Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

CMS na zdieľanie a spracovanie vedeckých článkov

Diplomová práca

2025
Bc. Samuel Slávik

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

CMS na zdieľanie a spracovanie vedeckých článkov

Diplomová práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	RNDr. Kristína Malinovská, PhD.

Bratislava, 2025
Bc. Samuel Slávik

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Samuel Slávik
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: CMS na zdieľanie a spracovanie vedeckých článkov
CMS for sharing and processing research papers

Anotácia: Systémy na evidovanie a zdieľanie študovanej literatúry, napr. Mendeley, sú často používané vo vedeckých skupinách a pri medzinárodnej spolupráci. Umožňujú udržovať databázu vedeckých článkov aj s pdf originálmi, kategorizovať a tagovať články a kolaboratívne ich pridávať, združovať a zdieľať. Nevýhodou je, že články ukladajú na vzdialené úložisko a tým pádom sú spoplatnené veľkosti uložených dát a neumožňujú priamy prístup k dátam. V rámci bakalárskej práce "CMS pre kolaboratívnu správu vedeckých článkov" bol vytvorený prototyp takéhoto systému ako lokálneho doma hostovaného riešenia. Tento systém slúži ako dobrý základ pre rozšírenie jeho funkcionality, efektivity, používateľskej prístupnosti a praktickej hodnoty. Zároveň tematika spracovania textu vedeckých článkov je bohatá na možnosti využitia metód strojového učenia pre rozšírené funkcie navrhovaného systému ako napríklad rozšírené vyhľadávanie vo vedeckých článkoch, automatické hľadanie súvislostí, anotovanie a extrakcia dôležitých pojmov, generovanie sumárov textov a podobne.

Cieľ: Cieľom diplomovej práce je rozšíriť existujúci systém na evidovanie a zdieľanie vedeckých článkov, ktorý bol vyvinutý ako bakalárska práca o rozšírenú funkcionalitu a nástroje na báze strojového učenia. Vylepšenie bude obsahovať dôsledné otestovanie systému pre účely doladenia používateľskej priateľskosti a dizajnu aplikácie, implementácia angličtiny ako druhého jazyka alebo prepínania jazykov. Z hľadiska používateľov a ich spätnej väzby rozšírenie možností skupinovej spolupráce s anotáciami a diskusiami. V rámci vylepšení existujúceho softvéru preskúmať možnosti implementácie full text vyhľadávania. Ďalším cieľom práce je navrhnúť a implementovať nové nástroje pre komplexnú správu vedeckých článkov vrátane kategorizácie, anotácie, odporúčaní a vyhľadávania s využitím moderných metód strojového učenia, ako sú napríklad umelé neurónové siete. Pre implementáciu týchto funkcii tiež preskúmať a porovnať možnosti strojového učenia a vybrať najvhodnejšie metódy.

Literatúra:
[1] Django: the web framework for perfectionists with deadlines https://www.djangoproject.com/
[2] Liu, Y. and Zhang, M., 2018. Neural network methods for natural language processing.

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

[3] Lynch, P.J. and Horton, S., 2016. Web style guide: Foundations of user experience design. Yale University Press.

Vedúci: RNDr. Kristína Malinovská, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 08.12.2024

Dátum schválenia: 09.12.2024 prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie:

Abstrakt

Názov práce: CMS na zdieľanie a spracovanie vedeckých článkov

Univerzita: Univerzita Komenského v Bratislave

Fakulta: Fakulta matematiky, fyziky a informatiky

Pracovisko: Katedra aplikovanej informatiky

Študijný program: Aplikovaná informatika

Autor: Bc. Samuel Slávik

Vedúci práce: RNDr. Kristína Malinovská, Phd.

Dátum odovzdania:

Táto diplomová práca sa zaoberá návrhom a implementáciou systému na evidovanie a zdieľanie vedeckých článkov, ktorý využíva metódy strojového učenia na zlepšenie správy vedeckej literatúry v akademických a výskumných prostrediach. Cieľom práce je rozšíriť existujúci prototyp, ktorý bol vytvorený v rámci bakalárskej práce, o nové funkcionality, ako sú automatické vyhľadávanie v plnom texte, extrakcia kľúčových slov a sumarizácia textov. Systém taktiež zahŕňa možnosti skupinovej spolupráce prostredníctvom anotácií a diskusií, čím podporuje kolaboráciu medzi používateľmi a zdieľanie vedomostí. Implementácia systému využíva moderné metódy strojového učenia, konkrétne neurónové siete, na analýzu a spracovanie vedeckých textov, čo umožňuje zlepšiť efektivitu vyhľadávania a kategorizácie článkov. Práca ďalej skúma rôzne prístupy k strojovému učeniu a vyhodnocuje ich účinnosť pri extrakcii dôležitých informácií zo študovaných článkov. Testovanie a porovnanie výsledkov rôznych modelov strojového učenia pomáha identifikovať najvhodnejšie metódy pre aplikáciu na spracovanie vedeckých textov. Konečným výsledkom práce je systém, ktorý poskytuje flexibilné a efektívne riešenie pre organizáciu, spravovanie a zdieľanie vedeckých článkov v prostredí, ktoré podporuje nielen individuálny výskum, ale aj spoluprácu a kolektívne zdieľanie poznatkov.

Kľúčové slová: machine learning, neural net, semantic tagging, natural language processing, keywords extraxtion

Abstract

Thesis title: CMS for sharing and processing scientific articles

University: Comenius University Bratislava

Faculty: Faculty of Mathematics, Physics and Informatics

Department: Department of Applied Informatics

Degree course: Applied Informatics

Author: Bc. Samuel Slávik

Supervisor: RNDr. Kristína Malinovská, Phd.

Submission date::

This thesis deals with the design and implementation of a system for registering and sharing scholarly articles that uses machine learning methods to improve the management of scholarly literature in academic and research environments. The aim of the work is to extend the existing prototype, which was developed as part of the bachelor thesis, with new functionalities such as automatic full-text search, keyword extraction and text summarization. The system also includes group collaboration capabilities through annotations and discussions, thus promoting collaboration between users and knowledge sharing. The system implementation uses modern machine learning methods, specifically neural networks, to analyse and process scientific texts, allowing for improved efficiency in searching and categorising articles. The thesis further explores different machine learning approaches and evaluates their effectiveness in extracting relevant information from the articles under study. Testing and comparing the results of different machine learning models helps to identify the most appropriate methods for application to scientific text processing. The end result of the work is a system that provides a flexible and efficient solution for organizing, managing, and sharing scholarly articles in an environment that supports not only individual research, but also collaboration and collective knowledge sharing.

