45]. Song a kol. [68] ve své studii využívají metodu NMF pro porozumění vývoje funkcí u dronových technologií.

Moderní přístup k modelování témat představují neuronové modely. Tento přístup byl aplikován například ve studii [69], která analyzovala patentové dokumenty věnující se příbojové energii.

3.5.1 Latentní sémantická analýza

Latentní sémantické analýza (LSA) představuje vůbec první metodu, která se začala využívat pro analýzu témat v textu [41]. LSA vychází z hypotézy, že význam slov je obsažen v kontextu, ve kterém byly použity [14]. Algoritmus této metody lze popsat ve 3 krocích [39]:

1) vytvoření dokument-term matice;
2) přiřazení vah jednotlivým prvkům matice;
3) dekompozice matice na singulární hodnoty (SVD).

LSA na vstupu předpokládá dokument-term matici. Druhým krokem algoritmu je přiřazení váhy pro každý prvek matice, a to za využití metody TF-IDF. Takto vytvořená matice A je poté za pomoci SVD rozložena na součin 3 matic U, Σ a V^T [39]:

$$A = U \times \Sigma \times V^T \tag{3.2}$$

Matice U je tvořena dokumenty a koncepty, matice V^T je tvořena koncepty a termíny a matice Σ je diagonální matice, která popisuje sílu jednotlivých konceptů, na základě čehož se určuje počet témat, které budou vyhodnocovány [14].

3.5.2 Nezáporná maticová faktorizace

Metoda Nezáporné maticové faktorizace (NMF) je dalším zástupcem algebraických modelů témat. Tato metoda vychází z předpokladu, že vícerozměrnou matici lze rozložit na součin nezáporných matic (3.3) nižšího rozměru [41]:

$$V \approx WH \tag{3.3}$$

Matice V o rozměru $n \times m$, kde n představuje počet atributů a m počet záznamů, je aproximována na matici W o rozměru $n \times r$ a matici H o rozměru $r \times m$ [70]. Matice V je DTM s hodnotami, které byly kódovány pomocí metody TF-IDF [71]. NMF tedy předpokládá stejný vstup jako LSA. Vzniklá matice W představuje matici slov a témat, matice H poté vyjadřuje rozložení témat v jednotlivých dokumentech [71].

Hledání matic W a H je iterativní proces, během kterého se postupně pomocí multiplikativních algoritmů nalézají stálé nové hodnoty těchto matic do doby nalezení lokálního minima ztrátové funkce [70]. Pro vyhodnocení kvality aproximace se využívají metriky, jako jsou eukleidovská vzdálenost nebo Kullbackova-Leiblerova divergence, které umožňují změřit vzdálenost mezi dvěma nezápornými maticemi [70].

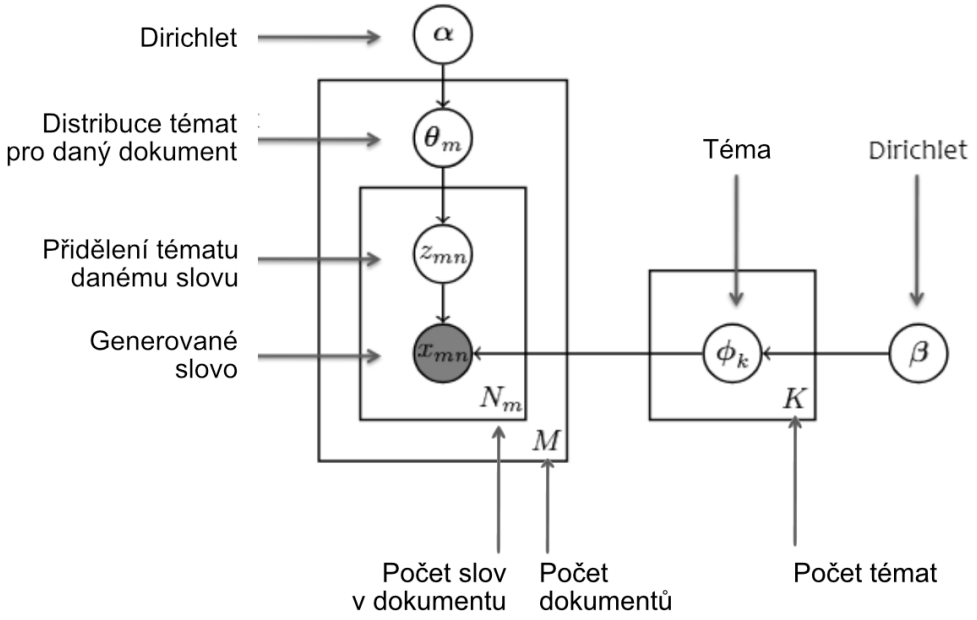
NMF se osvědčila pro nalézání témat v krátkých textech, kde se pravděpodobnostní modely jako LDA potýkají s nedostatkem informací pro vytvoření dostatečného statistického základu [71]. NMF se zároveň uplatňuje i v jiných oblastech, než je analýza témat v textu. NMF lze využít také pro segmentaci dat, redukci dimenzionality, zpracování obrazu, rozpoznávání vzorů apod. [39].

3.5.3 Latentní Dirichletova alokace

PLSA je metoda, která byla představena v roce 1999 s úmyslem napravit některé nedostatky LSA [40]. Se záměrem nahradit tuto metodu pak vznikla Latentní Dirichletova alokace [40].

Blei a kol. [72] ve své studii popisují tři entity, se kterými metoda LDA pracuje. První entitou jsou slova, druhou entitou dokumenty, které lze chápat jako sekvenci slov, a třetí entitou je korpus, který je možné popsat jako soubor všech dokumentů.

Principem metody LDA je reprezentace tématu jako rozdělení pravděpodobnosti slov, které určuje, s jakou pravděpodobností se dané slovo v tématu vyskytuje, přičemž každý dokument je pak směsí těchto témat [40; 72].



Obr. 8 Grafická reprezentace modelu LDA
Zdroj: [73], upraveno

LDA lze popsat graficky pomocí diagramu (obr. 8), kde každý obdélník reprezentuje opakování algoritmu pro každého zástupce dané entity [72]. LDA představuje generativní proces, který předpokládá, že každý dokument lze vygenerovat pomocí následujícího procesu [39; 72]:

a) pro každý dokument *M* je vygenerována distribuce témat *θ*;
b) pro každé slovo *N* v dokumentu je poté přiřazeno téma *z* distribuce *θ,* přičemž výsledné slovo *w* je vybráno z distribuce slov pro dané téma *z*.

Nevýhody modelu jsou nestabilita produkovaných výsledků, která je závislá na pořadí, ve kterém model zpracovává vstupní dokumenty, nebo nutnost předurčení počtu témat, které má model odhalit [41].

3.5.4 Top2Vec

Top2Vec [58] je neuronový model témat, který byl vytvořen Dimem Angelovem v roce 2020. Top2Vec v prvním kroku vytváří vektorovou reprezentaci dokumentů pomocí metody Doc2Vec, která vychází z algoritmu Word2Vec.

V dalších krocích dochází k redukci dimenzionality vytvořených vektorových reprezentací skrze model UMAP. Následně pak Top2Vec využívá model HDBSCAN pro vytvoření shluků s podobnými tématy.

Na základě identifikovaných shluků jsou pomocí výpočtu aritmetického průměru vektorů dokumentů patřících k danému shluku nalezeny vektory reprezentující jednotlivá témata. Poslední krok představuje nalezení slov reprezentující dané téma. Tato slova jsou vybrána na základě vektorů slov v sémantickém prostoru, které se nachází nejblíže k vektoru daného tématu.

3.5.5 BERTopic

BERTopic je model, který představil Maarten Grootendorst v roce 2022 [59]. Název modelu odkazuje na velký jazykový model BERT, který je v rámci modelu využíván pro vytvoření vektorové reprezentace dokumentů.

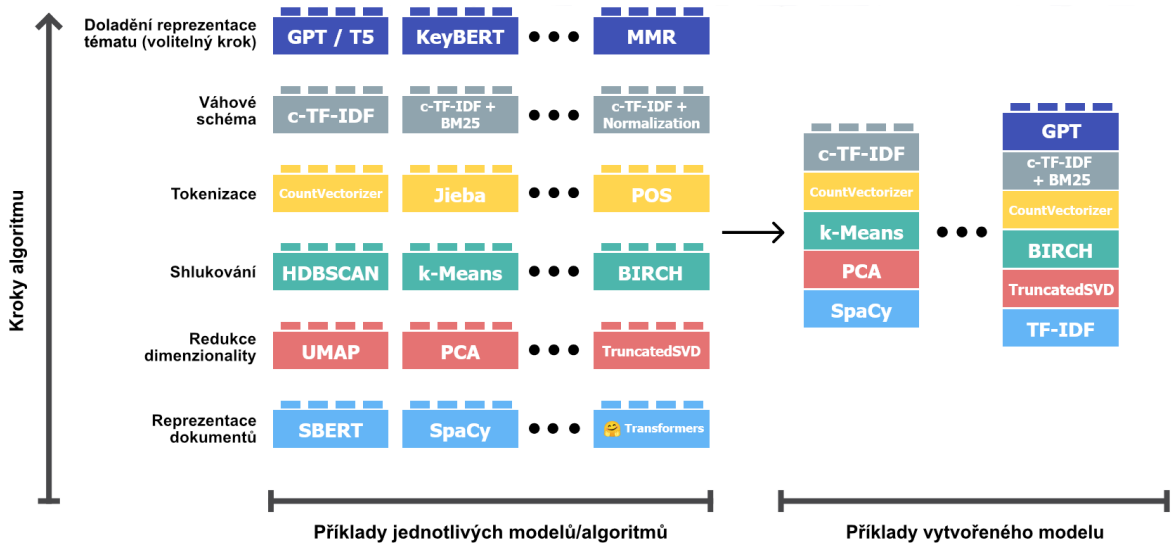
Celý algoritmus modelu BERTopic je možné rozdělit do tří základních kroků [59]. Model nejprve vytvoří vstupní vektorové reprezentace dokumentů. Následujícím krokem je redukce počtu dimenzí vytvořené reprezentace a shlukování dokumentů. Vytvořené shluky jsou popsány za pomoci c-TF-IDF.

Vektorová reprezentace jednotlivých dokumentů je vytvořena, jak již bylo uvedeno, skrze velký jazykový model BERT. BERTopic pak implementuje konkrétně model SBERT, ale díky modularitě modelu je možné využít i jiné modely, což umožňuje přizpůsobit model aktuálním trendům v rámci vývoje NLP [59]. Pro redukci dimenzionality a shlukování BERTopic využívá, stejně jako v případě modelu Top2Vec, modely UMAP a HDBSCAN [59].

Témata jsou v rámci modelu BERTopic reprezentována na základě důležitosti jednotlivých slov v rámci třídy, přičemž třídu tvoří všechny dokumenty odpovídajícího shluku, které jsou pro účely výpočtu sjednoceny do jediného dokumentu. BERTopic k tomu využívá metodu c-TF-IDF (3.4), což je modelu vlastní verze metody TF-IDF. Tato metoda namísto četnosti termínu (TF) v dokumentu počítá četnost termínu t ve třídě c. Inverzní četnost termínů napříč dokumenty (IDF) pak BERTopic nahrazuje výpočtem inverzní četnosti termínů napříč všemi třídami, kde A vyjadřuje průměrný počet slov v dokumentu v rámci jedné třídy a tf_t četnost termínu t napříč všemi třídami [59]:

$$W_{t,c} = tf_{t,c} \times \log\left(1 + \frac{A}{tf_t}\right) \tag{3.4}$$

Jak lze vidět na obrázku č. 9, model BERTopic je značně modulární. V rámci jednotlivých kroků algoritmu tak lze využít různé modely. Například pro redukci dimenzionality vstupní reprezentace dokumentů pomocí algoritmu UMAP lze využít analýzu hlavních komponent (Principal Component Analysis – PCA).



Obr. 9 Algoritmus modelu BERTopic
Zdroj: [74], upraveno

Mimo to BERTopic umožňuje také modelování vývoje témat v čase (DTM), hierarchické modelování nebo propojení modelu s velkými jazykovými modely, jako jsou ChatGPT, Llama 2 nebo Mistral, které mohou usnadnit interpretaci nalezených témat [74].

4 IDENTIFIKACE TECHNOLOGICKÝCH TRENDŮ

Cílem CI, jak bylo uvedeno v úvodní kapitole, je usnadnit manažerům a dalším pracovníkům společnosti rozhodování tak, že vždy budou mít všechny potřebné informace týkající se konkurenčního prostředí. Jedním z významných zdrojů těchto informací jsou patentová data, která reflektují aktuální stav technologického vývoje. Tato data je možné analyzovat různými metodami, přičemž jednou z metod je analýza témat, která umožňuje identifikovat patentové dokumenty popisující stejné technologie, a tak identifikovat aktuální technologické trendy.

Konkrétní využití techniky analýzy témat v textu v kontextu CI bude v této kapitole ilustrováno na identifikaci technologický trendů v oblasti fotografické techniky. Pro analýzu bude využit model BERTopic, který je zástupcem kategorie neuronových modelů témat. Model BERTopic využívá kontextuální reprezentaci textu založenou na transformer modelech, a tak představuje moderní přístup k analýze témat v textu.

Kapitola je rozdělena do čtyř podkapitol. První podkapitola je věnována vymezení konkrétních otázek, na které bude analýza hledat odpovědi. Druhá podkapitola je věnována sběru dat. Navazující podkapitola se věnuje analýze témat v textu a identifikaci technologických trendů. Poslední podkapitola představuje vizualizaci výsledků patentové analýzy pomocí dashboardu.

4.1 Vymezení oblasti zájmu a klíčových otázek

V rámci analýzy bude pracováno se scénářem hypotetické společnosti, jejíž hlavním produktem jsou fotografická zařízení – fotoaparáty. Tato společnost vyvíjí nové produkty, a proto její vedení potřebuje informace o tom, které technologie představují aktuální trendy a jaké produkty vyvíjí konkurence.

Odvětví fotografické techniky představuje rychle se rozvíjející odvětví, které v minulých letech prošlo výraznými změnami. Analogová zařízení byla nahrazena digitálními zařízeními, digitální zrcadlovky čím dále více nahrazují bezzrcadlové přístroje a namísto kompaktních fotoaparátů jsou dnes využívány především mobilní telefony. To lze ilustrovat na datech organizace CIPA (Camera & Imaging Products Association), která vytváří každoroční statistiky o počtu vyprodukovaných fotoaparátů výrobci, mezi které patří aktuálně 9 společností jako Canon, Sony, Nikon, Panasonic, Fujifilm apod. [75]. Zatímco v roce 2014 bylo podle dat CIPA vyprodukováno 42 768 140 digitálních fotoaparátů, přičemž 68 % z tohoto počtu tvořily kompaktní fotoaparáty, 24 % DSLR a 7 % bezzrcadlové fotoaparáty

s výměnnými objektivy [76]. V roce 2024 bylo vyprodukováno pouhých 8 365 303 digitálních fotoaparátů, kdy kompaktní fotoaparáty tvořily 23 %, DSLR 12 % a bezzrcadlové fotoaparáty s výměnnými objektivy 66 % [77].

Digitální fotoaparáty začaly postupně získávat i funkce pro nahrávání videa. Například studie z roku 2018 [78] poukazuje na to, že tzv. DSLR revoluce, která započala v roce 2008 s příchodem fotoaparátů jako Nikon D90 a Canon 5D Mark II, měla vliv na zpřístupnění technologií pro tvorbu dokumentárních a etnografických filmů. V souvislosti s touto funkcionalitou digitálních fotoaparátů došlo k rozvoji např. technologií pro stabilizaci obrazu nebo automatického ostření. Odvětví fotografických zařízení tak představuje proměnné odvětví, kde dochází k neustálému rozvoji nových technologií, a proto bylo zvoleno jako vhodné pro analýzu témat patentů v této práci.

Vymezenou oblastí zájmu uvažované společnosti je technologický vývoj v odvětví fotografických zařízení. Záměrem analýzy je odpovědět na otázky týkající se budoucího směřování odvětví, a tak usnadnit společnosti rozhodování v otázkách vývoje nových produktů. V návaznosti na vymezené cíle analýzy, byly formulovány tyto klíčové otázky:

- Jaké technologie představují aktuální trendy v oblasti fotografických zařízení?
- Jak se tyto trendy vyvíjely?
- Které společnosti jsou nejaktivnější v podávání patentových přihlášek?

