Univerzita Pardubice

Fakulta ekonomicko-správní

Analýza vědeckých textů pro téma data-driven business

2025 Bc. Jan Štuchal

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

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Jan Štuchal**
Osobní číslo:	**E23055**
Studijní program:	**N0613A140041 Aplikovaná informatika – Data Science pro business**
Téma práce:	**Analýza vědeckých textů pro téma data-driven business**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování

Cílem práce je analyzovat vědecké texty pro téma data-driven business s využitím textové analytiky a souvisejících metod.
Osnova:

- Charakteristika data-driven business
- Systematický postup pro analýzu vědeckých textů
- Sběr, zpracování a analýza získaných dat textovou analytikou pomocí vhodného nástroje
- Implementace a vytvoření dashboardu pomocí PowerBI

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

Seznam doporučené literatury:

MCKINNEY, Wes**.** *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 3rd ed. Sebastopol: O'Reilly Media, 2022. ISBN 978-1-09-810403-0.
MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. ISBN 978-0-521-86571-5.
GARLA, Satish. *Text Mining and Analysis*. Cary: SAS Institute, 2013. ISBN 978-1-61290-551-8.
KOTOROV, Rado. *Data-Driven Business Models for the Digital Economy*. New York: Business Expert Press, 2020. ISBN 978-1-95152-780-1.
PROVOST, Foster; FAWCETT, Tom. *Data Science for Business*. Sebastopol: O'Reilly Media, 2013. ISBN 978-1-4493-6132-7.

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 vědeckých textů pro téma data-driven business 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 pozdnějších 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 30.4. 2025

Jan Štuchal v. r

PODĚKOVÁNÍ

V první řadě bych rád poděkoval mému vedoucímu práce prof. Petrovi Hájkovi, Ing. za jeho přístup a cenné rady při zpracování. Dále bych rád poděkoval mé rodině a blízkým za jejich neustálou podporu při studiu.

ANOTACE

Cílem práce je analyzovat vědecké texty pro téma data-driven business s využitím textové analytiky a souvisejících metod. Práce začíná teoretický základem data-driven business. Dále je popsán systematický postup vyhledávání a výběru literatury podle metodiky PRISMA a hlavní přístupy pro textovou a bibliometrickou analýzu. Další část se zabývá praktickém zpracováním dat v Pythonu a aplikaci vybraných metod. Nakonec jsou výsledky shrnuty prostřednictvím dashboardů pro vizualizaci.

KLÍČOVÁ SLOVA

data-driven business, textová analýza, python, bibliometrická analýza

TITLE

Analysis of scientific texts for the topic of data-driven business

ANNOTATION

The aim of this thesis is to analyze scientific texts for the topic of data-driven business using text analytics and related methods. The thesis starts with explaining the theoretical basis of data-driven business. In the next chapter, a systematic procedure for literature search and selection according to the PRISMA methodology and the main approaches for textual and bibliometric analysis were described. The third chapter deals with practical data processing in Python and the application of the selected methods. The final chapter summarizes the results through dashboards for visualization.

KEYWORDS

data-driven business, text analysis, python, bibliometric analysis

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM POUŽITÝCH ZKRATEK

CSV – Comma Separated Value

LDA – Latent Dirichlet Allocation

NMF – Non-Negative Matrix Factorization

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

TF-IDF – Term Frequency-Inverse Document Frequency

XLSX – Excel Open XML Spreadsheet

ÚVOD

V současné době analýza dat je jednou z klíčových rolí v rozhodovacích procesech a strategickém plánování podniků. Koncept podnikání řízeného daty neboli data-driven business, je stále častěji implementován ve firmách po celém světě jako nezbytný nástroj pro udržení konkurenceschopnosti, zvyšování efektivity a inovací. Díky analýze dat je umožněno podnikům identifikovat nové příležitosti, optimalizovat existující procesy a reagovat na dynamické změny na trhu. Tento přístup je rovněž podporován rychlým vývojem technologií, jako jsou business intelligence, big data či strojové učení. Přestože se tento koncept dostává do popředí, v dostupné literatuře zatím chybí souhrnné analýzy, které by přehledně zachycovaly směry tohoto konceptu. To poukazuje na potřebu systematického přehledu, který by přispěl k lepšímu porozumění současného stavu a vývoje v této oblasti.

Cílem této diplomové práce je provést analýzu vědeckých textů zaměřených na téma data-driven business.

Tato diplomová práce je strukturována do čtyřech částí. V první části je prezentován teoretický přehled, který se zaměřuje na charakteristiku pojmu a další spojené aspekty. Druhá kapitola se zaměřuje na popis systematického sběru vědeckých textů, jednotlivých typů, kde pozornost je věnována metodice PRISMA. Poté navazuje teoretický základ pro textovou a bibliometrickou analýzu, ve kterých jsou definovány klíčové metody a metriky, ze kterých se vychází. Třetí část se věnuje nejprve popisu databází a následně sběru dat, dále formulaci výzkumných otázek a následné aplikaci metod textové a bibliometrické analýzy pomocí programovacího jazyka Python. Ve čtvrté části je popsána implementace a tvorba dashboardů v programu Power BI, díky kterému lze vizualizovat získané výsledky a poskytnou tak přehledný pohled na data.

1 CHARAKTERISTIKA DATA-DRIVEN BUSINESS

Tato kapitola se věnuje charakteristice konceptu data-driven business, a to se zaměřením na definici, související aspekty a přístupy.

1.1 Definice data-driven business

Tento pojem je definován jako moderní přístup k podnikání, který klade důraz na data jako klíčový prvek všech rozhodovacích procesů a strategií [1]. Tento koncept vychází z předpokladu, že efektivní sběr, analýza a interpretace dat umožňují organizacím inovovat byznys model, učinit správné rozhodnutí, optimalizovat procesy a získat konkurenční výhodu [2]. Tradiční podniky obvykle generují zisk prodejem produktů nebo služeb. Společnosti jako Google, Facebook nebo Airbnb nacházejí hodnotu v analýze uživatelského chování a efektivním propojení nabídky s poptávkou, a to vše bez potřeby vlastnit fyzická aktiva.

1.2 Modely data-driven business

Data-driven business modely představují inovativní přístupy v rámci podnikání, ve kterých hrají data klíčovou roli při vytváření hodnoty [3]. Společnosti, které tyto modely aplikují se zaměřují na shromažďování, analýzu a interpretaci dat s cílem optimalizovat své služby, personalizovat nabídky a generovat nové zdroje příjmů [4]. Význam těchto modelů se zvyšuje v rámci digitalizace, a to s rozvojem například big data, umělé inteligence a pokročilé analytiky. Tyto technologie umožňují organizacím nejen reagovat na aktuální trendy, ale také předpovídat budoucí vývoj trhu a chování zákazníků.

Mezi jednotlivé prvky data-driven business modelů patří [3] [4]:

- **Data jako zdroj** – organizace využívají data k tomu, aby získali konkurenční výhodu;
- **Monetizace dat** – organizace obchodují, sdílejí nebo využívají jiné způsoby k vytváření příjmů;
- **Pokročilá analytika a umělá inteligence** – zahrnuje umělou inteligenci a analytické modely, které umožňují získávat hodnotu z dat.

1.3 Definice dat

Data lze definovat jako soubor nashromážděných informací, které mohou být v podobě čísel, slov a dalších forem. Tato data však nemusí mít žádný konkrétní význam, dokud nejsou dále zpracována, analyzována a interpretována. Prostřednictvím správného zpracování a analýzy daných dat je umožněno organizacím transformovat jednotlivé informace na cenné poznatky, které napomáhají k lepšímu rozhodování a přispívají k vylepšení obchodních výsledků [5] [6].

1.4 Typy dat

Data se dělí do dvou základních kategorií, a to kvantitativní a kvalitativní. Každý z těchto typů se vyznačuje odlišnými vlastnostmi a využitím.

Kvantitativní data

Tato data jsou číselného charakteru, což znamená, že je lze snadno měřit a zpracovávat pomocí statistických a matematických metod [6] [7]. Kvantitativní data se následně dělí na diskrétní data jako je například počet prodaných produktů, a kontinuální data, která zahrnují například výši příjmu. Obvykle jsou tato data uspořádána ve strukturované podobě, což znamená, že se systematicky organizují v tabulkách nebo databázích.

Kvalitativní data

Tato data se vyznačují tím, že nejsou číselného charakteru, ale slouží k tomu, aby zachytila vlastnosti, zkušenosti nebo subjektivní názory. Kvalitativní data poskytují hlubší kontext a význam na rozdíl od kvantitativních dat [6]. Tato data lze získat například prostřednictvím rozhovorů, otevřených otázek v dotaznících, zpětné vazby od zákazníků nebo komentářů na sociálních sítích [8].

1.5 Typy dat podle struktury

Data lze rozdělit podle jejich struktury. Každý z těchto typů vykazuje specifické vlastnosti a nachází různorodé uplatnění.

Strukturovaná data

Data mají jasně definovaný formát a jsou systematicky uspořádána v tabulkách nebo databázích, kde jednotlivé hodnoty jsou rozděleny do předem stanovených kategorií [6]. Díky této organizaci lze informace snadno vyhledat a spravovat. Tento typ dat lze efektivně zpracovávat pomocí databázových dotazů pomocí jazyka SQL, což značně usnadňuje jejich vyhledávání, třídění, a analýzu [9].

Polostrukturovaná data

Data se skládají z kombinace strukturovaných a nestrukturovaných dat [6]. Na rozdíl od klasických databázových záznamů nemají pevně stanovenou tabulkovou strukturu, ale jsou organizována pomocí značek, štítků nebo metadat, což usnadňuje jejich zpracování a vyhledávání [9].

Nestrukturovaná data

Data nemají jasně vymezený formát nebo pevnou strukturu, což ztěžuje jejich uložení do tabulek nebo relačních databází [6]. Tato data zahrnují různé typy informací, jako jsou textové dokumenty, obrázky nebo multimediální soubory, a pro jejich zpracování je obvykle potřeba využít pokročilé analytické metody [9].

Metadata

Představují doplňkové informace o datech, které usnadňují jejich organizaci, vyhledávání a správu [10]. Zahrnují různé údaje, jako je název souboru, autor, datum vytvoření nebo formát. Metadata hrají klíčovou roli v efektivním fungování například informačních systémů, databází a digitálních archivů [11].

Big data

Jsou označovány jako velká a rychle rostoucí datové soubory, které pocházejí z různých digitálních zdrojů, jako jsou sociální sítě, senzory, online transakce nebo multimediální obsah [12]. Díky pokroku v technologiích je možné tato data snad uchovávat, sdílet a analyzovat, což otevírá nové možnosti pro jejich využití [13].

1.6 Vyhledávání informací z dat

Při práci s velkým množstvím dat, je důležité porozumění informačního vyhledávání. Tento proces usnadňuje efektivní vyhledávání relevantních informací v nestrukturovaných textech a zahrnuje několik navazujících kroků [13] [15] [16].

Celý proces vyhledávání lze rozdělit do základních kroků [13] [14] [15]:

- **Indexace** – dokumenty jsou připraveny a strukturovány pro efektivní vyhledávání;
- **Zpracování** – dotaz je upraven a přizpůsoben pro porovnání s dokumenty;
- **Porovnání** – hodnotí se, jak dokumenty odpovídají zadanému dotazu nebo tématu;
- **Řazení** – dokumenty jsou uspořádány podle míry jejich relevance;
- **Vyhodnocení** – hodnotí se kvalita výsledků.

1.7 Datová analýza

Tato oblast zahrnuje systematický proces, jehož cílem je transformovat data na cenné poznatky, které přispívají k lepšímu rozhodování. Prvním krokem je jasně vymezit klíčové otázky a cíle naší analýzy. Následně se shromažďují relevantní interní a externí datové zdroje, které se potom podrobí důkladnému čištění a předzpracování. Během této fáze se odstraňují chyby, duplicitní

záznamy, chybějící hodnoty, které by mohly ovlivnit výsledky [17]. Teprve poté je možné provádět analýzu dat, jejímž výstupem jsou přehledy, modely nebo vizualizace, které pomáhají odhalit skryté vzorce a souvislosti [18].

Typy datové analýzy

Datovou analýzu lze rozdělit do několika kategorií, a to na základě stanovených cílů a použitých metod [17] [19]:

- **Textová analýza** – zaměřuje se na zpracování volných textů, přičemž z nich dokáže extrahovat důležitá témata, klíčová slova a sentiment;
- **Deskriptivní analýza** – zaměřuje na shrnutí událostí prostřednictvím souhrnných statistik a vizuálních grafů;
- **Diagnostická analýza** – zaměřuje se na vysvětlení, proč k určitým jevům došlo, a zkoumá příčiny a vzorce;
- **Inferenční analýza** – zaměřuje se na hypotézy na základě vzorku a odhaduje vlastnosti širšího souboru dat;
- **Prediktivní analýza** – zaměřuje se na pravděpodobné budoucí výsledky s využitím strojového učení nebo regresních metod;
- **Preskriptivní analýza** – zaměřuje se na doporučení konkrétních kroků a optimalizační strategie, které pomáhají dosáhnout stanovených cílů.

Rámec datové analýzy

Jedná se o rámec, který zdůrazňuje propojení cílů s jednotlivými kroky a zaručuje, že každý krok přináší hodnotu [20]. Jednotlivé kroky jsou pravidelně přezkoumány a opatřeny tak, aby eliminovaly zbytečné činnosti a maximalizoval se přínos celého procesu [21].

Rámec datové analýzy se skládá z několika kroků [20] [21]:

- **Porozumění problému** – jasné stanovení cíle a klíčových otázek, na jejichž základě se bude odvíjet průběh celého procesu;
- **Porozumění datům** – zahrnuje podrobný průzkum dostupných zdrojů, kontrolu kvality dat, identifikace anomálií a posouzení, zda jsou data vhodná pro daný úkol;
- **Příprava dat** – zahrnuje jejich čištění, uspořádání, transformaci proměnných a vytvoření nových ukazatelů pro podporu modelování;

- **Modelování** – zahrnuje výběr odpovídajícího algoritmu podle charakteru úlohy, trénink modelu, ladění jednotlivých parametrů a porovnání výkonu různých modelů s cílem vybrat to nejefektivnější řešení;
- **Hodnocení** – zahrnuje posouzení výkonnosti prostřednictvím vybraných metrik a vyhodnocení přínosu v porovnání s dosavadní praktikami nebo základním řešením;
- **Nasazení** – implementace vybraného řešení do produkčního prostředí, nastavení monitoringu výkonu a vypracování plánu pro pravidelné aktualizace, které zajistí, že model zůstane relevantní.

1.8 Nástroje pro práci s daty

Pro práci s daty a následné analýzy závisí na využití vhodných softwarových nástrojů a doplňků, které usnadňují práci v různých fázích zpracování dat, a to od jejich získávání a přípravy až po analýzu, vizualizaci a interpretaci výsledků. V současnosti existuje široká škála nástrojů, které se liší nejen svou funkcionalitou, ale také uživatelskou přívětivostí i schopností se přizpůsobit různým metodám práce s daty [22] [23].

Programovací jazyky

Mezi nejpoužívanější programovací jazyky se řadí Python a R, které hrají klíčovou roli především práci s daty. Python je oblíbený pro svou jednoduchou syntaxi, širokou komunitu uživatelů a širokou nabídku knihoven, jako jsou „*pandas*“, „*NumPy*“, „*matplotlib*“, „*scikit-learn*“ nebo „*TensorFlow*“. Tyto nástroje pokrývají široké spektrum od základních datových operací až po pokročile techniky strojového učení a vizualizaci dat [22] [23]. Na druhé straně R se zaměřuje především na statistické výpočty a datovou vizualizaci, přičemž nabízí rozsáhlé množství balíčků pro analýzu a modelování, které usnadňují práci s daty.

Vizualizace a reporting

Prezentace výsledků analýzy představuje nezbytný prvek celého procesu [23]. K tomu se využívají různé nástroje, jako jsou Tableau, Power BI a tradiční Excel. Tableau je populární pro svou schopnost vytvářet interaktivní dashboardy díky svému rozhraní, které usnadňuje práci. Jako dalším je Power BI, kde tento nástroj vyniká díky silné integraci s dalšími produkty Microsoftu a podpoře připojení k různým datovým zdrojům. Excel je často považován za základní nástroj, zůstává cenným pomocníkem pro rychlou analýzu a vizualizaci menších datových sad.

2 SYSTEMATICKÝ POSTUP PRO ANALÝZU VĚDECKÝCH TEXTŮ

Tato kapitola se věnuje systematickému postupu pro analýzu vědeckých textů, který usnadňuje získání uceleného přehledu pro danou oblast.

2.1 Definice systematického postupu

Systematický postup představuje strukturovaný a metodicky jasně vymezený proces, kde cílem je vyhledávání, hodnocení a syntéza dostupné literatury [24]. Tento přístup zajišťuje objektivitu, transparentnost a opakovatelnost analýzy, kde se minimalizuje riziko subjektivního zkreslení a umožňuje spolehlivé porovnání stávajících poznatků. Součástí tohoto procesu jsou předem stanovená kritéria pro výběr studií, metodické hodnocení jejich kvality a syntéza výsledků, kde se často využívají standardizované postupy [25].

2.2 Typy systematického postupu

Systematický postup zahrnuje různé typy, které umožňují důkladné prozkoumání existující literatury, identifikaci klíčových poznatků a syntézu výzkumných závěrů. Každý z těchto typů se liší svým zaměřením a hloubkou analýzy, kde volba konkrétního přístupu závisí na cílech studie [26] [28].

Mezi jednotlivé typy systematického postupu patří [26] [27] [28]:

- **Systematický přehled –** představuje literární rešerši, která se opírá o jasně stanovená kritéria a strukturovaný přístup k vyhledávání, výběru a hodnocení dostupných studií. Důležitým milníkem v procesu standardizace bylo vydání metodiky PRISMA, která přispěla k vylepšení kvality a transparentnosti systematického přehledu;
- **Přehled smíšených metod –** jedná se o kombinaci kvantitativních a kvalitativních dat umožňující hlubší porozumění zkoumaného tématu. Tento typ je užitečný například při hodnocení efektivity různých přístupů, strategií nebo nástrojů v oblasti jejich praktického uplatnění. Je nezbytné nejen změřit dosažené výsledky, ale také porozumět postojům, zkušenostem a motivacím osob;
- **Rychlý přehled –** jedná se o zkrácenou verzi systematického postupu, jehož cílem je co nejrychleji shrnout dostupné poznatky. Tento typ se využívá především v situacích, kdy je nutné rychle rozhodnout na základě důkazů, například během krizí nebo ve zdravotnictví v průběhu pandemii. Některé kroky, které jsou součástí běžného systematického postupu, mohou být zjednodušeny nebo dokonce vynechány;

- **Mapovací přehled –** jedná se o typ, který slouží pro úvodní orientaci v tématu, které může být velmi široké, nebo zatím nebylo systematicky zkoumáno. Tento přehled umožňuje získat základní přehled o dostupné literatuře, identifikovat výzkumné oblasti a posoudit vhodné prohloubení výzkumu prostřednictvím detailní systematické analýzy.

V této diplomové práce byl zvolen systematický přehled, který poskytuje strukturovaný způsob k důkladnému prozkoumání stávajících poznatků v dané oblasti. K zajištění přehlednosti a opakovatelnosti celého procesu byla využita metodika PRISMA, která umožňuje popsat postupy vyhledávání, výběru a zpracování zdrojů.

2.3 Metodika PRISMA

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) představuje standardizovaný rámec, který zajišťuje transparentnost a přesnost při realizaci systematických přehledů a metaanalýz. Tato metodika byla vyvinuta s cílem zlepšit kvalitu a replikovatelnost systematických přehledů, především v oblasti medicínského výzkumu. V posledních letech tato metodika nalezla široké uplatnění i v dalších vědeckých a technických oborech [25] [26] [30].

Tento rámec zahrnuje jasné pokyny pro každý krok systematického přehledu, a to od definování výzkumné otázky, přes výběr článků, hodnocení zkreslení, až po syntézu výsledků. PRISMA 2020 je nejnovější aktualizací této metodiky a přináší zásadní vylepšení oproti verzi z roku 2009, včetně podrobnějších požadavků na vykazování strategie vyhledávání [25] [29] [30].

2.4 Fáze metodiky PRISMA

Tato metodika je rozdělená do tří hlavních fází, které postupně vedou od prvotního vyhledávání až po finální výběr článků vhodných k analýze.

Mezi jednotlivé fáze patří [30] [31] [32] [33] [34]:

- **Identifikace (Identification)** – tato fáze představuje vyhledávání odborných článků v různých databázích a dalších zdrojích. Cílem této fáze je shromáždit co nejvíce relevantních studií na základě předem určených klíčových slov;
- **Třídění (Screening)** – tato fáze zahrnuje odstranění duplicit, tedy článků, které se objevují v různých databázích opakovaně. Následně se provádí základní třídění, například podle názvu článku či abstraktu, a vyřazují se záznamy, které nejsou relevantní k danému tématu;

- **Zařazení (Included)** – v této fázi jsou vybrány články, které prošly předchozími kroky. Tyto články jsou definitivně zařazeny a představují tak finální výběr, se kterým se následně provádí analýza.

2.5 Dotazy pro vyhledávání na Web of Science

V této části práce byly aplikovány předem formulované dotazy pro vyhledávání relevantních vědeckých článků v daných databázích. Tyto dotazy byly navrženy tak, aby co nejpřesněji reflektovaly zkoumané téma zaměřené na data-driven business a zahrnovaly jak široce zaměřené klíčové termíny, tak kombinace specifických konceptů a přístupů. Tento proces zahrnoval použití jednotlivých dotazů a jejich následnou kombinaci s cílem zajistit co nejširší pokrytí a zachycení relevantních výsledků.

V databázi Web of Science bylo dostupné uživatelské rozhraní pro vyhledávání. Konkrétně se jedná o sekci „*Documents*“, která je určena k vyhledávání vědeckých článků a dalších akademických zdrojů. V horní části rozhraní je možné vybrat specifickou kolekci, v níž lze provádět vyhledávání. V tomto případě to je například „*Web of Science Core Collection*“, která je předem nastavena. Tato kolekce zahrnuje různé edice jako jsou Science „*Citation Index*“, „*Social Sciences Citation Index*“ a „*Arts and Humanities Citation Index*“, čímž poskytuje přístup k vysoce kvalitním, recenzovaným a vědecky ověřeným zdrojům v různých oborech.

Na základě rozhraní, které je dostupné v příslušné databázi, bylo provedeno a zvoleno pokročilé vyhledávání. Veškeré příslušné nastavení bylo ponecháno jako výchozí, a to konkrétně „*Web of Science Core Collection*“, spolu s dostupnými edicemi. Při samotném vyhledávání byly použity pokročilé dotazy zadávané do sekce zvané jako „*Query Preview*“, kde je možné aplikovat a předem zkontrolovat správnost syntaxe a strukturu dotazu. Byly aplikovány dva základní typy dotazů a jejich kombinace klíčových slov za pomocí logických operátorů.

První dotaz

Byl navržen tak, aby zahrnoval širokou škálu relevantních termínů a umožnil tak zachytit, co nejrelevantnější články. Klíčová slova jako „*data-driven business*“, „*data-driven strategy*“ a „*business intelligence*“ odrážejí základní pojmy v oblasti daného tématu. Dotaz zahrnuje i výrazy jako „*data analytics for business*“ a „*data-driven decision-making*“. Dotaz byl formulován pomocí syntaktického prvku Topic Search, který v databázi specifikuje, že vyhledávání má být provedeno v tematickém poli článků. Dále byl v dotazu byl použitý booleovský operátor „OR“, což umožňovalo zahrnout všechny vyjmenovaná klíčová slova, což napomohlo k rozšíření vyhledávání článků, které obsahují jakýkoliv z těchto výrazů.

Druhý dotaz

Podobal se prvnímu dotazu, s tím rozdílem, že byla zvolena další klíčová slova, jako jsou „*data-driven*“, „*data-informed*“, „*business model*“, „*business strategy*, „*corporate strategy*“ a „*business process improvement*“. Jako další rozdíl byl zvolen booleovský operátor „AND“, který umožňuje propojit dvě různé sady klíčových slov. Tento operátor zajistil, že výsledky vyhledávání zahrnovaly pouze články obsahující kombinaci termínů z obou skupin.