Keywords: machine learning, neural net, semantic tagging, natural language processing, keywords extraxtion

Obsah

Zoznam použitých skratiek

Úvod

Kapitola 1

Východiská práce

1.1 Úvod do problematiky

1.1.1 Význam vedeckej literatúry a jej správy

Vedecká literatúra predstavuje jeden zo základných pilierov rozvoja poznania a akademickej komunikácie. Výskumné tímy, organizácie aj jednotlivci sa pri formulovaní nových hypotéz, hodnotení súčasného stavu poznania či overovaní experimentálnych výsledkov opierajú o existujúce publikované zdroje. Vedecká komunikácia pritom nadobúda formu neustále rastúcej siete článkov, preprintov, konferenčných príspevkov a technických správ, ktoré sú navzájom previazané cez citácie a tematické oblasti. Ako uvádza Borgman [1], moderná veda stojí na systematickej výmene informácií, pričom efektívna práca so zdrojmi je nevyhnutným predpokladom kvalitného výskumu.

Súčasný akademický ekosystém však čelí fenoménu známeho ako *information overload* alebo preťaženie informáciami. Počet publikácií rastie exponenciálne – podľa štatistík a bibliometrických analýz Bornmanna a Mutza [2] pribúdajú rok čo rok milióny nových vedeckých článkov naprieč odbormi. Tento rast spôsobuje, že výskumníci majú stále väčší problém orientovať sa v relevantných zdrojoch, identifikovať najdôležitejšie publikácie a efektívne filtrovať neaktuálne alebo nesúvisiace práce. Efektívna správa vedeckej literatúry sa preto stáva kľúčovou súčasťou výskumného procesu.

V praxi to znamená, že výskumníci musia zvládnuť nielen samotné vyhľadávanie literatúry, ale aj jej dlhodobú organizáciu, anotovanie a opätovné využitie. Tradičný prístup, založený na ad hoc uložení PDF súborov do adresárov v počítači či e-mailových príloh, je pri väčšom počte dokumentov neudržateľný. Často vedie k duplikáciám, strate prehľadu o prečítaných a neprečítaných článkoch a k opakovanému vyhľadávaniu už raz nájdených zdrojov. Ako upozorňuje Borgman [1], infraštruktúra digitálneho bádania musí podporovať celý životný cyklus vedeckého dokumentu – od jeho získania cez spracovanie až po citovanie v ďalších prácach.

Zvýšené nároky na správu literatúry viedli k rozšíreniu nástrojov, akými sú refe-

renční manažéri a osobné digitálne knižnice, ktoré umožňujú kategorizovať, tagovať a vyhľadávať dokumenty podľa viacerých kritérií. Tieto nástroje však často fungujú ako izolované ostrovy, ktoré síce uľahčujú prácu jednotlivca, ale len čiastočne riešia potrebu prepojenia na širšie ekosystémy vyhľadávania, zdieľania a spolupráce. Kvalita metaúdajov (názov, autori, rok vydania, kľúčové slová) a ich konzistentné používanie pritom priamo ovplyvňuje viditeľnosť článkov a možnosti ich automatizovaného spracovania [3].

Z toho vyplýva potreba systémov, ktoré dokážu:

- zhromažďovať vedecké dokumenty z rôznych zdrojov,
- extrahovať z nich kľúčové metaúdaje a textový obsah,
- umožniť pokročilé vyhľadávanie podľa obsahu, autorov či tematických kategórií,
- podporovať personalizované odporúčania pre jednotlivcov alebo výskumné skupiny,
- a integrovať sa do každodenného pracovného postupu výskumníkov.

Takéto systémy vytvárajú most medzi rastúcim množstvom digitálne dostupných publikácií a konkrétnymi informačnými potrebami používateľov. Predstavujú základ, na ktorom je možné stavať pokročilejšie analytické a odporúčacie mechanizmy, ktoré budú podrobnejšie rozobraté v nasledujúcich častiach práce.

1.1.2 Existujúce systémy a ich obmedzenia

S rastúcim množstvom publikácií vznikla široká škála systémov, ktoré poskytujú prístup k vedeckej literatúre, metaúdajom a bibliometrickým ukazovateľom. Medzi najpoužívanejšie patria služby ako Google Scholar, Semantic Scholar, arXiv či ResearchGate, ktoré sa však výrazne líšia architektúrou, spôsobom získavania dát aj možnosťami integrácie do komerčných či akademických softvérových riešení.

Google Scholar ponúka jednoduché rozhranie a vysokú dostupnosť, avšak jeho API nie je oficiálne podporované a väčšina integrácií využíva neformálne techniky alebo web scraping. Podľa analýz od Beela a Gipp [4] je jeho najväčším problémom nekonzistentná kvalita záznamov a absencia transparentného hodnotiaceho algoritmu. Semantic Scholar, ktorý vyvíja Allen Institute for AI, sa naopak zameriava na extrakciu informácií pomocou metód strojového učenia, avšak API poskytuje len limitované množstvo údajov o článkoch a často neumožňuje prácu so samotným PDF dokumentom [5].

Reposzitáre ako arXiv poskytujú otvorené prístupy k preprintom, avšak neobsahujú mechanizmy personalizácie ani odporúčacie systémy. ResearchGate a Academia.edu fungujú ako akademické sociálne siete, pričom umožňujú zdieľanie publikácií, avšak

obmedzenia plynú z uzavretého charakteru platformy a komerčných modelov, ktoré bránia širšej automatizácii a integrácii.

Kľúčovým spoločným problémom týchto existujúcich riešení je:

- nedostatok otvorenosti (*closed ecosystems*),
- slabá podpora personalizovaných odporúčaní,
- obmedzené možnosti skupinovej spolupráce,
- neúplná práca priamo s PDF súbormi a obsahovými reprezentáciami dokumentov,
- obmedzenia API pri budovaní vlastných nadstavieb a systémov.

