Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie

Extrahovanie kľúčových slov z dokumentov

Semestrálne zadanie
Manažment znalostí
Zimný semester 2025/2026

Bc. Boris Daňko
Bc. Roland Palgut
Bc. Marek Puškáš

Košice, 28. októbra 2025

Obsah

1 Popis problému

V súčasnosti dochádza k neustálemu rastu objemu digitálnych dát vo forme textových dokumentov, vedeckých článkov, správ, e-mailov či príspevkov na sociálnych sieťach. Táto expanzia textových dát vyvoláva potrebu efektívnych metód na ich automatizované spracovanie, analýzu a porozumenie obsahu. Jednou z kľúčových úloh v oblasti spracovania prirodzeného jazyka (Natural Language Processing - NLP) a dolovania textu (Text Mining) je extrakcia kľúčových slov (Keyword Extraction - KE).

Cieľom extrakcie kľúčových slov je identifikovať tie termíny, ktoré najlepšie vystihujú obsah a tému daného dokumentu. Kľúčové slová majú zásadný význam pri úlohách, ako sú sumarizácia textu, klasifikácia dokumentov, vyhľadávanie informácií či automatické značkovanie obsahu [8]. Automatizovaná extrakcia kľúčových slov zároveň znižuje časovú náročnosť manuálneho anotovania a umožňuje spracovávať rozsiahle textové korpusy bez zásahu človeka.

Podľa štúdie [8] sú hlavné výzvy v tejto oblasti spojené s presnosťou a spoľahlivosťou extrakcie. Tradičné štatistické prístupy (napr. Log-Likelihood Test - LLT) často vykazujú závislosť od referenčných korpusov, čím sa znižuje ich robustnosť. Problémom býva aj nedostatočné filtrovanie neplnovýznamových slov, ako sú predložky či spojky, ktoré dosahujú vysokú frekvenciu, no nenesú významovú hodnotu.

Ďalšou významnou výzvou je rozpoznávanie viacslovných termov (multi-word expressions) a doménovo špecifických kľúčových pojmov, ktoré sa nemusia vyskytovať v štandardných lexikálnych databázach. Moderné prístupy preto čoraz viac využívajú sémantické modely, ktoré umožňujú zachytiť významové súvislosti medzi slovami, nielen ich frekvenciu výskytu.

Problém extrakcie kľúčových slov je úzko spätý aj s inými úlohami spracovania textu, ako je sumarizácia dokumentov či automatizované odpovedanie na otázky (Question Answering - QA). Moderné extraktory kľúčových slov a sumarizačné systémy využívajú modely typu BERT alebo KeyBERT [4], ktoré umožňujú systémom lepšie pochopiť kontext používateľského dopytu a prispôsobiť mu výslednú sumarizáciu či extrakciu informácií.

Celkovo možno povedať, že problém extrakcie kľúčových slov spočíva v potrebe vytvoriť metódu, ktorá bude presná, doménovo prispôsobiteľná a nezávislá od manuálneho zásahu človeka. Takáto metóda by mala zohľadniť nielen frekvenčné charakteristiky slov, ale aj ich sémantické a štylistické väzby v rámci textu.

2 Popis existujúcich metód

Extrakcia kľúčových slov sa historicky vyvíjala od jednoduchých štatistických prístupov až po moderné modely založené na hlbokých neurónových sieťach. V tejto časti sú predstavené najpoužívanejšie prístupy, ktoré sa vyskytujú v literatúre a prezentáciách k text miningu.