Kombinace dotazů

V rámci spojení dotazů byl použit booleovský operátor „OR“, který umožňuje rozšířit výsledky obou dotazů. Oba dotazy byly zachovány v identické podobě a každý z nich obsahoval svá daná klíčová slova. Díky této kombinaci obou dotazů byly pomocí operátorů získány výsledky.

Po provedení se lišil počet výsledků v závislosti na konkrétním dotazu. Při prvním vyhledávacím dotazu v databázi Web of Science bylo nalezeno 9 730 záznamů. Při druhém dotazu bylo nalezeno pouze 252 záznamů. Po kombinaci obou dotazů bylo dosaženo nejvíce, a to 9 890 záznamů.

Mezi filtry a omezení byly zvoleny následující:

- **Web of Science Categories** – zahrnuje možnost výběru jednotlivých kategorií. Pro tento případ byly vybrány tyto kategorie jako „*Computer Science Information Systems*“, která se specializuje na systémy pro zpracování dat. Dalším důležitým zaměřením byla kategorie „*Management*“, jež se zabývá řízením firem a strategiemi, které s tímto tématem úzce souvisejí. Kategorie „*Business*“ byla zařazena, protože přináší články soustředící se na obchodní strategie. Výběr také zahrnoval „*Computer Science Artificial Intelligence*“, neboť umělá inteligence hraje klíčovou roli v oblasti datové analytiky a strategií založených na datech. Na závěr byla zvolena kategorie „*Operations Research Management Science*“, která se soustředí na optimalizaci rozhodovacích procesů;
- **Languages** – byl zvolen anglický jazyk, jelikož se jedná o nejpoužívanější jazyk, který zahrnuje většinu dostupných článků.

Po aplikaci omezení a filtrů bylo pro první dotaz nalezeno celkem 4 938 výsledků. Druhý dotaz dosáhl 147 záznamů. Při použití kombinovaného dotazu bylo nalezeno 5 026 záznamů.

Export výsledků

Dalším krokem byl proveden export výsledků, které byly provedeny pomocí Excelu. Tento formát usnadňuje práci s daty, jejich třídění a následné další zpracování. Formát Excelu je užitečný pro přehledné uspořádání informací, jako jsou názvy článků, autoři, abstrakty, citační údaje a klíčová slova. Také umožňuje snadné propojení s nástroji pro textovou analýzu.

Při samotném exportu bylo možné přizpůsobit nastavení výběru dat dle specifických požadavků. Rozhraní zahrnovalo různé kategorie, které bylo možné zaškrtnout. Mezi základní volby patří informace o autorovi, název článku a zdroje, což poskytuje klíčové údaje o publikaci. Dále byly zahrnuty abstrakty, klíčová slova a typ dokumentu, což napomáhá k procesu textové analýzy dat.

2.6 Dotazy pro vyhledávání na Scopus

Jako dalším pro vyhledávání relevantních vědeckých článků byla využita databáze, která nabízí široké spektrum jednotlivých článků a dalších akademických zdrojů. Stejně jako v případě databáze Web of Science bylo nezbytné nejprve zajistit přístup prostřednictvím institucionálního přihlášení. Poté bylo možné dle uživatelského rozhraní využít jednotlivé prvky k vyhledávání.

Při vyhledávání v rámci prvního definovaného dotazu byla použita stejná klíčová slova, kde hlavním rozdílem byla odlišná syntaxe, která je charakteristická pro databázi Scopus. Dotaz byl formulován ve specifickém formátu „*TITLE-ABS-KEY*“, což znamená, že vyhledávání probíhá v názvech dokumentů, abstraktech a klíčových slovech.

Stejným způsobem jako u prvního dotazu byla použita syntaxe pro druhý dotaz a následnou kombinaci dotazů. Druhý dotaz obsahoval stejná klíčová slova jako dotaz provedený v databázi Web of Science. Po provedení druhého dotazu byla provedena kombinace dotazů pomocí booleovského operátoru OR. Tato kombinace umožnila rozšířit výsledky vyhledávání a zahrnout všechny dokumenty, které splnily kritéria obou dotazů. Po provedení se lišil počet výsledků v závislosti na konkrétním dotazu. Při prvním vyhledávacím dotazu v databázi Scopus bylo nalezeno 17 190 záznamů. Při druhém dotazu bylo nalezeno pouze 977 záznamů. Po kombinaci obou dotazů bylo dosaženo nejvíce, a to 17 817 záznamů.

Stejně jako u databáze Web of Science bylo potřeba soubor výsledků dále zpřesnit pomocí specifických filtrů.

Mezi filtry a omezení byly zvoleny následující:

- **Subject Area –** byly vybrány relevantní oblasti, jako je „*Computer Science*“, „*Business, Management and Accounting*“, „*Decision Science*“ a „*Economics, Econometrics and*

Finance“. Tyto oblasti zahrnují témata související s datovou analytikou, řízení podniků, rozhodovací procesy a další aspekty v rámci data-driven business;

- **Jazyk –** byl zvolen anglický jazyk, jelikož se jedná o nejpoužívanější jazyk, který zahrnuje většinu dostupných článků.

Po aplikaci filtrů a omezení první dotaz dosáhl 12 951 záznamů. Druhý dotaz dosáhl 779 záznamů, zatímco kombinovaný dotaz dosáhl celkových 13 442 záznamů.

Export výsledků

Následně byl proveden export výsledků za účelem následného zpracování a analýzy dat. Tato databáze dává uživatelům možnost vybrat si specifická nastavení exportu podle svých potřeb. Pro tento případ byla zvolena varianta exportu ve formátu CSV, který je vhodný po další práci. V rámci exportu byla nastavena specifická kritéria pro výběr dat, aby byly zahrnuty pouze relevantní informace. Pro analýzu byla vybrána kombinace dotazů, jelikož tento dotaz dosáhl nejvyššího počtu výsledků. Tímto způsobem byl zajištěn co nejširší přehled o dané oblasti, což je zásadní pro dosažení reprezentativních a spolehlivých výsledků.

Po provedení veškerých kroků, včetně získání celkových záznamů, odstranění duplicit, aplikace kritérií, byl na základě získaných hodnot vyplněn PRISMA diagram (viz **Příloha A**).

2.7 Textová analýza

Představuje metodu, která slouží k získávání smysluplných informací z nestrukturovaných textových dat [36]. Tento proces využívá techniky z oblasti zpracování přirozeného jazyka nebo strojového učení. Hlavním cílem je porozumět obsahu textů, identifikovat opakující se vzorce, klíčová slova, témata a významové souvislosti [37]. Na rozdíl od kvantitativních přístupů, které se zaměřují například na počet dokumentů nebo výskyt konkrétních autorů, textová analýza klade důraz na obsahovou stránku textu.

Předzpracování textů

Pro dosažení uvedeného cíle se využívá několik klíčových metod, které umožňují převod nestrukturovaného textu do formátu vhodného pro další zpracování a analýzu. Tyto metody se zaměřují především na přípravu textových dat, eliminaci šumu a zvýraznění významových prvků.

Mezi kroky předzpracování textů patří [36] [46]:

- **Tokenizace** – jedná se o počáteční krok, při kterém se text dělí na jednotlivé jednotky, obvykle slova nebo fráze, které jsou označovány jako tokeny. Tento krok umožňuje analyzovat text na úrovni jednotlivých jazykových prvků;
- **Odstranění stop slov** – jedná se například o výrazy, které jsou z jazykového hlediska nezbytné, avšak pro analýzu postrádají významnou informační hodnotu. Jedná se například o česká slova („*a*“, „*že*“) nebo anglická slova („*the*“, „*is*“). Vyloučením těchto výrazů se snižuje šum v datech a zvyšuje se kvalita následných analytických výstupů;
- **Lemmatizace** – jedná se o metodu, která slouží ke zjednodušení a sjednocení jazyka tím, že různé tvary toho samého slova převádí na jejich společný základ.

TF-IDF (Term Frequency-Inverse Document Frequency)

Jedná se o klíčovou metodu v oblasti textové analýzy, která slouží k hodnocení významu slov nejen v rámci jednotlivého dokumentu, ale i kontextu celého korpusu [36]. Uplatňuje se především při automatickém zpracování textu, kde slouží k zvýraznění významných slov a potlačení těch, která se v dokumentech objevují příliš často a postrádají dostatečnou informační hodnotu [37]. Díky tomu umožňuje odlišit klíčové pojmy, které jsou relevantní pro jednotlivé dokumenty [39].

Tato metoda spojuje dvě zásadní metriky [36] [37]:

- **TF (Term Frequency)** – vyjadřuje frekvenci výskytu určitého termínu v konkrétním dokumentu. Čím častěji se termín objevuje, tím větší důležitost to naznačuje v daném textu;
- **IDF (Inverse Document Frequency) –** tato metrika zohledňuje, v kolika dokumentech se tento termín vyskytuje napříč celým korpusem. Čím více dokumentů obsahuje daný výraz, tím menší informační hodnotu má jeho frekvence.

Výpočet výsledného skóre TF-IDF se obvykle provádí podle následujícího vzorce [36] [37]:

$$TF - IDF(t, d) = TF(t, d) * \log\left(\frac{N}{1 + DF(t)}\right)$$

- t – označení pro termín (slovo)
- d – označení pro dokument
- N – označení pro celkový počet dokumentů v korpusu
- $DF(t)$ – označeno pro počet dokumentů, ve kterých se termín (t) objevuje

NMF (Non-Negative Matrix Factorization)

Představuje metodu, která se využívá k odhalování skrytých témat ve skupině dokumentů. Tato metoda spočívá v rozkladu původní matice, která obsahuje nezáporné hodnoty, například TF-IDF skóre slov v jednotlivých dokumentech na dvě menší matice. Tyto matice odhalují vztahy mezi dokumenty a tématy, a také mezi tématy a slovy. Na rozdíl od některých jiných metod tematického modelování, jako je LDA, je NMF dobře interpretovatelná, protože všechny její výstupní hodnoty zůstávají nezáporné [37].

Základní výpočet pro metodu NMF se provádí podle následujícího vzorce [37]:

$$V \approx W * H$$

- V – označení pro původní matici výskytu slov v dokumentech
- W – označení pro matici vah dokumentů k tématům
- H – označení pro matici vah témat ke slovům

Rekonstrukční chyba

Tato metrika slouží k posouzení, jak dobře model dokáže rekonstruovat původní data na základě zvoleného počtu témat. Nižší hodnota rekonstrukční chyby naznačuje lepší schopnost modelu zachytit vzory v datech, a proto se často používá při určování optimálního počtu témat v tematickém modelování pomocí NMF [38].

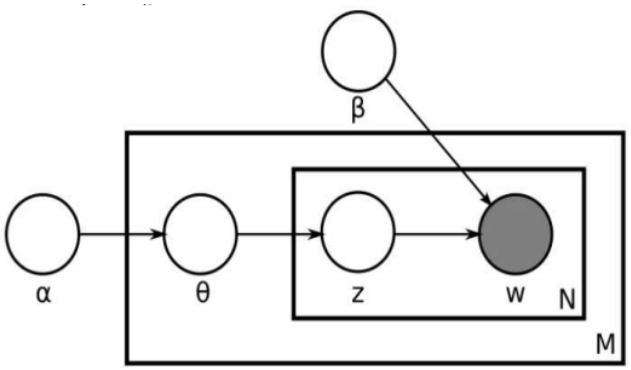
Bag of Words

Tato metoda je založena na principu reprezentace textu jako samostatné množiny slov, což se ukazuje jako efektivní přístup zejména při využití v tematickém modelování, například prostřednictvím modelu LDA [40]. V rámci této metody je každý dokument převeden na vektor, jehož prvky odrážejí výskyt konkrétních slov, aniž by zohledňovaly jejich pořadí nebo gramatické vztahy. Díky své jednoduchosti je metoda výpočetně nenáročná a snadno se implementuje [41].

LDA (Latent Dirichlet Allocation)

Představuje model, který slouží k automatické identifikaci témat v textových datech [37]. Jeho hlavním cílem je odhalit skrytou tematickou strukturu ve velkém sběru dokumentů, aniž by bylo nutné je předem označovat nebo znát jejich obsah [42]. Tento model vychází z předpokladu, že každý dokument je složen z kombinace několika témat, přičemž každé téma je reprezentováno určitou pravděpodobnostní distribucí slov (viz schéma Obrázek 1) [43].

Jedná se o to, že LDA předpokládá, že při vzniku každého dokumentu došlo k výběru několika témat, přičemž jednotlivá slova v textu byla náhodně generována na základě těchto témat a jejich slovní zásoby [37]. Tento model tak zpětně odhaluje, která témata se v dokumentu vyskytují a která slova k nim pravděpodobně patří [42].



Obrázek 1: Schéma LDA

Zdroj: [43]

- M – označení pro počet dokumentů v kolekci
- N – označení pro počet slov v jednotlivých dokumentech
- α – označení pro parametr Dirichletova rozdělení, který určuje, jak se témata rozdělují mezi jednotlivé dokumenty
- β – označení pro parametr Dirichletova rozdělení pro rozdělení slov v rámci jednotlivých témat
- θ – označení pro konkrétní rozdělení témat pro daný dokument
- w – označení pro skutečná slova v dokumentu
- z – označení pro skryté téma přiřazené jednotlivým slovům

Koherence témat

Tato metrika slouží pro hodnocení kvality výsledků při tematickém modelování, zejména pro metodu LDA. Díky této metrice je posuzováno, jak dobře spolu významově souvisejí slova uvnitř jednotlivých témat [44]. Vyšší koherence svědčí o lepší interpretaci daného tématu. Často se využívá při určování optimálního počtu témat, protože umožňuje zjistit, kdy jsou témata dostatečně přehledná a srozumitelná [45].

2.8 Bibliometrická analýza

Jedná se o metodu, která umožňuje numerické zkoumání vědeckých publikací a odhalení trendů v odborné literatuře [50]. Tato metoda slouží k identifikaci nejcitovanějších článků,

nejaktivnějších autorů, vysoce hodnocených výzkumných institucí a dalších. Tato analýza se běžně využívá v rámci systematických přehledů literatury, které přinášejí přehled o vývoji dané oblasti výzkumu v průběhu času [51].

Fáze bibliometrické analýzy

K tomu, aby byly získány relevantní výsledky je nezbytné provést několik klíčových fází. Tyto fáze zajistí správné zpracování a interpretaci dat. Každá tato fáze má svou důležitost a přispívá k vytvoření uceleného přehledu v rámci zkoumané oblasti. Celý proces lze rozdělit do hlavních fází, která na sebe logicky navazují a umožňují efektivní práci se získanými daty.

Proces zahrnuje tyto jednotlivé fáze [51] [55]:

- **Definování cíle** – určení cíle analýzy, jelikož je nutné jasně vymezit zaměření a jakých výsledků lze dosáhnout. Pokud by nebyly cíle dostatečně jasné, mohlo by to vést k tomu, že analýza nebude systematická a nebude dosaženo relevantních výsledků;
- **Sběr dat z databází** – je důležité pracovat s kvalitními a relevantními publikacemi, které mohou poskytnout přesné výsledky. Nesprávně zvolený dataset by mohl vést k chybným závěrům a zkreslené analýze;
- **Čištění a příprava dat** – jedná se o sjednocení formátu, odstranění duplicitních záznamů a úprava nesrovnalostí v údajích. Touto fází lze zajistit, že dataset bude přehledný a připravený k dalšímu využití;
- **Výběr a aplikace metod** – je klíčový krok, jelikož každá analýza vyžaduje vhodné nástroj a techniky k dosažení potřebných informací. Díky těmto metodám lze identifikovat nejcitovanější studie, zmapovat spolupráci mezi autory a sledovat vývoj klíčových témat v čase;
- **Vizualizace výsledků** – jedná se o grafy a další vizuální prvky, které usnadňují práci s daty a poskytují jasný přehled o dosažených výsledcích.

Metody bibliometrické analýzy

Tyto metody zahrnují různé přístupy, které umožňují porozumět vývoji konkrétní oblasti, jako je identifikovat klíčové autory, instituce a další. Tyto metody lze obecně rozdělit do dvou hlavních kategorií, a to výkonové analýzy a vědeckého mapování [48]. Výkonové analýzy se zaměřují na hodnocení výstupů autorů, časopisů nebo institucí. Sledují různé ukazatele, jako je počet publikací, počet citací, h-index, g-index, které pomáhají posoudit produktivitu a vliv jednotlivých autorů. Vědecké mapování se zaměřuje na odhalování vztahů v rámci vědecké produkce.

Používá techniky jako analýzu spoluautorství, ko-citační analýzu nebo analýzu výskytu klíčových slov. Tyto techniky umožňují sledovat vývoj tématu, strukturu spolupráce mezi autory a vznik nových výzkumných směrů [52].

Analýza klíčových slov

Představuje metodu, která slouží k porozumění souvislostí mezi tématy v určité oblasti. Principem je sledování pojmů, které se často objevují společně, například v názvech článků, klíčových slovech nebo abstraktech [53]. Na základě těchto souvislostí vytváří vizuální mapu vztahů, která znázorňuje, jak jsou jednotlivá slova propojena. Tato metoda je velmi užitečná pro odhalení struktury výzkumného pole, avšak její úspěšnost závisí na kvalitě získaných dat. Pokud články postrádají klíčová slova nebo jsou označena nekonzistentně, může to vést ke zkreslení výsledků [56].

Síť spoluautorství

Tato metoda se zaměřuje na zkoumání spolupráce mezi autory. Jde o mapování vědecké spolupráce, kdy dvě nebo více osob společně pracují na publikaci [53]. Tato metoda se často využívá například k identifikaci propojení mezi jednotlivými autory, výzkumnými institucemi nebo dokonce mezi státy. Díky této metodě je možné odhalit výzkumné týmy, klíčové spolupracovníky a místa s vysokou intenzitou spolupráce. Vizualizace těchto vztahů, například prostřednictvím síťových grafů, poskytuje jasný pohled na fungování výzkumné komunity [56].

h-index

Jedná se o metriku, která slouží jako snadný způsob, jak odhadnout vliv a produktivitu autora. Tento index vyjadřuje, že autor má h-index h, pokud publikoval alespoň h článků, přičemž každý z těchto článků byl citován minimálně h-krát [53]. Tento index funguje tak, že spojuje počet publikací s jejich citovaností, což zajišťuje, že hodnocení není zkresleno pouze jediným vysoce citovaným článkem ani množstvím méně citovaných textů [54].

g-index

Slouží jako alternativní metrika, která byla vyvinuta jako rozšíření h-indexu. Tento index zohledňuje nejen počet publikovaných článků, ale také celkový počet citací nejcitovanějších článků [53]. Autor má g-index g, pokud jeho g nejcitovanějších článků dohromady získalo alespoň g^2 citací [54].

2.9 Volba vhodného nástroje

K provedení dané analýzy je vhodný výběr nástroje, který umožní efektivní zpracování a vizualizaci dat. Výběr vhodného nástroje závisí na rozsahu analýzy a požadovaných výsledcích. V této práci byla analýza provedena pomocí programovacího jazyka Python, který zahrnuje širokou škálu knihoven a nástrojů pro práci s daty, vizualizaci a pokročilé analýzy. Tento programovací jazyk je dostupný zdarma a umožňuje flexibilní úpravy a kombinace různých analytických metod.

3 SBĚR, ZPRACOVÁNÍ A ANALÝZA ZÍSKANÝCH DAT

Tato část se zaměřuje na proces shromažďování, zpracování a analýzu dat na základě získaných vědeckých textů s využitím textové analytiky a použití vhodného nástroje. Cílem je ukázat, jak lze zpracovat získaná data z daných zdrojů, transformovat je do formy, která je praktická, a použít analytické metody k získaní cenných informací. Klade se důraz na jednotlivé kroky, a to přípravu dat, jejich analýzu a interpretaci k porozumění určeného tématu.

3.1 Sběr dat

Jedná se o první krok v rámci procesu a následného použití v analýze. Tento krok zahrnuje identifikaci, vyhledávání a získání relevantních zdrojů. v rámci zaměření této diplomové práce byla shromážděna data z citačních databází, jako jsou Web of Science a Scopus, které poskytují přístup k rozsáhlé škále odborných článků a dalších vědeckých textů. Zároveň pokrývají produkci nejvýznamnějších světových vydavatelů, čímž garantují komplexní a důvěryhodné pokrytí odborné literatury relevantní k danému tématu. Doplňkově byl zvažován také zdroj Google Scholar pro rozšíření o šedou literaturu, ale kvůli omezeným možnostem filtrování, chybějícímu strukturovanému exportu a různorodým bibliografickým údajům nebyl do konečného výběru zahrnut.

Citační databáze

Jedná se o specializovanou databázi, která indexuje vědeckou literaturu na základě citací mezi různými publikacemi. Tato databáze umožňuje sledovat vzájemné odkazy mezi články, což usnadňuje vyhledávání relevantních studií a měření citačního dopadu konkrétních publikací [47]. Každý záznam obsahuje bibliografické údaje o zdrojovém článku a seznam citovaných prací [48].

Web of Science

Představuje databázi, spravovanou společností Clarivate, která poskytuje přístup k recenzovaným článkům, různým příspěvkům a dalším akademickým zdrojům napříč různými vědními obory. Tato databáze umožňuje efektivní vyhledávání aktuálních a relevantních informací. Díky své spolehlivosti a rozsáhlému obsahu se stává zásadním zdrojem pro širokou část vědců, odborníků a studentů. Pro přístup k této databázi je nezbytné se přihlásit prostřednictvím institucionálního účtu, například prostřednictvím univerzity, nebo si založit vlastní účet.

Scopus

Představuje databázi, spravovanou vydavatelstvím Elsevier, která poskytuje podobné zdroje jako Web of Science, a to recenzované články, knihy, příspěvky z konferencí a další akademické zdroje napříč různými vědními obory. Díky své spolehlivosti, kvalitě zdrojů a jednotlivým možnostem představuje nástroj pro vědce, odborníky či studenty. Přístup do databáze je možný stejným způsobem jako v předchozím případě prostřednictvím institucionálního účtu nebo registrací vlastního účtu.

3.2 Výzkumné otázky

Tento krok byl zásadním v rámci procesu analýzy vědeckých textů [48]. Díky tomu bylo možné formulovat specifické předpoklady, které mohly být ověřeny různými analytickými metodami [49]. Výzkumné otázky byly zaměřeny na klíčové aspekty zkoumané problematiky, jako například sledování v čase identifikace autorů nebo mapování spoluprací a další. K dosažení výsledků, bylo nezbytné, aby byly výzkumné otázky jasně formulovány a plně odpovídaly cílům analýzy.

Formulace výzkumných otázek

Pro lepší porozumění v rámci dané analýzy zaměřené na téma data-driven business a souvisejících tématech, byly formulovány následující výzkumné otázky. Každá z těchto výzkumných otázek se zaměřuje na specifický aspekt daného tématu.

- **(O1):** Jaká jsou klíčová slova v oblasti „*data-driven business*“?

Výzkumná otázka je zaměřena na sledování vývoje trendů v publikacích zaměřených na téma „*data-driven business*“ v určeném časovém období. Výzkumná otázka předpokládá, že určitá slovní spojení jako jsou „*business intelligence*“, „*machine learning*“ nebo „*data-driven decision making*“, se v daném období budou vyskytovat častěji než jiná.

- **(O2):** Jak se mění průměrný počet citací publikací v závislosti na roce jejich vydání?

Výzkumná otázka je zaměřena na zjištění, jak se průměrný počet citací publikací v oblasti vyvíjí v závislosti na roce jejich vydání, a to za období posledních pěti let 2020 až 2024.

- **(O3):** Které články byly nejčastěji citovány v oblasti „*data-driven business*“?

Výzkumná otázka je zaměřena na identifikaci nejčastěji citovaných článků v dané oblasti v období roku 2020 do roku 2024. Cílem bude zjistit, které odborné publikace získaly nejvíce citací a lze je považovat za klíčové.

- **(O4):** Jak se mění zaměření v oblasti „*data-driven business*“ v průběhu let 2020 až 2024?

Výzkumná otázka je zaměřena na to, jak se mění témata, které jsou v oblasti nejvíce zkoumána v průběhu let. Na základě analýzy výskytu klíčových slov v publikacích z období 2020 až 2024, kde bude zkoumáno, zda se některé oblasti stávají populárnějšími, zatímco jiná témata mohou naopak ztrácet na významu.

- **(O5):** Kteří autoři publikovali nejvíce článků v oblasti „*data-driven business*“?