Tieto nedostatky tvoria hlavný motív pre návrh a realizáciu nového systému, ktorý využíva moderné postupy analýzy textu, odporúčacie mechanizmy a podporuje efektívnu spoluprácu výskumných tímov pri práci s vedeckou literatúrou.

1.2 Strojové učenie a skupinová spolupráca pri správe vedeckých článkov

Rastúci objem vedeckej literatúry a komplexita moderného výskumného ekosystému vytvárajú tlak na vznik systémov, ktoré dokážu efektívne podporovať vyhľadávanie, organizáciu a hodnotenie publikácií. Bibliometrické analýzy ukazujú, že počet vedeckých článkov neustále rastie a tradičné manuálne prístupy k prehľadávaniu literatúry sú čoraz menej udržateľné [2]. Zároveň sa mení aj charakter vedeckej komunikácie, ktorá je čoraz viac digitálna, distribuovaná a založená na prepojených informačných infraštruktúrach [1].

V tejto súvislosti možno formulovať niekoľko hlavných cieľov, ktoré sú typické pre moderné systémy na správu vedeckej literatúry:

- zefektívniť vyhľadávanie a filtrovanie článkov pomocou metód spracovania textu a strojového učenia,
- obohatiť metaúdaje o článkoch o automaticky extrahované informácie (kľúčové slová, tematické oblasti, vzťahy medzi dokumentmi),
- podporiť personalizované odporúčanie relevantných publikácií,
- vytvoriť priestor pre skupinovú prácu a zdieľanie znalostí v rámci výskumných tímov.

Nasledujúce podkapitoly sa zameriavajú na dve kľúčové oblasti, ktoré zohrávajú v takýchto systémoch centrálne postavenie: využitie strojového učenia pri správe vedeckých článkov a podpora skupinovej spolupráce.

1.2.1 Strojové učenie a jeho aplikácia v správe vedeckých článkov

Strojové učenie a hlboké učenie sa stali integrálnou súčasťou moderných informačných systémov, ktoré pracujú s veľkými objemami dát. Janiesch, Zschech a Heinrich poukazujú na to, že tieto metódy zásadným spôsobom menia spôsob, akým systémy spracúvajú dáta a podporujú rozhodovanie [6]. V oblasti vedeckej literatúry sa strojové učenie uplatňuje najmä pri:

- reprezentácii textu v numerickej podobe (TF–IDF, embeddingy),
- automatickej extrakcii kľúčových slov a fráz,
- klasifikácii a tematickom zaraďovaní článkov,
- odporúčaní príbuzných publikácií na základe obsahovej a citačnej podobnosti.

Klasickým východiskom pre reprezentáciu textu je metóda TF–IDF, ktorá váhovo zvýrazňuje termíny charakteristické pre konkrétny dokument a zároveň potláča bežné slová [7]. Na takto vytvorených vektoroch je potom možné počítať mieru podobnosti medzi článkami, napríklad pomocou cosine similarity. Nuri a Senyürek demonštrujú, že samotné abstrakty článkov stačia na to, aby bolo možné pomocou TF–IDF úspešne identifikovať tematicky príbuzné publikácie a radiť ich podľa podobnosti [8]. Podobné prístupy sú zaujímavé najmä ako ľahko implementovateľné a výpočtovo relatívne nenáročné riešenia pre menšie až stredne veľké repozitáre.

Na obsahové reprezentácie dokumentov nadväzujú aj odporúčacie systémy. Li napríklad navrhuje odporúčací systém pre vedecké články, ktorý spája TF–IDF reprezentáciu textu s konvolučnými neurónovými sieťami a link prediction v citačnej sieti [9]. Takýto hybridný prístup kombinuje informácie z textu s informáciami o citačných vzťahoch a ukazuje, že strojové učenie možno využiť nielen na spracovanie obsahu, ale aj na modelovanie vzťahov medzi publikáciami.

Ďalšou líniou výskumu sú systémy, ktoré spájajú klasické textové reprezentácie s modernými embedovacími modelmi. Rahman a kol. porovnávajú rôzne spôsoby reprezentácie textu (TF–IDF, Count Vectorizer, Sentence-BERT, USE, Mirror-BERT) pri klasifikácii vedeckých článkov z arXiv-u a ukazujú, že TF–IDF v kombinácii s logistickou regresiou môže byť aj napriek nástupu hlbokých modelov veľmi konkurencieschopné [10]. Nad takto klasifikovanými dokumentmi potom budujú odporúčací modul založený na cosine similarity, ktorý podporuje vyhľadávanie tematicky príbuzných článkov.

Strojové učenie zohráva dôležitú úlohu aj pri extrakcii kľúčových slov a fráz. Sarkar a kol. formulujú túto úlohu ako klasifikačný problém riešený neurónovou sieťou, ktorá rozlišuje medzi kandidátnymi frázami a skutočnými kľúčovými slovami [7]. V literatúre sa popri tom objavujú aj prístupy založené na hlbokých jazykových modeloch, ako je

BERT [11], ktoré umožňujú vytvárať kontextové reprezentácie slov a viet a aplikovať ich na úlohy, ako sú extrakcia kľúčových fráz, sumarizácia či klasifikácia. Projekty typu Semantic Scholar následne kombinujú tieto metódy s grafovým modelovaním literatúry [5], čím vznikajú bohaté „literatúrne grafy“ pre pokročilé vyhľadávanie a odporúčanie.

Celkovo možno konštatovať, že strojové učenie vytvára teoretické aj praktické východiská pre systémy, ktoré dokážu spracovať veľké množstvá vedeckých textov, automaticky z nich získavať štruktúrované informácie a prostredníctvom odporúčacích mechanizmov podporovať prácu výskumníkov.

1.2.2 Skupinová spolupráca a zdieľanie v systéme

Okrem individuálneho vyhľadávania a štúdia literatúry zohráva v modernej vede kľúčovú úlohu aj skupinová spolupráca. Rast počtu spoluautorských publikácií a tímovo orientovaných projektov odráža širšie trendy v globalizácii vedy a internacionalizácii výskumných sietí [2]. Vedecké tímy sa často skladajú z členov pôsobiacich na rôznych pracoviskách a v rôznych krajinách, ktorí potrebujú zdieľať nielen výsledky, ale aj samotnú literatúru, z ktorej pri svojej práci vychádzajú.