2.1 Tradičné modely pre vyhľadávanie informácií

V oblasti vyhľadávania informácií (Information Retrieval - IR) existujú tri základné klasické modely [6]:

1. **Boolovský model** - predstavuje najjednoduchší prístup, v ktorom sú dokumenty a dopyty reprezentované ako množiny termov. Vyhľadávanie je založené na logických operátoroch AND, OR, NOT. Tento model je efektívny, ale neumožňuje určiť mieru relevancie dokumentu.
2. **Vektorový model** - dokumenty aj dopyty sú reprezentované ako vektory termov v n-dimenzionálnom priestore. Miera podobnosti sa počíta pomocou kosínusovej podobnosti, pričom váhovanie termov je najčastejšie realizované pomocou metódy TF-IDF.
3. **Pravdepodobnostný model** - odhaduje pravdepodobnosť, že používateľ považuje dokument za relevantný pre svoj dopyt. Tento prístup predstavuje základ pre moderné re-ranking algoritmy v systémoch vyhľadávania informácií.

2.2 Štatistické metódy - LLT a TF-IDF

Tradičné korpusovo orientované metódy, ako napríklad **Dunningov Log-Likelihood Test (LLT)**, patria medzi najstaršie prístupy ku kľúčovosti slov [3]. LLT hodnotí štatistickú významnosť rozdielov vo frekvencii slov medzi cieľovým a referenčným korpusom. Slová s výrazne vyššou frekvenciou v cieľovom korpuse sú považované za kľúčové. Nevýhodou LLT je jeho závislosť od kvality referenčného korpusu a neschopnosť pracovať s významovou redundanciou.

Naopak, metóda **Term Frequency-Inverse Document Frequency (TF-IDF)** [9] je jednou z najpoužívanejších techník v NLP. Hodnotí dôležitosť termu na základe jeho relatívnej frekvencie v rámci dokumentu a zriedkavosti v celom korpuse. TF-IDF umožňuje jednoduché, ale účinné určenie relevantných termov bez potreby referenčného korpusu. Využíva sa nielen v extrakcii kľúčových slov, ale aj vo vyhľadávačoch, odporúčacích systémoch či textovej klasifikácii.

Štúdia [7] navrhuje vylepšenú verziu TF-IDF, ktorá zohľadňuje distribúciu slov naprieč dokumentmi a filtruje gramatické slová s nízkou informačnou hodnotou. Táto metóda využíva vnútorné štatistiky korpusu a eliminuje potrebu externých dát, čím sa zvyšuje jej prenositeľnosť a univerzálnosť.

2.3 Latentné sémantické a vektorové modely

Moderné NLP techniky rozširujú pôvodné frekvenčné metódy o sémantické chápanie textu. **Latentné sémantické indexovanie (LSI)** využíva singulárnu dekompozíciu matíc (SVD) na zistenie skrytých významových vzťahov medzi slovami a dokumentmi [1]. Umožňuje redukciu dimenzie a potláča problémy synonymie a polysémie, ktoré sú typické pre prirodzený jazyk.

Ďalším krokom boli modely založené na distribučných reprezentáciách, ako **Word2Vec**, **GloVe** a neskôr **BERT**, ktoré umožňujú zachytiť význam slov na základe ich kontextu.

Takéto modely sú základom moderných extrakčných nástrojov, ako napríklad **KeyBERT**, ktorý využíva vektorové reprezentácie a kosínusovú podobnosť na určenie najrelevantnejších kľúčových slov v texte [4].

2.4 Transformátorové a hybridné metódy

Najnovšie prístupy v oblasti extrakcie kľúčových slov využívajú **transformátorové architektúry** (napr. BERT, T5, PEGASUS), ktoré umožňujú pochopiť kontext textu na hlbšej úrovni [2, 5, 10]. Tieto modely dokážu kombinovať informácie o sémantike, gramatike aj pragmatike jazyka a sú vhodné pre úlohy, kde je dôležité pochopiť zámer a význam celého dokumentu.

V oblasti sumarizácie a odpovedania na otázky (Question Answering) sa objavujú metódy, ktoré využívajú extrahované kľúčové slová na navádzanie modelu počas generovania sumarizovaného textu [4]. Takéto prístupy spájajú extrakciu a generáciu a predstavujú prechod k hybridným systémom schopným integrovať viaceré NLP techniky.