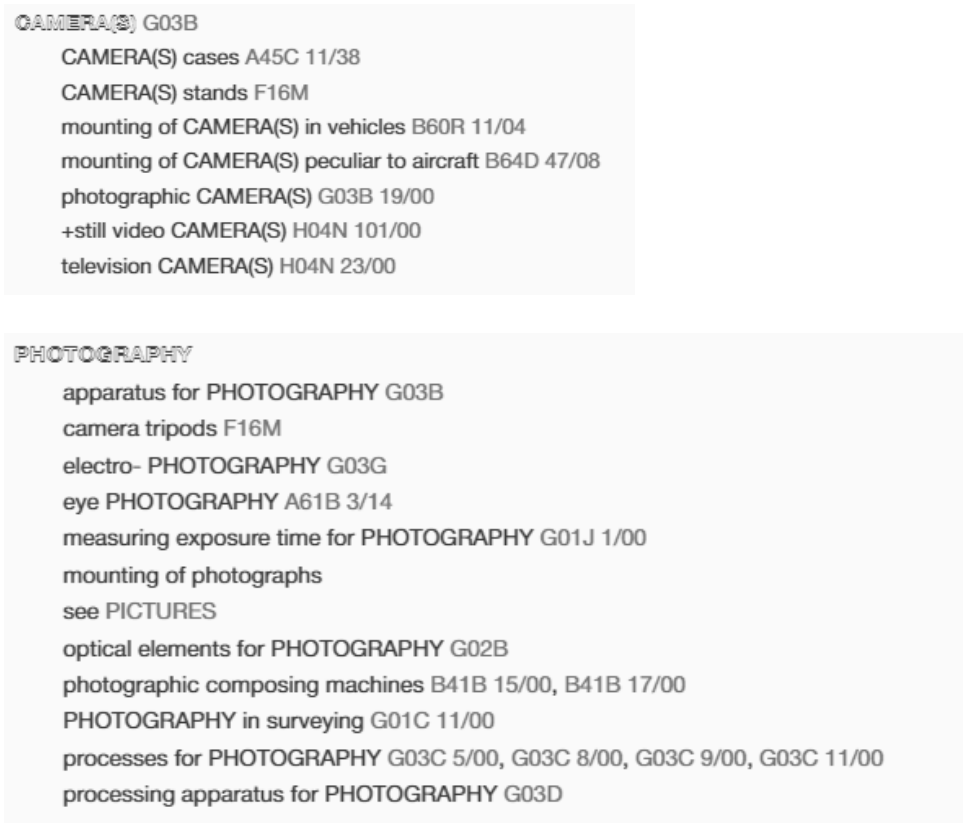
4.2 Sběr dat a jejich předzpracování

Data využitá v této práci pocházejí z patentové databáze Patentscope, kterou spravuje WIPO. V procesu získávání dat byl nejprve definován dotaz pro vyhledání relevantních patentových dokumentů. Následně pak byla vybrána nejvhodnější patentová databáze. Stažená data byla předzpracována, kdy byly podniknuty kroky, jako je sjednocení stažených souborů, odebrání nepotřebných údajů, odstranění duplicitních záznamů apod.

4.2.1 Sestavení dotazu

Vyhledávání pomocí třídy patentové klasifikace lze požadovat za základní způsob vyhledávání v patentové databázi (viz kapitola 2.3.3). Relevantní kód třídy nebo tříd klasifikace je možné volit na základě různých klasifikačních systémů. Vzhledem k tomu, že v rámci této práce budou analyzována patentová data pro celé vybrané odvětví, bylo zvoleno Mezinárodní patentové třídění (IPC), které se řadí mezi nejrozšířenější patentové klasifikační systémy.

Vhodné kódy klasifikace lze vypozorovat u relevantních patentů. Výběr vhodné třídy ale usnadňují také nástroje obsažené v třídníku WIPO [79], který na svých webových stránkách publikuje WIPO. Jak ukazuje obrázek č. 10, třídník WIPO obsahuje seznam tzv. záchytných slov. S ohledem na vymezené cíle analýzy byly vybrána slova „camera“ a „photography“, v důsledku čehož byl zvolena třída G03B.



Obr. 10 Záchytná slova v třídníku IPC
Zdroj: [79]

Třída G03, která se v Mezinárodním patentovém třídění nachází o úroveň výše, zahrnuje fotografii, kinematografii a další obdobné postupy. Do této třídy se řadí například materiály citlivé na světlo nebo elektrografie a magnetografie. Skupiny, které se v hierarchii IPC nachází o úroveň níže než podtřída G03B, odkazují na konkrétní typy fotografických přístrojů a zařízení, nebo na specifické části těchto přístrojů.

Podtřída G03B, která zahrnuje „*přístroje nebo zařízení pro zhotovování fotografických snímků nebo promítání nebo prohlížení snímků; přístroje nebo zařízení využívající obdobné postupy s jinými než optickými vlnami a jejich příslušenství*“ [30], proto vyhovuje vymezeným cílům analýzy nejvíce.

Podtřída G03B zahrnuje například i promítací přístroje, dotaz byl proto doplněn několika klíčovými slovy, a to konkrétně slovem „camera“, slovem „mirrorless“ a zkratkou DSLR, která se vztahuje k jednookým digitálním zrcadlovkám. Výsledný dotaz ve formě, který vyžaduje databáze Patentscope, byl vymezen jako „G03B AND (camera OR DSLR OR mirrorless)“.

4.2.2 Výběr databáze a stažení dat

Data, která jsou v této práci analyzována, pochází z databáze Patentscope. Tato databáze (v roce 2025) shromažďuje více něž 122 milionů patentů z více než 160 patentových úřadů, přičemž přibližně 5,1 milionu patentů z celkového počtu tvoří mezinárodní (PCT) přihlášky [80].

Kritérii pro výběr patentové databáze byly dostupnost potřebných údajů, respektive polí, nebo pokrytí databáze. V návaznosti na vymezené KIQs je nezbytné, aby zvolená databáze umožňovala přístup k abstraktům, datu podání patentu, informacím o přihlašovatelích patentu a informacím patentových úřadech, kde byl patent přihlášen.

Jiang a Goetz [38] poukazují na to, že většina studií pracuje s kratšími texty (názvy nebo abstrakty patentů), které mohou být příliš obecné, pravděpodobně kvůli tomu, že v minulosti neexistovali tak výkonné jazykové modely jako dnes. Ačkoli by tedy bylo možné analyzovat celé popisy patentů, budou upřednostněny abstrakty. Patentové databáze často neposkytují celé popisy, nebo je poskytují pouze jako sken dokumentů, zpracování dat by tak vyžadovalo zapojení dalších nástrojů.

Pro analýzu vývoj témat v čase bylo upřednostněno datum podání přihlášky před datem uveřejnění. Proces zpracování přihlášky může trvat i několik let, modelování na základě data uveřejnění patentu tak ukazuje trendy se zpožděním, jak upozorňuje Oldham [23].

Dalšími kritérii bylo množství patentů, které lze hromadně stáhnout, možnost vyhledávání pomocí kombinace IPC a klíčových slov a funkce pro filtrování výsledků podle patentových rodin, díky které byly odfiltrovány patenty popisující stejný vynález.

Patentscope umožňuje po registraci hromadné stažení 10 000 tisíc dokumentů. Zároveň nabízí pokročilé filtrování záznamů, včetně omezení vyhledaných záznamů pouze na jednoho člena patentové rodiny.

4.2.3 Předzpracování dat

Neuronové modely témat umožňují vynechat některé tradiční kroky předzpracování textových dat. Například odstranění stop slov není u modelu BERTopic doporučováno z toho důvodu, že modely založené na transformer architektuře, jako je BERT, vyžadují pro dosažení optimálních výsledků úplný kontext [74]. Model BERTopic v tomto ohledu přináší výhodu, že oddělení jednotlivých kroků algoritmu umožňuje využít různé přístupy předzpracování dat pro vytvoření kontextuální reprezentace dokumentů a pro finální reprezentaci témat [59].

Naopak je ovšem vhodné z dat odstranit například HTML tagy, které nejsou pro pochopení textu nijak významné [74], což je případ patentových dat, která jsou v této práci zpracována. Kromě HTML tagů některé abstrakty obsahovaly také copyright přihlašovatele patentu, který byl rovněž odstraněn.

Model BERTopic je založen na programovacím jazyce Python, a proto byly pro předzpracování dat i pro následnou pro vizualizaci výsledků využity nástroje založené právě na této platformě. S daty bylo pracováno v prostředí Jupyter Notebooku. V rámci procesu předzpracování dat byla využita především knihovna Pandas, která poskytuje nástroje pro práci a manipulaci s daty. Dále byly využity také knihovny glob pro načítání datových souborů, chardet pro detekci kódování textu, os pro interakci s operačním systémem a knihovna re umožňující práci s regulárními výrazy.

Načtení a sjednocení souborů

Data byla stažena ve formátu .xls, a proto byla nejprve převedena do formátu .csv, přičemž bylo z jednotlivých souborů odstraněno záhlaví obsahující informace o datu stažení, použitém dotazu apod.

Jak již bylo uvedeno v předchozí kapitole, databáze Patentscope je limitována stažením maximálně 10 000 záznamů na jednou. Dalším krokem proto bylo sjednocení stažených souborů. Všechna data byla uložena do jednoho adresáře, přičemž pro načtení všech relevantních souborů z adresáře byly využita knihovna glob.

Data byla postupně načítána do objektu dataframe pomocí funkce „read_csv“ z knihovny Pandas. Vzhledem k tomu, že některé záznamy obsahovaly specifické znaky, byla využita také funkce „detect“ z knihovny chardet pro detekci správného kódování daného textu. Identifikované kódování bylo využito pro nastavení parametru zmiňované funkce „read_csv“.

Selekce atributů a odstranění chybějících záznamů.

Z databáze Patentscope byla stažena kompletní data obsahující všechny dostupné atributy. Pomocí funkce „drop“ z knihovny Pandas tak byly odebrány sloupce odpovídající nepotřebným atributům. Ponecháno bylo identifikační číslo patentu, datum podání přihlášky, datum zveřejnění patentu, označení země příslušného patentového úřadu, abstrakty a jména přihlašovatelů. Názvy vybraných atributů byly sjednoceny na malá písmena.

Výchozí soubor obsahoval 105 514 záznamů. Ne všechny záznamy obsahovaly kompletní údaje, neúplné záznamy proto byly odebrány pomocí funkce „dropna“, což vedlo ke snížení počtu 104 420.

Odstranění duplicit

Ze vstupních dat byly rovněž odstraněny duplicitní záznamy, a to pomocí funkce „drop_duplicates“. Tato funkce byla nejprve využita pro odstranění případných duplicitních záznamů, jež se shodují ve všech atributech.

Následně byly odebrány záznamy se shodným identifikačním číslem přihlášky, jelikož se jedná o unikátní označení každého patentu, a také záznamy se shodnými abstrakty. Počet patentů tak byl redukován z 104 420 na 102 376.

Odstranění HTML značek a prohlášení autorských práv

V úvodu kapitoly bylo poukázáno na skutečnost, že přítomnost HTML značek, které neposkytují významné informace pro pochopení textu, může negativně ovlivnit kvalitu kontextuální reprezentace dokumentů. Pro odstranění těchto značek byla vytvořena funkce „clean_html“, která vrací očištěný text. Funkce využívá funkci. sub z knihovny re pro nahrazení regulárního výrazu zvoleným výrazem. Regulární výraz představuje vzor ve formě textového řetězce pro identifikaci odpovídajících výrazů v textu [81]. HTML značky byly identifikovány pomocí výrazu „<.*?>“. Takto vytvořená funkce byla aplikována na jednotlivé abstrakty pomocí funkce „map“.

Pro odstranění copyrightu byl zvolen obdobný přístup. Nejprve byly s využitím funkce „findall“ z knihovny re a funkce set identifikovány všechny unikátních varianty prohlášení. Teprve na základě těchto výsledků byl vytvořen výsledný regulární výraz, který lze vidět na obrázku č. 11.

```python
def clean_copyright(data):
    cleaned_coypright = re.compile("copyright:.*|copyright kipo.*?(?=Reference|$)|copyright .... KIPO", re.IGNORECASE)
    return cleaned_coypright.sub("", data)

#odstraněné copyrightu v abstraktech
patents["abstract"] = patents["abstract"].map(clean_copyright)
```

Obr. 11 Funkce pro vyhledání a vyčištění prohlášení autorských práv z textu
Zdroj: vlastní zpracování

Úprava názvů přihlašovatelů

Posledním krok procesu předzpracování dat, který byl v této práci využit byla úprava jmen osob nebo názvů společností, které daný patent přihlásili. Některé názvy a jména byly v návaznosti na patentový úřad, kde byl daný patent přihlášen, v datech uvedeny ve více jazykových mutacích. V těchto případech byl textový řetězec rozdělen pomocí funkce „split“ na dílčí části, přičemž následně bylo ponecháno pouze označení přihlašovatele v anglickém jazyce. Tato označení byla rovněž sjednocena z hlediska velikosti písma tak, že všechny názvy byly převedeny na velká písmena (verzálky).

Omezení data přihlášení patentu

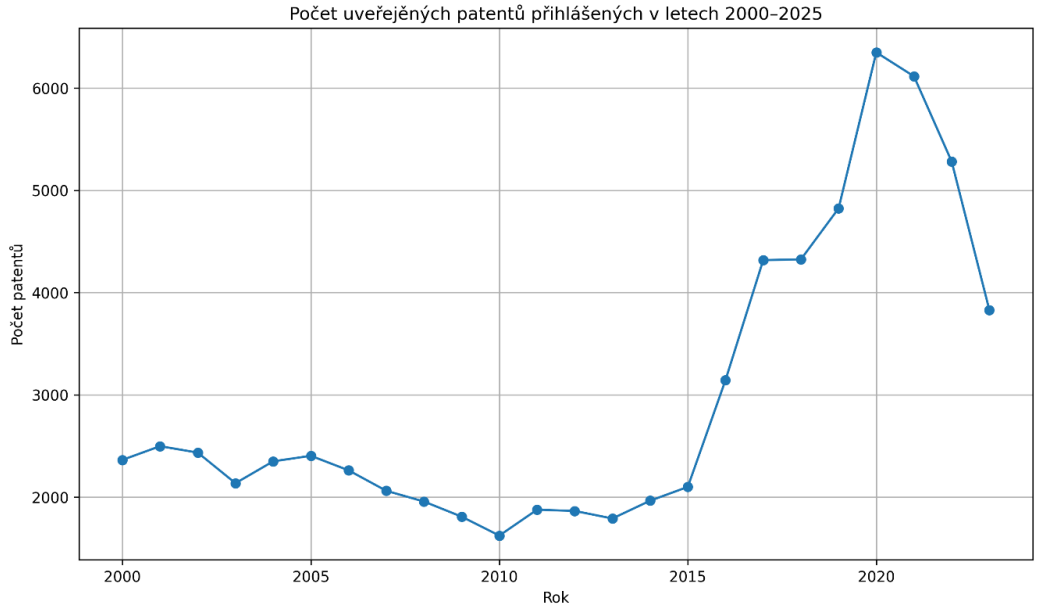
Odstraněním neúplných a duplicitních záznamů se počet patentů snížil z původních 105 514 patentů na 102 376. Pro účely analýzy trendů byly dále odfiltrovány patenty přihlášené před rokem 2000 a také patenty přihlášené v letech 2024 a 2025.

Roky 2024 a 2025 byly odebrány, jelikož data z posledních let bývají zkreslena v důsledku délky procesu schvalování patentové přihlášky. Oldham [28] uvádí, že mezi datem podání patentové přihlášky a schválením patentu obvykle uplyne 18 měsíců až 2 roky. Tento problém je podle Oldhama v patentové statistice označován jako „timeliness“. Pro ilustraci lze poukázat na studii Van Rijna a Timmise [82], kteří z tohoto důvodu ve své analýze, která byla provedena v polovině roku 2020, vynechali data z let 2019 s 2020.