Digitálne nástroje a platformy preto čoraz častejšie integrujú funkcionality podporujúce:

- zdieľanie zoznamov článkov v rámci výskumných skupín,
- spoločné anotovanie dokumentov (poznámky, komentáre, zvýraznenia),
- kolektívne hodnotenie relevancie článkov pre konkrétny projekt,
- koordináciu čítania a rozdelenie literatúry medzi členov tímu.

Borgman zdôrazňuje, že infraštruktúra digitálneho bádania by mala podporovať celý cyklus vedeckej práce, od zbierania dát a literatúry až po ich zdieľanie a opätovné využitie [1]. Skupinová spolupráca v tejto infraštruktúre neznamená iba spoločné písanie článkov, ale aj kolektívne budovanie znalostnej základne, ktorá stojí za jednotlivými výstupmi.

Význam skupinovej spolupráce sa odráža aj v návrhu odporúčacích mechanizmov. Systémy môžu pri generovaní odporúčaní zohľadňovať nielen individuálne preferencie používateľov, ale aj agregované správanie celých skupín. Interakcie, ako sú hodnotenia článkov, ukladanie do zoznamov či frekvencia otvárania dokumentov, môžu slúžiť ako signály o kolektívnom záujme o určité témy. V kombinácii s obsahovými metódami a citačnými grafmi tak vzniká priestor pre hybridné odporúčacie prístupy, ktoré reflektujú potreby tímovo orientovaného výskumu [3, 5].

Prínosom takto koncipovaných systémov je, že nepracujú len s individuálnym používateľom, ale podporujú aj skupinovú dynamiku a zdieľanie znalostí. Tým prispievajú

k efektívnejšiemu využívaniu vedeckej literatúry a k lepšiemu prepájaniu výskumníkov naprieč disciplínami a inštitúciami.

1.3 Technologické východiská

1.3.1 Strojové učenie v praxi

Strojové učenie a hlboké učenie sa v posledných rokoch stali kľúčovými technológiami pri spracovaní veľkých objemov dát v rôznych doménach, od podnikových informačných systémov až po odporúčacie platformy a analytické nástroje. Janiesch, Zschech a Heinrich zdôrazňujú, že tieto metódy umožňujú transformovať surové dáta na znalosti a podporovať rozhodovanie na základe dátových analýz, čo zásadným spôsobom mení charakter moderných informačných systémov [6].

V oblasti správy vedeckej literatúry sa strojové učenie využíva najmä pri:

- spracovaní a reprezentácii textu vedeckých článkov v numerickej podobe,
- automatickej extrakcii kľúčových slov a fráz,
- klasifikácii a tematickom zaraďovaní publikácií,
- modelovaní podobnosti medzi dokumentmi,
- odporúčaní príbuzných článkov na základe obsahových alebo citačných vzťahov [3, 5].

Tradičné prístupy k reprezentácii textu vychádzajú z jednoduchých štatistických modelov, ako je *Bag of Words*, ktoré počítajú výskyty slov v dokumentoch a ignorujú ich poradie či kontext. Na tieto reprezentácie nadväzujú váhovacie schémy typu TF–IDF a klasické algoritmy strojového učenia (napríklad logistická regresia alebo SVM). Aj napriek nástupu hlbokých jazykových modelov ostávajú tieto metódy v praxi veľmi rozšírené vďaka svojej jednoduchosti, interpretovateľnosti a výpočtovej nenáročnosti [10].

Paralelne s tým sa rozvíjajú aj prístupy založené na hlbokom učení a kontextových embeddingoch, ktoré dokážu zachytiť jemnejšie významové vzťahy medzi slovami a vetami. Modely typu BERT [11] a ich nadstavby (napríklad Sentence-BERT) sa využívajú pri úlohách, ako sú klasifikácia textu, vyhľadávanie podobných dokumentov, extrakcia kľúčových fráz či sumarizácia. V prostredí vedeckej literatúry sú tieto prístupy často kombinované s citačnými a grafovými modelmi, čím vznikajú komplexné systémy na analýzu a odporúčanie publikácií [9].

1.3.2 Reprezentácia textu pomocou TF–IDF

Jedným z najpoužívanejších prístupov k reprezentácii textu je metóda *Term Frequency–Inverse Document Frequency* (TF–IDF). Jej cieľom je priradiť jednotlivým termínom (slovám alebo frázam) také váhy, ktoré odrážajú ich dôležitosť v rámci konkrétneho dokumentu aj celého korpusu. Základnou myšlienkou je, že termíny, ktoré sa v dokumente vyskytujú často, ale sú zároveň relatívne zriedkavé v celom korpuse, by mali mať vyššiu váhu než veľmi bežné slová.

Formálne sa TF–IDF váha pre termín t v dokumente d často definuje ako:

$$\text{tfidf}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t), \tag{1.1}$$

kde $\text{tf}(t, d)$ označuje frekvenciu termínu t v dokumente d a

$$\text{idf}(t) = \log \frac{N}{n_t}, \tag{1.2}$$

pričom N je celkový počet dokumentov v korpuse a n_t je počet dokumentov, v ktorých sa termín t vyskytuje. Výsledkom je vektorová reprezentácia dokumentu, v ktorej každá zložka zodpovedá jednému termínu zo slovníka a jej hodnota vyjadruje váhu daného termínu v dokumente.

Takto získané TF–IDF vektory sa využívajú v rôznych úlohách:

- pri klasifikácii dokumentov pomocou tradičných algoritmov strojového učenia,
- pri vyhľadávaní relevantných dokumentov na základe dotazu,
- pri výpočte podobnosti medzi dokumentmi,
- ako vstupné znaky pre algoritmy extrakcie kľúčových slov [7].