Výzkumná otázka je zaměřena na identifikaci nejaktivnějších autorů, kteří publikovali nejvíce odborných článků na dané téma v období od roku 2020 do roku 2024. Cílem bude vyhledat autory, kteří se v této oblasti objevují opakovaně a svým přínosem podporují rozvoj dané oblasti.

- **(O6):** Které publikační zdroje nejčastěji publikují články k tématu „*data-driven business*“?

Výzkumná otázka je zaměřena na identifikaci publikačních zdrojů, které nejčastěji zveřejňují články zaměřené na téma „*data-driven business*“. Cílem bude zjistit, které publikační zdroje se této oblasti věnují nejvíce.

- **(O7):** Kteří autoři spolu nejčastěji spolupracují na publikacích v oblasti „*data-driven business*“?

Výzkumná otázka je zaměřena na identifikaci autorů, kteří nejčastěji spolupracují na odborných publikacích. Jejím cílem bude zjistit, jaké vztahy existuji mezi jednotlivými autory a zda se v rámci produkce objevují stabilní autorské dvojice nebo skupiny.

- **(O8):** Kteří autoři dosahují nejvyššího citačního dopadu v oblasti podle metrik jako h-index, g-index a celkový počet citací?

Výzkumná otázka je zaměřena na identifikaci autorů s největším citačním dopadem v dané oblasti. Cílem bude zjistit, kteří autoři publikovali nejvíce článků, ale také zjistit pomocí citačních metrik, jako jsou celkový počet citací, h-index a g-index.

3.3 Proces analýzy a její aplikace

V rámci procesu dané analýzy vědeckých textů bylo nezbytné provést jednotlivé kroky, a to předzpracování, jednotlivé kroky analýzy a vizualizaci dat. Veškeré kroky byly provedeny již

ve zvoleném nástroji, a to pomocí jazyka Python pomocí programového prostředí Google Collab.

Instalace a import knihoven

K provedení bibliometrické analýzy bylo nejprve nezbytné nainstalovat a naimportovat jednotlivé knihovny, které Python zahrnuje. Tyto knihovny zahrnují rozsáhlou škálu funkcí pro zpracování textu a zajištění funkčnosti daných příkazů.

V rámci této analýzy byly nainstalovány a importovány následující knihovny [57] [58] [59] [60] [61] [62] [63] [64] [65] [66] [67]:

- **pandas** – představuje knihovnu, která umožňuje práci s daty v tabulkové podobě. Umožňuje snadné načítání souborů, jejich úpravu a čištění, což usnadnilo organizaci dat, odstranění chyb, duplicit a dalších, tak aby byla data připravena na další zpracování;
- **numpy** – představuje knihovnu, která usnadňuje práci s číselnými daty a rychlé provedení matematických operací. v rámci této analýzy byla využitá zejména na zpracování velkého objemu textových dat, jako je například analýza frekvence slov nebo příprava textů;
- **pybibx** – představuje knihovnu, která zahrnuje nástroje pro bibliometrickou analýzu, jako je analýza počtu publikací, citační index, výpočet indexu a zkoumání vztahů mezi autory a institucemi. Tato knihovna je vhodná pro podrobné vyhodnocování citačních trendů;
- **networkx** – představuje knihovnu, která zahrnuje nástroje pro tvorbu, manipulaci a analýzu komplexních sítí a grafů. Tato knihovna usnadňuje proces definování uzlů a hran, hledání optimálních cest a vizualizaci síťových struktur;
- **matplotlib a seaborn** – představuje knihovnu, která slouží k vizualizaci dat a snadnější přehlednosti výsledků analýzy. Matplotlib zahrnuje možnost vytvářet různé druhy grafů a přizpůsobit jejich vzhled. Dále Seaborn, který vychází z Matplotlibu, který zahrnuje lepší a přehlednější grafy;
- **wordcloud** – představuje knihovnu, která zahrnuje grafické zobrazení frekvence slov, kde velikost jednotlivých výrazů odpovídá jejich četnosti v textu. Toto umožňuje identifikovat nejvýznamnější pojmy v datasetu a poskytuje tak přehled o nejčastějších tématech v analyzovaných textech;

- **spacy** – představuje knihovnu, která obsahuje nástroje pro rozdělení textu na jednotlivá slova, odstranění běžně se vyskytujících bezvýznamových výrazů a převod slov na jejich základní tvary;
- **re** – představuje modul, který obsahuje nástroje pro definici a překlad regulárních výrazů. Tyto nástroje umožňují rozdělení textu podle vzorů, vyhledávání a nahrazování podřetězců a také extrakci specifických částí textu;
- **gensim** – zahrnuje knihovnu, která je určena k vytváření a aplikaci modelů vektorových prostorů na velké textové korpusy. Dále je zaměřena na modelování témat (topic modelling), indexaci dokumentů a vyhledávání podobností;
- **xslxwrite**r – jedná se o modul určený pro tvorbu excelových souborů ve formátu XLSX, bez nutnosti nainstalovaného programu Microsoft Excel.

Základní operace

Tato část byla zaměřena na proces jednotlivých kroků, které zahrnovaly načtení a úpravy exportovaných výsledků z vybraných databází, s cílem zajistit jejich jednotnou strukturu a přípravu pro další analýzu. Různé databáze, jako jsou Scopus a Web of Science, využívají odlišné formáty exportovaných dat, což vyžaduje sladění a sjednocení struktury dat před samotnou analýzou.

Načtení dat

V první řadě bylo nezbytné nejprve načíst oba datasety na základě prvního dotazu a provést, tak jejich základní kontrolu. Tento krok byl nutný pro ověření, že byla data importována správně a jejich struktura odpovídá.

Identifikace a kontrola dat

Po načtení datasetů z databází Scopus a Web of Science byla provedena základní kontrola, která byla zásadní pro ověření správnosti a přípravy dat k dalšímu zpracování. Tento krok klade důraz na získání přehledu o objemu a charakteru vstupních dat.

Při kontrole dat bylo zaměření se na tyto kroky:

- **Zobrazení počtu záznamů** – jednalo se o výpis celkového množství nalezených článků v jednotlivých databázích, což poskytlo základní přehled o velikosti jednotlivých datasetů;

- **Zobrazení typů dokumentů** – jedná se o výpis různých forem dokumentů, jako jsou články, recenze nebo konferenční příspěvky, což usnadnilo orientaci v obsahu souborů;
- **Kontrola datových typů** – ověření, zda jsou důležité proměnné, například počet citací, uvedeny ve správném formátu (například v číselném a ne textovém), což je nezbytné pro analýzu;
- **Kontrola chybějících hodnot** – ověření, zda některé záznamy neobsahovaly klíčové informace, jako je rok publikace, klíčová slova nebo jména autorů.

Na základě provedené kontroly dat z jednotlivých databází byl získán ucelený přehled o struktuře a kvalitě vstupních záznamů. Během této kontroly bylo potvrzeno, že databáze Scopus obsahuje celkem 17 414 záznamů a Web of Science obsahuje 9 793 záznamů. Následně bylo provedeno zobrazení typů dokumentů v těchto zdrojích. Ve Scopusu převažovaly tyto dokumenty, a to články a recenze. Naopak v databázi Web of Science převažovaly především články a sborníkové příspěvky. Dále byla provedena kontrola datových typů, které potvrdila konzistentní formát důležitých proměnných, jako je například počet citací nebo rok vydání, které jsou uloženy jako číselné hodnoty. Kontrola chybějících hodnot ukázala, že některé sloupce obsahovali značné množství prázdných polí, jako je například „*Author Keywords*“ nebo „*Affiliations*“. Naopak klíčové sloupce jako „*DOI*“ a „*Title*“ byly v obou datasetech téměř kompletní. Na základě sloučení dat ze Scopusu a Web of Science vznikl souhrnný datový rámec, který obsahoval celkem 27 207 záznamů. V rámci fáze identifikace bylo následně třeba odstranit duplicitní záznamy na základě shody v poli „*DOI*“. Celkem bylo vyřazeno 12 214 duplicitních záznamů. Po této deduplikaci zbývalo 14 993 záznamů, které byly následně předmětem fáze screeningu.

Pro provedení výše uvedených kroků byl aplikován odpovídající kód v Pythonu, který je uveden (viz **Příloha B**).

Screening a úprava dat

V dalším kroku bylo třeba připravit konzistentní datový soubor pro analýzu. Vstupní data z databází Scopus a Web of Science obsahovala různé nesrovnalosti, chybějící hodnoty, duplicitní záznamy nebo odlišnosti ve formátu záznamů. Nedílnou součástí bylo také filtrování záznamů podle typu dokumentu časového období, aby byly vybrány pouze relevantní publikace pro další zpracování.

V rámci tohoto kroku byly provedeny následující činnosti:

- **Výběr relevantních proměnných** – byly ponechány pouze ty sloupce, které byly nezbytné pro analýzu, a to „*Title*“, „*Authors*“, „Year“, „*Source title*“, „*Cited by*“, „*DOI*“, „*Abstract*“, „*Author Keywords*“ a „*Document Type*“. Nepotřebné sloupce byly odstraněny;
- **Oprava chybějících hodnot** – v datasetech se nacházely chybějící hodnoty, které by mohly negativně ovlivnit přesnost analýzy. Bylo proto nutné tyto hodnoty identifikovat a následně je vhodně doplnit nebo odstranit, aby nedošlo ke zkreslení výsledků analýzy;
- **Sjednocení formátu klíčových polí** – v daných databázích se název článku, „*DOI*“, citační počty a další mohou lišit svým zápisem. Některé texty mohou obsahovat velká písmena, diakritiku nebo speciální znaky, což může vést k situaci, kdy stejné údaje nejsou rozpoznávány. Proto bylo nezbytné převést všechny textové hodnoty do jednoho formátu;
- **Filtrování podle typu dokumentu** – v rámci analýzy byly zachovány pouze typy dokumentu „*Article*“ a „*Review*“. Ostatní typy dokumentů byly vyloučeny;
- **Filtrování podle časového omezení** – v rámci analýzy byly zařazeny pouze publikace vydané v období roku 2020 do roku 2024, které zahrnuje nejnovější poznatky a reflektuje aktuální vývoj v oblasti data-driven business.

3.4 Textová analýza v Pythonu

Po dokončení základních kroků, zahrnujících načtení, kontrolu, sloučení a očištění dat z daných zdrojů bylo možné přejít k textové analýze. Tato analýza byla zaměřena na porozumění obsahové stránce vědeckých článků a vytvořit základ pro podrobnější zkoumání klíčových témat, pojmů a jazykových vzorců v oblasti data-driven business.

Textová analýza byla zaměřena na klíčové části článků, které nejlépe vystihují jejich odborné zaměření, a to konkrétně na názvy, abstrakty a klíčová slova uvedená autory. Tyto prvky byly sloučeny do jednoho textového pole a následně podrobeny standardnímu předzpracování. Tato fáze zahrnovala odstranění speciálních znaků, převod textu na malá písmena, vyloučení stop slov a lemmatizaci, která sjednotila různé jazykové tvary. Výsledkem byl očištěný a sjednocený textový celek, který byl připraven na další kroky v rámci analýzy. Na následující ukázce (viz Obrázek 2) byl zobrazen přehled částí těchto upravených textů, který zahrnoval příklady několika náhodně vybraných záznamů po dokončeném předzpracování.

	Title	Processed_Text
1291	machine learning-based techniques for land sub...	machine learning base technique land subsidenc...
6469	revenue optimization for less-than-truckload c...	revenue optimization truckload carrier physica...
4774	parallel learning of koopman eigenfunctions an...	parallel learning koopman eigenfunction invari...
5518	the role of nontechnical skills in providing v...	role nontechnical skill provide value analytic...
4703	the convergence of big data and accounting: in...	convergence big accounting innovative research...

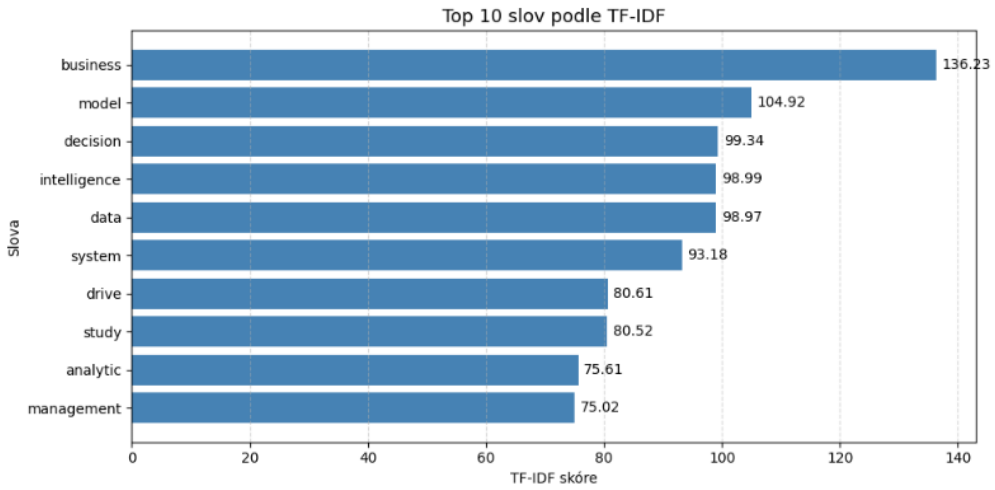
Obrázek 2: Předzpracování textů

Zdroj: Vlastní zpracování

Nejčastější slova dle TF-IDF

Po předzpracování textových dat, během kterého byly sloučeny názvy článků, abstrakty a klíčová slova do jednoho pole a odstraněny nadbytečné prvky jako speciální znaky nebo běžná slova bez významu, byl postup zaměřen na analýzu nejčastěji používaných slov s využitím metody TF-IDF. Cílem této analýzy bylo zjistit, která slova se v článcích objevují nejvýznamněji, tedy nejen často, ale také s ohledem na specifikaci jednotlivých článků. Dalším krokem bylo nutné transformovat texty na číselné hodnoty, kde pro každé slovo v celém datasetu byla spočítána jeho váha, a následně byl pro přehled vytvořen souhrn všech TF-IDF hodnoty pro jednotlivá slova. Díky tomuto bylo snadné určit, která slova mají největší význam v kontextu všech článků.

Aby bylo možné výsledky lépe interpretovat, byla vytvořena vizualizace deseti nejvýznamnějších slov podle jejich TF-IDF skóre. Tato slova nejlépe vystihují tematické zaměření analyzovaných článků. Vizualizace zahrnovala přehled na klíčová slova, která se v článcích objevují nejčastěji a mají největší význam pro dané téma. Mezi nejvýše hodnocená slova byla identifikována pomocí TF-IDF skóre patří například „*business*“ s hodnotou 136,23, „*model*“ s hodnotou 104,92 a „*decision*“ s hodnotou 99,34. Výsledky tak potvrzují, že výzkumná oblast se zaměřuje například na oblasti rozhodování, modelování, práci s daty a další (viz Obrázek 3).

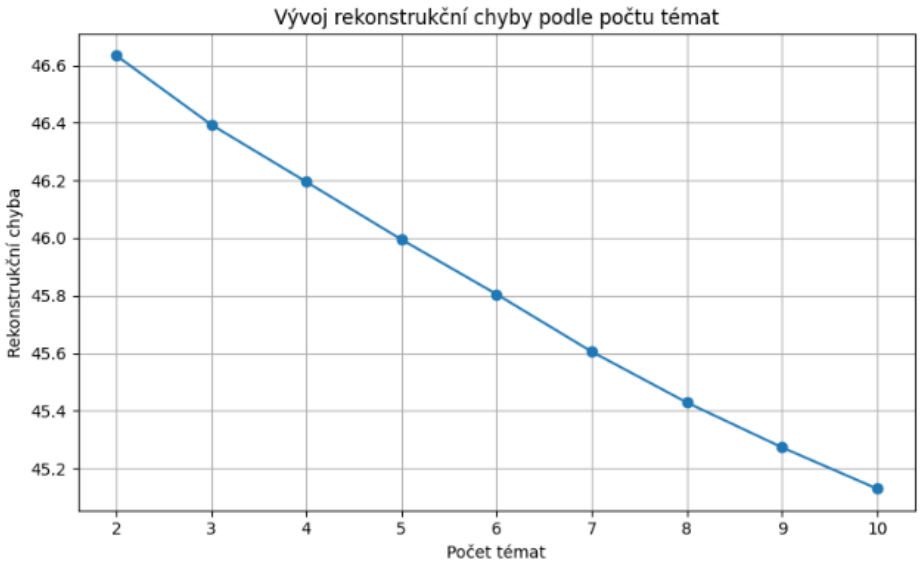


Obrázek 3: TF-IDF skóre

Zdroj: Vlastní zpracování

Tvorba témat pomocí metody NMF

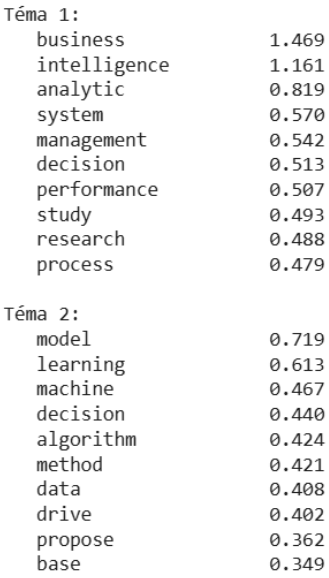
Na základě vypočtené TF-IDF matice byl proveden další krok, jehož cílem bylo určit optimální počet témat pro následné tematické modelování metodou NMF. Pro určení byla použita rekonstrukční chyba, která vyjadřuje rozdíl mezi původní maticí a jejím přiblížením modelu. Čím nižší hodnota této chyby, tím lépe model vystihuje strukturu dat. Během analýzy byla rekonstrukční chyba vypočtena pro různé počty témat v rozmezí od 2 do 10. Výsledek byl poté vizualizován, což umožnilo identifikovat bod, v němž další zvyšování počtu témat nepřináší zlepšení (viz Obrázek 4). Tento bod byl zohledněn při rozhodování a na jeho základě byl zvoleno 5 témat jako kompromis mezi srozumitelností a deskriptivní schopností modelu.



Obrázek 4: NMF – rekonstrukční chyba

Zdroj: Vlastní zpracování

Podle vybraného počtu témat, který byl stanoven pomocí rekonstrukční chyby, bylo přistoupeno k jejich vytváření pomocí metody NMF. Tato metoda umožnila odhalit skryté vzorce v textových datech a seskupit články podle podobnosti jejich slovní zásoby. Model byl nastaven tak, aby rozdělil soubor článků do pěti hlavních témat. Každé téma bylo charakterizováno sadou slov, která v daném kontextu měla nejvyšší význam. Dále byl pro každé z těchto slov zaznamenán váhový koeficient, který odráží míru důležitosti daného výrazu v kontextu konkrétního tématu. Tyto váhy umožnily lépe porozumět, jak výrazně jednotlivá slova přispívají k celkovému tematickému profilu. Pro každé téma bylo vybráno deset klíčových slov, která posloužila jako základ pro následnou interpretaci (viz Obrázek 5).



Obrázek 5: NMF – váhy slov

Zdroj: Vlastní zpracování

Po aplikaci metody NMF, kterou se rozdělil soubor článků do pěti hlavních tematických oblastí, bylo každému článku přiřazeno jedno konkrétní téma. Následoval krok, který pro každý řádek matice určil index tématu s nejvyšší vahou. Pro přehlednost byl index zvýšen o jedna, aby číslování témat začínalo od jedničky. Tímto postupem byl každému dokumentu ve finálním datasetu přiděleno nové pole, které obsahuje informaci o tématu, které nejlépe odpovídá na základě jeho textové reprezentace. Ukázka pěti náhodně vybraných článků indikuje, že byly rozřazeny do jednotlivých tematických oblastí, například druhý článek patří do Téma 4, zatímco poslední článek se nachází v Tématu 1 (viz Obrázek 6).

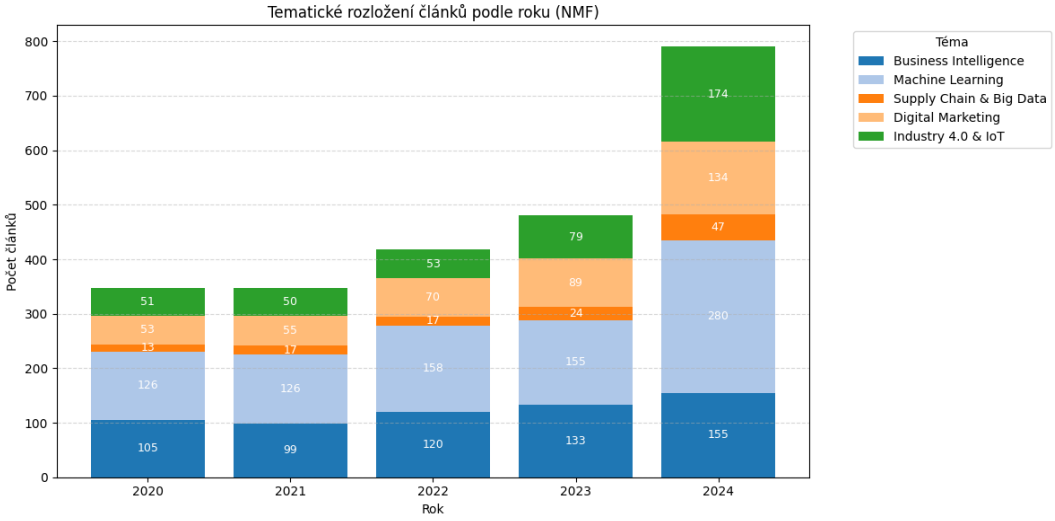
Na základě nejdůležitějších klíčových slov, která měla nejvyšší váhu, byla jednotlivým tématům přiřazena výstižná pojmenování. Téma 1 bylo nazváno jako „*Business Intelligence*“, elikož dominovala slova jako business, intelligence a analytic. Téma 2 nese název „*Machine Learning*“, díky častému výskytu výrazů jako „*learning*“, „*algorithm*“ a „*model*“. Téma 3 bylo označeno jako „*Supply Chain and Big Data*“, protože zahrnovalo výrazy jako „*supply*“, „*chain*“ a „*industry*“. Téma 4 dostalo název „*Digital Marketing*“ s ohledem na výrazy „*marketing*“, „*digital*“ a „*platform*“. Téma 5 bylo pojmenováno jako „*Industry 4.0 and IoT*“, jelikož se v něm často objevovala slova jako „*iot*“, „*technology*“ a „*smart*“.

	Title	Topic	Topic_Label
500	designing of data-driven strategies for the on...	2	Machine Learning
503	iot-based fish recommendation system: a machin...	5	Industry 4.0 & IoT
504	understanding the sharing economy: exploring c...	4	Digital Marketing
506	enhancing environmental policy decisions in ko...	2	Machine Learning
509	effect of business intelligence on organizatio...	1	Business Intelligence

Obrázek 6: Ukázka článků dle témat

Zdroj: Vlastní zpracování

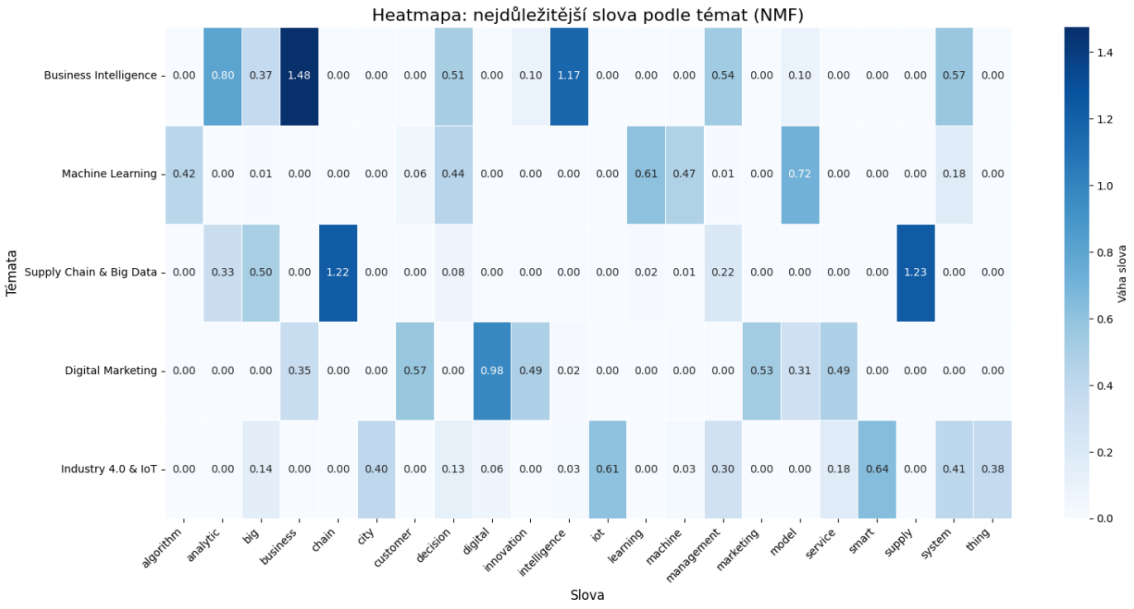
K lepšímu porozumění vývoje publikační činnosti v dané oblasti byla vytvořena vizualizace počtu článků, které byly rozděleny podle jednotlivých témat a roku. Výsledný graf (Obrázek 7) vykazuje roční aktivitu v různých tematických oblastech, které byly identifikovány pomocí metody NMF. Z daného grafu (Obrázek 7) je patrné, že celkový počet článků v daném období postupně vzrůstal. Nejvýraznější nárůst byl zaznamenán v roce 2024, kdy došlo ke zvýšení publikační aktivity ve všech tématech. Mezi hlavní témata, která dominovala v průběhu let, patřily zejména „*Business Intelligence*“ a „*Machine Learning*“. Téma „*Industry 4.0 and IoT*“ mělo nižší počet publikací, avšak v roce 2024 byl zaznamenán nárůst, což odráží rostoucí zájem tématu.