4.2.4 Datový soubor

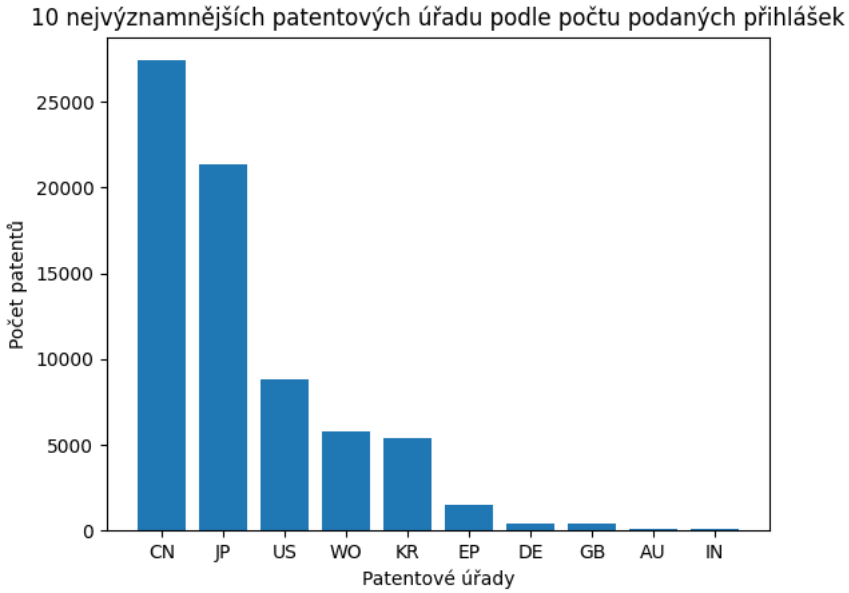
Datovou sadu po odstranění všech neúplných a duplicitních záznamů, a vyfiltrování relevantních záznamů tvořilo 71 679 patentů přihlášených v letech 2000–2023. Datová sada obsahovala 6 atributů, kterými byly identifikační kód daného patentu, datum přihlášení a datum publikování patentu, země patentového úřadu, název, abstrakt a přihlašovatel.

Počet patentů v datové sadě zaznamenal nárust po roce 2015 (obr. 12). V roce 2021 pak začíná počet patentů klesat.



Obr. 12 Počet uveřejněných přihlášek v jednotlivých letech

Zdroj: vlastní zpracování



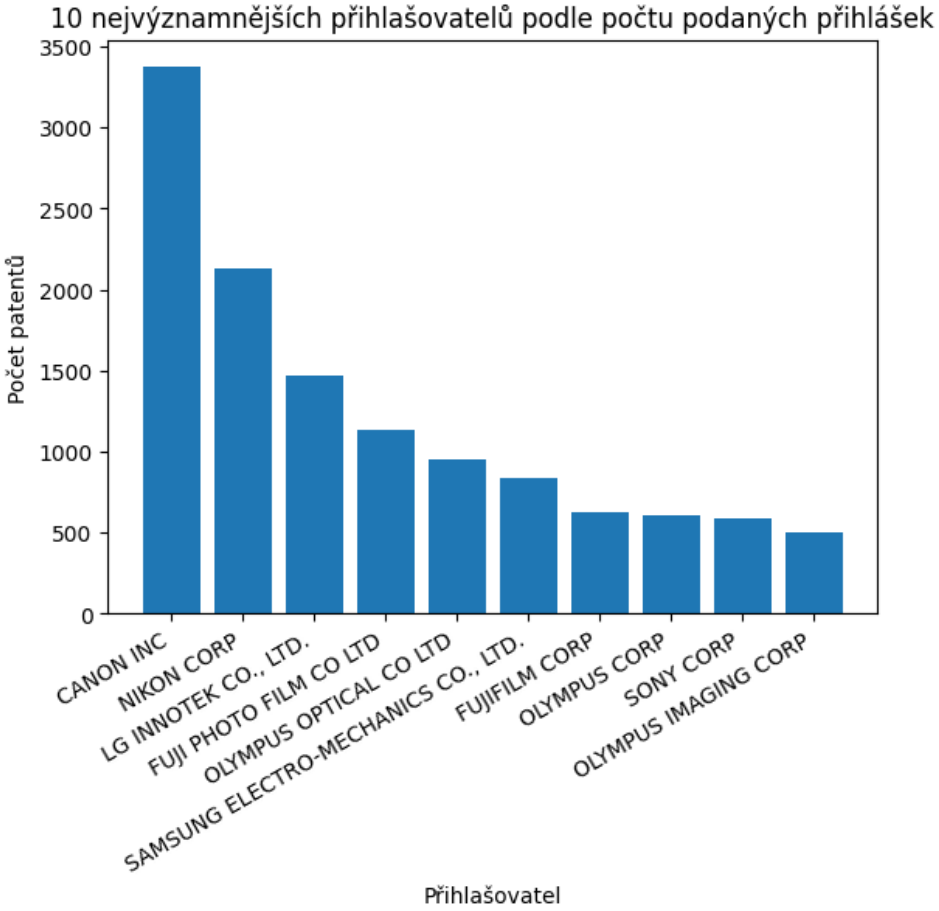
Obr. 13 10 nejvýznamnějších patentových úřadů

Zdroj: vlastní zpracování

Datová sada obsahuje patenty, které byly přihlášeny u 41 patentových úřadů. Jak lze vidět na obrázku č. 13. Nejvýznamnějšími patentovými úřady podle počtu přihlášených patentů byly patentové úřady Číny, Japonska a Spojených Států amerických, WIPO a patentový úřad Jižní Koreje. Těchto 5 úřadů pokrývá 95,9 % patentů v datové sadě.

Společnost Canon Inc. představuje s 3 372 patenty nejvýznamnější společnost z hlediska patentové aktivity. Společnost Canon je následována korporacemi Nikon Corp. (2 126 patentů) a LG Innotek Co., Ltd. (1 471 patentů). Dalšími významnými přihlašovateli (obr. 14) jsou společnosti Fuji Photo Film Co., Ltd., Olympus Co., Ltd., Samsung Electro-Mechanics Co., Ltd., Co. nebo Sony Corp.

Patenty v datové sadě lze přiřadit celkem k 25 263 přihlašovatelům. Je nutné upozornit, že některé společnosti jsou zde uvedeny vícekrát. Název přihlašovatele souvisí se zemí patentového úřadu, kde byl daný patent přihlášen.



Obr. 14 10 nejvýznamnějších přihlašovatelů

Zdroj: vlastní zpracování

4.3 Analýza témat v patentových datech

Model použitý pro analýzu patentových dat byl navržen primárně na základě dokumentace modelu BERTopic [74]. Popis modelu a možností jeho nastavení v této kapitole rovněž vychází z dokumentace modelu [74], ale také z původní studie [59].

4.3.1 Model

V procesu modelování byla nejprve vytvořena reprezentace dokumentů. Následně bylo optimalizováno nastavení shlukovacího modelu. Závěrečným krokem byla tvorba finálních reprezentací témat, kdy byly využity různé přístupy s cílem usnadnit interpretaci identifikovaných témat.

Reprezentace dokumentů

Vytvoření kontextuální reprezentace jednotlivých dokumentů, respektive patentů v případě této práce, představuje první krok algoritmu modelu BERTopic. Pro vytvoření reprezentace lze využít různé modely, které lze získat prostřednictvím Python modulu Sentence Transformers, případně lze využít i jiné modely, jako je například Model2Vec.

Pro vytvoření reprezentace byl zvolen dokumentací doporučovaný model all-MiniLM-L6-v2. Model all-MiniLM-L6-v2 představuje univerzální model, který vytváří reprezentace o rozměru 384 dimenzí a jeho trénovací dataset tvořila více než 1 miliarda trénovacích párů [83]. Tento model se s velikostí 80 MB řadí k nejmenším modelům, zároveň však v porovnání s většími modely dosahuje dobrých výsledků při tvorbě textové reprezentace [83].

Reprezentace představuje výpočetně náročný úkol, vytvořená reprezentace proto byla pomocí knihovny Pickle uložena do samostatného souboru, ze kterého byla v dalších krocích načítána.

Redukce dimenzionality a shlukování

Pro redukci dimenzionality a k následnému shlukování byly využity doporučované metody UMAP a HDBSCAN. Implementace UMAP v BERTopic umožňuje nastavit parametry „n_neighbors“, „n_components“, „metric“ a „low_memory“. U HDBSCAN lze nastavit hodnoty parametrů „min_cluster_size“, „min_samples“ a „metric“.

Nastavení parametru „n_components“ má vliv na počet dimenzí ve výstupní matici. Nastavení „n_neighbors“ ovlivňuje počet sousedních bodů, které jsou během výpočtu brány v úvahu, přičemž zvýšení hodnoty tohoto parametru vede k tvorbě větších shluků.

Většina parametrů byla ponechána na doporučovaném výchozím nastavení. Parametr, jehož nastavení bylo optimalizováno je „min_cluster_size“. Tento parametr má přímý vliv na výsledný počet témat tím způsobem, že ovlivňuje minimální počet dokumentů v jednom shluku. Optimální nastavení parametru bylo vyhodnoceno na základě výpočtu koherence témat (viz kapitola 4.3.2).

Reprezentace témat

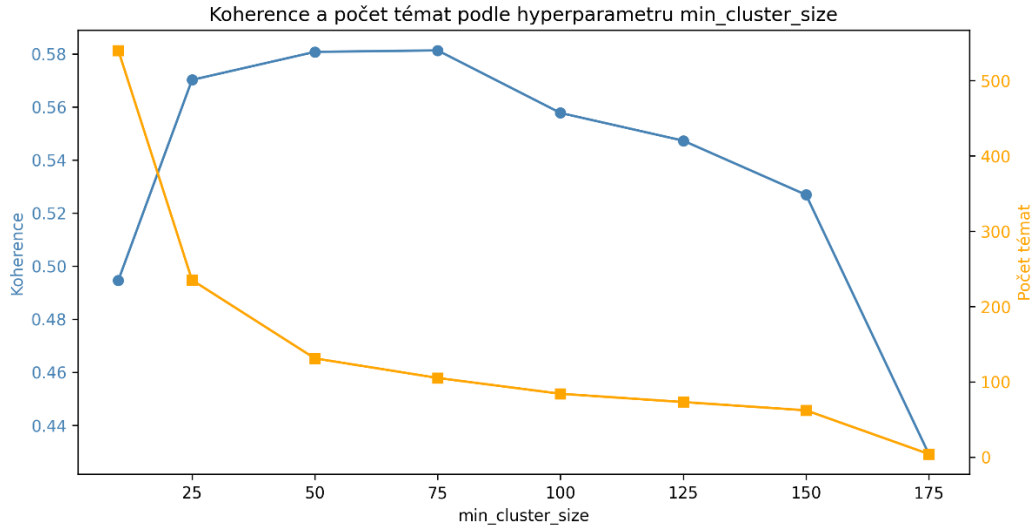
BERTopic vytváří reprezentace témat na základě již popsané metody c-TF-IDF. Modularita modelu BERTopic umožňuje tyto reprezentace optimalizovat za pomoci metody CountVectorizer. Zároveň lze využít i další způsoby reprezentace, jako jsou reprezentace vytvořené pomocí jazykových modelů.

CountVectorizer představuje metodu z knihovny sklearn pro tvorbu DTM, která zároveň umožňuje odstranit stop slova, vypustit slova, která se v korpusu vyskytují s vysokou nebo naopak s nízkou frekvencí nebo pomocí parametru „ngram_range“ ovlivnit velikost tokenů. V kontextu BERTopic ovlivňuje výsledné reprezentace témat. Tato metoda byla nastaveny tak, aby z reprezentací byla odstraněna stop slova. Parametr „min_df“ byl nastaven na doporučovanou hodnotu 2 a parametr „ngram_range“ na hodnoty 1, 2 tak, aby slova mohla být sloučena i do bigramů.

Posledním způsobem reprezentace témat, který byl využit je reprezentace za pomoci jazykových modelů. BERTopic lze propojit například s modely společnosti OpenAI nebo s modelem Llama. Využití těchto modelů však může vyžadovat zaplacení poplatku nebo vysoký výpočetní výkon. V této práci proto byl využit open-source model FLAN-T5 od společnosti Google. Pomocí promptu, jenž je uveden v dokumentaci modelu BERTopic, bylo otevřeno označení pro každé z nalezených témat.

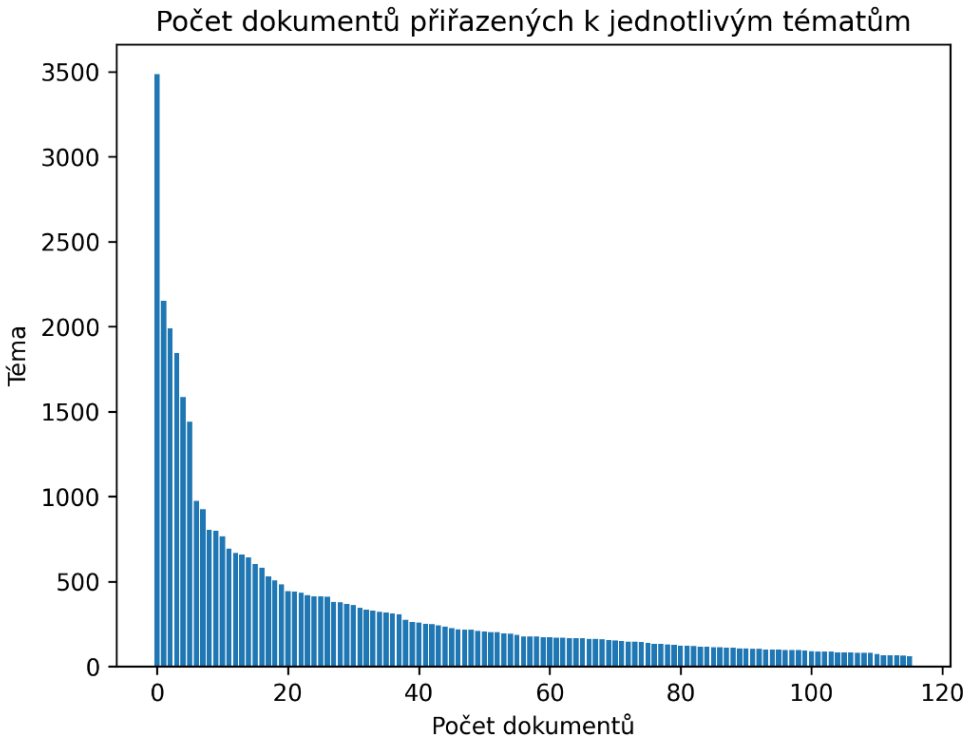
4.3.2 Vyhodnocení modelu

V rámci procesu modelování témat bylo experimentováno s nastavením parametru „min_cluster_size“. Pro nalezení nejvhodnějšího nastavení parametru byla využita metrika koherence témat. Tato metrika představuje průměr nebo medián podobnosti párů slov, které jsou tvořeny slovy reprezentující dané témat [84]. Pro výpočet koherence byla využita knihovna Gensim, která obsahuje funkci „coherence“.