Na porovnávanie podobnosti dvoch dokumentov reprezentovaných vektormi TF–IDF sa často používa *cosine similarity*. Ide o mieru podobnosti založenú na kosíne uhla medzi dvoma vektormi v n-rozmernom priestore:

$$\text{cosine\_sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}, \tag{1.3}$$

kde $\mathbf{a}$ a $\mathbf{b}$ sú vektory reprezentujúce dva dokumenty. Hodnota blízka 1 znamená vysokú podobnosť, zatiaľ čo hodnota blízka 0 signalizuje, že dokumenty sú si obsahovo vzdialené. Táto kombinácia TF–IDF a cosine similarity sa bežne využíva pri odporúčaní podobných článkov alebo pri radení výsledkov vyhľadávania podľa relevancie [3].

Nuri a Senyürek demonštrujú, že samotné abstrakty vedeckých článkov postačujú na to, aby bolo možné pomocou TF–IDF a podobnostných mier identifikovať príbuzné publikácie a radiť ich podľa podobnosti k referenčnému článku [8]. Rahman a kol. ukazujú, že TF–IDF v kombinácii s logistickou regresiou môže byť aj v porovnaní s

modernejšími embeddingovými metódami konkurencieschopným riešením pri klasifikácii vedeckých článkov, pričom nad získanými reprezentáciami budujú aj odporúčací modul [10].

V prostredí správy vedeckej literatúry tak TF–IDF predstavuje dôležitý referenčný bod a často slúži ako základný model, voči ktorému sa porovnávajú pokročilejšie prístupy.

1.3.3 Kontextové embeddingy a model SBERT

Hoci TF–IDF a príbuzné štatistické metódy sú jednoduché a účinné, neberú do úvahy poradie slov ani ich kontextové vzťahy. Moderné prístupy k spracovaniu prirodzeného jazyka preto využívajú *kontextové embeddingy*, ktoré každému slovu alebo vete priradzujú vektorovú reprezentáciu závislú od jeho kontextu v texte. Kľúčovú úlohu v tomto smere zohrali modely založené na Transformer architektúre, najmä BERT (Bidirectional Encoder Representations from Transformers) [11].

BERT je predtrénovaný na veľkých textových korpusoch pomocou samoučiacich úloh a následne je možné ho jemne doladiť (*fine-tuning*) pre konkrétne úlohy, ako sú klasifikácia viet, rozpoznávanie pomenovaných entít alebo odpovedanie na otázky. Štandardný BERT je však primárne navrhnutý pre úlohy, kde sa hodnotí celá veta alebo sekvencia, nie pre efektívne vyhľadávanie podobných viet či dokumentov na základe vektorovej vzdialenosti.

Modely typu Sentence-BERT (SBERT) túto medzeru riešia tak, že na základe BERT architektúry konštruujú sieť, ktorá priamo produkuje vektorové reprezentácie viet vhodné na porovnávanie pomocou cosine similarity. Formálne možno takýto model chápať ako funkciu

$$f_\theta : S \to R^d, \tag{1.4}$$

kde S je množina viet alebo dokumentov a $f_\theta(s)$ je d-rozmerný embedding vety s naučený tak, aby vety s podobným významom mali vektorovo blízke reprezentácie. Podobnosť dvoch viet alebo dokumentov sa potom opäť počíta pomocou cosine similarity medzi ich embeddingmi:

$$\mathrm{sim}(s_1, s_2) = \mathrm{cosine\_sim}(f_\theta(s_1), f_\theta(s_2)). \tag{1.5}$$

Takéto embeddingy sa dajú využiť pri vyhľadávaní podobných článkov, klasifikácii dokumentov alebo zoskupovaní tematicky príbuzných textov. Rahman a kol. porovnávajú pri klasifikácii a odporúčaní vedeckých článkov viacero typov reprezentácií textu vrátane TF–IDF, Sentence-BERT, Universal Sentence Encoder a Mirror-BERT a ukazujú, že aj keď TF–IDF ostáva silnou bázovou metódou, kontextové embeddingy poskytujú ďalšie možnosti najmä pri jemnejšom rozlišovaní významových rozdielov [10].

V iných prácach sa embeddingy kombinujú s grafovými modelmi citačných sietí a metódami link prediction, čím vznikajú hybridné odporúčacie systémy pre vedecké články [5, 9].

Kontextové embeddingy tak predstavujú dôležitý smer vývoja v oblasti správy vedeckej literatúry, ktorý dopĺňa a rozširuje možnosti tradičných štatistických reprezentácií.

1.3.4 Výzvy implementácie v reálnom prostredí

Pri praktickej implementácii systémov na správu a odporúčanie vedeckých článkov je potrebné zohľadniť viaceré technické a dátové výzvy. Jednou z kľúčových je kvalita vstupných dokumentov, ktoré sú vo väčšine prípadov dostupné vo formáte PDF. Nie všetky PDF sú vytvorené ako „čistý“ text; v mnohých prípadoch ide o naskenované dokumenty, hybridné súbory alebo články s komplexným rozložením (viacstĺpcové formátovanie, množstvo tabuliek a obrázkov). To môže sťažovať proces extrakcie textu a viesť k chybám, ako sú:

- nezachytenie časti textu,
- pomiešané poradie viet alebo odstavcov,
- zlúčenie viacerých stĺpcov do jednej textovej línie,
- chýbajúce alebo skreslené znaky v dôsledku OCR.

Tieto problémy majú priamy dopad na kvalitu textových reprezentácií (či už ide o TF–IDF, alebo embeddingy) a následne aj na kvalitu vyhľadávania a odporúčania dokumentov.

Ďalšou dôležitou výzvou sú neúplné alebo nekonzistentné metaúdaje. Analýzy akademických vyhľadávačov ukazujú, že záznamy o článkoch často obsahujú chýbajúcich autorov, skrátené názvy, nejednoznačné roky vydania, chýbajúce kľúčové slová alebo viaceré duplicity toho istého článku [3, 4]. Podobné problémy sa vyskytujú aj pri veľkých vedeckých repozitároch a citačných databázach, kde sa dáta agregujú z rôznych zdrojov a v rôznej kvalite [5]. Neúplné metaúdaje komplikujú nielen vyhľadávanie, ale aj korektné prepojenie článkov, tvorbu citačných grafov a hodnotenie vplyvu jednotlivých publikácií.