Obrázek 7: NMF – tematické rozložení podle roku

Zdroj: Vlastní zpracování

Pro hlubší porozumění obsahu jednotlivých témat byla vytvořena heatmapa (Obrázek 8), která znázorňuje význam jednotlivých slov pro každé z pěti témat, která byla identifikována pomocí metody NMF. Zobrazené hodnoty nepředstavují frekvenci výskytu slov, ale jejich významovou váhu, tedy míru, jakou jednotlivé slovo přispívá k definování konkrétního tématu.



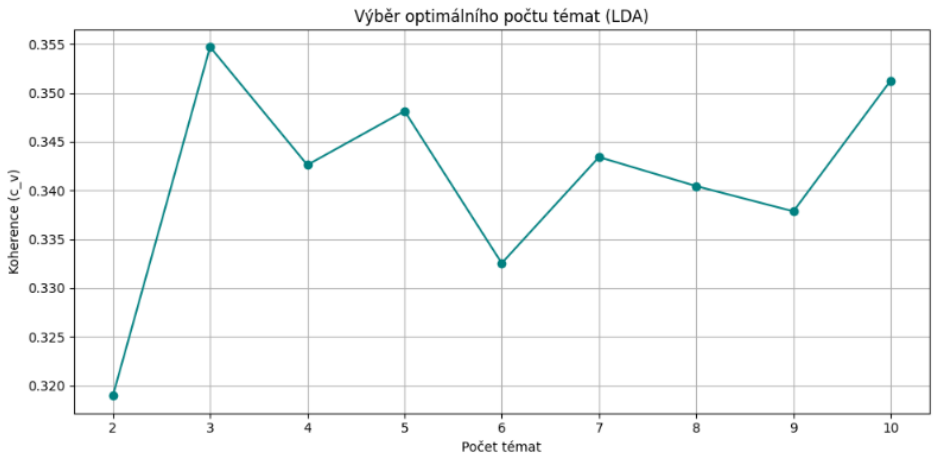
Obrázek 8: NMF – heatmapa slov podle témat

Zdroj: Vlastní zpracování

Z vytvořené heatmapy (Obrázek 8) vyplývá, že téma „*Business Intelligence"* je nejvíce formovaná slovy jako „business", „*intelligence*" a „*analytic*", která vykazují výrazně vyšší váhy než ve všech ostatních tématech. To naznačuje, že právě tato slova jsou klíčová a nejvíce ho charakterizují. Naopak „*Machine Learning*" vykazuje rozptýlenější váhové rozložení. i přesto, že se zde objevují výrazy jako „*learning*" nebo „*machine*" s vyšší váhou, jejich postavení není tak výrazné jako u předchozího tématu. To může naznačovat větší tematickou různorodost, nebo naopak, že dané téma není tak silně zaměřené na jednu konkrétní oblast, ale zahrnuje více souvisejících pojmů.

Tvorba témat pomocí metody LDA

Po provedení metody NMF, která poskytla první přehled o tématech na základě váhových vztahů mezi slovy, byla analýzy rozšířena o metodu LDA. Tato metoda vychází z předpokladu, že každý dokument je složen z kombinace různých témat. Každé téma představuje určitou skupinu slov, která se v textu objevují s konkrétní pravděpodobností. Před vytvořením modelu bylo nezbytné určit optimální počet témat. k tomu byla využita metrika koherence, která hodnotí soudržnost témat na základě shody mezi nejvýznamnějšími slovy. Koherence byla vypočtena pro různé počty témat v rozmezí od 2 do 10 a výsledky byly zobrazeny v grafu (viz Obrázek 9). Na základě dosažených hodnot byla zvolena jako nejlepší varianta tři témata, která vykazovala nejvyšší úroveň.



Obrázek 9: LDA – koherence

Zdroj: Vlastní zpracování

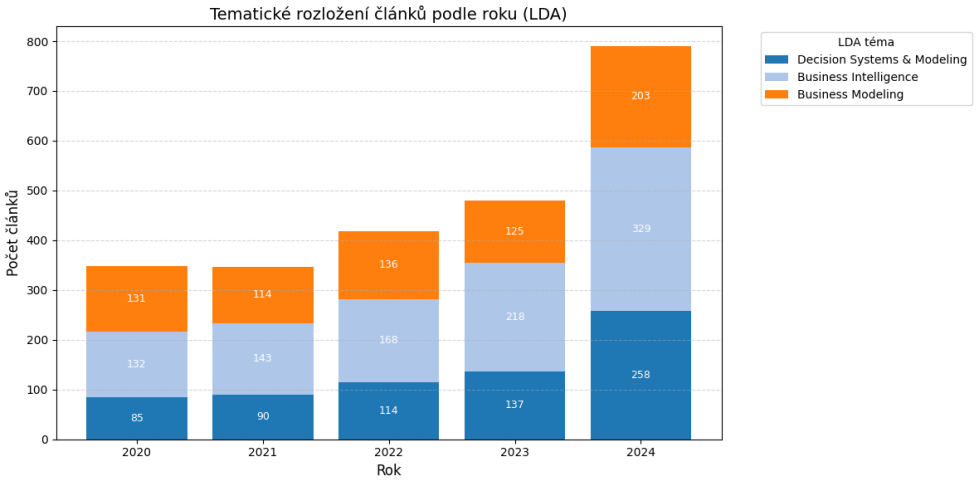
Na základě určitého počtu témat byla provedena analýza pomocí metody LDA. Model byl nastaven na tři témata (s nejvyšší koherencí) a pro každé z těchto témat bylo extrahováno deset slov s nejvyššími váhovými koeficienty, které reflektovaly význam jednotlivých výrazů v tematickém profilu (viz Obrázek 10). Každému článku bylo poté přiděleno téma s nejvyšší pravděpodobností a číslování témat bylo upraveno tak, aby začínalo od jedničky. Témata byla poté pojmenována na základě nejvýznamnějších klíčových, a to „*Decision Systems and Modeling*“, „*Business Intelligence*“ a „*Business Modeling*“, což usnadnilo interpretaci a porovnání zaměření.

```
Téma 1:
  model        0.0123
  system       0.0119
  drive        0.0112
  decision     0.0104
  base         0.0081
  propose      0.0056
  management   0.0055
  time         0.0050
  approach     0.0048
  iot          0.0047

Téma 2:
  business     0.0250
  intelligence 0.0149
  study        0.0143
  research     0.0116
  analytic     0.0107
  decision     0.0099
  management   0.0096
  model        0.0083
  system       0.0083
  drive        0.0078
```

Obrázek 10: LDA – váhy slov

Zdroj: Vlastní zpracování

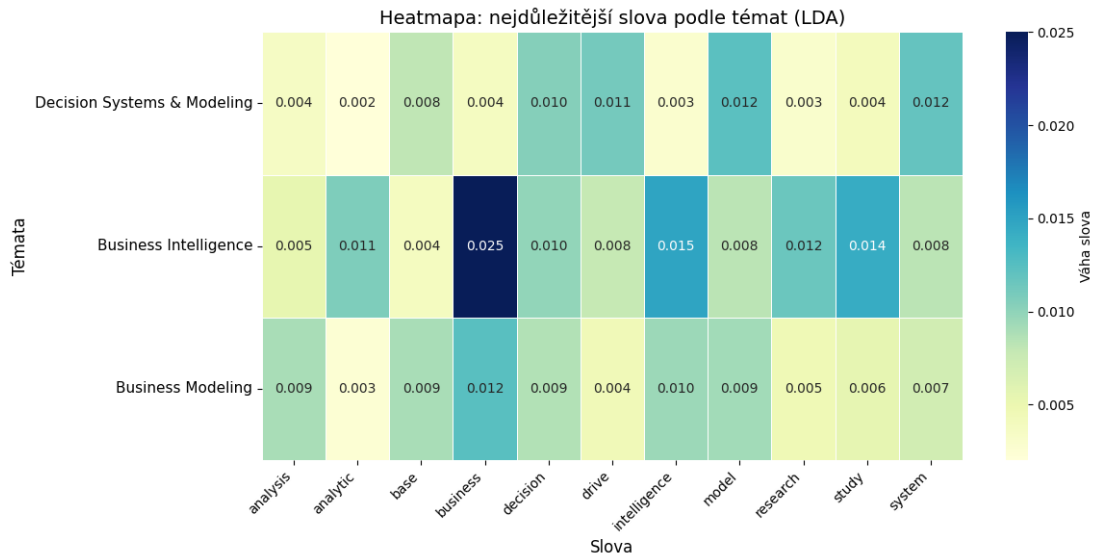
Pro tento případ byla vytvořena stejná vizualizace jako u NMF, která zobrazuje počet článků rozdělených podle témat a let. Výsledný graf (Obrázek 11) zachycuje roční aktivitu ve třech hlavních tematických oblastech, jež byly odhaleny pomocí metody LDA. Z grafu je patrné, že dominantní postavení si udržovala zejména témata „*Business Intelligence*“ a „*Business Modeling*“, zatímco nejmenší počet článků byl konzistentně zaznamenáván v oblasti „*Decision Systems & Modeling*“.



Obrázek 11: LDA – tematické rozložení podle roku

Zdroj: Vlastní zpracování

Z heatmapy (viz Obrázek 12) je patrné, že téma „*Decision Systems and Modeling*“ zahrnuj pojmy jako „*model*“, „*system*“, „*intelligence*“ a „*decision*“. Oproti tomu je téma „*Business Intelligence“* formováno slovy jako „*business*“, „*intelligence*“, „*study*“ a „*research*“. Další téma „*Business Modeling*“ vykazuje vyváženější spektrum klíčových slov, jako jsou „*business*“, „*model*“, „*base*“ a „*drive*“, což svědčí o jeho širokém tematickém záběru.



Obrázek 12: LDA – heatmapa slov podle témat

Zdroj: Vlastní zpracování

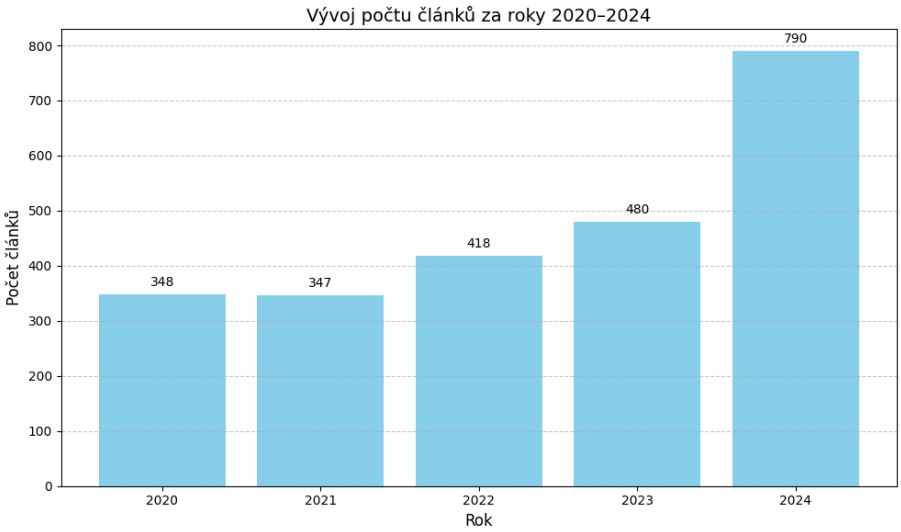
Pro provedení výše uvedených kroků byl aplikován odpovídající kód v Pythonu, který je uveden (viz **Příloha C**).

3.5 Bibliometrická analýza v Pythonu

Díky bibliometrické analýze bylo možné prozkoumat obsah těchto článků, nejcitovanější články, zjištění aktivních autorů a klíčová témata spolu s jejich vývojem v čase. Analýza byla zaměřena zejména na názvy článků, abstrakty a klíčová slova, která byla definována autory, protože právě tyto části textu nejlépe zachycovali jednotlivé poznatky daného tématu. Následující kroky jasně popisují celý proces bibliometrické analýzy, a to od importu potřebných knihoven, provedení analýzy frekvence slov, až po následnou vizualizaci.

Vývoj počtu článků za sledované období

Po sjednocení obou datasetů a provedení filtrování časového období v předchozích krocích. následovala analýza vývoje počtu článků v čase. Cílem bylo zjistit, zda se v průběhu let měnila publikační aktivita v dané oblasti, zda došlo k nárůstu či k poklesu nebo zda existovala období s větším nebo menším počtem článků. Jelikož v předchozím kroku bylo zahrnuto filtrování časového období, ke zjištění celkového trendu počtu publikací byl proveden výpočet, který spočítal počet článků ve všech jednotlivých letech. Následně byla data seřazena, což zajistilo jejich správné uspořádání. Výsledek dat byl následně vizualizován pomocí sloupcového grafu (viz Obrázek 13), kde každý sloupec znázorňoval počet článku v daném roce. Graf byl doplněn o popisky os, název a mřížku pro lepší přehlednost.



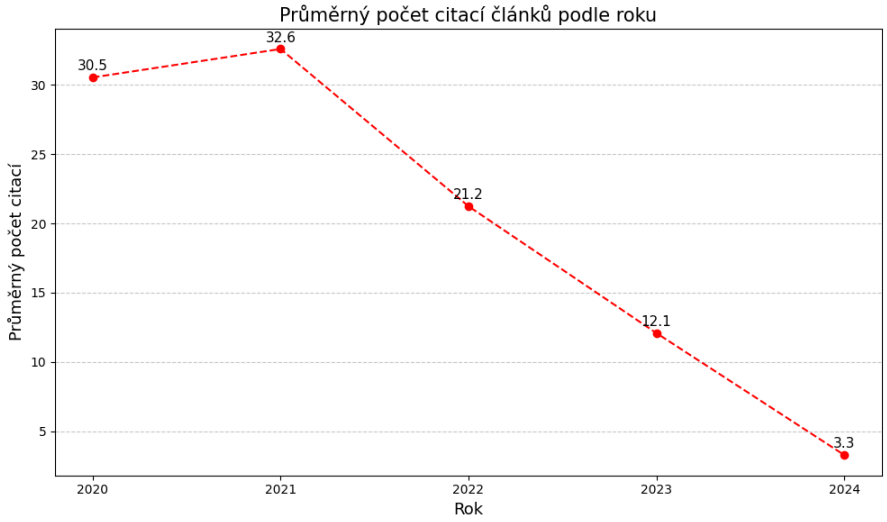
Obrázek 13: Vývoj počtu článků za rok 2020 až 2024

Zdroj: Vlastní zpracování

Dle grafu je patrný vývoj v počtu článků vztahujících se k téma data-driven business. V roce 2020 bylo publikováno 348 článků, avšak následující rok přinesl menší pokles na 347 článků. V roce 2022 došlo opět k nárůstu na 418 článků, nicméně ve všech dalších letech se aktivita opět zvýšila. V roce 2023 bylo dosaženo 480 článků a nejvýraznější nárůst se projevil v roce 2024, kdy bylo publikováno až 790 článků, což téměř dvojnásobně převyšuje počty z roku 2020. Tento trend naznačuje, že téma data-driven business získává na významu, a vědecká komunita mu věnuje zvýšenou pozornost.

Průměrný počet citací za sledované období

Po analýze vývoje článků za jednotlivé roky následovalo provedení průměrného počtu citací článků podle jednotlivých roků. Cílem bylo zjistit, jak se vliv publikovaných článků mění v průběhu času, zda starší studie získávají více citací díky delší dostupnosti, nebo naopak, jestli rostoucí počet publikovaných článků ovlivňuje průměrnou citovanost. Následně byl proveden výpočet průměrného počtu citací v jednotlivých letech. Pro lepší přehlednost byl vytvořen liniový graf, který zobrazuje vývoj průměrného počtu citací v průběhu času. Graf byl vykreslen červenou přerušovanou čárou a jednotlivé body byly opatřeny označením, což usnadňovalo sledování změn mezi jednotlivými roky. Kromě popisků os a názvů grafu byly také zahrnuty číselné hodnoty umístěné přímo nad jednotlivými body (viz Obrázek 14).



Obrázek 14: Průměrný počet citací podle roku

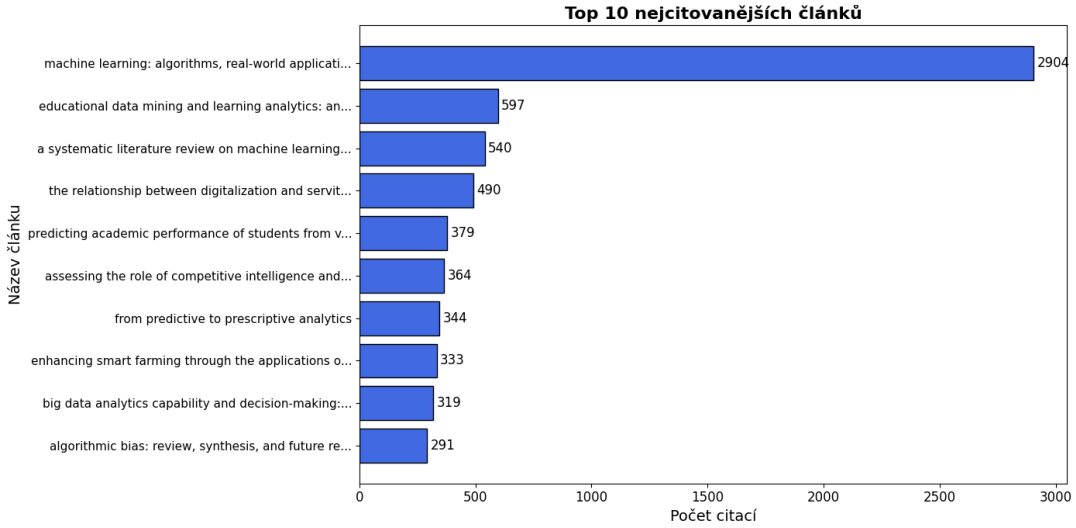
Zdroj: Vlastní zpracování

Z grafu je patrné (Obrázek 14), že počet citací postupně klesá, což je přirozené, jelikož starší články měly více času, aby byly citovány. V letech 2020 a 2021 se průměrná citovanost článků pohybovala přibližně 30 citací na článek, což naznačovalo že starší publikované články měly značný vědecký dopad. v roce 2021 došlo k nárůstu, kde tento průměr stoupl na hodnotu 32,6 citace na článek a v roce 2022 se snížil až na hodnotu 21,2. Tento trend pokračoval i v následujících letech, kdy průměr citací v roce 2023 klesl na hodnotu 12,1 a v roce 2024 dokonce na hodnotu 3,3. Toto lze vysvětlit tím, že novější články zatím nedostaly tolik příležitostí na získání citací, a zároveň s rostoucím počtem publikovaných článků se citovanost rozděluje mezi více studií. Starší články si udržují vyšší průměr, jelikož byly citovány po delší dobu, zatímco novější publikované články teprve čekají na citace.

Nejcitovanější články

Po analýze vývoje publikací a jejich průměrné citovanosti byl postup zaměřen na nejcitovanější články, tedy publikace, které měly největší dopad v rámci oblasti data-driven business. Cílem bylo identifikovat nejvýznamnější publikované články a zjistit, která témata jsou v této oblasti nejvíce sledována.

Nejprve byly články seřazeny podle počtu citací v sestupném pořadí, kde následně bylo vybráno 10 nejcitovanějších článků.



Obrázek 15: Top 10 nejcitovanějších článků

Zdroj: Vlastní zpracování

Z grafu (Obrázek 15) je patrné, že nejcitovanější články vykazují výrazné odlišnosti v počtu citací. Nejvýše postavené studie dosáhla až 2904 citací, zatímco desátý nejcitovanější článek dosáhl 291 citací, což vykazuje značný pokles citací mezi nejvýznamnějšímu publikovanými články. Nejcitovanější články se především zaměřují na témata jako je strojové učení, analýza vzdělávacích dat, digitalizace služeb a využití digitálních dvojčat. Dále se objevují studie zaměřené například predikci akademického výkonu studentů, využití konkurenční inteligence nebo přechod od prediktivní k preskriptivní analytice. Toto poukazuje, že oblast data-driven business není omezena pouze na technické aspekty, ale zasahuje i do širšího společenského a organizačního kontextu.

Analýza klíčových slov

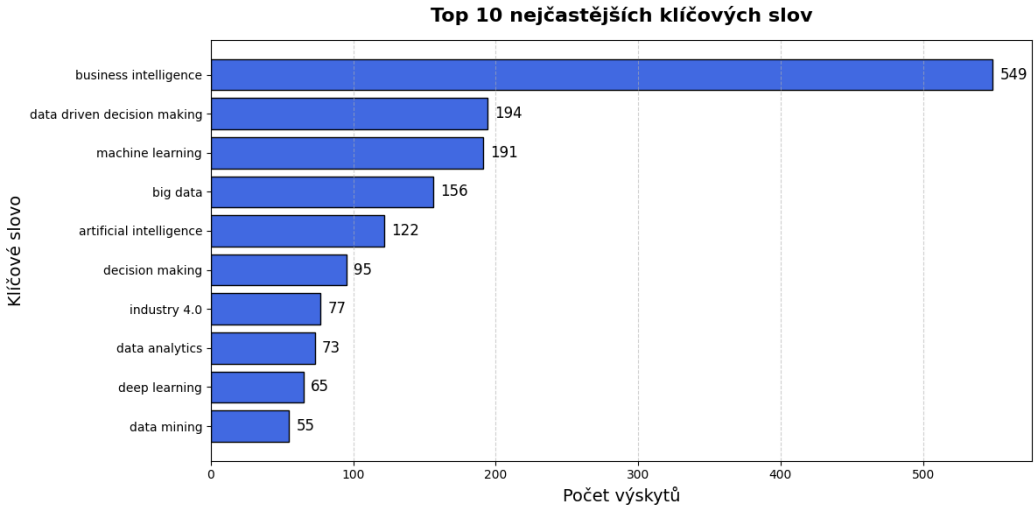
Pro analýzu bylo nezbytné importovat specifické knihovny. Pro práci s regulárními výrazy byla využita knihovna re, která umožnila efektivní rozdělení klíčových slov na základě různých oddělovačů, jako jsou čárky nebo středníky, a také jejich následnou normalizaci. Po vyčištění klíčových slov následovala jejich normalizace. Všechny výrazy byly převedeny na malá písmena a byly odstraněny mezery na začátku a konci. Klíčová slova, která byla v původních datech oddělena čárkami nebo středníky, byla rozdělena na jednotlivé výrazy, což umožnilo přesnost jejich četností. Následně došlo k odstranění speciálních znaků a sjednocení různých variant zápisu slov, například spojovníky a podtržítka byly nahrazeny mezerou, čímž byl zajištěn jednotný formát dat. Po jednotlivých úpravách následovala analýza frekvence jednotlivých klíčových slov. Dále byl vypočítán počet výskytů každého klíčového slova v datasetu a na základě těchto dat bylo vybráno 10 nejčastějších výrazů.