2.5 Zhrnutie

Klasické metódy (LLT, TF-IDF) sú výpočtovo jednoduché a interpretovateľné, no často ignorujú významové súvislosti. Na druhej strane, moderné vektorové a transformátorové prístupy umožňujú modelom pochopiť kontext a význam slov, ale vyžadujú veľké množstvo dát a výpočtových zdrojov. V praxi sa preto často používajú **hybridné riešenia**, ktoré kombinujú štatistické váhovanie s vektorovou reprezentáciou textu, čím sa dosahuje kompromis medzi efektivitou a presnosťou extrakcie.

3 Návrh metódy a výber testovacích dát

Cieľom navrhovanej metódy je umožniť flexibilnú a modulárnu extrakciu kľúčových slov z textových dokumentov rôznych typov, ako sú PDF alebo TXT súbory. Metóda kombinuje tri prístupy: TF-IDF, YAKE a KeyBERT, pričom každý z nich poskytuje odlišný spôsob hodnotenia významu termov v dokumente.

- **TF-IDF** vyhodnocuje relatívnu dôležitosť termov na základe ich frekvencie v dokumente a zriedkavosti naprieč korpusom.
- **YAKE** je nezávislý, korpusovo orientovaný algoritmus založený na štatistike n-gramov a váh slov.
- **KeyBERT** využíva moderné transformátorové modely (BERT), ktoré zachytávajú kontext a význam slov vo vete, vrátane možnosti použitia Maximal Marginal Relevance (MMR) pre diverzifikované kľúčové slová.

Všetky metódy umožňujú upravovať parametre, ako sú počet extrahovaných kľúčových slov, rozsah n-gramov alebo diverzita v KeyBERT, čo umožňuje testovanie na rôznych typoch textov a dopytoch.

Pre testovanie metódy boli použité dva zdroje dát:

1. **Vlastné nahrané dokumenty** - používateľ môže nahrať ľubovoľný súbor vo formáte TXT alebo PDF a overiť spracovanie reálnych textových dát.
2. **Vzorky zo zložky DATA/** - pripravené textové súbory s rôznou dĺžkou a obsahom, slúžiace na porovnanie výkonu a presnosti jednotlivých metód.

Pred spracovaním textu sa vykonáva predspracovanie, ktoré odstraňuje nadbytočné medzery, čísla a špeciálne znaky, čím sa zvyšuje kvalita extrahovaných kľúčových slov.

4 Experimentovanie a vyhodnotenie

Po implementácii metódy boli vykonané experimenty s cieľom porovnať kvalitu a relevantnosť kľúčových slov generovaných rôznymi prístupmi. Pre každý dokument sa extrahovali kľúčové slová pomocou TF-IDF, YAKE a KeyBERT a výsledky sa vizualizovali prostredníctvom barových grafov a wordcloudov.

Hlavné parametre testované počas experimentov boli:

- **Počet kľúčových slov** (top_n)
- **Rozsah n-gramov**
- **Diverzita v KeyBERT** pri použití MMR

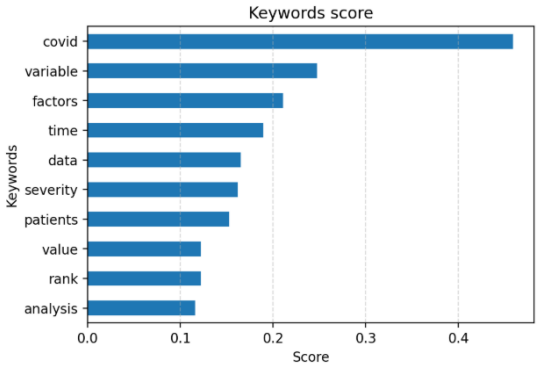
4.1 Zhrnutie experimentov

- TF-IDF je rýchly a účinný pre kratšie dokumenty s jasne definovanými termami.
- YAKE eliminuje bežné stopwords a irelevantné výrazy.
- KeyBERT poskytuje najlepší kontextový opis, hlavne pri viacslovných termoch; MMR zvyšuje diverzitu výsledkov.