Z pohľadu návrhu informačných systémov pre vedeckú literatúru preto vzniká potreba:

- robustných nástrojov na extrakciu textu a metaúdajov z PDF rôznej kvality,
- mechanizmov na detekciu a konsolidáciu duplicitných záznamov,

- procesov na validáciu a dopĺňanie metaúdajov (či už automatizovane, alebo s podporou používateľov),
- a metód, ktoré dokážu pracovať aj s neúplnými alebo čiastočne nekvalitnými dátami.

Ako upozorňuje Borgman, kvalita a integrita dátovej infraštruktúry je kľúčová pre spoľahlivú vedeckú komunikáciu v digitálnom prostredí [1]. V praxi to znamená, že technické riešenia musia počítať nielen s ideálnymi, ale aj s problematickými vstupmi a poskytovať mechanizmy na ich ošetrenie. Tieto výzvy formujú technologické východiská pre systémy spracúvajúce vedeckú literatúru a zároveň otvárajú priestor pre ďalší výskum v oblasti robustných metód spracovania textu, správy metaúdajov a odporúčacích algoritmov.

1.4 Súčasný stav výskumu a súvisiace práce

1.4.1 Význam a metódy extrakcie kľúčových slov

Kľúčové slová zohrávajú v oblasti správy a vyhľadávania vedeckej literatúry zásadnú úlohu. Predstavujú kompaktný slovný opis hlavného tematického zamerania článku a slúžia ako most medzi plnotextovým obsahom dokumentu a používateľskými dopytmi v informačných systémoch. Správne zvolené kľúčové slová uľahčujú indexáciu a zlepšujú presnosť vyhľadávania. Zároveň umožňujú efektívnejšie prepájanie príbuzných publikácií v rámci databáz a citačných sietí [3].

Vzhľadom na rast počtu článkov a rôznorodosť publikačných platforiem je manuálne prideľovanie kľúčových slov často nedostatočné, nejednotné alebo úplne chýba. To vedie k nekvalitným metaúdajom a znižuje účinnosť vyhľadávania a odporúčania literatúry [5]. Preto sa v posledných desaťročiach intenzívne rozvíjajú metódy automatickej extrakcie kľúčových slov (keyphrase extraction), ktoré majú za cieľ identifikovať najdôležitejšie pojmy v texte bez priameho zásahu človeka.

Metódy automatickej extrakcie kľúčových slov sa dajú rámcovo rozdeliť do niekoľkých kategórií:

- *štatistické a heuristické prístupy*, ktoré využívajú frekvenciu slov, ich rozloženie v texte, dĺžku fráz a jednoduché jazykové pravidlá (napríklad TF–IDF, RAKE či algoritmy založené na grafoch typu TextRank),
- *dozorované metódy*, v ktorých sa model strojového učenia učí rozlišovať medzi kľúčovými a nekľúčovými frázami na základe anotovaných dát,
- *neurónové a hlboké modely*, ktoré pri extrakcii kľúčových slov využívajú kontextové reprezentácie textu a sekvenčné modely.

Sarkar a kol. formulujú extrakciu kľúčových fráz ako klasifikačný problém riešený pomocou neurónových sietí [7]. Kandidátne frázy sú najprv odvodené z textu na základe syntaktických a štatistických kritérií a následne sú hodnotené modelom, ktorý sa učí rozlišovať medzi relevantnými a nerelevantnými kľúčovými frázami. Tento typ prístupu ukazuje, že kombinácia tradičných štatistických znakov (napríklad frekvencia, pozícia v texte či dĺžka frázy) s modelmi strojového učenia môže výrazne zlepšiť kvalitu automaticky generovaných kľúčových slov oproti čisto heuristickým metódam.

Výskum zároveň poukazuje na to, že kvalita kľúčových slov a metaúdajov má priamy vplyv na viditeľnosť článkov v akademických vyhľadávačoch a na ich citačný potenciál [3]. Projekt Semantic Scholar demonštruje, že automatizované spracovanie textu a metaúdajov, vrátane extrakcie kľúčových pojmov a vzťahov, je kľúčové pre vytváranie tzv. literatúrnych grafov, ktoré spájajú články na základe ich obsahovej a citačnej príbuznosti [5]. V takýchto grafoch zohrávajú extrahované kľúčové pojmy dôležitú rolu pri identifikácii tém, komunitných štruktúr a významných publikácií v rámci odboru.

Automatická extrakcia kľúčových slov sa tak stala jednou z ústredných úloh v oblasti správy vedeckej literatúry. Výstupy týchto metód slúžia ako vstup pre ďalšie procesy, ako sú tvorba indexov pre vyhľadávanie podľa tém, rozšírenie metaúdajov článkov a výpočet obsahovej podobnosti medzi dokumentmi.

1.4.2 Hlboké učenie a jeho aplikácie v NLP

Rozvoj hlbokého učenia zásadne ovplyvnil oblasť spracovania prirodzeného jazyka (NLP). Tradičné prístupy sa spoliehali na ručne navrhované znaky a štatistické modely nad jednoduchými reprezentáciami typu Bag of Words alebo TF–IDF. Moderné metódy naopak využívajú neurónové siete na automatické učenie reprezentácií slov, viet a dokumentov z veľkých textových korpusov [6]. Takto získané reprezentácie, známe ako embeddingy, zachytávajú kontext a sémantické vzťahy medzi slovami, vďaka čomu umožňujú presnejšie modelovanie významu textu.

K zásadnému posunu v NLP prispel nástup Transformer architektúr a modelov ako BERT (Bidirectional Encoder Representations from Transformers). Devlin a kol. ukazujú, že predtrénovaný model BERT dokáže vďaka obojsmernému spracovaniu kontextu dosiahnuť špičkový výkon v širokej škále úloh, vrátane klasifikácie viet, odpovedania na otázky či rozpoznávania pomenovaných entít [11]. Kľúčovou myšlienkou je, že model je najprv natrénovaný na veľkom všeobecnom korpuse prostredníctvom samoučiacich úloh a následne je jemne doladený (*fine-tuning*) na konkrétne úlohy.

Hlboké modely na báze Transformerov majú priamu relevanciu aj pre spracovanie vedeckej literatúry. Umožňujú:

- vytvárať kontextové vektorové reprezentácie abstraktov a plných textov článkov,

- identifikovať tematické podobnosti medzi publikáciami aj pri odlišnej terminológii,
- podporovať extrakciu kľúčových pojmov a vzťahov medzi nimi,
- zlepšovať kvalitu odporúčaní článkov nad rámec jednoduchých štatistických metód.