Obr. 15 Koherence a počet témat podle nastavení parametru „min_cluster_size“
Zdroj: vlastní zpracování

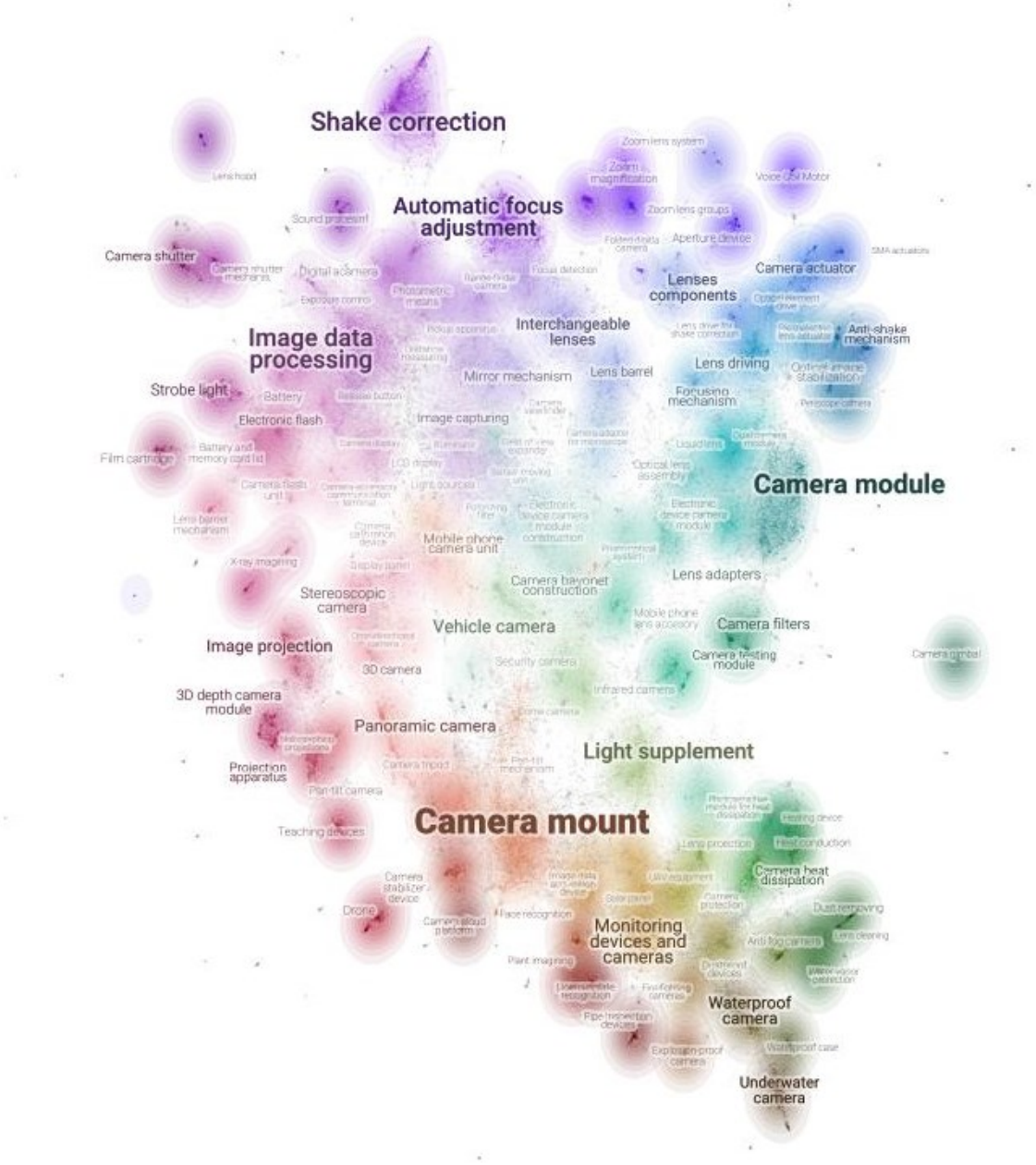
Na obrázku č. 15 je možné vidět, že nejvyšší koherence témat model dosahoval při nastavení parametru „min_cluster_size“ na hodnoty 50 a 70. Při nastavení parametru na hodnotu 50 byla koherence témat 0,5807 a model dokumenty rozdělil do 131 shluků. Při nastavení parametru „min_cluster_size“ na hodnotu 75 pak koherence témat byla 0,5812 a model dokumenty rozdělil do 105 shluků. Ve finálním nastavení modelu byla zvolena hodnota parametru 60, model tak dokumenty rozdělil do 116 shluků s koherencí témat 0,585.



Obr. 16 Počet dokumentů přiřazených k jednotlivým tématům

Zdroj: vlastní zpracování

Z celkového počtu 71 679 patentů, které byly analyzovány, bylo 39 997 přiřazeno k jednomu ze 116 témat. 31 682 dokumentů z celkového počtu pak nebylo přiřazeno k žádnému tématu. K tématu č. 1 bylo přiřazeno 3474 patentů a k tématu č. 116 61 patentů. Počet dokumentů, jež byly přiřazeny k jednotlivým tématům vyjadřuje graf na obrázku č. 16.



Obr. 17 Mapa identifikovaných témat
Zdroj: vlastní zpracování

Obrázek č. 17 představuje mapu všech témat a dokumentů, která byla vytvořena za využití Python knihovny datamapplot. Jednotlivá témata jsou popsána označeními, které vychází z klíčových slov popisující dané téma, ale také z reprezentace, která byla vytvořena pomocí modelu FLAN-T5, a reprezentativních dokumentů daného tématu.

12 nejzastoupenějších témat je vyobrazeno na obrázku č. 18. Interpretaci těchto témat kromě reprezentativních slov usnadňují také štítky vytvořené pomocí modelu FLAN-T5 a reprezentativní dokumenty, které k jednotlivým tématům model BERTopic přiřadil. Model BERTopic identifikoval mimo těchto 12 témat dalších 104 témat, mezi které patří například 3D kamery, ovládání fotoaparátu pomocí hlasu nebo fotoaparáty v mobilních telefonech. Přehled všech identifikovaných témat v podobě reprezentativních slov se nachází v příloze této práce.

Obr. 18 12 nejzastoupenějších témat
Zdroj: vlastní zpracování

K tématu 0, které je popsáno slovy „support“, „rod“, „plate“, „supporting“, „mounting“, bylo přiřazeno celkem 3 485 patentů. Model Flan-T5 k tomuto tématu vygeneroval označení „camera mount“, které koresponduje s reprezentativními dokumenty, ze kterých vyplývá, že toto téma popisuje různá zařízení pro upevnění a stabilizaci kamer a fotoaparátů.

Téma 1 je popsáno slovy „module“, „camera module“, „housing“, „lens“ a „image senzor“, přičemž k tomuto tématu náleží celkem 2 151 patentů. Toto téma je věnováno konstrukci kamerových modulů. Ty se využívají pro integraci kamer v zařízeních, jako jsou mobilní telefony, automobily nebo produkty chytré domácnosti [85].

Téma 2 popisuje systémy pro detekci a redukci třesu využívající různé gyroskopické, optické a elektronické mechanismy. Téma je charakterizováno slovy shake, correction, shake correction, blur, vibration apod. Z celkového počtu 71 679 patentů se tomuto tématu věnuje celkem 1 990 patentů.

Další téma je zastoupeno 1 847 patenty a popisuje funkce digitálních fotoaparátů pro ukládáním, pokročilé třídění a další práci s daty. Příkladem je funkce pro automatické spuštění spouště při rozpoznání specifického vzoru a následné zatřídění fotografie do správné složky.

Téma 4 je věnováno mechanismům a systémům pro automatické zaostřování. Téma 5 poté popisuje světelné zdroje, jako jsou například LED lampy a zábleskova světla pro osvětlování obrazu. K tématu č. 6, které zahrnuje 974 patentů, se řadí různá monitorovací zařízení a kamery, jako jsou kamery pro monitorování budov a kamery pro monitorování silničního provozu. Téma 7 přísluší vynálezům, které popisují kamery, jež jsou součástí vozidel. Téma 8 je reprezentováno pomocí slov, jako jsou projection nebo screen, a náleží vynálezům kombinující projekční systémy s kamerami pro zlepšení kvality výsledného promítaného obrazu. Mezi 12 nejzastoupenějšími tématy se nachází také téma reprezentující panoramatické kamery, téma zahrnující voděodolné kamery a mechanismy zajišťující voděodolnost zařízení. Téma 11 patří k 692 patentům popisující výměnné objektivy a mechanismy pro komunikaci mezi fotoaparátem a objektivem.

4.3.3 Vývoj témat v čase

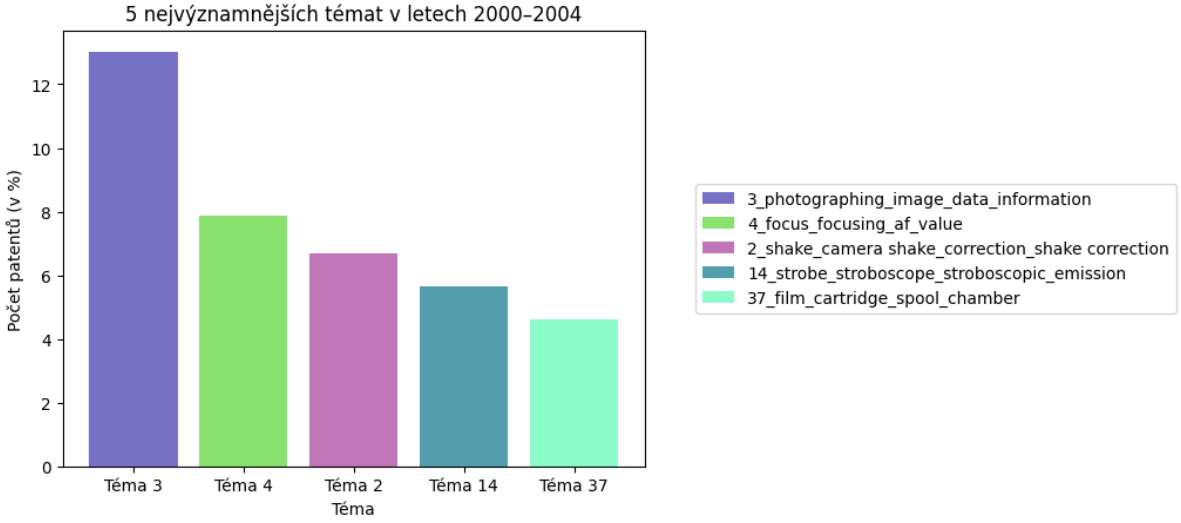
Jeden ze způsobů, jak sledovat vývoj témat a technologických trendů v čase, využívají ve své studii Teshome a kol. [86]. Studie ukazuje, že technologické trendy lze sledovat na základě počtu patentů přiřazených k danému tématu ve vybraných obdobích. Na základě těchto počtů lze určit i podíl, který dané téma ve vybraném období zaujímá. Díky relativním hodnotám je pak možné porovnávat významnost témat v jednotlivých období.

Tento přístup bude využit i v této práci, data proto byla rozdělena do čtyř 5letých období a jednoho neúplného 4letého období. První období zahrnuje patenty z let 2000 až 2004, druhé období poté patenty z let 2005 až 2009, třetí z let 2010 až 2014 a čtvrté z let 2015 až 2019. Poslední neúplné období zahrnuje data z let 2020 až 2023. Jak již bylo uvedeno, data z let 2024 a 2025 nejsou brány v potaz, jelikož zde dochází ke zkreslení vlivem délky procesu uveřejnění patentové přihlášky. Následující podkapitoly popisují vývoj 5 nejvýznamnějších témat v každém z vymezených období.

2000–2004

Datová sada obsahuje 6 402 patentů, které byly publikovány v letech 2000–2004. Nejvýznamnějším tématem v těchto letech (obr. 19) bylo téma 3, které je popisováno slovy „photographing“, „image“, „data“, „information“, „user“, „image data“. Tímto tématem se zabývalo 13,03 % patentů přihlášených v daném období. Jak bylo uvedeno výše, toto téma je věnováno funkcím pro ukládání, pokročilé třídění a další práci s daty.

Druhému nejvýznamnějšímu tématu, které je věnováno automatickému zaostřování, se ve vymezeném období věnovalo 7,87 % patentů. Další patenty publikované v letech 2000–2004 popisují vynálezy spojené s řešením problému nechtěného třesu, který snižuje kvalitu výsledných fotografií (6,7 % patentů), záblesková zařízení (5,64 % patentů) a mechanismy manipulace s kazetami uchovávající filmový materiál (4,61 %). Lze tedy vidět, že v popředí se v tomto období nacházely i technologie související s analogovými fotografickými zařízeními.

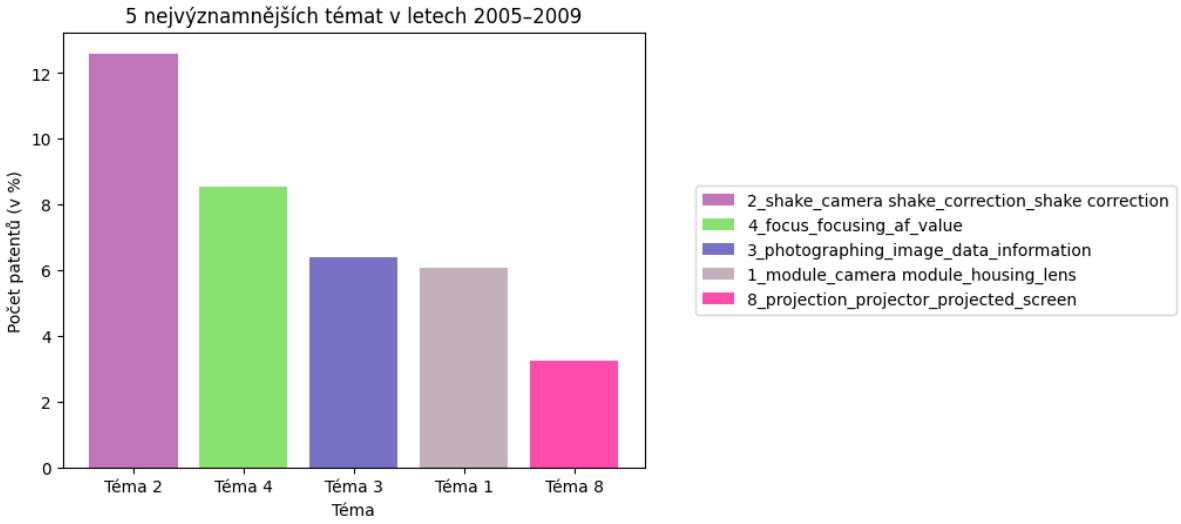


Obr. 19 5 nejvýznamnějších témat v letech 2000–2004
Zdroj: vlastní zpracování

2005–2009

Mezi nejvýznamnějšími tématy patentů v letech 2005 až 2009 (obr. 20) zůstaly na prvních 3 příčkách stejná témata jako v předchozím období. Došlo zde ovšem k výměně pořadí. Na první místo se dostala redukce třesu, kterou se zabývalo 12,6 % patentů publikovaných v tomto období. Na druhém místě zůstalo automatické zaostřování (8,56 % patentů) a na třetí příčku se posunuly funkce pro zpracování dat (6,39 % patentů).

Další významné skupiny v tomto období představovaly patenty věnující se konstrukci kamerových modulů (5,22 % patentů) a patenty věnující se vylepšení projekcí obrazu (3,26 % patentů).

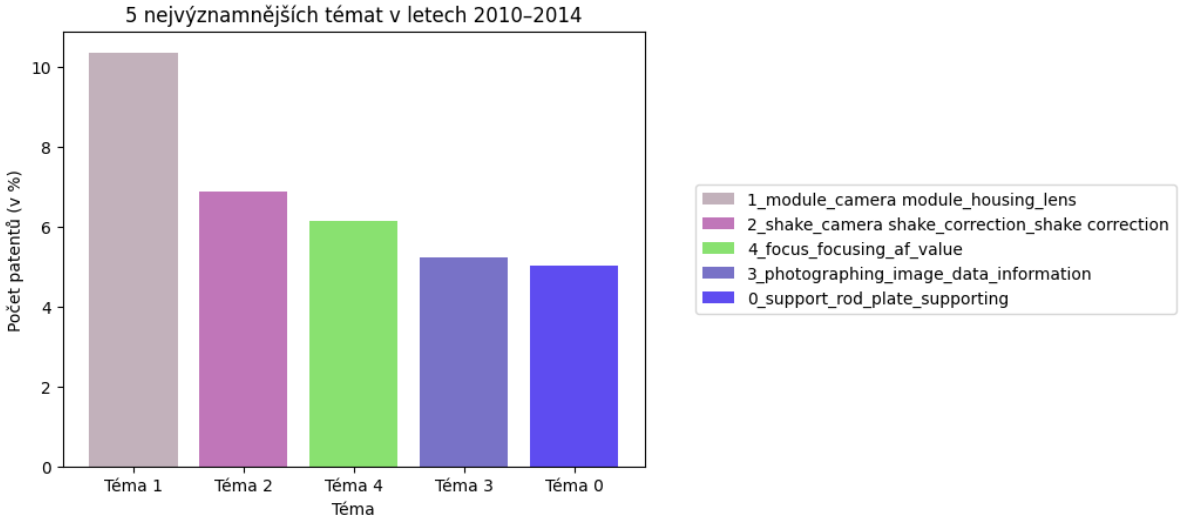


Obr. 20 5 nejvýznamnějších témat v letech 2005–2009
Zdroj: vlastní zpracování

2010–2014

Nejvýznamnější téma patentů v letech 2010 až 2014 (obr. 21) představovaly kamerové moduly. Zatímco v letech 2010–2014 se tomuto tématu věnovalo 5,22 % patentů, v tomto období to bylo 10,37 %.

Na dalších příčkách se poté opět systémy pro redukci chvění a třesu, umístily systémy pro automatické zaostřování (6,15 % patentů) a technologie zaměřené na zpracování textových dat (5,22 % patentů). Významné téma představovaly také patenty popisující různé způsoby upevnění a montáže fotoaparátů a kamer (5,03 % patentů).



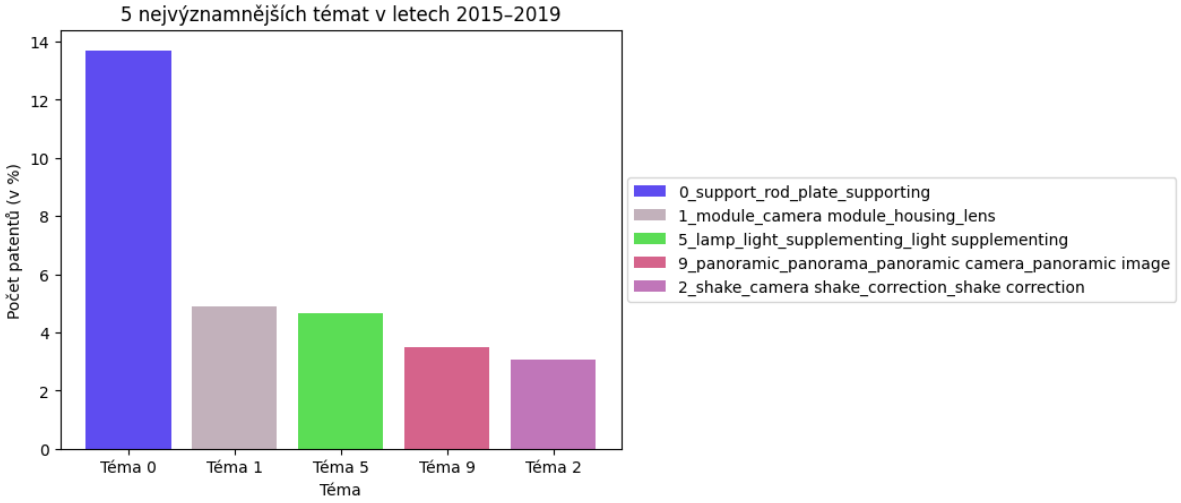
Obr. 21 5 nejvýznamnějších témat v letech 2010–2014

Zdroj: vlastní zpracování

2015–2019

V období zahrnující roky 2015 až 2019 (obr. 22) se nejvíce patentů (13,7 %) věnovalo upevnění a montáži kamer. Toto téma zaznamenalo nárust oproti předchozímu období, kdy se tomuto tématu věnovalo 5,03 % patentů. Absolutní počet patentů věnujících se tomuto tématu vzrostl z 259 na 1 444 patentů.

4,89 % patentů se zabývalo konstrukcí kamerových modulů. Další významná témata představovaly patenty popisující konstrukci osvětlovacích zařízení (4,68 % patentů), panoramatické kamery (3,15 % patentů) a technologie pro redukci třesu (3,05 % patentů).



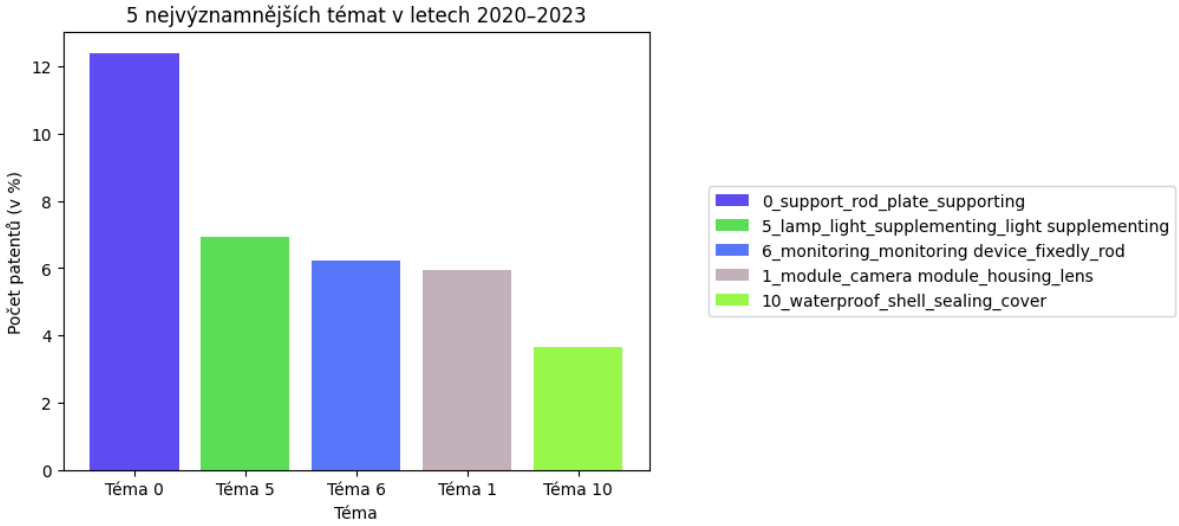
Obr. 22 5 nejvýznamnějších témat v letech 2015–2019

Zdroj: vlastní zpracování

2020–2023

Poslední analyzované období tvořily roky 2020–2023 (obr. 23). Jak již bylo uvedeno, jedná se o neúplné období a data z let 2024 a 2025 nebyla analyzována. Z tohoto důvodu může dojít ke zkreslení výsledků v porovnání s předchozími obdobími.

Nejvýznamnější téma v tomto období představovaly upevnění a montáž kamer (12,4 % patentů), osvětlovací zařízení (6,93 % patentů), monitorovací zařízení (6,23 % patentů), kamerové moduly (5,94 % patentů) a technologie zaměřené na voděodolnost kamer (3,66 % patentů).



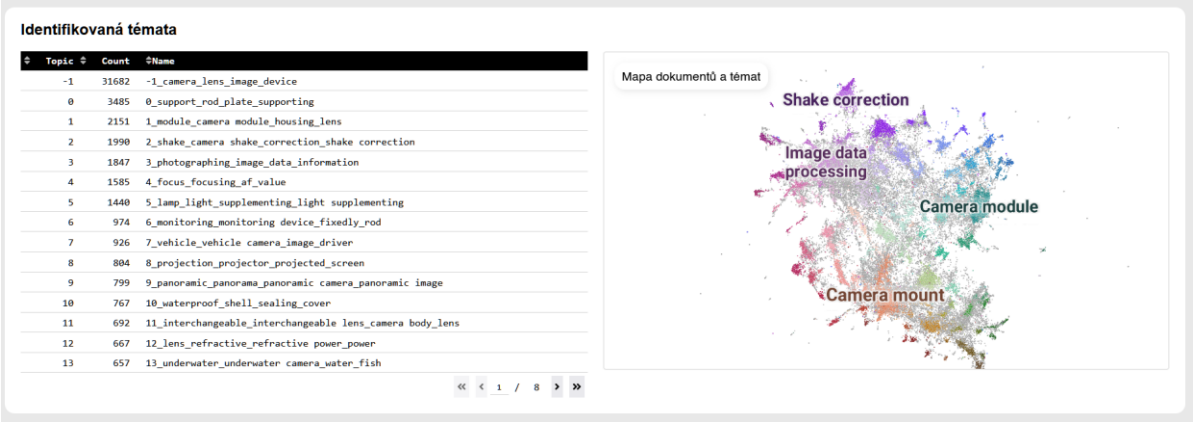
Obr. 23 5 nejvýznamnějších témat v letech 2020–2023
Zdroj: vlastní zpracování

4.4 Prezentace výsledků

Často využívanými nástroji v rámci patentové analýzy jsou nástroje založené na vizualizaci dat, jako jsou například patentové mapy napomáhající porozumět technologickým trendům [17]. V úvodní kapitole práce poté bylo uvedeno, že výstup z procesu CI vždy musí odpovídat požadavkům zadavatele. Jak ovšem ukazují odborné studie, právě vizualizační nástroje představují účinný způsob prezentace informací. Pandey a kol. [87] poukazují na to, že vizualizační nástroje a dashboardy obecně představují nástroj pro vypořádání se s neustále narůstajícím objemem dat, přičemž napomáhají činit správná rozhodnutí.

Model BERTopic sám poskytuje několik prostředků pro vizualizaci výsledků modelování. Tyto nástroje využívají knihovny Plotly, která je zaměřena na vytváření interaktivních grafů a diagramů, a DataMapPlot, jež slouží k vytváření vizualizací ve formě datových map. Společnost Plotly zároveň vyvíjí knihovnu Dash, která slouží k tvorbě interaktivních dashboardů.

Knihovna Dash tak poskytuje z hlediska této práce možný nástroj, pro prezentaci výsledků analýzy témat v patentových datech pro účely konkurenčního zpravodajství. Výsledný dashboard, který byl vytvořen na základě dokumentace knihoven Dash [88] a DataMapPlot [89], byl rozdělen na tři sekce. Úvodní sekce (obr. 24) představuje identifikovaná témata a skládá se z tabulky témat mapy dokumentů a témat. Tato mapa byla vytvořena pomocí funkce „create_interactive_plot“ z knihovny DataMapPlot.

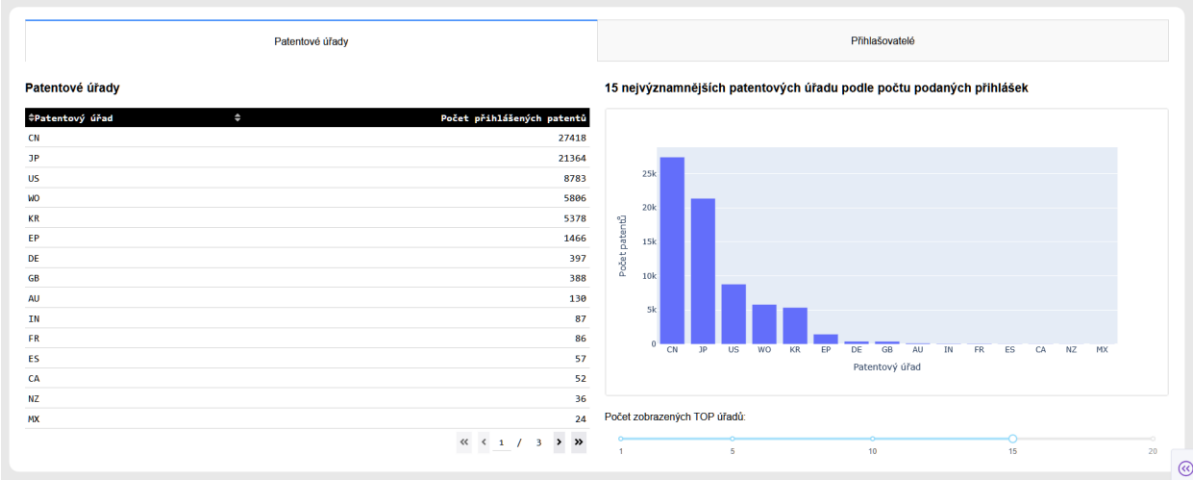


Obr. 24 Dashboard vyobrazující identifikovaná témata
Zdroj: vlastní zpracování

Druhá část dashboardu (obr. 25) nabízí přehled nejvýznamnějších témat ve vybrané období a umožňuje tak vhled na technologické trendy. Plotly Dash umožňuje využít tzv. „callback“ dekorátory, díky kterým lze využít ovládací prvky pro uživatelskou interakci s dashboardem. Tato funkcionalita knihovny Dash tak byly využita pro vytvoření ovládacích prvků, které umožňují zvolit vybrané období a počet zobrazených nejvýznamnějších témat.

Obr. 25 Dashboard vyobrazující nejvýznamnější témata ve zvoleném období
Zdroj: vlastní zpracování

Třetí část dashboardu umožňuje vyobrazit základní údaje o patentových úřadech a přihlašovatelích patentů. Sekce obsahuje seznam úřadů a přihlašovatelů s odpovídajícími počty přihlášených patentů a příslušný graf (obr. 26).



Obr. 26 Dashboard vyobrazující údaje o patentových úřadech a přihlašovatelích
Zdroj: vlastní zpracování

ZÁVĚR

Tématem diplomové práce byla analýza témat patentů pro účely konkurenčního zpravodajství. Práce si kladla za cíl charakterizovat analytické metody používané v konkurenčním zpravodajství a představit problematiku patentové analýzy a analýzy témat v textu. Cílem práce byla identifikace témat patentových dokumentů v odvětví fotografických zařízení a následná identifikace technologických trendů.

Úvodní část práce představuje teoretický úvod do problematiky konkurenčního zpravodajství. V této části jsou vysvětleny základní pojmy z této oblasti a charakterizovány analytické metody používané v konkurenčním zpravodajství. Navazující část se podrobněji věnuje jedné z metod, a to patentové analýze. Další kapitola představuje úvod do problematiky analýzy témat v textových datech. V této kapitole jsou objasněny základní pojmy a představeny principy nejrozšířenějších modelů témat.

Témata byla v rámci praktické čísti práce identifikována pomocí neuronového modelu témat BERTopic, který využívá kontextuální reprezentaci dokumentů založenou na neuronových sítí. Tento model tak představuje moderní přístup k analýze témat v textu. Technologické trendy byly následně vyhodnoceny na základě relativní četnosti daného tématu ve zvoleném období.

Z výsledků analýzy vyplývá, že patenty se nejčastěji zabývali konstrukcí mechanismů pro upevnění kamera, kamerovými moduly, technologiemi pro redukci třesu, zpracováním obrazových dat a technologiemi automatického zaostřování. Pro identifikaci technologických trendů data byla rozdělena do 5 období. Na základě výsledků lze konstatovat, že nejvýznamnější témata v jednotlivých období vychází z nejvýznamnější témat celkově. Zatímco v dřívějších obdobích byly nejvýznamnějšími tématy automatické zaostřování, redukce třesu a zpracování obrazových dat, v novějších období se do popřední dostaly konstrukce kamerových modulů, které se využívají například v mobilních zařízeních, a mechanismy pro upevnění kamer.

Výsledky analýzy byly prezentovány pomocí dashboardu, který byl vytvořen pomocí Python knihovny Dash. Tato knihovna tak umožňuje přímé propojení s nástroji pro analýzu dat, jako může představovat například knihovna Matplotlib, a představuje tak zajímavý nástroj pro prezentaci výsledků analýzy v rámci konkurenčního zpravodajství.

Tato diplomová práce tak představuje úvod do problematiky analýzy témat patentových data, přičemž získané poznatky by bylo možné využít v praxi. Téma práce zároveň nabízí několik možností rozšíření. Model BERTopic je stále aktualizován a přichází s novými funkcionalitami. Jednou z nich je například možnost propojení více modelů, kterou je možné podle dokumentace modelů využít pro porovnání nových dat se dřívějšími výsledky. Tato funkce by tak mohla představovat další způsob, jak identifikovat technologické trendy v patentových datech.

POUŽITÁ LITERATURA

[1] RODENBERG, Josèph H. A. M. *Competitive intelligence and senior management: "the best solution to where to place the office of competitive intelligence is on a par with functions that report directly to the Board"*. Delft: Eburon, 2007. ISBN 978-90-5972-192-0.

[2] VELLA, Carolyn M. a MCGONAGLE, John J. A case for competitive intelligence. Online. *Information Management Journal*. 2002, roč. 36, č. 4, s. 35-40. Dostupné z: https://www.proquest.com/docview/227729242/fulltextPDF/E2266BEA5DB64F66PQ/1?accountid=17239&sourcetype=Scholarly%20Journals. [cit. 2025-04-09].

[3] MOLNÁR, Zdeněk. *Competitive intelligence, aneb, Jak získat konkurenční výhodu*. Odborná kniha s vědeckou redakcí. V Praze: Oeconomica, 2012. ISBN 978-80-245-1908-1. Dostupné také z: http://www.digitalniknihovna.cz/mzk/uuid/uuid:887e4610-142f-11ef-951a-005056827e51.

[4] BARTES, František. *Konkurenční zpravodajství: tvorba podkladů pro strategické rozhodování podniku*. Praha: Grada Publishing, 2022. ISBN 978-80-271-3504-2. Dostupné také z: https://www.bookport.cz/kniha/konkurencni-zpravodajstvi-11912/.

[5] RANJIT, Bose. Competitive intelligence process and tools for intelligence analysis. Online. *Industrial Management & Data Systems*. 2008, roč. 108, č. 4, s. 510–528. Dostupné z: https://doi.org/10.1108/02635570810868362. [cit. 2025-04-20].

[6] PELLISSIER, Rene a NENZHELELE, Tshilidzi E. Towards a universal definition of competitive intelligence. Online. *SA Journal of Information Management*. 2013, roč. 15, č. 2. ISSN 1560-683X. Dostupné z: https://doi.org/10.4102/sajim.v15i2.559. [cit. 2025-04-20].

[7] KAHANER, Larry. *Competitive intelligence: how to gather, analyze, and use information to move your business to the top*. A Touchstone book. New York: Touchstone, 1997. ISBN 06-848-4404-4.

[8] PELLISSIER, Rene a NENZHELELE, Tshilidzi E. Towards a universal competitive intelligence process model. Online. *SA Journal of Information Management*. 2013, roč. 15, č. 2. ISSN 1560-683X. Dostupné z: https://doi.org/10.4102/sajim.v15i2.567. [cit. 2025-04-20].

[9] DU TOIT, Adeline. Understanding key intelligence needs (KINs). Online. *Managing Strategic Intelligence*. 2007, s. 111-121. ISBN 9781599042435. Dostupné z: https://doi.org/10.4018/978-1-59904-243-5.ch007. [cit. 2025-04-20].

[10] MULLER, Marie-Luce. Key intelligence needs. Roadmap of your competitive intelligence capability and activities. Online. *SA Journal of Information Management*. 2004, roč. 6, č. 1. ISSN 1560-683X. Dostupné z: https://doi.org/10.4102/sajim.v6i1.289. [cit. 2025-04-20].

[11] ČERNÝ, Jan; POTANČOK, Martin a MOLNÁR, Zdeněk. Using open data and Google search data for competitive intelligence analysis. Online. *Journal of Intelligence Studies in Business*. 2019, roč. 9, č. 2, s. 72-81. ISSN 2001-015X. Dostupné z: https://doi.org/10.37380/jisib.v9i2.470. [cit. 2025-04-23].

[12] BARTES, František. Intelligence analysis – the royal discipline of Competitive Intelligence. *Acta univ. agric. et silvic. Mendel. Brun*. 2011, roč. 59.7, s. 39–56.

[13] OLSZAK, Celina M. OLSZAK, Celina M. An overview of information tools and technologies for competitive intelligence building: theoretical approach. *Issues in Informing Science and Information Technology*. 2014, roč. 11.1, s. 139-153.

[14] HENDL, Jan. *Big data: Věda o datech - základy a aplikace*. Online. Průvodce. Praha: Grada Publishing, 2021. ISBN 978-80-271-3031-3. Dostupné z: https://www.bookport.cz/kniha/big-data-10604/. [cit. 2025-04-13].

[15] MORENO, Antonio a REDONDO, Teófilo. Text Analytics: The Convergence of Big Data and Artificial Intelligence. Online. *International Journal of Interactive Multimedia and Artificial Intelligence*. 2016, roč. 3, č. 6, s. 57–64. ISSN 1989-1660. Dostupné z: https://doi.org/10.9781/ijimai.2016.369. [cit. 2025-04-24].

[16] DAI, Yue; KAKKONEN, Tuomo a SUTINEN, Erkki. MinEDec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis methods. Online. *International Journal of Computer Information Systems and Industrial Management Applications*. 2011, roč. 3, s. 165–173. ISSN ISSN 2150-7988. Dostupné z: https://cspub-ijcisim.org/index.php/ijcisim/article/view/82. [cit. 2025-04-24].

[17] ABBAS, Assad; ZHANG, Limin a KHAN, Samee U. A literature review on the state-of-the-art in patent analysis. Online. *World Patent Information*. 2014, roč. 37, s. 3-13. ISSN 01722190. Dostupné z: https://doi.org/10.1016/j.wpi.2013.12.006. [cit. 2025-04-24].

[18] WIPO. *WIPO Guide to Using Patent Information*. Online. Geneva: World Intellectual Property Organization, 2021. Dostupné z: https://www.wipo.int/edocs/pubdocs/en/wipo-pub-rn2021-1e-en-wipo-guide-to-using-patent-information.pdf. [cit. 2025-03-26].

[19] JEONG, Byeongki a YOON, Janghyeok. Competitive intelligence analysis of augmented reality technology using patent information. Online. *Sustainability*. 2017, roč. 9, č. 4. ISSN 2071-1050. Dostupné z: https://doi.org/10.3390/su9040497. [cit. 2025-04-24].

[20] JEON, Eunji; YOON, Naeun a SOHN, So Young. Exploring new digital therapeutics technologies for psychiatric disorders using BERTopic and PatentSBERTa. Online. *Technological Forecasting and Social Change*. 2023, roč. 186. ISSN 00401625. Dostupné z: https://doi.org/10.1016/j.techfore.2022.122130. [cit. 2025-04-24].

[21] WORLD INTELLECTUAL PROPERTY ORGANIZATION. *Patents*. Online. World Intellectual Property Organization. Dostupné z: https://www.wipo.int/en/web/patents. [cit. 2025-03-23].

[22] ČESKO. Zákon č. 527/1990 Sb., o vynálezech, průmyslových vzorech a zlepšovacích návrzích. In: *Sbírka zákonů*. 1990. Dostupné také z: https://www.e-sbirka.cz/sb/1990/527/2022-02-01?f=527%2F1990%20Sb.&zalozka=text.

[23] OLDHAM, Paul. *The WIPO Manual on Open Source Patent Analytics*. Online. 2nd edition. Geneva: World Intellectual Property Organization, 2022. Dostupné z: https://wipo-analytics.github.io/manual/. [cit. 2025-03-16].

[24] TRIPPE, Anthony. *Guidelines for Preparing Patent Landscape Reports*. Online. Geneva: World Intellectual Property Organization, 2015. ISBN 978-92-805-2529-8. Dostupné z: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_946.pdf. [cit. 2025-03-16].

[25] LUPU, Mihai; MAYER, Katja; KANDO, Noriko a ANTHONY, Trippe J. (ed.). *Current Challenges in Patent Information Retrieval*. Online. Second Edition. Heidelberg: Springer, 2017. ISBN 978-3-662-53817-3. ISSN 1387-5264. Dostupné z: https://doi.org/10.1007/978-3-662-53817-3. [cit. 2025-03-23].

[26] WIPO. *WIPO Regional Training Course on Intellectual Property for Developing Countries of Asia and the Pacific: Patent Information and Documentation*. PDF dokument. In: WORLD INTELLECTUAL PROPERTY ORGANIZATION. WIPO. Dostupné z: https://www.wipo.int/edocs/mdocs/sme/en/wipo_ip_cm_99/wipo_ip_cm_99_16.pdf. [cit. 2025-04-24].

[27] *A detailed look at a patent*. Online. In: VEREENIGDE OCTROOIBUREAUX N.V. V.O. Patents & Trademarks. Dostupné z: https://www.vo.eu/a-detailed-look-at-a-patent/. [cit. 2025-04-24].

[28] OLDHAM, Paul. *The WIPO Patent Analytics Handbook*. Online. Geneva: World Intellectual Property Organization, 2022. Dostupné z: https://wipo-analytics.github.io/handbook/. [cit. 2025-03-18].

[29] *About the International Patent Classification*. Online. World Intellectual Property Organization. Geneva: World Intellectual Property Organization. Dostupné z: https://www.wipo.int/en/web/classification-ipc/preface. [cit. 2025-03-20].

[30] *Číselník mezinárodního patentového třídění*. Online. Úřad průmyslového vlastnictví. 2025. Dostupné z: https://isdv.upv.gov.cz/webapp/webapp.hxmptn. [cit. 2025-03-18].

[31] *Mezinárodní patentové třídění: Návod k MPT*. Online. Vydání 2024. Praha: Úřad průmyslového vlastnictví, 2024. ISBN 80-728-2050-8. Dostupné z: https://upv.gov.cz/files/uploads/PDF_Dokumenty/tridniky/vynalezy/mpt_2024_navod.pdf. [cit. 2025-03-20].

[32] *Cooperative Patent Classification (CPC)*. Online. Espacenet. 2018. Dostupné z: https://cz.espacenet.com/help?locale=cz_CZ&method=handleHelpTopic&topic=cpc. [cit. 2025-04-25].

[33] SINGH, Vikram; CHAKRABORTY, Kajal a VINCENT, Lavina. Patent database: Their importance in prior art documentation and patent search. *Journal of Intellectual Property Rights*. 2016, roč. 2016, č. 21, s. 42-56.

[34] JÜRGENS, Björn a VICTOR, HERRERO-SOLANA. Espacenet, Patentscope and Depatisnet: A comparison approach. Online. *World Patent Information*. roč. 2015, č. 42, s. 4-12. Dostupné z: https://doi.org/10.1016/j.wpi.2015.05.004. [cit. 2025-03-16].

[35] *Databáze patentů a užitných vzor*. Online. Úřad průmyslového vlastnictví. Dostupné z: https://upv.gov.cz/informacni-zdroje/narodni-databaze/databaze-patentu-a-uzitnych-vzoru. [cit. 2025-04-25].

[36] CLARKE, Nigel S. The basics of patent searching. Online. *World Patent Information*. 2018, roč. 54, s. S4-S10. ISSN 01722190. Dostupné z: https://doi.org/10.1016/j.wpi.2017.02.006. [cit. 2025-03-26].

[37] TSENG, Yuen-Hsien; LIN, Chi-Jen a LIN, Yu-I. Text mining techniques for patent analysis. Online. 2007, roč. 43, č. 5, s. 1216-1247. ISSN 03064573. Dostupné z: https://doi.org/10.1016/j.ipm.2006.11.011. [cit. 2025-04-24].

[38] LEKANG, Jiang a GOETZ, Stephan. Natural language processing in patents: A survey. Online. *ArXiv preprint arXiv:2403.04105*. 2024. Dostupné z: https://doi.org/10.48550/arXiv.2403.04105. [cit. 2025-05-31].

[39] KHERWA, Pooja a BANSAL, Poonam. Topic Modeling: A Comprehensive Review: A Comprehensive Review. Online. *ICST Transactions on Scalable Information Systems*. 2020, roč. 7, č. 24. ISSN 2032-9407. Dostupné z: https://doi.org/10.4108/eai.13-7-2018.159623. [cit. 2025-04-13].

[40] ALGHAMDI, Rubayyi a ALFALQI, Khalid. A survey of topic modeling in text mining. Online. *International Journal of Advanced Computer Science and Applications*. 2015, roč. 6, č. 1, s. 147-153. ISSN 21565570. Dostupné z: https://doi.org/10.14569/IJACSA.2015.060121. [cit. 2025-04-01].

[41] ABDELRAZEK, Aly; EID, Yomna; GAWISH, Eman; MEDHAT, Walaa a HASSAN, Ahmed. Topic modeling algorithms and applications: A survey. Online. *Information Systems*. 2023, roč. 112. ISSN 03064379. Dostupné z: https://doi.org/10.1016/j.is.2022.102131. [cit. 2025-04-01].

[42] GHAFFARI, Mohsen; ALIAHMADI, Alireza; KHALKHALI, Abolfazl; ZAKERY, Amir; DAIM, Tugrul U. et al. Exploring the technological leaders using tire industry patents: A topic modeling approach. Online. *Technology in Society*. 2024, roč. 78. ISSN 0160791X. Dostupné z: https://doi.org/10.1016/j.techsoc.2024.102664. [cit. 2025-04-14].

[43] YUN, Junghwan a GEUM, Youngjung. Automated classification of patents: A topic modeling approach. Online. *Computers & Industrial Engineering*. 2020, roč. 147. ISSN 03608352. Dostupné z: https://doi.org/10.1016/j.cie.2020.106636. [cit. 2025-04-14].

[44] RASHID, Junaid; SHAH, Syed Muhammad Adnan a IRTAZA, Aun. Fuzzy topic modeling approach for text mining over short text. Online. *Information Processing and Management*. 2019, roč. 56, č. 6. ISSN 03064573. Dostupné z: https://doi.org/10.1016/j.ipm.2019.102060. [cit. 2025-04-25].

[45] ZENGUL, Ferhat; BULUT, Aysegul; ONER, Nurettin; AHMED, Abdulaziz; YADAV, Manju et al. A practical and empirical comparison of three topic nodeling methods using a COVID-19 corpus: LSA, LDA, and Top2Vec. Online. *Hawaii International Conference*

on System Sciences. 2023. Dostupné z: https://doi.org/10.24251/HICSS.2023.116. [cit. 2025-04-25].

[46] WU, Xiaobao; NGUYEN, Thong a LUU, Anh Tuan. A survey on neural topic models: methods, applications, and challenges. Online. *Artificial Intelligence Review*. 2024, roč. 57, č. 2. ISSN 1573-7462. Dostupné z: https://doi.org/10.1007/s10462-023-10661-7. [cit. 2025-04-24].

[47] *IBM SPSS Modeler CRISP-DM Guide*. PDF. In: IBM. Dostupné z: https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf. [cit. 2025-06-20].

[48] DENNY, Matthew J. a SPIRLING, Arthur. Text Preprocessing For Unsupervised Learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*. 2018, roč. 26.2, s. 168–189.

[49] VIJAYARANI, S.; ILAMATHI, M. a NITHYA, M. Preprocessing techniques for text mining. Online. *International Journal of Computer Science & Communication Networks*. 2015, roč. 5, č. 1, s. 7-16. ISSN 2249-5789. Dostupné z: https://www.researchgate.net/profile/Vijayarani-Mohan/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview/links/5e57a0f7299bf1bdb83e7505/Preprocessing-Techniques-for-Text-Mining-An-Overview.pdf. [cit. 2025-04-04].

[50] ANADAKUMAR, K. a PADMAVATHY, V. A survey on preprocessing in text mining. Online. *International Journal of Advanced Research in Computer Scienc*. 2013, roč. 4, č. 9, s. 79-91. Dostupné z: https://ijarcs.info/index.php/Ijarcs/article/download/1832/1820/3639. [cit. 2025-04-04].

[51] KARL, Andrew; WISNOWSKI, James a RUSHING, W. Heath. A practical guide to text mining with topic extraction. Online. *WIREs Computational Statistics*. 2015, roč. 7, č. 5, s. 326-340. ISSN 1939-5108. Dostupné z: https://doi.org/10.1002/wics.1361. [cit. 2025-04-05].

[52] SIINO, Marco; TINNIRELLO, Ilenia a LA CASCIA, Marco. Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Online. *Information Systems*. 2024, roč. 121, s. 102342. ISSN 03064379. Dostupné z: https://doi.org/10.1016/j.is.2023.102342. [cit. 2025-04-05].

[53] STROSSA, Petr. *Počítačové zpracování přirozeného jazyka*. Odborná kniha s vědeckou redakcí. Praha: Oeconomica, 2011. ISBN 978-80-245-1777-3.

[54] PATIL, Rajvardhan; BOIT, Sorio; GUDIVADA, Venkat a NANDIGAM, Jagadeesh. A survey of text representation and embedding techniques in NLP. Online. *IEEE Access*. 2023, roč. 11, s. 36120-36146. ISSN 2169-3536. Dostupné z: https://doi.org/10.1109/ACCESS.2023.3266377. [cit. 2025-04-25].

[55] MITCHELL, Melanie. *Artificial intelligence: a guide for thinking humans*. Pelican Book. London: A Penguin Book, 2020. ISBN 978-0-241-40483-6.

[56] ZHANG, Tao; CUI, Wenbo; LIU, Xiaoli; JIANG, Lei a LI, Jinling. Research on Topic evolution path recognition based on LDA2vec symmetry model. Online. *Symmetry*. 2023, roč. 15, č. 4. ISSN 2073-8994. Dostupné z: https://doi.org/10.3390/sym15040820. [cit. 2025-04-28].

[57] EGGER, Roman a YU, Joanne. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. Online. *Frontiers in Sociology*. 2022, roč. 7. ISSN 2297-7775. Dostupné z: https://doi.org/10.3389/fsoc.2022.886498. [cit. 2025-04-25].

[58] ANGELOV, Dimo. Top2Vec: Distributed representations of topics. Online. *ArXiv preprint arXiv:2008.09470*. 2020. Dostupné z: https://doi.org/10.48550/arXiv.2008.09470. [cit. 2025-04-25].

[59] GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Online. *ArXiv preprint arXiv:2203.05794*. Dostupné z: https://doi.org/10.48550/arXiv.2203.05794. [cit. 2025-04-14].

[60] BERKA, Petr. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9. Dostupné také z: http://krameriusndk.nkp.cz/search/handle/uuid:7e218600-062d-11e6-a611-005056827e51.

[61] Text Mining: Use of TF-IDF to examine the relevance of words to documents. Online. *International Journal of Computer Applications*. 2018, roč. 181, č.1, s. 25–29. ISSN 0975 – 8887. Dostupné z: https://www.researchgate.net/profile/Shahzad-Qaiser/publication/326425709_Text_Mining_Use_of_TF-IDF_to_Examine_the_Relevance_of_Words_to_Documents/links/5b4cd57fa6fdcc8dae245aa3/Text-Mining-Use-of-TF-IDF-to-Examine-the-Relevance-of-Words-to-Documents.pdf. [cit. 2025-04-16].

[62] MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg a DEAN, Jeffrey. Efficient estimation of word representations in vector space. Online. In: *ArXiv preprint arXiv:1301.3781*. 2013. Dostupné z: https://doi.org/10.48550/arXiv.1301.3781. [cit. 2025-05-31].

[63] DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton a TOUTANOVA, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. Online. *ArXiv:1810.04805*. 2019. Dostupné z: https://doi.org/10.48550/arXiv.1810.04805. [cit. 2025-05-31].

[64] VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion a kol. Attention is all you need. Online. *Advances in Neural Information Processing Systems*. 2017, roč. 30. Dostupné z: https://doi.org/10.48550/arXiv.1706.03762. [cit. 2025-06-06].

[65] KIM, Mujin; PARK, Youngjin a YOON, Janghyeok. Generating patent development maps for technology monitoring using semantic patent-topic analysis. *Computers & Industrial Engineering*. Online. 2016, roč. 98, s. 289-299. ISSN 03608352. Dostupné z: https://doi.org/10.1016/j.cie.2016.06.006. [cit. 2025-04-24].

[66] TIAN, Chen; ZHANG, Junyan; LIU, Dayong; WANG, Qing a LIN, Shen. Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model. Online. *Technology Analysis & Strategic Management*. 2024, roč. 36, č. 9, s. 2084-2099. ISSN 0953-7325. Dostupné z: https://doi.org/10.1080/09537325.2022.2130039. [cit. 2025-04-24].

[67] GHAFFARI, Mohsen; ALIAHMADI, Alireza; KHALKHALI, Abolfazl; ZAKERY, Amir; DAIM, Tugrul U. et al. Topic-based technology mapping using patent data analysis: A case study of vehicle tires. Online. *Technological Forecasting and Social Change*. 2023, roč. 193. ISSN 00401625. Dostupné z: https://doi.org/10.1016/j.techfore.2023.122576. [cit. 2025-04-24].

[68] SONG, Hyeonik; SELVA, Daniel a MCADAMS, Daniel A. Patent mining to understand functional evolution of engineered products. Online. *Volume 3A: 48th Design Automation Conference*. 2022, s. -. ISBN 978-0-7918-8622-9. Dostupné z: https://doi.org/10.1115/DETC2022-89405. [cit. 2025-04-25].

[69] PAZHOUHAN, Mohamadreza; KARIMI MAZRAESHAHI, Amin; JAHANBAKHT, Mohammad; REZANEJAD, Kourosh a ROHBAN, Mohammad Hossein. Wave and tidal

energy: A patent landscape study. Online. *Journal of Marine Science and Engineering*. 2024, roč. 12, č. 11. ISSN 2077-1312. Dostupné z: https://doi.org/10.3390/jmse12111967. [cit. 2025-04-25].

[70] LEE, Daniel D. a SEUNG, H. Sebastian. Algorithms for non-negative matrix factorization. Online. *Advances in Neural Information Processing Systems*. 2001, roč. 13, s. 556-562. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf. [cit. 2025-06-04].

[71] CHEN, Yong; ZHANG, Hui; LIU, Rui; YE, Zhiwen a LIN, Jianying. Experimental explorations on short text topic mining between LDA and NMF based Schemes. Online. *Knowledge-Based Systems*. 2019, roč. 163, s. 1-13. ISSN 09507051. Dostupné z: https://doi.org/10.1016/j.knosys.2018.08.011. [cit. 2025-06-04].

[72] BLEI, David M.; NG, Andrew Y. a JORDAN, Michael I. Latent dirichlet allocation. Online. *Journal of Machine Learning Research*. 2003, roč. 3, s. 993-1022. Dostupné z: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf. [cit. 2025-06-04].

[73] *Dirichlet process and Dirichlet process mixtures*. PDF. In: CMU School of Computer Science. C2025. Dostupné z: https://www.cs.cmu.edu/~epxing/Class/10708-16/note/10708_scribe_lecture18.pdf. [cit. 2025-06-20].

[74] GROOTENDORST. *BERTopic*. Online. C2024. Dostupné z: https://maartengr.github.io/BERTopic/index.html. [cit. 2025-06-02].

[75] *List of Members participating in CIPA Statistical Research*. Online. In: Camera & Imaging Products Association. C2025. Dostupné z: https://www.cipa.jp/stats/documents/e/list_e.pdf. [cit. 2025-06-01].

[76] *Production, Shipment of Digital Still Cameras: January-December in 2014*. Online. In: Camera & Imaging Products Association. C2025. Dostupné z: https://www.cipa.jp/stats/documents/e/d-2014_e.pdf. [cit. 2025-06-01].

[77] *Production,Shipment of Digital Still Camera: January-December in 2024*. Online. In: Camera & Imaging Products Association. C2025. Dostupné z: https://www.cipa.jp/stats/documents/e/d-2024_e.pdf. [cit. 2025-06-01].

[78] NUSKA, Petr. The DSLR revolution and its impact on documentary and ethnographic filmmaking. Online. *Visual Ethnography*. 2018, roč. 7, č. 2, s. 24–44. Dostupné z: https://doi.org/10.12835/ve2018.1-0111. [cit. 2025-06-01].

[79] *IPC Publication*. Online. World Intellectual Property Organization. 2025. Dostupné z: https://ipcpub.wipo.int/?notion=scheme&version=20250101&symbol=none&menulang=en&lang=en&viewmode=f&fipcpc=no&showdeleted=yes&indexes=no&headings=yes&notes=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart. [cit. 2025-06-01].

[80] *National Collections - Data Coverage*. Online. WIPO. World Intellectual Property Organization, 2025. Dostupné z: https://patentscope.wipo.int/search/en/help/data_coverage.jsf. [cit. 2025-06-02].

[81] *Re — Regular expression operations*. Online. PYTHON SOFTWARE FOUNDATION. Python. C2021-2025. Dostupné z: https://docs.python.org/3/library/re.html. [cit. 2025-06-03].

[82] VAN RIJN, Tomas a TIMMIS, James Kenneth. Patent landscape analysis—Contributing to the identification of technology trends and informing research and innovation funding policy. Online. *Microbial Biotechnology*. 2023, roč. 16, č. 4, s. 683-696. ISSN 1751-7915. Dostupné z: https://doi.org/10.1111/1751-7915.14201. [cit. 2025-06-18].

[83] *Pretrained Models*. Online. Sentence Transformers Documentation. C2025. Dostupné z: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. [cit. 2025-06-04].

[84] ROSNER, Frank; HINNEBURG, Alexander; RÖDER, Michael; NETTLING, Martin a BOTH, Andreas. Evaluating topic coherence measures. Online. *RXiv preprint arXiv:1403.6397*. 2014. Dostupné z: https://doi.org/10.48550/arXiv.1403.6397. [cit. 2025-06-23].

[85] *Camera Module*. Online. Samsung Electro-Mechanics. Dostupné z: https://www.samsungsem.com/global/product/module/camera-module.do. [cit. 2025-06-23].

[86] TESHOME, Mehari Beyene; PODRECCA, Matteo a ORZES, Guido. Technological trends in mountain logistics: A patent analysis. Online. *Research in Transportation Business & Management*. 2024, roč. 57. ISSN 22105395. Dostupné z: https://doi.org/10.1016/j.rtbm.2024.101202. [cit. 2025-06-20].

[87] PANDEY, Aryamaan. Comparative study of data visualization tools in BigData analysis for business intelligence. Online. *International Journal for Research in Applied Science*

and Engineering Technology. 2022, roč. 10, č. 6, s. 2591-2600. ISSN 23219653. Dostupné z: https://doi.org/10.22214/ijraset.2022.44400. [cit. 2025-06-04].

[88] PLOTLY. *Dash Python User Guide*. Online. C2025. Dostupné z: https://dash.plotly.com/. [cit. 2025-06-26].

[89] MCINNES, Leland. *DataMapPlot: Creating beautiful plot of data maps*. Online. C2023. Dostupné z: https://datamapplot.readthedocs.io/en/latest/. [cit. 2025-06-26].

SEZNAM PŘÍLOH

Příloha A Tabulka identifikovaných témat

Téma	Počet	Název	Reprezentace tématu
-1	31682	-1_camera_lens_image_device	['camera', 'lens', 'image', 'device', 'light', 'body', 'unit', 'arranged', 'second', 'module']
0	3485	0_support_rod_plate_supporting	['support', 'rod', 'plate', 'supporting', 'mounting', 'fixing', 'rotating', 'connecting', 'adjusting', 'sliding']
1	2151	1_module_camera module_housing_lens	['module', 'camera module', 'housing', 'lens', 'image sensor', 'sensor', 'circuit board', 'disposed', 'board', 'printed circuit']
2	1990	2_shake_camera shake_correction_shake correction	['shake', 'camera shake', 'correction', 'shake correction', 'blur', 'vibration', 'correcting', 'means', 'image', 'imaging']
3	1847	3_photographing_image_data_information	['photographing', 'image', 'data', 'information', 'means', 'user', 'image data', 'photographed', 'problem solved', 'solution']
4	1585	4_focus_focusing_af_value	['focus', 'focusing', 'af', 'value', 'focus detection', 'detection', 'position', 'evaluation', 'focus lens', 'subject']
5	1440	5_lamp_light_supplementing_light supplementing	['lamp', 'light', 'supplementing', 'light supplementing', 'light source', 'source', 'led', 'supplementing lamp', 'supplement', 'light supplement']
6	974	6_monitoring_monitoring device_fixedly_rod	['monitoring', 'monitoring device', 'fixedly', 'rod', 'monitoring camera', 'connected', 'plate', 'mounting', 'end', 'fixedly connected']
7	926	7_vehicle_vehicle camera_image_driver	['vehicle', 'vehicle camera', 'image', 'driver', 'camera', 'view', 'housing', 'rear', 'windshield', 'unit']
8	804	8_projection_projector_projected_screen	['projection', 'projector', 'projected', 'screen', 'image', 'projection image', 'image projection', 'projecting', 'image projected', 'projection surface']
9	799	9_panoramic_panorama_panoramic camera_panoramic image	['panoramic', 'panorama', 'panoramic camera', 'panoramic image', 'image', 'images', '360', 'panorama camera', 'camera', 'shooting']
10	767	10_waterproof_shell_sealing_cover	['waterproof', 'shell', 'sealing', 'cover', 'water', 'rain', 'ring', 'rainwater', 'groove', 'arranged']
11	692	11_interchangeable_interchangeable lens_camera body_lens	['interchangeable', 'interchangeable lens', 'camera body', 'lens', 'communication', 'information', 'body', 'means', 'unit', 'control']
12	667	12_lens_refractive_refractive power_power	['lens', 'refractive', 'refractive power', 'power', 'positive', 'negative', 'surface', 'fourth lens', 'convex', 'optical']
13	657	13_underwater_underwater camera_water_fish	['underwater', 'underwater camera', 'water', 'fish', 'camera', 'device', 'housing', 'waterproof', 'body', 'end']
14	641	14_strobe_stroboscope_stroboscopic_emission	['strobe', 'stroboscope', 'stroboscopic', 'emission', 'light emission', 'light', 'strobe light', 'emitting', 'light emitting', 'stroboscopic device']
15	601	15_filter_optical filter_optical_switching	['filter', 'optical filter', 'optical', 'switching', 'infrared', 'ring', 'filter switching', 'filtering', 'filters', 'light filter']
16	580	16_stereoscopic_stereo_stereoscopic image_left	['stereoscopic', 'stereo', 'stereoscopic image', 'left', 'image', 'right', 'stereo camera', 'eye', 'images', 'cameras']
17	530	17_ring_lens_camera lens_connecting	['ring', 'lens', 'camera lens', 'connecting', 'cone', 'lens cone', 'utility model', 'utility', 'model', 'fixing']
18	506	18_lens driving_driving_driving device_coil	['lens driving', 'driving', 'driving device', 'coil', 'magnet', 'bobbin', 'lens', 'carrier', 'disposed', 'base']
19	484	19_laser_structured light_light_structured	['laser', 'structured light', 'light', 'structured', 'depth', 'depth camera', 'projection', 'beam', 'source', 'light source']

20	444	20_board_substrate_member_connector	['board', 'substrate', 'member', 'connector', 'accessory', 'contact', 'flexible', 'cover', 'solution', 'problem solved']
21	441	21_actuator_camera actuator_disposed_mover	['actuator', 'camera actuator', 'disposed', 'mover', 'direction', 'housing', 'carrier', 'second', 'driving', 'magnet']
22	432	22_mobile_terminal_mobile terminal_phone	['mobile', 'terminal', 'mobile terminal', 'phone', 'portable', 'mobile phone', 'photographing', 'telephone', 'mobile device', 'smartphone']
23	421	23_mirror_finder_flux_luminous flux	['mirror', 'finder', 'flux', 'luminous flux', 'reflex', 'lens reflex', 'single lens', 'luminous', 'optical', 'reflex camera']
24	413	24_focusing_automatic focusing_automatic_focusing device	['focusing', 'automatic focusing', 'automatic', 'focusing device', 'motor', 'gear', 'assembly', 'focusing mechanism', 'lens', 'driving']
25	413	25_blade_shutter_member_blades	['blade', 'shutter', 'member', 'blades', 'shutter blade', 'plane shutter', 'rotor', 'driving', 'base plate', 'focal plane']
26	410	26_heat_dissipation_heat dissipation_heat conduction	['heat', 'dissipation', 'heat dissipation', 'heat conduction', 'conduction', 'fan', 'shell', 'cooling', 'air', 'plate']
27	380	27_barrel_lens barrel_mount_lens	['barrel', 'lens barrel', 'mount', 'lens', 'member', 'optical axis', 'cam', 'optical', 'axis', 'problem solved']
28	377	28_projecting apparatus_projector_projection_projecting	['projecting apparatus', 'projector', 'projection', 'projecting', 'apparatus', 'camera lens', 'utility', 'model', 'utility model', 'projector body']
29	366	29_image_information_unit_processing	['image', 'information', 'unit', 'processing', 'apparatus', 'capture', 'capturing', 'images', 'captured', 'information processing']
30	359	30_anti shake_anti_shake_optical anti	['anti shake', 'anti', 'shake', 'optical anti', 'driving', 'assembly', 'movable', 'coil', 'shake structure', 'magnet']
31	344	31_test_camera module_testing_module	['test', 'camera module', 'testing', 'module', 'inspection', 'socket', 'unit', 'camera', 'present invention', 'invention']
32	335	32_emission_flash_light emission_light	['emission', 'flash', 'light emission', 'light', 'flash light', 'emitting', 'light emitting', 'quantity', 'electronic flash', 'light quantity']
33	328	33_3d_dimensional_dimensional image_3d image	['3d', 'dimensional', 'dimensional image', '3d image', 'image', 'shooting', 'images', 'cameras', '3d camera', 'object']
34	321	34_battery_power_voltage_supply	['battery', 'power', 'voltage', 'supply', 'residual', 'consumption', 'power supply', 'mode', 'capacity', 'power consumption']
35	317	35_drone_aerial_aircraft_flight	['drone', 'aerial', 'aircraft', 'flight', 'unmanned', 'ground', 'present invention', 'present', 'photographing', 'image']
36	312	36_ois_stabilization_image stabilization_optical image	['ois', 'stabilization', 'image stabilization', 'optical image', 'optical', 'module', 'direction', 'image stabilizer', 'sensor', 'image']
37	309	37_film_cartridge_spool_chamber	['film', 'cartridge', 'spool', 'chamber', 'film cartridge', 'cartridge chamber', 'feeding', 'film cassette', 'rewinding', 'cassette']
38	273	38_aperture_iris_variable aperture_blades	['aperture', 'iris', 'variable aperture', 'blades', 'variable', 'blade', 'driving', 'hole', 'light', 'assembly']
39	261	39_digital_digital camera_image_unit	['digital', 'digital camera', 'image', 'unit', 'digital image', 'display', 'data', 'sensor', 'image sensor', 'signal']
40	257	40_teaching_teacher_students_education	['teaching', 'teacher', 'students', 'education', 'rod', 'learning', 'table', 'connected', 'fixedly', 'sliding']
41	252	41_lens_lens assembly_optical lens_second lens	['lens', 'lens assembly', 'optical lens', 'second lens', 'assembly', 'optical', 'lens barrel', 'barrel', 'second', 'surface']
42	248	42_surveillance_surveillance camera_housing_cover	['surveillance', 'surveillance camera', 'housing', 'cover', 'case', 'portion', 'housing portion', 'camera', 'enclosure', 'includes']

43	243	43_zoom_zooming_magnification_zoom magnification	['zoom', 'zooming', 'magnification', 'zoom magnification', 'zoom lens', 'electronic zoom', 'zoom operation', 'operation', 'image', 'means']
44	234	44_flash_camera flash_flash unit_light	['flash', 'camera flash', 'flash unit', 'light', 'flash module', 'emitting', 'unit', 'flash light', 'flash device', 'light emitting']
45	225	45_electronic device_electronic_housing_module	['electronic device', 'electronic', 'housing', 'module', 'portion', 'camera module', 'second', 'disposed', 'portable electronic', 'includes']
46	217	46_light_illumination_source_light source	['light', 'illumination', 'source', 'light source', 'lighting', 'emitting', 'light emitting', 'sources', 'color', 'illuminating']
47	214	47_pan_tilt_pan tilt_tilt camera	['pan', 'tilt', 'pan tilt', 'tilt camera', 'motor', 'shaft', 'rotating', 'connected', 'head', 'connecting']
48	214	48_dust_foreign_dust proof_removing	['dust', 'foreign', 'dust proof', 'removing', 'element', 'imaging', 'member', 'cleaning', 'imaging element', 'filter']
49	210	49_phone_mobile phone_mobile_cell phone	['phone', 'mobile phone', 'mobile', 'cell phone', 'cell', 'camera lens', 'lens', 'phone camera', 'clamp', 'phone lens']
50	206	50_stabilizer_motor_camera stabilizer_stabilizing	['stabilizer', 'motor', 'camera stabilizer', 'stabilizing', 'arm', 'stabilization', 'balance', 'damping', 'platform', 'handheld']
51	203	51_protection_lens protection_camera lens_protective	['protection', 'lens protection', 'camera lens', 'protective', 'lens', 'cover', 'sleeve', 'protective cover', 'ring', 'shell']
52	201	52_photometric_luminance_photometry_value	['photometric', 'luminance', 'photometry', 'value', 'exposure', 'brightness', 'area', 'photometric value', 'color', 'means']
53	195	53_sound_voice_microphone_noise	['sound', 'voice', 'microphone', 'noise', 'speaker', 'audio', 'recording', 'speech', 'data', 'shutter']
54	193	54_barrier_lens barrier_closing_position	['barrier', 'lens barrier', 'closing', 'position', 'opening', 'barrier member', 'lens', 'member', 'opening closing', 'open']
55	184	55_fog_anti fog_anti_heating	['fog', 'anti fog', 'anti', 'heating', 'temperature', 'fogging', 'air', 'shell', 'glass', 'anti fogging']
56	177	56_battery_lid_card_memory card	['battery', 'lid', 'card', 'memory card', 'lid body', 'contact', 'memory', 'opening', 'battery lid', 'closing']
57	175	57_tripod_leg_tripod head_legs	['tripod', 'leg', 'tripod head', 'legs', 'head', 'camera tripod', 'supporting', 'rod', 'foot', 'connecting']
58	175	58_photosensitive_camera module_module_reflecting	['photosensitive', 'camera module', 'module', 'reflecting', 'assembly', 'module electronic', 'electronic', 'lens assembly', 'lens', 'electronic equipment']
59	173	59_infrared_infrared camera_night_night vision	['infrared', 'infrared camera', 'night', 'night vision', 'vision', 'shell', 'connected', 'infrared lamp', 'cover', 'arranged']
60	171	60_explosion_explosion proof_proof_proof camera	['explosion', 'explosion proof', 'proof', 'proof camera', 'proof shell', 'anti explosion', 'shell', 'cover', 'heat', 'arranged']
61	170	61_pan_tilt_universal head_universal	['pan', 'tilt', 'universal head', 'universal', 'rotation', 'head', 'pan tilt', 'panning', 'motor', 'worm']
62	168	62_heat_radiation_heat radiation_imaging	['heat', 'radiation', 'heat radiation', 'imaging', 'heat generated', 'imaging element', 'element', 'temperature', 'heat transfer', 'cooling']
63	165	63_shutter_blade_shutter blade_shutter device	['shutter', 'blade', 'shutter blade', 'shutter device', 'camera shutter', 'magnet', 'electromagnet', 'drive', 'shutter mechanism', 'coil']
64	165	64_voice coil_coil_coil motor_voice	['voice coil', 'coil', 'coil motor', 'voice', 'motor', 'carrier', 'magnetic', 'base', 'magnet', 'elastic']
65	164	65_pipeline_pipe_inspection_cable	['pipeline', 'pipe', 'inspection', 'cable', 'pipeline detection', 'detection', 'robot', 'water', 'pipe inspection', 'push']
66	163	66_dust_dustproof_cleaning_air	['dust', 'dustproof', 'cleaning', 'air', 'dust removal', 'cover', 'monitoring', 'removal', 'fixedly', 'plate']

67	161	67_zoom_zoom lens_lens group_lens	['zoom', 'zoom lens', 'lens group', 'lens', 'group', 'zooming', 'optical zoom', 'optical', 'second lens', 'driving']
68	158	68_display_display panel_screen_area	['display', 'display panel', 'screen', 'area', 'panel', 'display screen', 'display area', 'display device', 'layer', 'backlight']
69	157	69_license_license plate_parking_plate recognition	['license', 'license plate', 'parking', 'plate recognition', 'recognition', 'plate', 'parking lot', 'lot', 'recognition camera', 'supplementing']
70	154	70_cloud_cloud platform_platform_cloud deck	['cloud', 'cloud platform', 'platform', 'cloud deck', 'deck', 'axle', 'utility', 'utility model', 'motor', 'model']
71	149	71_gimbal_gimbal camera_camera gimbal_axis	['gimbal', 'gimbal camera', 'camera gimbal', 'axis', 'roll', 'control', 'vehicle', 'configured', 'axis motor', 'gimbal device']
72	147	72_lens group_group_zoom lens_positive	['lens group', 'group', 'zoom lens', 'positive', 'zoom', 'lens', 'group having', 'power', 'having positive', 'refractive']
73	146	73_calibration_camera calibration_calibrating_parameter	['calibration', 'camera calibration', 'calibrating', 'parameter', 'calibration device', 'reference', 'method', 'calibration method', 'points', 'chart']
74	141	74_distance_measuring_distance measuring_object	['distance', 'measuring', 'distance measuring', 'object', 'subject', 'measurement', 'subject image', 'distance measurement', 'image', 'correlation']
75	138	75_liquid_liquid lens_crystal_electrode	['liquid', 'liquid lens', 'crystal', 'electrode', 'liquid crystal', 'lens', 'cavity', 'disposed', 'conductive liquid', 'voltage']
76	133	76_liquid crystal_crystal_liquid_crystal display	['liquid crystal', 'crystal', 'liquid', 'crystal display', 'display', 'crystal panel', 'panel', 'crystal monitor', 'display device', 'finder']
77	133	77_face recognition_recognition_face_recognition camera	['face recognition', 'recognition', 'face', 'recognition camera', 'recognition device', 'arranged', 'shell', 'machine', 'fixedly', 'connected']
78	130	78_waterproof_case_waterproof case_provide waterproof	['waterproof', 'case', 'waterproof case', 'provide waterproof', 'lid', 'solution waterproof', 'member', 'waterproof camera', 'case body', 'body']
79	127	79_dome_dome cover_cover_dome camera	['dome', 'dome cover', 'cover', 'dome camera', 'dome type', 'spherical', 'dome shaped', 'type', 'camera', 'camera unit']
80	123	80_mirror_image_second_array	['mirror', 'image', 'second', 'array', 'mirrors', 'view', 'images', 'light', 'imaging', 'beam']
81	123	81_ray_framing_pulse_framing camera	['ray', 'framing', 'pulse', 'framing camera', 'ultrafast', 'time', 'terahertz', 'resolution', 'high', 'image intensifier']
82	119	82_button_release_release button_switch	['button', 'release', 'release button', 'switch', 'operation', 'operation button', 'operating', 'hand', 'solution', 'problem solved']
83	116	83_polarization_polarized_polarizing_polaroid	['polarization', 'polarized', 'polarizing', 'polaroid', 'filter', 'polarizer', 'polarized light', 'polarizing filter', 'polarization filter', 'light']
84	115	84_protection_protective_camera protection_cover	['protection', 'protective', 'camera protection', 'cover', 'protection frame', 'plate', 'shell', 'protection device', 'body', 'buffer']
85	114	85_optical element_element_element driving_driving device	['optical element', 'element', 'element driving', 'driving device', 'optical', 'driving', 'movable', 'device', 'optical member', 'drive']
86	113	86_prism_light_reflection_path	['prism', 'light', 'reflection', 'path', 'surface', 'light path', 'incident', 'optical', 'splitting', 'reflecting']
87	110	87_hood_lens hood_light shielding_shielding	['hood', 'lens hood', 'light shielding', 'shielding', 'lens', 'hood body', 'shielding hood', 'light', 'light shield', 'barrel']
88	109	88_plant_crop_root_cultivation	['plant', 'crop', 'root', 'cultivation', 'tobacco', 'growth', 'leaf', 'acquisition', 'plants', 'greenhouse']

89	106	89_acquisition_acquisition device_data acquisition_box	['acquisition', 'acquisition device', 'data acquisition', 'box', 'rod', 'information acquisition', 'plate', 'fixedly', 'data', 'arranged']
90	105	90_piezoelectric_piezoelectric element_piezoelectric actuator_driving	['piezoelectric', 'piezoelectric element', 'piezoelectric actuator', 'driving', 'piezoelectric driving', 'piezo', 'actuator', 'element', 'lens', 'second piezoelectric']
91	103	91_dual_dual camera_camera module_second	['dual', 'dual camera', 'camera module', 'second', 'module', 'second camera', 'relates dual', 'magnet', 'disposed', 'double']
92	103	92_viewfinder_electronic viewfinder_camera viewfinder_eyepiece	['viewfinder', 'electronic viewfinder', 'camera viewfinder', 'eyepiece', 'optical viewfinder', 'optical', 'viewfinder assembly', 'electronic', 'display', 'view']
93	101	93_focus_distance_contrast_focus lens	['focus', 'distance', 'contrast', 'focus lens', 'focusing', 'focal', 'focus detection', 'contrast data', 'point', 'image']
94	101	94_opfe_tele_folded_lens elements	['opfe', 'tele', 'folded', 'lens elements', 'folding', 'optical path', 'folding element', 'lens element', 'path', 'path folding']
95	101	95_water_lens 13_droplet_lens	['water', 'lens 13', 'droplet', 'lens', 'hydrophilic', '13', 'water droplet', 'droplets', 'lens space', 'lens 14']
96	97	96_exposure_sensitivity_value_exposure time	['exposure', 'sensitivity', 'value', 'exposure time', 'exposure control', 'time', 'shutter speed', 'shutter', 'exposure value', 'control']
97	97	97_aerial_unmanned_unmanned aerial_aerial vehicle	['aerial', 'unmanned', 'unmanned aerial', 'aerial vehicle', 'vehicle', 'vehicle body', 'protective', 'mounting', 'protection', 'shell']
98	96	98_lens drive_coil_coils_drive device	['lens drive', 'coil', 'coils', 'drive device', 'direction', 'magnet', 'drive', 'lens support', 'provide lens', 'optical axis']
99	92	99_range finding_finding_range_range finder	['range finding', 'finding', 'range', 'range finder', 'finder', 'distance', 'ranging', 'finding device', 'means', 'light receiving']
100	91	100_infrared_infrared light_visible_visible light	['infrared', 'infrared light', 'visible', 'visible light', 'near infrared', 'light', 'near', 'wavelength', 'infrared camera', 'infrared ray']
101	88	101_fighting_monitoring_smoke_flame	['fighting', 'monitoring', 'smoke', 'flame', 'fireproof', 'water', 'box', 'alarm', 'forest', 'heat insulation']
102	86	102_heat_dissipation_heat dissipation_photosensitive	['heat', 'dissipation', 'heat dissipation', 'photosensitive', 'chip', 'circuit board', 'photosensitive chip', 'module', 'board', 'circuit']
103	86	103_omnidirectional_omnidirectional image_omnidirectional camera_mirror	['omnidirectional', 'omnidirectional image', 'omnidirectional camera', 'mirror', 'omni', 'omniazimuth', 'image', 'directional', 'cameras', 'omnidirectional imaging']
104	84	104_terminal_accessory_clock_clock signal	['terminal', 'accessory', 'clock', 'clock signal', 'terminals', 'data signal', 'plurality terminals', 'signal', 'terminal second', 'second clock']
105	83	105_solar_solar panel_panel_photovoltaic	['solar', 'solar panel', 'panel', 'photovoltaic', 'energy', 'solar cell', 'solar energy', 'cell panel', 'cell', 'rod']
106	82	106_heating_heating device_temperature_heating element	['heating', 'heating device', 'temperature', 'heating element', 'heating sheet', 'lens', 'vehicle mounted', 'module', 'heat', 'vehicle']
107	81	107_light_illuminating_illumination_illuminator	['light', 'illuminating', 'illumination', 'illuminator', 'light source', 'emitting', 'source', 'irradiation', 'lighting', 'led']
108	80	108_pickup_image pickup_pickup apparatus_image	['pickup', 'image pickup', 'pickup apparatus', 'image', 'apparatus', 'pickup device', 'unit', 'ranging', 'member', 'operation']
109	79	109_substrate_moving_disposed_sensor	['substrate', 'moving', 'disposed', 'sensor', 'unit', 'camera device', 'second moving', 'moving unit', 'image sensor', 'connecting substrate']

110	72	110_holographic_holographic projection_projection_hologram	['holographic', 'holographic projection', 'projection', 'hologram', 'holographic imaging', 'screen', 'display', 'holographic image', 'interactive', 'holographically projected']
111	68	111_periscopic_periscopic camera_prism_periscope	['periscopic', 'periscopic camera', 'prism', 'periscope', 'assembly', 'module', 'camera module', 'light', 'reflection', 'photosensitive']
112	65	112_display_monitor_lcd_display surface	['display', 'monitor', 'lcd', 'display surface', 'display unit', 'body', 'camera body', 'digital camera', 'digital', 'display device']
113	65	113_cleaning_wiper_nozzle_fluid	['cleaning', 'wiper', 'nozzle', 'fluid', 'washing', 'device cleaning', 'cleaning device', 'discharge', 'discharge port', 'cleaning lens']
114	64	114_adapter_microscope_mount section_mount	['adapter', 'microscope', 'mount section', 'mount', 'adaptor', 'camera adapter', 'eyepiece', 'eyepiece lens', 'second mount', 'attached']
115	61	115_sma_sma actuator_actuator_sma wire	['sma', 'sma actuator', 'actuator', 'sma wire', 'wire', 'support structure', 'movable element', 'wires', 'actuating', 'elastic']