Pro lepší vizualizaci výsledků byla vytvořena grafika (Obrázek 16) ve formě wordcloud, která zobrazuje nejčastěji vyskytující se klíčová slova v datasetu. Čím větší je dané klíčové slovo zobrazeno, tím častěji se v datech nachází. Na základě dané grafiky je jasné, že dominuje klíčové slovo „*business intelligence*“, „*decision making*“, „*data driven*“ a další. Tento způsob vizualizace poskytuje rychlý a názorný přehled o tématech, které se v analyzovaných článcích vyskytují nejčastěji.

Obrázek 16: Nejčastější klíčová slova

Zdroj: Vlastní zpracování

Následující graf (Obrázek 17) zobrazuje deset nejčastěji se objevujících klíčových slov v daných článcích. Na prvním místě dominuje pojem „*business intelligence*“, jehož počet výskytů dosahuje 549 výskytů. Následující pojmy jako „*machine learning*“, „*data driven decision making*“, a „*big data*“, které naznačují, že vědecký zájem směřuje k propojení datové analytiky s procesem rozhodování a pokročilými technologiemi.



Obrázek 17: Top 10 nejčastějších klíčových slov

Zdroj: Vlastní zpracování

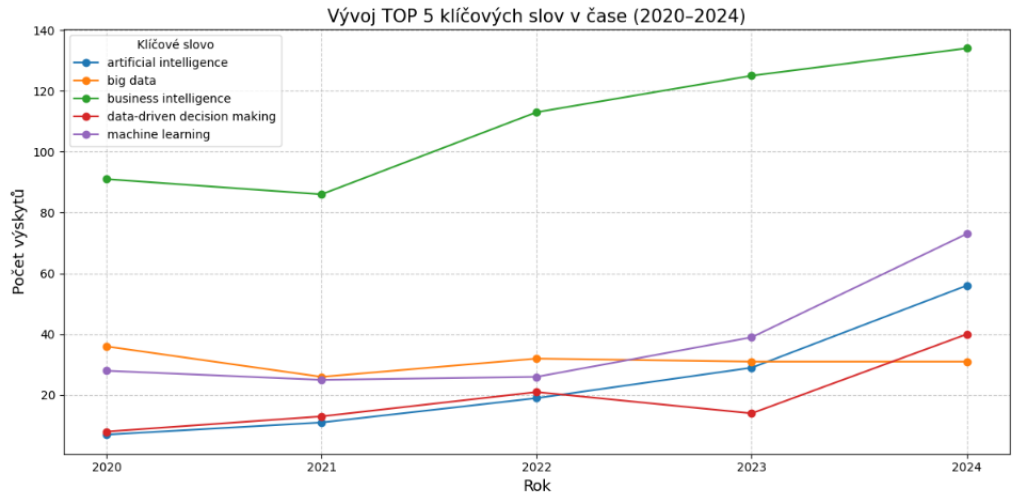
Vývoj klíčových slov v čase

Další část byla zaměřena na vývoj klíčových slov v čase, kde byly vybrány sloupce, které obsahovaly rok publikace a klíčová slova. Klíčová slova byla nejprve upravena, což zahrnovalo převod na malá písmena a odstranění nadbytečných mezer. Následně byla rozdělena pomocí regulárního výrazu podle čárek a středníků. Tento krok zajistil, že každé klíčové slovo bylo analyzováno samostatně, což umožnilo přesnější rozložení dat podle jednotlivých let.

Na základě četnosti výskytu bylo vybráno 5 nejčastějších klíčových slov, která byla dále analyzována z hlediska jejich výskytu v průběhu jednotlivých let. Každému záznamu byl přiřazen příslušný rok publikace a byla použita metoda seskupení podle roku a klíčového slova, aby byl spočítán počet výskytů. Následně byla výsledná data transformována do kontingenční tabulky, což umožnilo přehledně porovnat změny jednotlivých klíčových slov v průběhu času. Výsledky byly vizualizovány pomocí čárového grafu, který poskytl přehledný pohled na vývoj těchto pojmů v čase.

Na základě grafu (Obrázek 18) lze pozorovat, že klíčové slovo „*business intelligence*“ mělo v posledních letech jednoznačně největší výskyt. Toto klíčové slovo konstantně rostlo a v roce 2023 dosáhlo svého maxima. To naznačuje, že pojem „*business intelligence*“ zůstává v popředí zájmu a hraje klíčovou roli v oblasti data-driven business. Značný vývoj se také objevuje u „*machine learning“,* které od roku 2023 vykazuje značný nárůst. Naopak pojem „*big data*“ sice nadále udržuje stabilní pozici, přesto po svém maximu v roce 2020 začal jeho výskyt mírně klesat. Další pojmy „*artificial intelligence*“ a „*data-driven decision making*“

vykazovaly výraznější nárůst zejména v posledních letech, a to konkrétně v roce 2024, což může naznačovat, že se tyto pojmy dostaly více do popředí odborného zájmu.

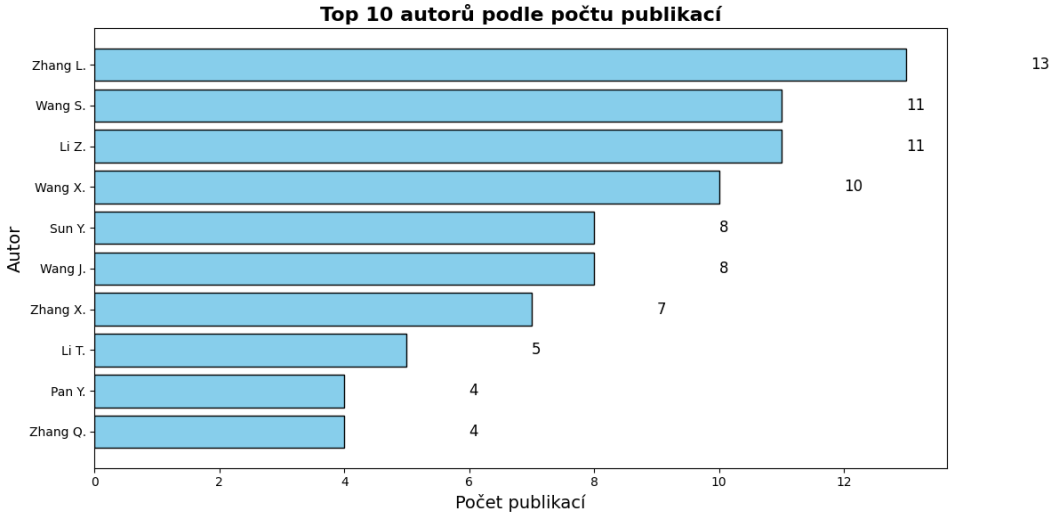


Obrázek 18: Vývoj klíčových slov v čase

Zdroj: Vlastní zpracování

Nejproduktivnější autoři

Tato část byla zaměřena na vytvoření přehledu deseti nejvýznamnějších autorů podle počtu publikací v dané oblasti. Cílem bylo identifikovat nejproduktivnější autory a vizuálně znázornit jejich hodnoty v této výzkumné oblasti. Pro analýzu byl vybrán sloupec s autory, který obsahuje seznam autorů uvedených u jednotlivých publikací. Na základě těchto informací byl sestaven slovník, který zaznamenává počet publikací pro každého autora.



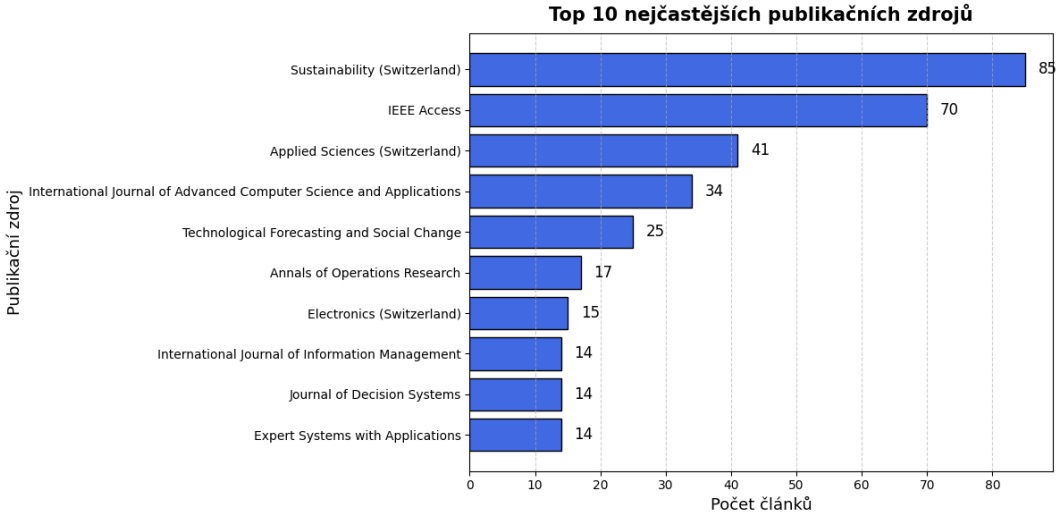
Obrázek 19: Nejproduktivnější autoři

Zdroj: Vlastní zpracování

Tento graf (Obrázek 19) zobrazuje deset nejvýznamnějších autorů v dané oblasti a období 2020 až 2024, seřazených podle počtu jejich publikací. Největší aktivitu prokázal autor „*Zhang L.*“ s 13 články, což prokazuje jeho aktivitu a vliv na tuto oblast. Dalšími autory jsou „*Wang S.*.“ a „*Li Z.*“, kteří publikovali stejný počet publikací. Poté následovali další autoři, kteří dosahovali různého či stejného počtu publikací. Tento přehled ukazuje, kteří autoři byli nejaktivnější a také nejvlivnější.

Nejčastější publikační zdroje

Po analýze nejproduktivnějších autorů následovala část zaměřená na identifikaci vědeckých časopisů, ve kterých byly články nejčastěji publikovány. Tento krok napomohl zjistit, které publikační zdroje nejvíce přispívají k šíření v rámci dané oblasti. Z celkového sjednoceného datasetu byl analyzován sloupec s názvem „*Source title*“, který zahrnuje názvy publikačních zdrojů. Následně byl proveden výpočet četnosti jednotlivých zdrojů a následně bylo vybráno deset nejčastějších. Díky přehledu byly identifikovány publikační zdroje s nejvyšší publikační aktivitou v rámci oblasti data-driven business, a tím zjistit, kde se o této oblasti nejčastěji publikuje. Získané výsledky byly vizualizovány pomocí sloupcového grafu (Obrázek 20), který přehledně zachycuje počet článků v různých publikačních zdrojích. Každý sloupec obsahuje přesnou číselnou hodnotu, což usnadňuje srovnání mezi jednotlivými zdroji a také, které publikační zdroje přispívají k šíření oblasti data-driven business.



Obrázek 20: Top 10 nejčastějších publikačních zdrojů

Zdroj: Vlastní zpracování

Graf (Obrázek 20) zobrazuje deset nejčastěji publikovaných zdrojů v oblasti data-driven business za období 2020 až 2024. Na prvním místě se umístil zdroj „*Sustainability*

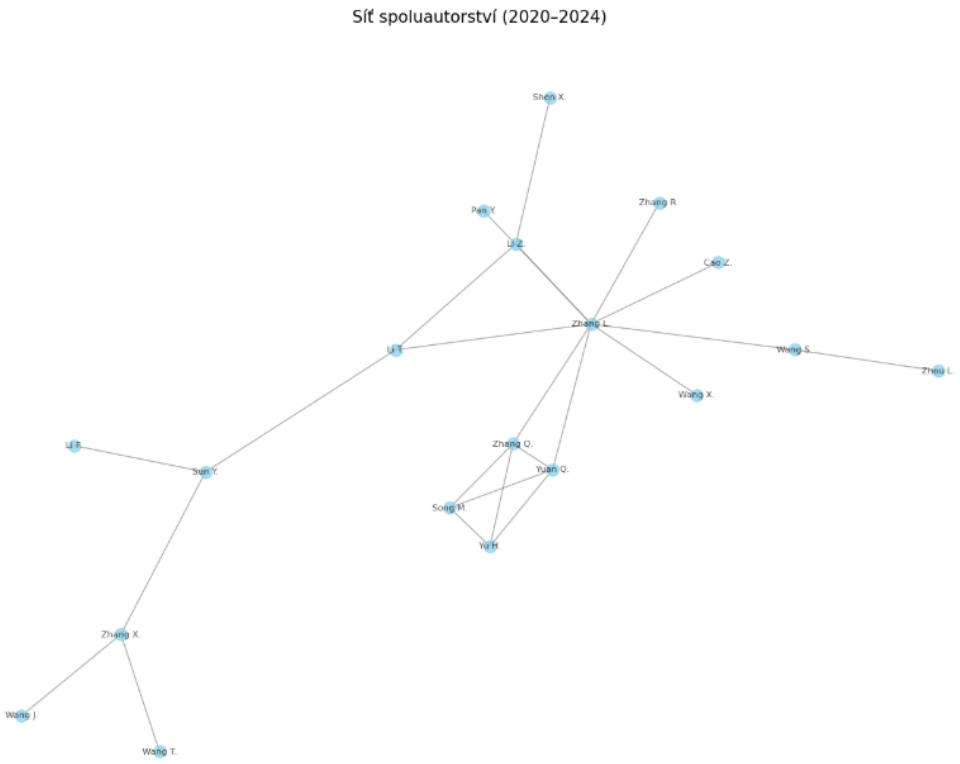
(Switzerland)“, který zahrnuje 85 článků, čímž se stal hlavním zdrojem v rámci daného datasetu. Na druhé pozici se umístil zdroj „*IEEE Access*“ s 70 články a poté zdroj „*Applied Science (Switzerland)“* s celkovým počtem článků, a to 41. Mezi další zdroje patří například „*International Journal of Advanced Computer Science and Applications*“, který publikoval 34 článků, a „*Technological Forecasting and Social Change*“ s 25 články. Umístily se zde další publikační zdroje, jejichž přítomnost v oblasti zasahuje do různorodých oblastí, a to od energetiky, přes aplikovanou matematiku a výpočetní techniku, až po otevřené multidisciplinární platformy.

Síť spoluautorství

V rámci další fáze bibliometrické analýzy byla provedena síťová analýza spoluautorství, jejímž cílem bylo identifikovat vzorce spolupráce mezi jednotlivými autory v publikacích za sledované období. Tato analýza umožnila odhalit strukturu výzkumné komunity zaměřené na oblast data-driven business a vizuálně znázornit klíčové autory spolu s jejich propojeními v této oblasti.

Z finálního datasetu byl využit sloupec s názvem „*Authors*“, který obsahuje jména všech autorů uvedených u jednotlivých publikací. Pro rozdělení těchto jmen na jednotlivé autory, bez ohledu na to, jaký oddělovač byl použit, byl použit regulární výraz. Na základě takto vytvořených seznamů autorů byla pro každou publikaci vytvořena hrana mezi všemi dvojicemi spoluautorů. Tímto způsobem každá publikace přispívá k vytvoření neorientovaného grafu, ve kterém uzly představují autory a hrany představují jejich vzájemnou spolupráci.

Dále byla vytvořena síť, která vizualizuje vztahy mezi jednotlivými uzly. Aby byla výsledná síť přehlednější a lépe odrážela významné vazby, byly dále ponechány pouze ty spoluautorské vztahy, které se v datasetu objevily minimálně dvakrát. Tímto způsobem byly vyloučeny jednorázové spolupráce, které mohly síť zahlcovat a komplikovat její interpretaci. z vytvořeného grafu byla následně extrahována největší souvislá komponenta, která představuje jádro nejaktivnějších autorů, kteří jsou propojeni silnými vazbami. Tato komponenta byla poté vizualizována a napomohla umístit uzly tak, aby graf byl přehledný a uspořádaný.



Obrázek 21: Síť spoluautorství

Zdroj: Vlastní zpracování

Výsledná graf (Obrázek 21) přehledně ukazuje, jak jsou mezi sebou propojeni autoři na základě jejich společných publikací. Tato vizualizace usnadňuje identifikaci úzce spolupracujících výzkumných týmů i klíčových autorů, kteří fungují jako spojovací článek mezi různými skupinami.

Výpočet a výstup metrik

Zde byly provedeny výpočty a výstupy dvou klíčových metrik pro hodnocení autorů v dané oblasti, a to h-index a g-index. Tyto metriky byly vypočítány pro autory, kteří jsou součástí největší souvislé komponenty spoluautorství, což znamená, že se jedná o autory s výraznou mírou spolupráce. Poté byl stanoven h-index pomocí funkce, která určuje, kolik článků bylo citováno minimálně tolikrát, kolik jich je. Následující g-index byl vypočítán, že byl ověřen součet citací u počátečního počtu článků. Tento index zohledňuje nejen samotný počet citací, ale také intenzitu citování těch nejvíce citovaných článků.

	Autor	Publikace	Celkový počet citací	h-index	g-index
0	Wang S.	11	154	7	11
1	Li Z.	11	312	7	11
2	Zhang X.	7	203	7	7
3	Wang J.	8	102	6	8
4	Zhang L.	13	220	6	13
5	Wang X.	10	98	6	9
6	Zhang Q.	4	203	3	4
7	Wang T.	3	112	3	3
8	Sun Y.	8	63	3	7
9	Song M.	3	135	3	3

Obrázek 22: Výstup metrik

Zdroj: Vlastní zpracování

Dle výstupu (Obrázek 22) bylo zobrazeno 10 nejvýznamnějších autorů v dané oblasti a jejich příslušné hodnoty. Na prvním místě se nachází autor „*Wang S.*“ s 11 publikacemi a 154 citacemi. Jeho h-index dosahuje hodnoty 7, což znamená, že nejméně 7 jeho publikací bylo citováno minimálně sedmkrát, a g-index dosahuje hodnoty 11. Dále autor „Li Z.“ s rovněž 11 publikacemi, ale vyšším počtem citací. Ostatní autoři dosahují h-indexu v rozmezí 3 až 7 a g-indexu v rozmezí 3 až 11, což vykazuje významnou publikační aktivitu a vliv v oblasti.

Veškeré předchozí kroky byly aplikovány pomocí kódu v Pythonu, který je uveden (viz **Příloha D**).

3.6 Omezení použitých metod a postupů

Ačkoli při použití systematického přístupu metodiky PRISMA a také textové a bibliometrické analýzy bylo možné poskytnout podrobný pohled na zkoumané téma, je důležité upozornit na určitá omezení, která mohou ovlivnit interpretaci výsledků.

Klasifikace dokumentů

Do analýzy byly zahrnuty pouze články typu „*Article*“ a „*Review*“. Konferenční příspěvky či jiné typy dokumentů byly vyloučeny z důvodu nízké četnosti v rámci datasetu, ale také kvůli obavám o jejich srovnatelnost a odbornou kvalitu, přičemž podobný postup je v bibliometrických analýzách obvykle preferovány recenzované vědecké články a přehledové studie. [68] Takové rozhodnutí může mít dopad na celkový přehled o dané oblasti.

Výběr databází a jazyk

Analýza byla omezena na záznamy z databází Scopus a Web of Science a na články psané v angličtině. Některé relevantní studie, které se nacházejí mimo tyto zdroje, tak nemusely být zahrnuty.

Shodná jména autorů

V některých případech, jako například u čínských příjmení Wang, Zhang, Li a dalších, se může pod jedním jménem skrývat více různých osob. Automatizované nástroje nedokážou vždy tuto skutečnost přesně odlišit. Vzhledem k tomu, že autoři mohou měnit své afiliace nebo používat různé varianty jména, není možné tento problém vždy spolehlivě vyřešit. Výsledky se proto v některých případech týkajících se autorství mohou mírně zkreslit.

Více autorů na článek

Jeden článek může mít několik autorů, avšak analýzy obvykle nerozlišují jejich individuální přínos. Nebyla použita vážená metrika, neboť tyto informace nejsou dostupné v původních datech.

Afiliace a formát dat

Formát dat o institucích se mezi různými databázemi liší a postrádá jednotnost. To komplikuje přesné zjištění spoluprací nebo přiřazení autorů ke konkrétním institucím.

Nerovnoměrné počty článků publikované ve zdrojích

Výsledky mohou dále ovlivnit i časopisy, které ročně publikují stovky a tisíce článků, a mají tak výhodu oproti časopisům, které ročně publikují jen desítky článků. S tím samozřejmě také souvisí i kvalita obsahu těchto článků, která ale nebyla cílem provedené analýzy.

Textová analýza

Metody jako TF-IDF, NMF nebo LDA se opírají o frekvenční analýzu slov, aniž by braly v úvahu jejich konkrétní význam v daném kontextu. Je proto nezbytné výsledky interpretovat s určitou rezervou a interpretovat je v kontextu odborných znalostí dané oblasti.

4 IMPLEMENTACE A VYTVOŘENÍ DASHBOARDU POMOCÍ POWER BI

Tato část byla zaměřena na převedení výsledků textové a bibliometrické analýzy do přehledné vizualizační podoby. Pro tento účel byl zvolen nástroj Microsoft Power BI, který umožňuje kombinovat různé datové zdroje a vytvářet grafy, filtry a interaktivní reporty. Výsledkem bylo vytvoření dashboardu, která nejen prezentuje výstupy z praktické části, ale také usnadňuje jejich porozumění.

4.1 Použité výstupy

Pro tvorbu dashboardu byly použité výstupy, které byly získány v předchozí části této práce. Konkrétně se jednalo o výsledky textové a bibliometrické analýzy, které poskytly přehled o publikační činnosti, struktuře dokumentů a jazykových vzorcích v analyzovaných textech. Jednotlivé výstupy textové analýzy byly exportovány do formátu XLSX pomocí provedeného kódu (viz **Příloha C**). Stejný způsob byl aplikován i pro výstupy bibliometrické analýzy (viz **Příloha D**). Následně získané soubory byly importovány do Power BI. Sloužily jako datový základ pro vytváření vizuálních prezentací a interaktivní zobrazení dosažených výsledků.

4.2 Struktura a komponenty dashboardu

Na základě výsledků textové a bibliometrické analýzy byly navrženy dva přehledné dashboardy. Hlavním cílem vytvoření bylo zajistit, aby vizualizace zahrnovala hlavní podstatné poznatky z analytické části práce a současně poskytovala uživatelský přívětivý a efektivní způsob interpretaci dat. Struktura dashboardů byla navržena přehledně a logicky, aby jednotlivé komponenty odpovídaly jednotlivým oblastem a byly snadno čitelné i pro uživatele bez hlubší znalosti datové analýzy. Pro lepší přehlednost byly v dashboardech implementovány základní filtry a provedeny drobné úpravy dat jako například úprava názvů sloupců, formáty a další pro jasnější a unikátní vizualizaci.

Dashboard textové analýzy

Tento dashboard byl sestaven tak, aby nabízel rychlý přehled o výsledcích textové analýzy. Na levé straně byly umístěny dvě tabulky, které zobrazují váhy klíčových slov podle dvou hlavních metod, a to NMF a LDA. Každá z těchto tabulek zahrnuje název sloupce, a to slovo, téma a váhu. Sloupec s váhami vykazuje důležitost jednotlivých slov v rámci přiřazeného tématu (viz Obrázek 23). Dále byly pravém horním rohu umístěny tři informační bloky, počet článků,

časové období a nejčastěji objevující se klíčové slovo. Pod těmito bloky byl umístěn horizontální sloupcový graf, který zobrazuje nejčastější klíčová slova pomocí TF-IDF.