V praktických systémoch sa hlboké modely často kombinujú s grafovými štruktúrami zachytávajúcimi citačné a autorovské vzťahy. Príkladom je konštrukcia „literatúrneho grafu“, v ktorom sú uzly reprezentované článkami a hrany rôznymi typmi vzťahov (citácie, spoloční autori, tematická podobnosť). Ammar a kol. opisujú, ako sa v takomto grafe kombinujú informácie z textu, metaúdajov a citačných vzťahov na podporu vyhľadávania a odporúčania v systéme Semantic Scholar [5]. Podobné prístupy sú využívané aj v hybridných odporúčacích systémoch, ktoré spájajú obsahovú analýzu textu s modelovaním štruktúry citačnej siete [9].

V literatúre sa objavujú aj práce, ktoré porovnávajú klasické štatistické reprezentácie textu (napríklad TF–IDF) s modernými embeddingovými modelmi pri úlohách klasifikácie a odporúčania vedeckých článkov. Rahman a kol. ukazujú, že TF–IDF v kombinácii s jednoduchými modelmi strojového učenia môže byť aj naďalej konkurencieschopným riešením, zatiaľ čo kontextové embeddingy poskytujú potenciál na zlepšenie výkonu v komplexnejších scenároch [10].

Celkový trend v oblasti NLP a správy vedeckej literatúry smeruje k integrácii hlbokých jazykových modelov do existujúcich informačných systémov. Tie poskytujú bohatšie reprezentácie textu, ktoré je možné využiť pri vyhľadávaní, odporúčaní, sumarizácii a ďalších pokročilých analytických funkciách. ::contentReference[oaicite:0]index=0

Kapitola 2

Návrh systému

2.1 Architektúra systému

2.1.1 Funkčné požiadavky

2.1.2 Návrh rozhrania

Kapitola 3

Implementácia systému

3.1 Implementácia backendu

3.1.1 Implementácia frontendovej časti

3.1.2 Integrácia strojového učenia

Kapitola 4

Výskumná časť a testovanie

4.1 Testovanie implementácie

4.1.1 Testovacie prostredie a dátová sada

Testovanie implementácie odporúčacieho systému prebiehalo v kontrolovanom prostredí, ktoré zodpovedá typickej konfigurácii servera pre menšie webové aplikácie. Backendová časť systému bola nasadená ako Django aplikácia využívajúca databázový server PostgreSQL, frontend bol implementovaný v technológii React ako jednostránková aplikácia (SPA). Všetky testy boli realizované na rovnakej verzii kódu, aby sa eliminoval vplyv zmien v implementácii počas meraní.

Na účely testovania bola pripravená databáza obsahujúca niekoľko desiatok vedeckých článkov z rôznych oblastí. Pri každom článku boli vyplnené metaúdaje:

- názov (title),
- abstrakt alebo hlavný text (content),
- kľúčové slová (keywords),
- tematické kategórie (categories),
- autori (authors).

Tieto polia sa používajú pri budovaní textového korpusu pre výpočet TF–IDF reprezentácií. Funkcia `build_corpus` vytvára pre každý článok text spojením názvu, obsahu, kľúčových slov, kategórií a autorov, ktorý následne vstupuje do vektorizačného procesu.

Samotné TF–IDF vektory sa počítajú pomocou knižnice `scikit-learn` s použitím triedy `TfidfVectorizer`. Funkcia `fit_tfidf` zodpovedá za natrénovanie vektorizačného modelu na celom korpuse článkov a vráti maticu reprezentácií, ktorá je ďalej spracovaná v indexovacom module. Pre každý článok je následne vygenerovaný hustý

vektor a uložený do databázy ako číselný zoznam, ktorý je možné neskôr použiť pri výpočte podobnosti.

4.1.2 TF–IDF index a príprava vektorových reprezentácií

Základom výskumnej časti je hotový model založený na TF–IDF reprezentácii článkov. Všetky dostupné články v databáze sa periodicky spracujú dávkovým spôsobom: najprv sa pre každý článok vytvorí textová reprezentácia spojením názvu, obsahu, kľúčových slov, kategórií a autorov, následne sa tieto texty vektorujú pomocou TF–IDF a vzniknuté vektory sa uložia ako samostatná vrstva dát.

Výsledkom je:

- jednotný slovník termínov vytvorený na základe celého korpusu,
- matica TF–IDF vektorov, kde každý riadok zodpovedá jednému článku,
- tabuľka s uloženými vektorovými reprezentáciami (napríklad vo forme poľa reálnych čísel).

Takto pripravené vektory tvoria podklad pre ďalší výskum. Keďže sú uložené v databáze a oddelené od aplikačnej logiky, je možné ich exportovať a analyzovať mimo samotnej webovej aplikácie. Typickým postupom bude:

1. export TF–IDF vektorov a súvisiacich metaúdajov (ID článku, názov, kľúčové slová) do formátu CSV alebo JSON,
2. načítanie týchto dát v samostatnom skripte alebo v prostredí Jupyter Notebook,
3. vykonanie experimentov, pri ktorých sa porovnávajú rôzne spôsoby výpočtu podobnosti a rôzne prahy či stratégie radenia výsledkov.

V rámci výskumnej časti sa plánuje pracovať s fixnou verziou TF–IDF indexu (napríklad „tfidf-v1“), aby boli výsledky experimentov reprodukovateľné. Zmena parametrov vektorizačného modelu (maximálny počet znakov, rozsah n-gramov, stopslová) môže byť predmetom samostatných doplnkových experimentov.