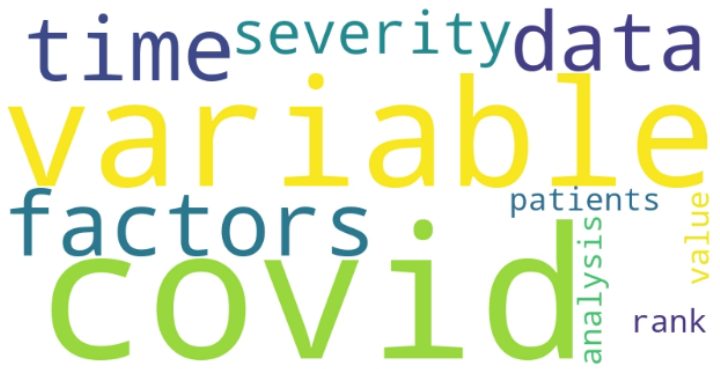
Kombinácia týchto metód poskytuje robustný nástroj na extrakciu kľúčových slov a umožňuje prispôsobiť parametre podľa typu a dĺžky dokumentu.

	Keyword	Score	Document
0	covid	0.4587	A composite ranking of risk factors for COVID-19 time-t
1	variable	0.2477	A composite ranking of risk factors for COVID-19 time-t
2	factors	0.211	A composite ranking of risk factors for COVID-19 time-t
3	time	0.1896	A composite ranking of risk factors for COVID-19 time-t
4	data	0.1651	A composite ranking of risk factors for COVID-19 time-t
5	severity	0.1621	A composite ranking of risk factors for COVID-19 time-t
6	patients	0.1529	A composite ranking of risk factors for COVID-19 time-t
7	rank	0.1223	A composite ranking of risk factors for COVID-19 time-t
8	value	0.1223	A composite ranking of risk factors for COVID-19 time-t
9	analysis	0.1162	A composite ranking of risk factors for COVID-19 time-t

(a) Tabuľka kľúčových slov TF-IDF



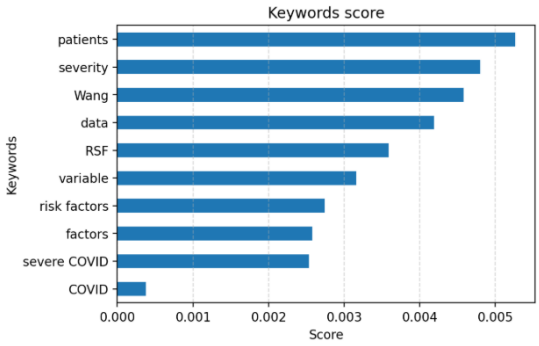
(b) Barh graf TF-IDF



Obr. 1: Výsledok extrakcie kľúčových slov metódou TF-IDF - tabuľka, barh graf a wordcloud

	Keyword	Score	Document
0	patients	0.0053	A composite ranking of risk factors for COVID-19 ti
1	severity	0.0048	A composite ranking of risk factors for COVID-19 ti
2	Wang	0.0046	A composite ranking of risk factors for COVID-19 ti
3	data	0.0042	A composite ranking of risk factors for COVID-19 ti
4	RSF	0.0036	A composite ranking of risk factors for COVID-19 ti
5	variable	0.0032	A composite ranking of risk factors for COVID-19 ti
6	risk factors	0.0027	A composite ranking of risk factors for COVID-19 ti
7	factors	0.0026	A composite ranking of risk factors for COVID-19 ti
8	severe COVID	0.0025	A composite ranking of risk factors for COVID-19 ti
9	COVID	0.0004	A composite ranking of risk factors for COVID-19 ti

(a) Tabuľka kľúčových slov YAKE

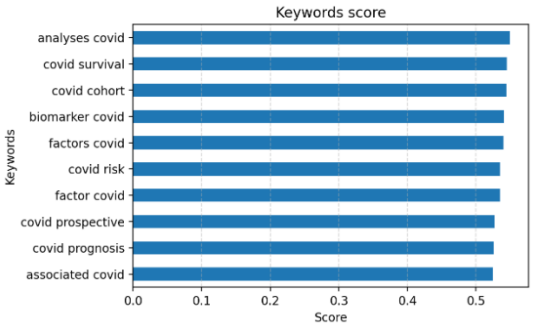


(b) Barh graf YAKE

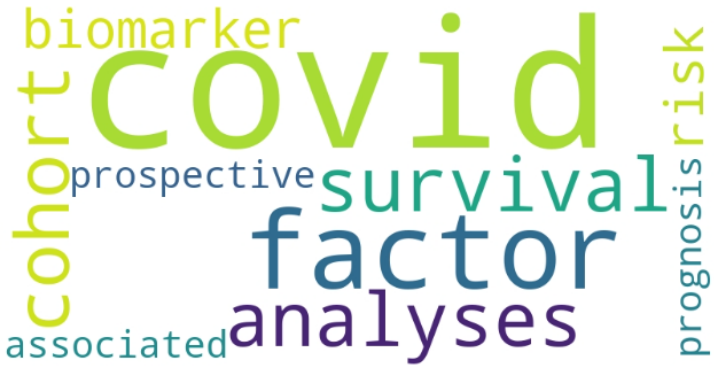
Obr. 2: Výsledok extrakcie kľúčových slov metódou YAKE

	Keyword	Score	Document
0	analyses covid	0.5492	A composite ranking of risk factors for COVIE
1	covid survival	0.5453	A composite ranking of risk factors for COVIE
2	covid cohort	0.5443	A composite ranking of risk factors for COVIE
3	biomarker covid	0.5407	A composite ranking of risk factors for COVIE
4	factors covid	0.5401	A composite ranking of risk factors for COVIE
5	covid risk	0.5353	A composite ranking of risk factors for COVIE
6	factor covid	0.5348	A composite ranking of risk factors for COVIE
7	covid prospective	0.5268	A composite ranking of risk factors for COVIE
8	covid prognosis	0.5256	A composite ranking of risk factors for COVIE
9	associated covid	0.5243	A composite ranking of risk factors for COVIE

(a) Tabuľka kľúčových slov KeyBERT



(b) Barh graf KeyBERT



Obr. 3: Výsledok extrakcie kľúčových slov metódou KeyBERT

5 Referencie

[1] Scott Deerwester et al. „Indexing by latent semantic analysis“. In: *Journal of the American society for information science* 41.6 (1990), s. 391–407.

[2] Jacob Devlin et al. „BERT: Pre-training of deep bidirectional transformers for language understanding“. In: *NAACL-HLT* (2019), s. 4171–4186.

[3] Ted Dunning. „Accurate methods for the statistics of surprise and coincidence“. In: *Computational Linguistics* 19.1 (1993), s. 61–74.

[4] Maarten Grootendorst. „KeyBERT: Minimal keyword extraction with BERT“. In: *arXiv preprint arXiv:2009.10849* (2020).

[5] Mike Lewis et al. „BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension“. In: *arXiv preprint arXiv:1910.13461* (2020).

[6] Christopher D. Manning, Prabhakar Raghavan a Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[7] Stuart Rose. „An extended TF-IDF method for automatic keyword extraction“. In: *Proceedings of the International Conference on Data Mining*. 2011, s. 45–52.

[8] Stuart Rose et al. „Automatic keyword extraction from individual documents“. In: *Text Mining: Applications and Theory* (2010), s. 1–20.

[9] Gerard Salton a Christopher Buckley. „Term-weighting approaches in automatic text retrieval“. In: *Information Processing & Management* 24.5 (1988), s. 513–523.

[10] Jingqing Zhang, Yao Zhao a Mohammad Saleh. „PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization“. In: *arXiv preprint arXiv:1912.08777* (2020).