NMF - váhy slov

Slovo	Téma	Váha
analytic	Supply Chain & Big Data	0,33
bda	Supply Chain & Big Data	0,20
big	Supply Chain & Big Data	0,50
capability	Supply Chain & Big Data	0,16
data	Supply Chain & Big Data	0,22
firm	Supply Chain & Big Data	0,15
chain	Supply Chain & Big Data	1,22
industry	Supply Chain & Big Data	0,20
management	Supply Chain & Big Data	0,22
supply	Supply Chain & Big Data	1,23
algorithm	Machine Learning	0,42
base	Machine Learning	0,35
data	Machine Learning	0,41
decision	Machine Learning	0,44
drive	Machine Learning	0,40
learning	Machine Learning	0,61
machine	Machine Learning	0,47
method	Machine Learning	0,42
model	Machine Learning	0,72
propose	Machine Learning	0,36
city	Industry 4.0 & IoT	0,40
internet	Industry 4.0 & IoT	0,37
iot	Industry 4.0 & IoT	0,61

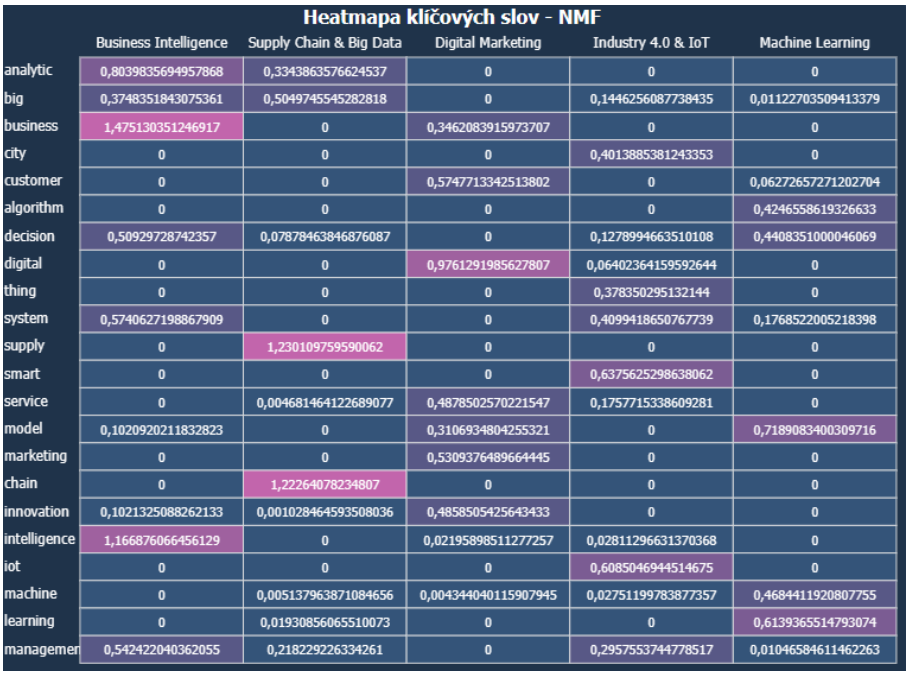
LDA - váhy slov

Slovo	Téma	Váha
business	Business Intelligence	0,03
intelligence	Business Intelligence	0,01
study	Business Intelligence	0,01
business	Business Modeling	0,01
model	Decision Systems & Modeling	0,01
system	Decision Systems & Modeling	0,01
research	Business Intelligence	0,01
drive	Decision Systems & Modeling	0,01
analytic	Business Intelligence	0,01
decision	Decision Systems & Modeling	0,01
decision	Business Intelligence	0,01
management	Business Intelligence	0,01
intelligence	Business Modeling	0,01
model	Business Modeling	0,01
base	Business Modeling	0,01
analysis	Business Modeling	0,01
decision	Business Modeling	0,01
model	Business Intelligence	0,01
system	Business Intelligence	0,01
base	Decision Systems & Modeling	0,01
drive	Business Intelligence	0,01
information	Business Modeling	0,01

Obrázek 23: Ukázka textové analýzy – váhy

Zdroj: Vlastní zpracování

Druhá část zahrnovala výsledky tematického modelování dvou metod NMF a LDA. Na levé straně byla umístěna vizualizace článků uspořádaná podle pěti NMF témat, kde každý daný pruh byl označen počtem článků za konkrétní rok a téma. Vedle této vizualizace byla znázorněna heatmapa klíčových slov v rámci NMF témat (viz Obrázek 24), která barevně a přesně vyjadřovala váhu každého slova v souvislosti s daným tématem. Stejný postup byl uplatněn i ve spodní části, kde byly prezentovány další vizualizace pro LDA. Toto uspořádání umožnilo snadno porovnat obě metody a odhalit doplňující i jedinečné výsledky, které byly získány (viz **Příloha E**).



Heatmapa klíčových slov - NMF

	Business Intelligence	Supply Chain & Big Data	Digital Marketing	Industry 4.0 & IoT	Machine Learning
analytic	0,8039835694957868	0,3343863576624537	0	0	0
big	0,3748351843075361	0,5049745545282818	0	0,1446256087738435	0,01122703509413379
business	1,475130351246917	0	0,3462083915973707	0	0
city	0	0	0	0,4013885381243353	0
customer	0	0	0,5747713342513802	0	0,06272657271202704
algorithm	0	0	0	0	0,4246558619326633
decision	0,50929728742357	0,07878463846876087	0	0,1278994663510108	0,4408351000046069
digital	0	0	0,9761291985627807	0,06402364159592644	0
thing	0	0	0	0,378350295132144	0
system	0,5740627198867909	0	0	0,4099418650767739	0,1768522005218398
supply	0	1,230109759590062	0	0	0
smart	0	0	0	0,6375625298638062	0
service	0	0,004681464122689077	0,4878502570221547	0,1757715338609281	0
model	0,1020920211832823	0	0,3106934804255321	0	0,7189083400309716
marketing	0	0	0,5309376489664445	0	0
chain	0	1,22264078234807	0	0	0
innovation	0,1021325088262133	0,001028464593508036	0,4858505425643433	0	0
intelligence	1,166876066456129	0	0,02195898511277257	0,02811296631370368	0
iot	0	0	0	0,6085046944514675	0
machine	0	0,005137963871084656	0,004344040115907945	0,02751199783877357	0,4684411920807755
learning	0	0,01930856065510073	0	0	0,6139365514793074
managemer	0,542422040362055	0,218229226334261	0	0,2957553744778517	0,01046584611462263

Obrázek 24: Ukázka textové analýzy – heatmapa slov

Zdroj: Vlastní zpracování

Dashboard bibliometrické analýzy

V rámci tohoto dashboardu byly v horní části zvoleny a umístěny klíčové ukazatele, které poskytují rychlý přehled o rozsahu zpracovaných dat. Mezi zvolené ukazatele patří celkový počet analyzovaných článků, časové období a nejvyšší počet citací. Tyto ukazatele slouží jako vstupní informace, které umožňují uživateli rychle porozumět kontextu analýzy. Samostatné vizuály vycházely ze získaných výsledků jako je vývoj publikační aktivity v jednotlivých letech (viz Obrázek 25), průměrný počet citací v závislosti na roce nebo rozložení klíčových slov. Následující část zahrnuje další výstupy.

Obrázek 25: Ukázka bibliometrické analýzy – vývoj počtu článků

Zdroj: Vlastní zpracování

V poslední části dashboardu byly zahrnuty vizualizace, které rozšiřují celkový přehled o struktuře autorství a použitých klíčových slov v analyzovaných článcích. Výskyt klíčových pojmů byl zobrazen pomocí wordcloudu (viz Obrázek 26), kde velikost jednotlivých slov vizuálně odráží jejich frekvenci. Druhá vizualizace zobrazuje síť vztahů mezi autory na základě společných publikací. Tato síť spoluautorství poskytuje pohled do míry spolupráce, identifikuje propojené skupiny autorů a zvýrazňuje autory s publikační aktivitou v rámci daného období (viz **Příloha F**).

Obrázek 26: Ukázka bibliometrické analýzy – wordcloud

Zdroj: Vlastní zpracování

ZÁVĚR

Cílem této diplomové práce bylo provést analýzu vědeckých textů na téma data-driven business s využitím textové analytiky a dalších souvisejících metod.

V první kapitole byl vymezen pojem data-driven business, objasněn jeho význam a představeny související koncepty. Ve druhé kapitole byl popsán systematický postup analýzy textů, v němž byl definován rámec metodiky PRISMA. Jako další byly navrženy a otestovány pokročilé dotazy v databázích Web of Science a Scopus s využitím filtrů, popsány kroky předzpracování textu a vysvětleny metody TF-IDF a Bag of Words. Pro tematické modelování byly představeny metody NMF s optimalizací podle rekonstrukční chyby a LDA s hodnocením koherence. Doplněna byla bibliometrická analýza zahrnující fáze v rámci jejího provedení a popsaný jednotlivé metody. Ve třetí kapitole byl tento postup implementován v Pythonu, kde proběhlo předzpracování, vektorizace textů, tematické modelování a výpočty bibliometrických metrik.

Na základě získaných výsledků bylo možné zodpovědět na výzkumné otázky. Pro zodpovězení (**O1**) bylo provedeno předzpracování textů pomocí jednotlivých kroků, čímž vznikl čistý korpus. Na tento korpus byla následně aplikována metoda TF-IDF, která odhalila, že vyplynulo, že pojmy „*business*“, „*model*“, „*decision*“, „*intelligence*“ a „*data*“ dosáhly nejvyšších hodnot. Doplněná analýza četnosti a vizualizace wordcloud, pak zvýraznily, že mezi nejběžnější slova patří „*business intelligence*“, „*data driven decision making*“, „*machine learning*“, „*big data*“ a „*artificial intelligence*“. (**O2**) byla zaměřena na průměrný počet citací podle roku vydání. Analýza odhalila, že nejvyšší citovanost vykázaly publikace z roku 2021, těsně následované rokem 2020. V následujících letech pak průměr citací postupně klesal na hodnotu 21,2 v roce 2022, 12,1 v roce 2023 a 3,3 v roce 2024. Tento výsledek odráží to, že novější články dosud neměly dostatek času na citování. Pro zodpovězení (**O3**) byly všechny články seřazeny sestupně podle celkového počtu citací. Z této analýzy vyplynulo, že nejvýraznější dopad měla práce zabývající se strojovým učením a algoritmy, která dosahovala hodnoty až 2 904. Pro zodpovězení (**O4**) bylo použito tematické modelování pomocí NMF, které ukázalo, že v letech 2020 až 2021 převládalo téma „*Business Intelligence*“, zatímco od roku 2022 začalo silně růst „Machine Learning“ a zájem o „*Supply Chain & Big Data*“, „*Digital Marketing*“ a „*Industry 4.0 & IoT*“. Dále byla provedena analýza metodou LDA, která vykazovala, že mezi lety 2020 a 2024 se z původního zaměření na „*Decision Systems & Modeling“* postupně přesunul hlavní důraz na „*Business Intelligence*“ a dále na „*Business Modeling*“. Název každého tématu byl odvozen podle nejvýznamnějších slov, která se v daném tématu vyskytovalo. V rámci (**O5**) bylo analyzováno, kolik

článků každý autor přispěl mezi lety 2020 a 2024. Na prvním se umístil autor „*Zhang L.*“ s 13 publikacemi, hned za ním následovali „*Wang S.*“ a „*Li Z.“* se 11 příspěvky, dále „*Wang X.*“ s deseti články a několik dalších autorů s osmi až devíti články. Tento výsledek tak odhalil, kteří autoři patřili mezi nejvíce aktivní v oblasti. Pro zodpovězení (**O6**) byla nejprve spočítána četnost výskytu jednotlivých časopisů. Z těchto výsledků pak vyplynulo, že nejvíce publikací vyšlo pod „*Sustainability (Switzerland)*“, těsně následoval publikační zdroj „*IEEE Access*“, přičemž zbývající publikace se dělily mezi odborné periodikum zaměřená na různé obory. Pro získání odpovědi na (**O7**) byla sestavena síť spoluautorství, ve které zůstala jen propojení autorů se dvěma a více společnými články. Po odečtení slabších vazeb se síť spoluautorství rozdělila do tří hlavní klusterů. Nejsilnější seskupení se utvořilo kolem „*Zhanga L.*“, dále skupina „Sun Y.“ s „*Zhangem X.*“ a menší skupinu autorů „*Song M.*“, „*Yu H.*“ a „*Zhang Q.*“. Tento výstup ukázal, kdo tvořil jádro vědecké spolupráce v dané oblasti. Pro získání výsledků na (**O8**) byly nejprve vypočteny hodnoty h-indexu a g-indexu u autorů z hlavních klastrů, přičemž h-index se pohyboval mezi hodnotou 3 až 7 a g-index mezi hodnotou 3 až 13. Nejlepšími výsledky se vyznačovali „*Wang S.*“ a „*Li Z.*“, dále například „*Zhang X.*“ a „*Zhang L.*“, který dosáhl nejvyššího g-indexu 13 (viz Obrázek 22). Právě tito autoři byli díky svým metrikám identifikováni jako nejvlivnější autoři. Výzkumné otázky definované v kapitole 3.2 byly zodpovězeny.

Ve čtvrté kapitole byly výsledky exportovány do Power BI. Následně byly sestaveny jednotlivé dashboardy, které zahrnovaly veškeré hlavní výstupy z provedených analýz. Tím byl poskytnut ucelený přehled o celé provedené analýzy.

Hlavním přínosem práce je sestavení a praktická aplikace komplexního postupu, který propojil systematický postup, textová a bibliometrická analýza. Díky tomu je možné získat ucelený přehled například o klíčových oblastech, častých tématech i nejvlivnějších autorech v oblasti data-driven business a získané výsledky prezentovat v jednotlivých dashboardech. Na základě provedené analýzy a dosažených výsledků lze konstatovat, že cíl této diplomové práce byl naplněn.

SEZNAM POUŽITÉ LITERATURY

[1] KOTOROV, Rado. *Data-Driven Business Models for the Digital Economy*. New York: Business Expert Press, 2020. ISBN 978-1-95152-780-1.

[2] TRIPATHI, Shailesh; BACHMANN, Nadine; BRUNNER, Manuel a JODLBAUER, Herbert. Preparedness for Data-Driven Business Model Innovation: A Knowledge Framework for Incumbent Manufacturers. *Applied Sciences* [online]. 2024, roč. 14, č. 8, čl. 3454. ISSN 2076-3417. Dostupné z: https://doi.org/10.3390/app14083454 [cit. 2025-01-05].

[3] FIETL, Erwin; DESOUZA, Kevin C.; GABLE, Guy a WESTERVELD, Peter**.** Data-Driven Business Models and Professional Services Firms: A Strategic Framework and Transitionary Pathways. In: *Lecture Notes in Business Information Processing*. Santa Clara: Springer Verlag, 2019, sv. 357, s. 26–38. ISBN 978-3-030-22783-8. Dostupné z: https://doi.org/10.1007/978-3-030-22784-5_3. [cit. 2025-01-06].

[4] HARTMANN, P.M.; ZAKI, M.; FELDMANN, N. a NEELY, A., 2014 *Capturing Value from Big Data – A Taxonomy of Data-Driven Business Models used by Start-Up Firms* [online]. Cambridge: University of Cambridge, Cambridge Service Alliance, Dostupné z: https://cambridgeservicealliance.eng.cam.ac.uk/system/files/documents/2014_March_DataDrivenBusinessModels.pdf.pdf. [cit. 2025-02-15].

[5] OLSON, Karin. What Are Data? *Qualitative Health Research* [online]. 2021, roč. 31, č. 9, s. 1567–1569. Dostupné z: https://doi.org/10.1177/10497323211015960 [cit. 2025-01-05].

[6] BADMAN, Annie a KOSINSKI, Matthew. *What is data?* Online. Ibm.com. 2024. Dostupné z: https://www.ibm.com/think/topics/data. [cit. 2025-01-08].

[7] TIMONERA, Kaye. *What Is Quantitative Data? Characteristics & Examples*. Online. Datamation.com. C2023. Dostupné z: https://www.datamation.com/big-data/what-is-quantitative-data/. [cit. 2025-01-08].

[8] TIMONERA, Kaye. *What Is Qualitative Data? Characteristics & Examples*. Online. Datamation.com. 2024. Dostupné z: https://www.datamation.com/big-data/what-is-qualitative-data/. [cit. 2025-01-08].

[9] GEEKSFORGEEKS. *Difference between Structured, Semi-structured and Unstructured data*. Online. Geeksforgeeks.org. 2023. Dostupné z: https://www.geeksforgeeks.org/difference-between-structured-semi-structured-and-unstructured-data/. [cit. 2025-01-10].

[10] BADMAN, Annie; KOSINSKI, Matthew. *What is Metadata?* Online. Ibm.com. 2024. Dostupné z: https://www.ibm.com/think/topics/metadata. [cit. 2025-01-10].

[11] MAYERNIK, Matthew S. Metadata. *Knowledge Organization [online]*, 2020, roč. 47, č. 8, s. 696–713. ISSN 0943-7444. Dostupné z: https://doi.org/10.5771/0943-7444-2020-8-696. [cit. 2025-01-13].

[12] BADMAN, Annie a KOSINSKI, Matthew. *What is Big Data?* Online. Ibm.com. 2024. Dostupné z: https://www.ibm.com/think/topics/big-data. [cit. 2025-01-13].

[13] ŠEBALJ, Darjan; ŽIVKOVIĆ, Andreja a HODAK, Kristina. Big Data: Changes in Data Management. *Ekonomski vjesnik/Econviews – Review of Contemporary Business, Entrepreneurship and Economic Issues* [online]. 2016, roč. 29, č. 2, s. 487–499. ISSN 1848-9635. Dostupné z: https://hrcak.srce.hr/ojs/index.php/ekonomski-vjesnik/article/view/4208. [cit. 2025-01-13].

[14] MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

[15] HAMBARDE, K. A. a PROENÇA, H. Information Retrieval: Recent Advances and Beyond. *IEEE Access* [online]. 2023, roč. 11, s. 76581–76604. Dostupné z: https://doi.org/10.1109/ACCESS.2023.3295776. [cit. 2025-01-15].

[16] GARLA, Satish. *Text Mining and Analysis*. Cary: SAS Institute, 2013. ISBN 978-1-61290-551-8.

[17] ISLAM, Mohaiminul. Data Analysis: Types, Process, Methods, Techniques and Tools. *International Journal on Data Science and Technology* [online]. 2020, roč. 6, č. 1, s. 10–15. ISSN 2472-2236. Dostupné z: https://doi.org/10.11648/j.ijdst.20200601.12. [cit. 2025-01-15].

[18] ALEM, Dawit Dibekulu. An Overview of Data Analysis and Interpretations in Research [online]. *International Journal of Academic Research in Education and Review*, 2020, roč. 8, č. 1, s. 1–27. ISSN 2360-7866. Dostupné z: http://www.academicresearchjournals.org/IJARER/Index.htm. [cit. 2025-01-15].

[19] TAHERDOOST, Hamed. Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects. *International Journal of Academic Research in Management* [online]. 2022, roč. 9, č. 1, s. 1–9. Dostupné z: https://ssrn.com/abstract=4178680. [cit. 2025-01-18].

[20] PROVOST, Foster; FAWCETT, Tom. *Data Science for Business*. Sebastopol: O'Reilly Media, 2013. ISBN 978-1-4493-6132-7.

[21] SHIMAOKA, A. M.; FERREIRA, R. C. a GOLDMAN, A. The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks. *SBC Reviews on Computer Science* [online]. 2024, roč. 4, č. 1, s. 28–43. Dostupné z: https://doi.org/10.5753/reviews.2024.3757. [cit. 2025-01-18].

[22] MUJTHABA, G.M.; AL AMEEN, Abdalla; KOLHAR, Manjur; RAHMATH, Mohammed**.** Data Science Techniques, Tools and Predictions. *International Journal of Recent Technology and Engineering (IJRTE)* [online]. 2020, roč. 8, č. 6, s. 5661–5668. ISSN 2277-3878. Dostupné z: https://doi.org/10.35940/ijrte.F9887.038620. [cit. 2025-01-15].

[23] HAQ, Hafiz Burhan Ul; KAYANI, Haroon Ur Rashid; TOOR, Saba Khalil; ZAFAR, Sadia a KHALID, Imran. The Popular Tools Of Data Sciences: Benefits, Challenges and Applications. *IJCSNS International Journal of Computer Science and Network Security* [online]. 2020, roč. 20, č. 5, s. 64–70.

[24] STAHL, Bernd; HÄCKEL, Björn; LEUTHE, Daniel a RITTER, Christoph. Data or Business First? — Manufacturers' Transformation Toward Data-driven Business Models [online]. *Schmalenbach Journal of Business Research*, 2023, roč. 75, č. 3, s. 303–343. ISSN 2366-6153. Dostupné z: https://doi.org/10.1007/s41471-023-00154-2. [cit. 2025-02-19].

[25] JAHAN, Nusrat; NAVEED, Sadia; ZESHAN, Muhammad a TAHIR, Muhammad A. How to Conduct a Systematic Review: a Narrative Literature Review [online]. *Cureus*, 2016, roč. 8, č. 11, e864. Dostupné z: https://doi.org/10.7759/cureus.864. [cit. 2025-01-19].

[26] PAUL, Justin; LIM, Weng Marc; O'CASS, Aron; HAO, Andy Wei a BRESCIANI, Stefano. Scientific Procedures and Rationales for Systematic Literature Reviews (SPAR-4-SLR). *International Journal of Consumer Studies* [online]. 2021, roč. 45, č. 4, s. 1–16. ISSN 1470-6431. Dostupné z: https://doi.org/10.1111/ijcs.12695. [cit. 2025-01-20].

[27] SATALOFF, Robert T.; BUSH, Matthew L. a CHANDRA, Rakesh, et al. Systematic and other reviews: Criteria and complexities. *Journal of Otolaryngology – Head & Neck Surgery* [online]. 2021, roč.50, č.1, s. 41. ISSN 1916-0216. DOI: https://doi.org/10.1186/s40463-021-00527-9. [cit. 2025-01-20].

[28] AHN, EunJin a KANG, Hyun**.** Introduction to Systematic Review and Meta-analysis. *Korean Journal of Anesthesiology* [online]. 2018, roč. 71, č. 2, s. 103–112. ISSN 2005-6419. DOI: https://doi.org/10.4097/kjae.2018.71.2.103. [cit. 2025-01-20].

[29] PHILLIPS, Veronica a BARKER, Eleanor**.** Systematic Reviews: Structure, Form and Content. *Journal of Perioperative Practice* [online]. 2021, roč. 31. Dostupné z: https://doi.org/10.1177/1750458921994693. [cit. 2025-01-20].

[30] RETHLEFSEN, Melissa L. a PAGE, Matthew J. PRISMA 2020 and PRISMA-S: Common Questions on Tracking Records and the Flow Diagram [online]. *Journal of the Medical Library Association*, 2022, roč. 110, č. 2, s. 253–257. ISSN 1536-5050. Dostupné z: https://doi.org/10.5195/jmla.2022.1449. [cit. 2025-01-21].

[31] PAGE, Matthew J.; MOHER, David; BOSSUYT, Patrick M.; BOUTRON, Isabelle; HOFFMANN, Tammy C. a MULROW, Cynthia D. et al. PRISMA 2020 Explanation and Elaboration: Updated Guidance and Exemplars for Reporting Systematic Reviews [online]. *BMJ*, 2021, roč. 372, n160. Dostupné z: https://doi.org/10.1136/bmj.n160. [cit. 2025-01-21].

[32] LIVSCHITZ, Jennifer a SOPHIE, D. How to Write a Systematic Review [online]. *American Journal of Surgery*, 2023, roč. 226, č. 4, s. 553–555. ISSN 0002-9610. Dostupné z: https://doi.org/10.1016/j.amjsurg.2023.05.015. [cit. 2025-01-21].

[33] AMERICAN JOURNAL EXPERTS. *How to Create an Effective PRISMA Flow Diagram*. Online. Aje.com. C2025. Dostupné z: https://www.aje.com/arc/how-to-create-prisma-flow-diagram/. [cit. 2025-01-21].

[34] THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL. *Creating a PRISMA flow diagram: PRISMA 2020*. Online. Guides.lib.unc.edu. C2024. Dostupné z: https://guides.lib.unc.edu/prisma. [cit. 2025-01-23].

[35] Haddaway, N. R.; Page, M. J.; Pritchard, C. C. a McGuinness, L. A. (2022). PRISMA2020: An R Package and Shiny App for Producing PRISMA 2020-compliant Flow Diagrams, with Interactivity for Optimised Digital Transparency and Open Synthesis. *Campbell Systematic Reviews*, 18, e1230. https://doi.org/10.1002/cl2.1230. [cit. 2025-01-23].

[36] MITTAL, Mamta; BATTINENI, Gopi; BHIMAVARAPU, Usharani a LALIT**.** Text Analysis with Python: a Research Oriented Guide. [online]. *Sharjah: Bentham Science Publishers*, 2023. ISBN 978-981-5049-60-2. ISBN 978-981-5049-61-9.

[37] SARKAR, Dipanjan. *Text Analytics with Python: a Practitioner's Guide to Natural Language Processing*. *2*. vydání. [online]. Berkeley: Apress, 2019. ISBN 978-1-4842-4354-1.

[38] VANGARA, Rajashekar; SAHOO, Swagat Kumar; KUMAR, Shubham; SINGH, Sanjay Kumar a PATNAIK, Surya Narayan. Finding the Number of Latent Topics With Semantic

Non-Negative Matrix Factorization. *IEEE Access* [online]. 2021, roč. 9, s. 117217–117231. DOI: 10.1109/ACCESS.2021.3106879 [cit. 2025-01-25].

[39] MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

[40] WALLACH, Hanna M. Topic Modeling: Beyond Bag-of-Words [online]. *In: Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*. Pittsburgh: ACM, 2006, s. 977–984. Dostupné z: https://mimno.infosci.cornell.edu/info6150/readings/wallach06topic.pdf. [cit. 2025-01-25].

[41] JULURU, Krishna; SHIH, Hao-Hsin; KESHAVA, Krishna Nand a ELNAJJAR, Pierre. Bag-of-Words Technique in Natural Language Processing: a Primer for Radiologists. *RadioGraphics* [online]. 2021, roč. 41, č. 5, s. 1420–1426. ISSN 0271-5333. Dostupné z: https://doi.org/10.1148/rg.2021210025. [cit. 2025-02-03].

[42] JELODAR, Hamed; WANG, Yongli; YUAN, Chi; FENG, Xia; JIANG, Xiahui; LI, Yanchao a ZHAO, Liang. Latent Dirichlet allocation (LDA) and Topic Modeling: Models, Applications: A Survey. *Multimedia Tools and Applications* [online]. 2019, roč. 78, č. 11, s. 15169–15211. ISSN 1573-7721

[43] UGORJI, C. Calistus; ONYESOLU, Moses O.; ASOGWA, C. Doris a EGWU, Chukwudumebi V. Exploring Latent Dirichlet Allocation (LDA) in Topic Modeling: Theory, Applications, and Future Directions [online]. *Newport International Journal of Engineering and Physical Sciences*, 2024, roč. 4, č. 1, s. 9–16. DOI: 10.59298/NIJEP/2024/41916.1.1100

[44] BISMI, Iqra. *Topic Modelling using LDA*. Online. Medium.com. 2023. Dostupné z: https://medium.com/@iqra.bismi/topic-modelling-using-lda-fe81a2a806e0. [cit. 2025-02-07].

[45] AMOUALIAN, Hesam; LU, Wei; GAUSSIER, Eric; BALIKAS, Georgios; AMINI, Massih R. a CLAUSEL, Marianne. Topical Coherence in LDA-based Models through Induced Segmentation [online]. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: *Association for Computational Linguistics*, 2017, s. 1799–1809. Dostupné z: https://aclanthology.org/P17-1165/ [cit. 2025-02-22].

[46] RANJAN, Nihar a PRASAD, Rajesh. Text Analytics: An Application of Text Mining. *Journal of Data Mining and Management*, 2022, roč. 6. Dostupné z: https://doi.org/10.46610/JoDMM.2021.v06i03.001. [cit. 2025-02-15].

[47] BUCHANAN, Robert A**.** *Accuracy of Cited References: The Role of Citation Databases*. *College & Research Libraries* [online]. 2006, 67(4), s. 292–303. ISSN 0010-0870. Dostupné z: https://crl.acrl.org/index.php/crl/article/view/15806/19007. [cit. 2025-02-15].

[48] MEHO, Lokman I. a YANG, Kiduk. *A New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar* [online]. Ithaca: Cornell University Library, arXiv.org, 2006. Dostupné z: https://arxiv.org/abs/cs/0612132. [cit. 2024-02-15].

[49] KUMAR, R. Bibliometric Analysis: Comprehensive Insights into Tools, Techniques, Applications, and Solutions for Research Excellence. *Spectrum of Engineering and Management Sciences* [online]. 2025, roč. 3, č. 1, s. 45–62. Dostupné z: https://doi.org/10.31181/sems31202535k. [cit. 2024-03-11].

[50] PASSAS, Ioannis. Bibliometric Analysis: The Main Steps [online]. *Encyclopedia*, 2024, roč. 4, č. 2, s. 1014–1025. Dostupné z: https://doi.org/10.3390/encyclopedia4020065. [cit. 2025-03-11].

[51] HASSAN, Waseem a Antonia Eliene DUARTE, 2024. Bibliometric analysis: a few suggestions. *Current Problems in Cardiology* [online]. 49(8), 102640. ISSN 0146-2806. Dostupné z: https://doi.org/10.1016/j.cpcardiol.2024.102640. [cit. 2025-03-11].

[52] SONG, Guandong, WU, Jiying a WANG, Sihui. [Retracted] Text Mining in Management Research: a Bibliometric Analysis. *Security and Communication Networks* [online]. 2021, článek 2270276, 15 s. Dostupné z: https://doi.org/10.1155/2021/2270276. [cit. 2025-03-13].

[53] AHMI, Aidi**.** *Bibliometric Analysis for Beginners*. [online] Sintok: UUM Press, 2022. ISBN 978-967-0031-53-8.

[54] ALI, Mohammad J. Understanding the ‘g-index’ and the ‘e-index’. *Seminars in Ophthalmology* [online]. 2021, roč. 36, č. 4, s. 139. ISSN 0882-0538. Dostupné z: https://doi.org/10.1080/08820538.2021.1922975. [cit. 2025-03-13].

[55] ÖZTÜRK, Oğuzhan; KOCAMAN, Rıdvan a KANBACH, Dominik K. How to Design Bibliometric Research: An Overview and a Framework Proposal. *Review of Managerial Science* [online]. 2024, roč. 18, č. 11, s. 3333–3361. ISSN 1863-6691. Dostupné z: https://doi.org/10.1007/s11846-024-00738-0. [cit. 2025-03-15].

[56] DONTHU, Naveen, KUMAR, Satish, MUKHERJEE, Debmalya, PANDEY, Nitesh a LIM, Weng Marc. How to Conduct a Bibliometric Analysis: An Overview and Guidelines. *Journal of Business Research* [online]. 2021, roč. 133, s. 285–296. ISSN 0148-2963. Dostupné z: https://doi.org/10.1016/j.jbusres.2021.04.070. [cit. 2025-03-15].

[57] PYTHON SOFTWARE FOUNDATION. *Pandas*. Online. Pypi.org. C2025. Dostupné z: https://pypi.org/project/pandas/. [cit. 2025-03-16].

[58] NUMPY DEVELOPERS. What is NumPy?. [online]. Numpy.org. C2008-2024. Dostupné z: https://numpy.org/doc/2.2/user/whatisnumpy.html#whatisnumpy. [cit. 2025-03-16].

[59] PEREIRA, Valdecy; BASILIO, Marcio Pereira; SANTOS, Carlos Henrique Tarjano. PyBibX – A Python Library for Bibliometric and Scientometric Analysis Powered with Artificial Intelligence Tools [online]. *Data Technologies and Applications*, 2025, roč. 59, č. 2, s. 302–337. Dostupné z: https://arxiv.org/abs/2304.14516. [cit. 2025-03-16].

[60] NETWORKX DEVELOPERS. *Introduction*. [online]. Networkx.org. C2004-2024. Dostupné z: https://networkx.org/documentation/stable/reference/introduction.html. [cit. 2025-03-16].

[61] THE MATPLOTLIB DEVELOPMENT TEAM. *Pyplot Tutorial*. [online]. Matplotlib.org. C2012-2025. Dostupné z: https://matplotlib.org/stable/tutorials/pyplot.html. [cit. 2025-03-17].

[62] WASKOM, Michael. *An Introduction to Seaborn*. [online]. Seaborn.pydata.org. C2012-2024. Dostupné z: https://seaborn.pydata.org/tutorial/introduction.html. [cit. 2025-03-17].

[63] TOPCODER. *Word Cloud in Python*. Online. Topcoder.com. C2025. Dostupné z: https://www.topcoder.com/thrive/articles/word-cloud-in-python. [cit. 2025-03-17].

[64] EXPLOSION. *SpaCy 101: Everything you Need to Know*. [online]. spacy.io. C2016–2025. Dostupné z: https://spacy.io/usage/spacy-101. [cit. 2025-03-18].

[65] PYTHON SOFTWARE FOUNDATION. *Re — Regular Expression Operations* [online]. python.org. C2001–2025. Dostupné z: https://docs.python.org/3/library/re.html. [cit. 2025-03-18].

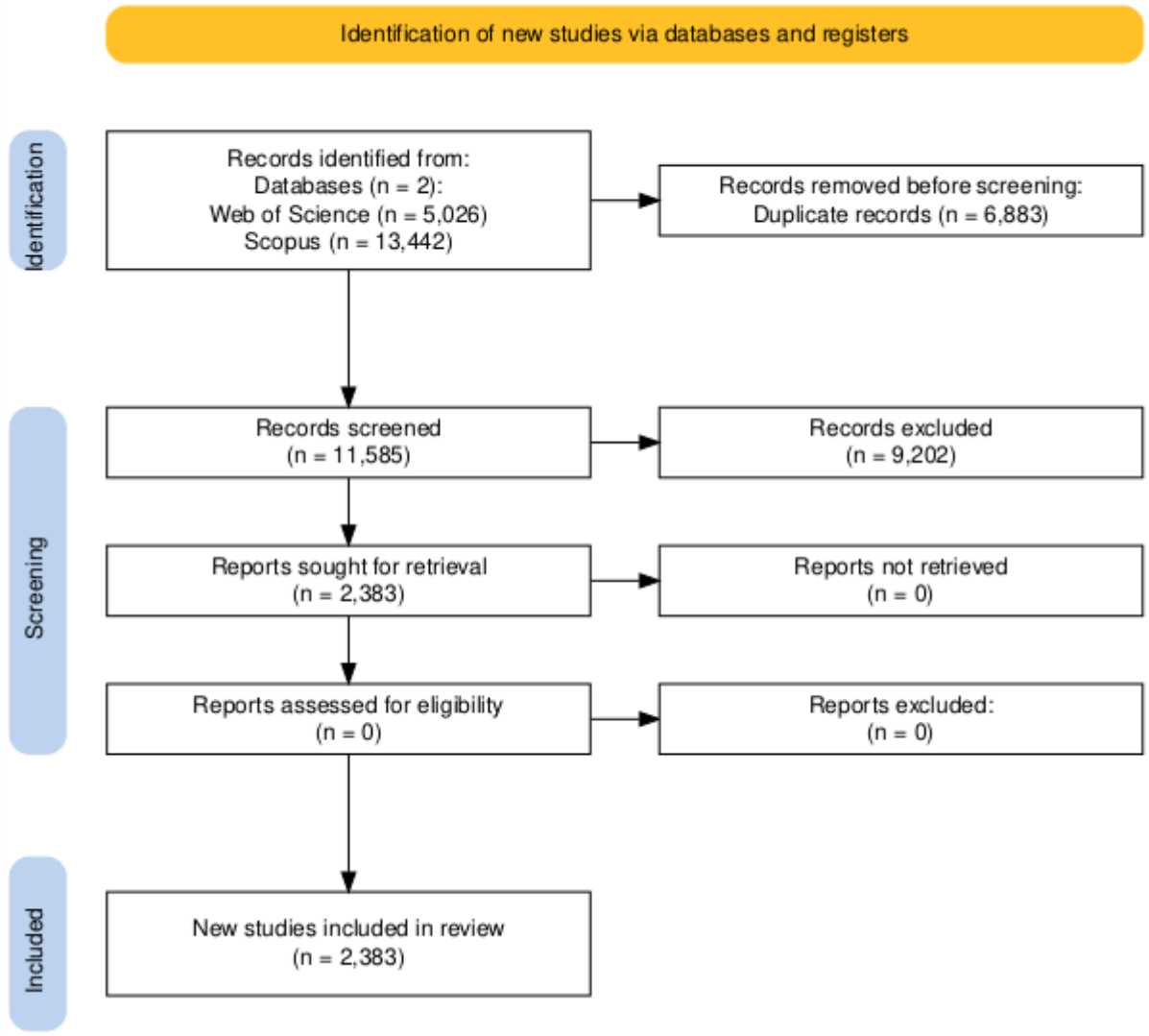
[66] PYTHON SOFTWARE FOUNDATION. *Gensim* [online]. python.org. C2025. Dostupné z: https://pypi.org/project/gensim/. [cit. 2025-03-18].

[67] PYTHON SOFTWARE FOUNDATION. *XlsxWriter* [online]. python.org. C2025. Dostupné z: https://pypi.org/project/gensim/. [cit. 2025-04-20].

[68] ELLEGAARD, Ole. *The Application of Bibliometric Analysis: Disciplinary and User Aspects*. Scientometrics, 2018, roč. 116, č. 1, s. 181–202. Dostupné z: https://doi.org/10.1007/s11192-018-2765-z. [cit. 2025-04-20].

SEZNAM PŘÍLOH

Příloha A: *PRISMA diagram*



Zdroj: [35]

Příloha B: *Kód načtení a zpracování*

```
# Import knihoven
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import networkx as nx
import pybibx as pbx
import spacy
import re
import matplotlib.ticker as mticker
from sklearn.feature_extraction.text import TfidfVectorizer
from wordcloud import WordCloud
from sklearn.decomposition import NMF
from collections import defaultdict, Counter
from gensim import corpora
from gensim.models import LdaModel
from gensim.models import CoherenceModel

# Načtení výsledků z databází (Scopus a Web of Science)
scopus_df = pd.read_csv("SCP-3.csv", delimiter=",", encoding="utf-8")
wos_df = pd.read_csv("WS-3.csv", delimiter=",", encoding="utf-8", low_memory=False)

rename_map = {
    'Authors': 'Authors',
    'Author(s)': 'Authors',
    'Document Title': 'Title',
    'Title': 'Title',
    'Source title': 'Source title',
    'Cited by': 'Cited by',
    'DOI': 'DOI',
    'Year': 'Year',
    'Document Type': 'Document Type',
    'Author Keywords': 'Author Keywords',
    'Affiliations': 'Affiliations',
    'Abstract': 'Abstract'
}

# Výpis počtu záznamů v obou databázích
n_scopus = len(scopus_df)
n_wos = len(wos_df)

print(f"Počet záznamů nalezených v databázi Scopus: {n_scopus}")
print(f"Počet záznamů nalezených v databázi Web of Science: {n_wos}")

# Výpis rozložení typů dokumentů ve Scopusu
print("\nTypy dokumentů v databázi Scopus:")
if 'Document Type' in scopus_df.columns:
```

```
    print(scopus_df['Document Type'].value_counts())
else:
    print("Sloupec 'Document Type' nebyl nalezen ve Scopus datasetu.")

# Výpis rozložení typů dokumentů ve Web of Science
print("\nTypy dokumentů v databázi Web of Science:")
if 'Document Type' in wos_df.columns:
    print(wos_df['Document Type'].value_counts())
else:
    print("Sloupec 'Document Type' nebyl nalezen ve WOS datasetu.")

# Kontrola datových typů
print("\nDatové typy ve Scopus datasetu:")
print(scopus_df.dtypes)

print("\nDatové typy ve Web of Science datasetu:")
print(wos_df.dtypes)

# Kontrola chybějících hodnot
print("\nChybějící hodnoty v datech Scopus:")
print(scopus_df.isnull().sum())

print("\nChybějící hodnoty v datech Web of Science:")
print(wos_df.isnull().sum())

# Sloučení dat ze Scopusu a Web of Science do jednoho datasetu
combined_raw = pd.concat([scopus_df, wos_df], ignore_index=True)

# Počet záznamů před deduplikací
records_identified = len(combined_raw)

# Odstranění duplicit podle DOI
combined_unique = combined_raw.drop_duplicates(subset='DOI')

# Počet záznamů po odstranění duplicit
records_screened = len(combined_unique)
duplicates_removed = records_identified - records_screened

# Výstup výsledků
print(f"Celkový počet identifikovaných záznamů: {records_identified}")
print(f"Počet odstraněných duplicit (pouze DOI): {duplicates_removed}")
print(f"Počet záznamů po deduplikaci (pro screening): {records_screened}")
```

Příloha C: *Kód textové analýzy*

```
# Načtení anglického lemmatizačního modelu spaCy
nlp = spacy.load("en_core_web_sm")

# Sloučení textu: Title + Abstract + Author Keywords
def merge_text_fields(row):
    title = row['Title'] if pd.notna(row['Title']) else ''
    abstract = row['Abstract'] if pd.notna(row['Abstract']) else ''
    keywords = row['Author Keywords'] if pd.notna(row['Author Keywords']) else ''
    return f"{title} {abstract} {keywords}"

# Kopie finálního datasetu
final_df = screened_df[
    screened_df['Abstract'].notna() &
    (screened_df['Abstract'].str.strip() != '') &
    screened_df['Year'].between(2020, 2024, inclusive='both') &
    screened_df['Document Type'].isin(['Article', 'Review'])
].copy()

# Vytvoření sloupce se sloučeným textem
final_df['Merged_Text'] = final_df.apply(merge_text_fields, axis=1)

# Funkce pro předzpracování textu – lemmatizace, odstranění stop slov
def preprocess(text):
    doc = nlp(text.lower())  # převod na malá písmena
    return ' '.join([
        token.lemma_ for token in doc
        if token.is_alpha and not token.is_stop and token.lemma_.lower() != 'datum' and len(to-
ken) > 2
    ])

# Aplikace předzpracování na sloupec s textem
final_df['Processed_Text'] = final_df['Merged_Text'].astype(str).apply(preprocess)

# Ukázka výsledku
final_df[['Title', 'Processed_Text']].sample(5, random_state=42)

# Výpočet TF-IDF matice z očištěného textu
vectorizer = TfidfVectorizer(max_features=1000)
tfidf_matrix = vectorizer.fit_transform(final_df['Processed_Text'])

# Získání všech termínů a jejich celkové váhy (součet přes všechny články)
words = vectorizer.get_feature_names_out()
score = tfidf_matrix.toarray().sum(axis=0)

# Vytvoření tabulky s výsledky
tfidf_df = pd.DataFrame({
    'word': words,
```

```
    'tfidf': score
}).sort_values(by='tfidf', ascending=False)

# Vybrání 10 slov s nejvyšším TF-IDF skóre
top10_tfidf = tfidf_df.head(10)

# Vykreslení horizontálního sloupcového grafu
plt.figure(figsize=(10, 5))
bars = plt.barh(top10_tfidf['word'][::-1], top10_tfidf['tfidf'][::-1], color='steelblue')

# Přidání hodnot k jednotlivým sloupcům
for bar in bars:
    width = bar.get_width()
    plt.text(width + 1, bar.get_y() + bar.get_height() / 2,
             f'{width:.2f}', va='center', fontsize=10)

# Popisky a formát
plt.xlabel('TF-IDF skóre')
plt.ylabel('Slova')
plt.title('Top 10 slov podle TF-IDF', fontsize=13)
plt.grid(axis='x', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()

# Vytvoření TF-IDF matice z předzpracovaného textu
vectorizer = TfidfVectorizer(max_features=1000)
tfidf_matrix = vectorizer.fit_transform(final_df['Processed_Text'])

# Pole pro ukládání chyb
errors = []

# Rozsah počtu témat
topic_range = range(2, 11)  # Témata 2 až 10

# Výpočet rekonstrukční chyby pro každý počet témat
for n_topics in topic_range:
    model = NMF(n_components=n_topics, random_state=42)
    model.fit(tfidf_matrix)
    errors.append(model.reconstruction_err_)

# Vykreslení výsledků
plt.figure(figsize=(8, 5))
plt.plot(topic_range, errors, marker='o')
plt.title('Vývoj rekonstrukční chyby podle počtu témat')
plt.xlabel('Počet témat')
plt.ylabel('Rekonstrukční chyba')
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
# Nastavení počtu témat
n_topics = 5

# Trénink NMF modelu na TF-IDF matici
nmf_model = NMF(n_components=n_topics, random_state=42)
W = nmf_model.fit_transform(tfidf_matrix)  # matice dokumentů vs. témata
H = nmf_model.components_                  # matice témat vs. slov

# Získání slov z TF-IDF vektoru
feature_names = vectorizer.get_feature_names_out()
n_top_words = 10  # počet nejdůležitějších slov na téma

# Výpis top slov pro každé téma
for topic_idx, topic in enumerate(H):
    top_words = [feature_names[i] for i in topic.argsort()[:-n_top_words - 1:-1]]
    print(f"\nTéma {topic_idx + 1}: {', '.join(top_words)}")

# Přiřazení tématu s nejvyšší vahou pro každý dokument
final_df['Topic'] = W.argmax(axis=1) + 1  # +1, aby témata začínala od 1

# Vytvoření názvů témat
topic_labels = {
    1: 'Business Intelligence',
    2: 'Machine Learning',
    3: 'Supply Chain & Big Data',
    4: 'Digital Marketing',
    5: 'Industry 4.0 & IoT'
}

# Přiřazení popisků podle tématu
final_df['Topic_Label'] = final_df['Topic'].map(topic_labels)

# Výpis prvních 5 záznamů
final_df[['Title', 'Topic', 'Topic_Label']].head()

# Přiřazení tématu s nejvyšší vahou pro každý dokument
final_df['Topic'] = W.argmax(axis=1) + 1  # +1, aby témata začínala od 1

# Kontrola – výpis prvních 5 záznamů s přiřazeným tématem
final_df[['Title', 'Topic']].head()

# Kontingenční tabulka: počet článků podle roku a tématu
topic_by_year = final_df.groupby(['Year', 'Topic']).size().unstack(fill_value=0)

# Názvy témat
topic_labels = {
    1: 'Business Intelligence',
    2: 'Machine Learning',
    3: 'Supply Chain & Big Data',
```

```
    4: 'Digital Marketing',
    5: 'Industry 4.0 & IoT'
}

# Vykreslení stacked bar chart

fig, ax = plt.subplots(figsize=(12, 6))
colors = plt.cm.tab20.colors
bottom = [0] * len(topic_by_year)

for i, column in enumerate(topic_by_year.columns):
    bars = ax.bar(topic_by_year.index, topic_by_year[column],
                  bottom=bottom, label=topic_labels.get(column, f'Téma {column}'),
color=colors[i])

    # Číselné hodnoty do sloupců
    for bar in bars:
        height = bar.get_height()
        if height > 0:
            ax.text(bar.get_x() + bar.get_width() / 2,
                    bar.get_y() + height / 2,
                    str(int(height)),
                    ha='center', va='center', fontsize=9, color='white')

    # aktualizace pro další pruhy
    bottom = [sum(x) for x in zip(bottom, topic_by_year[column])]

# Popisky grafu
ax.set_title('Tematické rozložení článků podle roku')
ax.set_xlabel('Rok')
ax.set_ylabel('Počet článků')
ax.legend(title='Téma', bbox_to_anchor=(1.05, 1), loc='upper left')
ax.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()

# Nastavení top slov
top_slova = 5
seznam_slov = []

# Výběr nejdůležitějších slov pro každé téma
for i, tema in enumerate(H):
    indexy = tema.argsort()[-top_slova:][::-1]
    slova = [feature_names[j] for j in indexy]
    seznam_slov.extend(slova)

# Unikátní slova pro výběr sloupců
slova_pro_heatmapu = sorted(set(seznam_slov))
```

```
topic_labels = {
    0: 'Business Intelligence',
    1: 'Machine Learning',
    2: 'Supply Chain & Big Data',
    3: 'Digital Marketing',
    4: 'Industry 4.0 & IoT'
}

# DataFrame pro heatmapu
df_heatmap = pd.DataFrame(H, columns=feature_names)
df_heatmap = df_heatmap[slova_pro_heatmapu]
df_heatmap.index = [topic_labels[i] for i in range(df_heatmap.shape[0])]

# Vykreslení heatmapy
plt.figure(figsize=(16, 8))
sns.heatmap(df_heatmap, annot=True, fmt=".2f", cmap="Blues", linewidths=0.5,
            cbar_kws={'label': 'Váha slova'}, annot_kws={"size": 10})

# Popisky
plt.title("Heatmapa: nejdůležitější slova podle témat (NMF)", fontsize=16)
plt.xlabel("Slova", fontsize=12)
plt.ylabel("Témata", fontsize=12)
plt.xticks(rotation=45, ha='right', fontsize=10)
plt.yticks(rotation=0, fontsize=10)
plt.tight_layout()
plt.show()

# Funkce pro výpočet coherence pro různý počet témat
def compute_coherence_values(dictionary, corpus, texts, start, limit, step):
    coherence_scores = []
    model_list = []
    for num_topics in range(start, limit + 1, step):
        model = models.LdaModel(corpus=corpus,
                                id2word=dictionary,
                                num_topics=num_topics,
                                random_state=42,
                                passes=10)
        model_list.append(model)
        coherencemodel = CoherenceModel(model=model, texts=texts, dictionary=dictionary,
coherence='c_v')
        coherence_scores.append(coherencemodel.get_coherence())
    return model_list, coherence_scores

# Spuštění výpočtu coherence pro témata od 2 do 10
start, limit, step = 2, 10, 1
model_list, coherence_values = compute_coherence_values(dictionary, corpus, tokeni-
zed_docs, start, limit, step)

# Vykreslení
```

```
plt.figure(figsize=(10,5))
plt.plot(range(start, limit + 1, step), coherence_values, marker='o', color='teal')
plt.xlabel("Počet témat")
plt.ylabel("Koherence (c_v)")
plt.title("Výběr optimálního počtu témat (LDA)")
plt.grid(True)
plt.tight_layout()
plt.show()

# Načtení anglického lemmatizačního modelu
nlp = spacy.load("en_core_web_sm")

# Funkce pro tokenizaci a lemmatizaci
def tokenize_lemmas(text):
    doc = nlp(text.lower())
    return [
        token.lemma_ for token in doc
        if token.is_alpha
          and not token.is_stop
          and len(token) > 2
          and token.lemma_ != 'datum'
    ]

# Tokenizace textu
tokenized_docs = final_df['Processed_Text'].apply(tokenize_lemmas).tolist()

# Vytvoření bag-of-words
dictionary = corpora.Dictionary(tokenized_docs)
corpus = [dictionary.doc2bow(doc) for doc in tokenized_docs if len(doc) > 0]

# Trénink LDA modelu
lda_model_final = LdaModel(
    corpus=corpus,
    id2word=dictionary,
    num_topics=3,
    random_state=42,
    passes=10
)

# Výpis top 10 slov s jejich vahami pro každé téma
for tid in range(lda_model_final.num_topics):
    print(f"Téma {tid+1}:")
    # get_topic_terms vrací seznam (word_id, váha)
    for word_id, weight in lda_model_final.get_topic_terms(topicid=tid, topn=10):
        word = dictionary[word_id]
        print(f"  {word:<12} {weight:.4f}")
    print()

doc_topics = [lda_model_final.get_document_topics(bow) for bow in corpus]
```

```
# Matice W_lda (dokumenty × témata)
import numpy as np
n_docs   = len(doc_topics)
n_topics = lda_model_final.num_topics
W_lda    = np.zeros((n_docs, n_topics))

for i, topic_list in enumerate(doc_topics):
    for topic_id, weight in topic_list:
        W_lda[i, topic_id] = weight

mask = [len(doc)>0 for doc in tokenized_docs]
final_df.loc[mask, 'Topic']       = W_lda.argmax(axis=1) + 1
final_df['Topic_Label'] = final_df['Topic'].map(topic_labels)

# Kontingenční tabulka: Počet článků podle roku a LDA-tématu
topic_by_year_lda = (
    final_df
    .groupby(['Year', 'Topic'])
    .size()
    .unstack(fill_value=0)
)

# Omezení LDA – témata
topic_by_year_lda = topic_by_year_lda.reindex(columns=[1, 2, 3], fill_value=0)

# Názvy LDA – témat
topic_labels_lda = {
    1: 'Decision Systems & Modeling',
    2: 'Business Intelligence',
    3: 'Business Modeling'
}

# Vykreslení stacked bar chart
fig, ax = plt.subplots(figsize=(12, 6))
colors = plt.cm.tab20.colors
bottom = [0] * len(topic_by_year_lda)

for i, col in enumerate(topic_by_year_lda.columns):
    counts = topic_by_year_lda[col]
    bars = ax.bar(
        topic_by_year_lda.index,
        counts,
        bottom=bottom,
        label=topic_labels_lda[col],
        color=colors[i]
    )
    for bar in bars:
        h = bar.get_height()
```

```
        if h > 0:
            ax.text(
                bar.get_x() + bar.get_width() / 2,
                bar.get_y() + h / 2,
                str(int(h)),
                ha='center', va='center', fontsize=9, color='white'
            )
    bottom = [b + c for b, c in zip(bottom, counts)]

# Popisky a legenda
ax.set_title('Tematické rozložení článků podle roku (LDA)', fontsize=14)
ax.set_xlabel('Rok', fontsize=12)
ax.set_ylabel('Počet článků', fontsize=12)
ax.legend(title='LDA téma', bbox_to_anchor=(1.05, 1), loc='upper left')
ax.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()

# Matice LDA – váhy slov pro každé téma
H_lda = lda_model_final.get_topics()

# Seznam všech slov ve správném pořadí
feature_names = [dictionary[i] for i in range(len(dictionary))]

# Nastavení top slov
top_words = 5
all_words = []

# Výběr nejdůležitějších slov pro každé téma
for i, topic_vector in enumerate(H_lda):
    top_idx = topic_vector.argsort()[-top_words:][::-1]
    for j in top_idx:
        all_words.append(feature_names[j])

# Unikátní slova pro výběr sloupců
words_for_heatmap = sorted(set(all_words))

topic_labels = {
    1: 'Decision Systems & Modeling',
    2: 'Business Intelligence',
    3: 'Business Modeling'
}

# DataFrame pro heatmapu
df_heatmap_lda = pd.DataFrame(H_lda, columns=feature_names)
df_heatmap_lda = df_heatmap_lda[words_for_heatmap]
df_heatmap_lda.index = [topic_labels[i+1] for i in range(df_heatmap_lda.shape[0])]

# Vykreslení heatmapy
```

```
plt.figure(figsize=(12, 6))
sns.heatmap(
    df_heatmap_lda,
    annot=True,
    fmt=".3f",
    cmap="YlGnBu",
    linewidths=0.5,
    cbar_kws={'label': 'Váha slova'},
    annot_kws={"size": 10}
)
plt.title("Heatmapa: nejdůležitější slova podle témat (LDA)", fontsize=14)
plt.xlabel("Slova", fontsize=12)
plt.ylabel("Témata", fontsize=12)
plt.xticks(rotation=45, ha='right', fontsize=10)
plt.yticks(rotation=0, fontsize=11)
plt.tight_layout()
plt.show()

# Export výsledků
output_path = 'Výstupy-textové-analýzy.xlsx'

# Příprava listů pro soubor
sheets = {
    'TFIDF Top 10':       top10_tfidf.reset_index(drop=True),
    'NMF Assignments':    final_df[['Title','Topic','Topic_Label','Year']],
    'NMF Topics×Year':    topic_by_year.reset_index(),
    'NMF Heatmap Data':   df_heatmap_nmf.reset_index().rename(columns={'index':'Topic'}),
    'NMF Topic Words':    nmf_df,
    'LDA Topics×Year':    topic_by_year_lda.reset_index(),
    'LDA Heatmap Data':   df_heatmap_lda.reset_index().rename(columns={'index':'Topic'}),
    'LDA Topic Words':    lda_df
}

with pd.ExcelWriter(output_path, engine='xlsxwriter') as writer:
    for sheet_name, df in sheets.items():
        df.to_excel(writer, sheet_name=sheet_name, index=False)

print('Všechny výstupy byly uloženy do souboru', output_path)
```

Příloha D: *Kód bibliometrické analýzy*

```
# Výpočet počtu článků podle roku a převod indexu na celé číslo
year_counts = screened_df['Year'].value_counts().sort_index()
year_counts.index = year_counts.index.astype(int)  # převod 2024.0 na 2024

# Vykreslení grafu
plt.figure(figsize=(10, 6))
bars = plt.bar(year_counts.index.astype(str), year_counts.values, color="skyblue")
plt.title("Vývoj počtu článků za roky 2020–2024", fontsize=14)
plt.xlabel("Rok", fontsize=12)
plt.ylabel("Počet článků", fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Popisky nad sloupci
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval + 10, int(yval), ha='center', va='bottom', font-
size=10)

plt.tight_layout()
plt.show()

# Výpočet průměrného počtu citací podle roku
avg_citations_per_year = screened_df.groupby('Year')['Cited by'].mean()

# Vykreslení grafu
plt.figure(figsize=(10, 6))
plt.plot(avg_citations_per_year.index.astype(int),
         avg_citations_per_year.values,
         marker='o', linestyle='dashed', color='red')

# Popisky a formátování
plt.title('Průměrný počet citací článků podle roku', fontsize=15)
plt.xlabel('Rok', fontsize=13)
plt.ylabel('Průměrný počet citací', fontsize=13)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Úprava osy X – pouze celá čísla
plt.gca().xaxis.set_major_locator(mticker.MaxNLocator(integer=True))

# Číselné hodnoty nad body
for x, y in zip(avg_citations_per_year.index, avg_citations_per_year.values):
    plt.text(x, y + 0.5, f'{y:.1f}', ha='center', fontsize=11, color='black')

plt.tight_layout()
plt.show()

# Výběr 10 nejcitovanějších článků z finálního datasetu
```

```
top_cited_articles = screened_df.nlargest(10, 'Cited by').copy()

# Zkrácení názvů článků pro lepší čitelnost
top_cited_articles['Short Title'] = top_cited_articles['Title'].apply(
    lambda x: x[:50] + '...' if isinstance(x, str) and len(x) > 50 else x
)

# Vykreslení grafu
plt.figure(figsize=(14, 7))
bars = plt.barh(top_cited_articles['Short Title'], top_cited_articles['Cited by'],
                color='royalblue', edgecolor='black')

# Přidání hodnot k jednotlivým sloupcům
for bar in bars:
    plt.text(bar.get_width() + 15, bar.get_y() + bar.get_height() / 2,
             int(bar.get_width()), ha='left', va='center', fontsize=12, color='black')

plt.gca().invert_yaxis()
plt.title('Top 10 nejcitovanějších článků', fontsize=16, fontweight='bold')
plt.xlabel('Počet citací', fontsize=14)
plt.ylabel('Název článku', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=11)
plt.tight_layout()
plt.show()

# Vyčištění a normalizace klíčových slov
keywords_series = screened_df['Author Keywords'].dropna().str.lower().str.strip()

# Rozdělení podle čárky nebo středníku, odstranění mezer
all_keywords = keywords_series.str.split(r';|,').explode().str.strip()

# Odstranění speciálních znaků a sjednocení spojovníků
all_keywords = all_keywords.str.replace(r'[-_]', ' ', regex=True)

# Počet výskytů jednotlivých klíčových slov
keyword_counts = Counter(all_keywords.dropna())

# Výběr TOP 10 nejčastějších klíčových slov
top_keywords = keyword_counts.most_common(10)

# Wordcloud vizualizace
wordcloud_text = ' '.join(all_keywords.dropna())

wordcloud = WordCloud(
    width=1200, height=600,
    background_color='white',
    colormap='viridis',
    max_words=150
```

```
).generate(wordcloud_text)

plt.figure(figsize=(12,6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title("Nejčastější klíčová slova (wordcloud)", fontsize=14)
plt.tight_layout()
plt.show()

# 7. Horizontální sloupcový graf pro TOP 10
top_words, top_counts = zip(*top_keywords)

plt.figure(figsize=(12,6))
bars = plt.barh(top_words[::-1], top_counts[::-1], color='royalblue', edgecolor='black')

# Přidání hodnot na konec pruhů
for bar in bars:
    plt.text(bar.get_width() + 5, bar.get_y() + bar.get_height()/2,
             int(bar.get_width()), ha='left', va='center', fontsize=12, color='black')

plt.xlabel('Počet výskytů', fontsize=14)
plt.ylabel('Klíčové slovo', fontsize=14)
plt.title('Top 10 nejčastějších klíčových slov', fontsize=16, fontweight='bold', pad=15)
plt.grid(axis='x', linestyle='--', alpha=0.6)
plt.tight_layout()
plt.show()

# Výběr sloupců a odstranění prázdných hodnot
keywords_df = screened_df[['Year', 'Author Keywords']].dropna()

# Úprava klíčových slov (malá písmena, odstranění mezer)
keywords_df['Author Keywords'] = keywords_df['Author Keywords'].str.lower().str.strip()

# Rozdělení klíčových slov a jejich rozbalení na samostatné řádky
keywords_df['Author Keywords'] = keywords_df['Author Keywords'].str.split(r';|,')
keywords_exploded = keywords_df.explode('Author Keywords')
keywords_exploded['Author Keywords'] = keywords_exploded['Author Keywords'].str.strip()

# Úprava let
keywords_exploded['Year'] = pd.to_numeric(keywords_exploded['Year'], errors='coer-
ce').astype('Int64')
keywords_exploded = keywords_exploded.dropna(subset=['Year'])

# Výběr 5 nejčastějších klíčových slov
top_keywords = keywords_exploded['Author Keywords'].value_counts().nlargest(5).index
top_keywords_df = keywords_exploded[keywords_exploded['Author
Keywords'].isin(top_keywords)]

# Kontingenční tabulka (rok × klíčové slovo)
```

```
trend_table = top_keywords_df.groupby(['Year', 'Author Keywords']).size().unstack(fill_va-
lue=0)

# Vykreslení vývoje klíčových slov
plt.figure(figsize=(12, 6))
for keyword in trend_table.columns:
    plt.plot(trend_table.index, trend_table[keyword], marker='o', label=keyword)

plt.title('Vývoj TOP 5 klíčových slov v čase (2020–2024)', fontsize=15)
plt.xlabel('Rok', fontsize=13)
plt.ylabel('Počet výskytů', fontsize=13)
plt.legend(title='Klíčové slovo')
plt.grid(True, linestyle='--', alpha=0.7)
plt.xticks(trend_table.index)
plt.tight_layout()
plt.show()

# Výběr sloupce se zdrojovým názvem časopisu
source_counts = screened_df['Source title'].dropna().value_counts().head(10)

# Vizualizace – TOP 10 zdrojů
plt.figure(figsize=(12, 6))
bars = plt.barh(source_counts.index[::-1], source_counts.values[::-1],
                color='royalblue', edgecolor='black')

# Přidání hodnot na konec pruhů
for bar in bars:
    plt.text(bar.get_width() + 2, bar.get_y() + bar.get_height()/2,
             int(bar.get_width()), ha='left', va='center', fontsize=12, color='black')

plt.xlabel('Počet článků', fontsize=13)
plt.ylabel('Publikační zdroj', fontsize=13)
plt.title('Top 10 nejčastějších publikačních zdrojů', fontsize=15, fontweight='bold', pad=10)
plt.grid(axis='x', linestyle='--', alpha=0.6)
plt.tight_layout()
plt.show()

# Výpočet počtu publikací na autora
publication_counts = defaultdict(int)

for authors in screened_df['Authors'].dropna():
    author_list = re.split(r';|,|\band\b', authors)
    author_list = [a.strip() for a in author_list if len(a.strip()) > 3]
    for author in author_list:
        publication_counts[author] += 1

# Vytvoření Counter objektu z výsledků
author_counts = Counter(publication_counts)
```

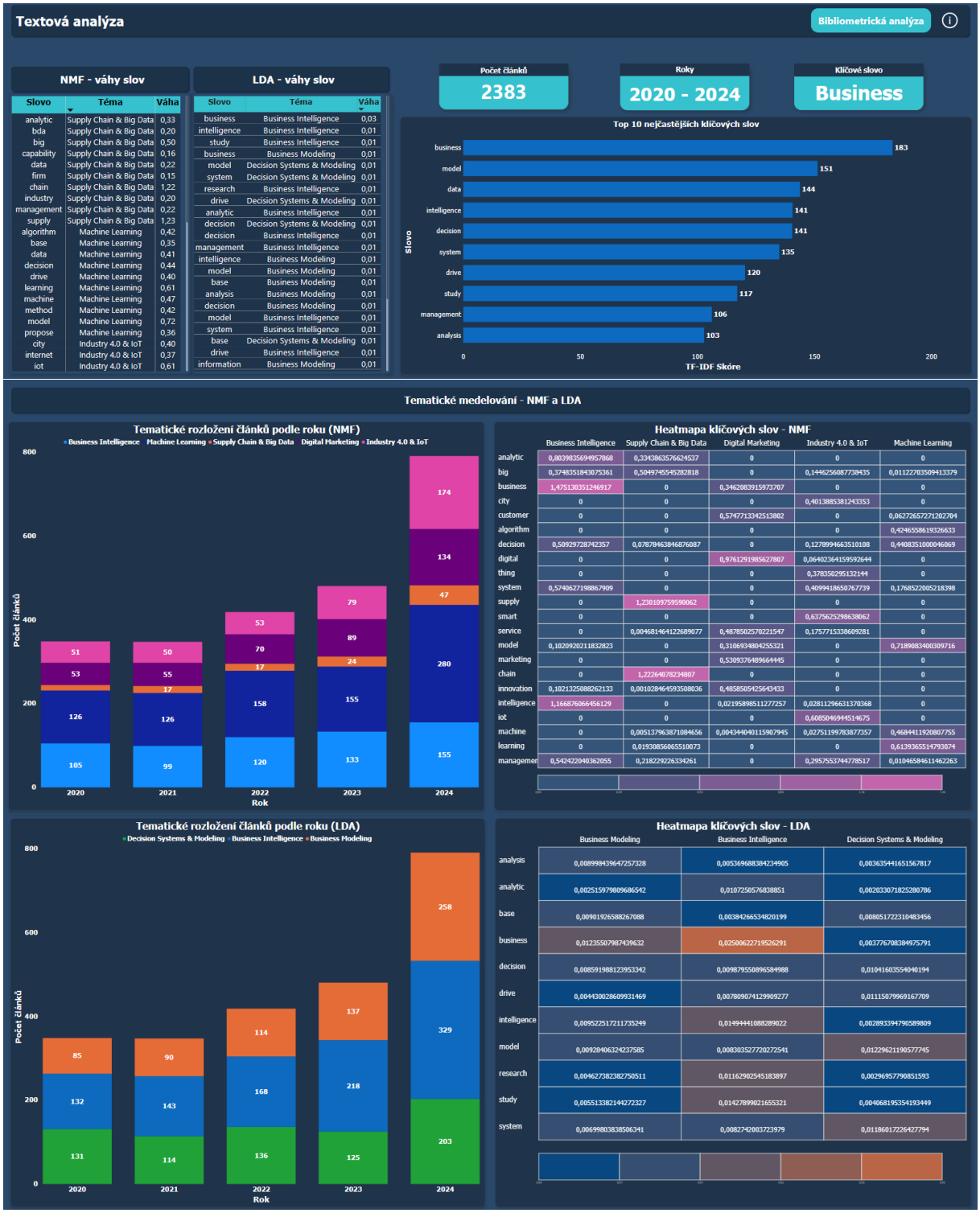
```
# Výběr TOP 10 autorů podle počtu publikací
top_authors = author_counts.most_common(10)
top_authors_names, top_authors_count = zip(*top_authors)

# Vykreslení grafu
plt.figure(figsize=(12, 6))
plt.barh(top_authors_names[::-1], top_authors_count[::-1], color='skyblue', edgecolor='black')

# Přidání hodnot k pruhům
for bar in plt.gca().patches:
    plt.text(bar.get_width() + 2, bar.get_y() + bar.get_height()/2,
             f'{int(bar.get_width())}', ha='left', va='center', fontsize=12, color='black')

plt.xlabel('Počet publikací', fontsize=14)
plt.ylabel('Autor', fontsize=14)
plt.title('Top 10 autorů podle počtu publikací', fontsize=16, fontweight='bold')
plt.tight_layout()
plt.show()

# Vytvoření prázdného grafu
G = nx.Graph()

# Procházení autorů a vytvoření hran mezi nimi
for row in screened_df['Authors'].dropna():
    authors = re.split(r';|,|\band\b', row)  # rozdělení podle běžných oddělovačů
    authors = [a.strip() for a in authors if len(a.strip()) > 3]  # odstranění mezer a krátkých jmen

    for i in range(len(authors)):
        for j in range(i + 1, len(authors)):
            G.add_edge(authors[i], authors[j], weight=G.get_edge_data(authors[i], authors[j],
{}).get('weight', 0) + 1)

# Filtrování: ponechání pouze spoluprací, které se opakovaly 2x a více
edges_to_keep = [(u, v) for u, v, d in G.edges(data=True) if d['weight'] >= 2]
G_filtered = G.edge_subgraph(edges_to_keep).copy()

# Vybrání největší propojené skupiny
largest_component = max(nx.connected_components(G_filtered), key=len)
G_main = G_filtered.subgraph(largest_component)

# Vizualizace sítě
plt.figure(figsize=(14, 10))
pos = nx.spring_layout(G_main, k=0.3, seed=42)

nx.draw(G_main, pos,
        with_labels=True,
        node_size=120,
        node_color='skyblue',
        font_size=8,
```

```
        edge_color='gray',
        alpha=0.7)

plt.title("Síť spoluautorství (2020–2024)", fontsize=15)
plt.axis('off')
plt.tight_layout()
plt.show()

# Počty publikací a citací
publication_counts = defaultdict(int)
author_citations = defaultdict(list)

# Získání počtu publikací a citací na autora
for _, row in screened_df[['Authors', 'Cited by']].dropna().iterrows():
    citations = int(row['Cited by']) if not pd.isna(row['Cited by']) else 0
    authors = re.split(r';|,|\band\b', row['Authors'])
    authors = [name.strip() for name in authors if len(name.strip()) > 3]

    for author in authors:
        publication_counts[author] += 1
        author_citations[author].append(citations)

# Vytvoření sítě spoluautorství
if 'G_main' not in locals():
    G = nx.Graph()
    for row in screened_df['Authors'].dropna():
        authors = re.split(r';|,|\band\b', row)
        authors = [a.strip() for a in authors if len(a.strip()) > 3]
        for i in range(len(authors)):
            for j in range(i + 1, len(authors)):
                G.add_edge(authors[i], authors[j], weight=G.get_edge_data(authors[i], authors[j],
{}).get('weight', 0) + 1)

    # Filtrování silnější spoluprací (2× a více)
    edges_to_keep = [(u, v) for u, v, d in G.edges(data=True) if d['weight'] >= 2]
    G_filtered = G.edge_subgraph(edges_to_keep).copy()

    # Výběr největší komponenty
    largest_component = max(nx.connected_components(G_filtered), key=len)
    G_main = G_filtered.subgraph(largest_component)

# Výpočet metrik pro autory
results = []

for author in G_main.nodes():
    publications = publication_counts.get(author, 0)
    citations = sorted(author_citations.get(author, []), reverse=True)
    total_citations = sum(citations)
```

```
    # h-index
    h_index = sum(1 for i, c in enumerate(citations) if c >= i + 1)

    # g-index
    g_index = 0
    for i in range(1, len(citations) + 1):
        if sum(citations[:i]) >= i**2:
            g_index = i
        else:
            break

    results.append({
        'Autor': author,
        'Publikace': publications,
        'Celkový počet citací': total_citations,
        'h-index': h_index,
        'g-index': g_index
    })

# Vytvoření a seřazení přehledové tabulky
metrics_df = pd.DataFrame(results)
metrics_df = metrics_df.sort_values(by='h-index', ascending=False).reset_index(drop=True)

# Zobrazení TOP 10 autorů podle h-indexu
metrics_df.head(10)


# Export výsledků
output_path = 'Výstupy-bibliometrické-analýzy.xlsx'

# Příprava listů pro soubor
sheets = {
    'Year Counts':       year_counts.rename_axis('Year').reset_index(name='Article Count'),
    'Avg Citations':     avg_citations_per_year.rename_axis('Year').reset_index(name='Avg Ci-
tations'),
    'Top Cited Articles': top_cited_articles,
    'Keyword Counts':    pd.DataFrame(keyword_counts.items(), columns=['Keyword',
'Count']),
    'Top Keywords':      top_keywords_df,
    'Keyword Trends':    trend_table.reset_index(),
    'Top Sources':       source_counts.rename_axis('Source').reset_index(name='Article
Count'),
    'Top Authors':       pd.DataFrame({
                             'Author':      top_authors_names,
                             'Publications': top_authors_count
                         }),
    'Author Metrics':    metrics_df,
    'Coauthorship Net':  pd.DataFrame(
                             [(u, v, d['weight']) for u, v, d in G_main.edges(data=True)],
```

```
                    columns=['Author 1', 'Author 2', 'Weight']
                )
}

with pd.ExcelWriter(output_path, engine='xlsxwriter') as writer:
    for sheet_name, df in sheets.items():
        df.to_excel(writer, sheet_name=sheet_name, index=False)

print('Všechny výstupy byly uloženy do souboru', output_path)
```

Příloha E: *Dashboard textové analýzy*



Příloha F: *Dashboard bibliometrické analýzy*