4.1.3 Plán porovnania TF–IDF a modelu SBERT

Na to, aby bolo možné posúdiť kvalitu TF–IDF reprezentácií v kontexte podobnosti vedeckých článkov, je potrebné vytvoriť experimentálny rámec, v ktorom sa budú porovnávať rôzne prístupy k reprezentácii textu. Plán počíta s dvoma hlavnými krokmi:

1. **Offline vyhodnotenie TF–IDF modelu mimo webovej aplikácie.**

2. **Budúce doplnenie modelu na báze SBERT a jeho porovnanie s TF–IDF.**

Pri offline vyhodnotení TF–IDF sa počíta s tým, že sa nad exportovanými vektormi z databázy vytvorí samostatný experimentálny skript, ktorý bude:

- pracovať s množinou ručne zvolených „kotviacich“ článkov,
- pre každý takýto článok vypočítať rebríček najpodobnejších článkov na základe cosine similarity,
- porovnať tieto rebríčky s očakávaniami (napríklad pomocou ručného hodnotenia alebo pripraveného zoznamu tematicky príbuzných článkov).

Pre formálnejšie vyhodnotenie je možné pre vybranú podmnožinu článkov pripraviť:

- binárne označené páry „podobný / nepodobný“, na ktorých sa bude sledovať, či má podobný pár vyššie podobnostné skóre než nepodobný,
- rebríčky relevantných článkov, pre ktoré sa dajú počítať metriky typu Precision@k, Recall@k alebo nDCG@k.

V ďalšej fáze sa plánuje doplniť druhá reprezentácia textu založená na modeli typu SBERT (Sentence-BERT). Tento model bude pre tie isté články generovať husté embeddingy s menším rozmerom, ktoré umožnia počítať podobnosť článkov opäť na základe cosine similarity. Celý experiment potom bude pozostávať z porovnania:

- TF–IDF podobnosti,
- SBERT podobnosti,

a to na tých istých dátach a tých istých hodnotiacich sadách. Výsledkom bude porovnanie dvoch prístupov, ktoré sa dá prezentovať v podobe tabuliek a grafov (napríklad hodnoty Precision@10 pre oba modely alebo porovnanie priemernej pozície očakávane relevantných článkov v rebríčku).

4.1.4 Testovanie modulov odporúčania a podobných článkov

Popri offline porovnávaní reprezentácií textu je dôležité overiť aj správanie jednotlivých modulov v rámci celého systému. V praxi ide predovšetkým o dve kategórie funkcionalít:

- modul odporúčania článkov na základe používateľského profilu („recommender“),
- modul zobrazovania podobných článkov k zvolenému článku („similar articles“).

Modul podobných článkov využíva vektorové reprezentácie článkov a pre zadaný článok vypočíta rebríček najpodobnejších dokumentov na základe cosine similarity. Testovanie tohto modulu sa môže sústrediť na:

- kvalitatívne vyhodnotenie top-k výsledkov pre niekoľko vybraných článkov (kontrola, či sú odporúčané články tematicky príbuzné),
- kontrolu, že referenčný článok sa nenachádza medzi odporúčanými,
- porovnanie rebríčkov pre TF–IDF a SBERT reprezentáciu na rovnakých vstupoch.

Modul odporúčania na základe používateľa pracuje s profilom, ktorý je odvodený z interakcií používateľa s článkami (napríklad hodnotenia, označenie ako obľúbené, pridanie do zoznamu). Tento profil je možné reprezentovať napríklad ako priemer vektorov článkov, s ktorými používateľ pracoval. Testovanie takéhoto modulu môže zahŕňať:

- vytvorenie niekoľkých syntetických používateľských profilov so zameraním na konkrétne témy,
- vyhodnotenie odporúčaných článkov pre každý profil (či sú v súlade s očakávanou témou),
- porovnanie odporúčaní generovaných na báze TF–IDF a na báze SBERT.

Popri kvalitatívnej analýze je možné definovať aj jednoduché kvantitatívne metriky, napríklad:

- podiel odporúčaných článkov, ktoré patria do očakávanej kategórie,
- priemerná pozícia manuálne označených relevantných článkov v odporúčanom rebríčku,
- miera prekrývania odporúčaní medzi TF–IDF a SBERT variantom.

Takto navrhnuté testovacie scenáre vytvárajú rámec pre ďalšie podkapitoly, v ktorých je možné podrobnejšie analyzovať výsledky, porovnávať správanie rôznych modelov a diskutovať ich výhody a obmedzenia.

4.1.5 Porovnanie metód strojového učenia

4.1.6 Vyhodnotenie výsledkov

Kapitola 5

Diskusia a zhodnotenie výsledkov

5.1 Výhody a nevýhody implementovaných funkcií

5.1.1 Obmedzenia a výzvy pri implementácii

5.1.2 Možnosti vylepšenia a rozšírenia

Kapitola 6

Zhrnutie hlavných prínosov práce

Záver

Literatúra

[1] Christine L. Borgman. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT Press, Cambridge, MA, 2007.

[2] Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015. Navštívené: 2025-11-26.

[3] Jöran Beel, Bela Gipp, and Erik Wilde. Academic search engine optimization (aseo): Optimizing scholarly literature for google scholar & co. *Journal of Scholarly Publishing*, 41(2):176–190, 2010. Navštívené: 2025-11-26.

[4] Jöran Beel and Bela Gipp. Google scholar's ranking algorithm: An introductory overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, pages 230–241, Rio de Janeiro, Brazil, 2009. International Society for Scientometrics and Informetrics. Navštívené: 2025-11-26.

[5] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans, Louisiana, 2018. Association for Computational Linguistics. Navštívené: 2025-11-26.

[6] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021. Navštívené: 2025-05-13.

[7] Kamal Sarkar, Mita Nasipuri, and Suranjan Ghose. A new approach to keyphrase extraction using neural networks. *IJCSI International Journal of Computer Science Issues*, 7(2, No 3):16–23, 2010. Navštívené: 2025-05-13.

[8] Yusra Nuri and Edip Senyürek. Research abstracts similarity implementation by using tf-idf algorithm. *IOSR Journal of Computer Engineering*, 27(1, Ser. 4):4–10, 2025. Navštívené: 2025-12-09.

[9] Weijuan Li. Scientific paper recommender system using deep learning and link prediction in citation network. *Heliyon*, 10(14):e34685, 2024. Navštívené: 2025-12-09.

[10] Shadikur Rahman, Hasibul Karim Shanto, Umme Ayman Koana, and Syed Muhammad Danish. Automated research article classification and recommendation using nlp and ml. *arXiv preprint arXiv:2510.05495*, 2025. Navštívené: 2025-12-09.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. Navštívené: 2025-05-13.

Príloha A: