

ČVUT, FAKULTA INFORMAČNÍCH TECHNOLOGIÍ  
VYHLEDÁVÁNÍ NA WEBU A V MULTIMEDIÁLNÍCH DATABÁZÍCH  
LETNÍ SEMESTR 2019/2020  
ZÁVĚREČNÁ ZPRÁVA K PROJEKTU

---

## LSI vektorový model

---

David Mašek a Kristýna Klesnilová

13. května 2020

## OBSAH

<b>1</b>	<b>Popis projektu</b>	<b>3</b>
<b>2</b>	<b>Způsob řešení</b>	<b>3</b>
2.1	Preprocessing dokumentů . . . . .	3
2.2	Výpočet vah termů . . . . .	3
2.3	Implementace LSI . . . . .	4
2.4	Vyhodnocení dotazu . . . . .	5
<b>3</b>	<b>Implementace</b>	<b>5</b>
<b>4</b>	<b>Příklad výstupu</b>	<b>5</b>
<b>5</b>	<b>Experimentální sekce</b>	<b>6</b>
5.1	Určení optimálního počtu konceptů . . . . .	6
5.2	Porovnání vlivu LSI na kvalitu výsledků vyhledávání s ohledem na výskyt synonym a homonym . . . . .	7
5.2.1	Synonyma . . . . .	7
5.2.2	Homonyma . . . . .	7
5.3	Porovnání průchodu pomocí LSI vektorového modelu se sekvenčním průchodem databáze s ohledem na čas vykonání dotazu . . . . .	8
5.4	Vliv různých vnitřních parametrů na výkon algoritmu (změna počtu konceptů, změna počtu extrahovaných termů, použití lemmatizace namísto stemmingu, odstranění číslovek při preprocessingu, použití jiného vzorce na výpočet vah termů) . . . . .	9
5.4.1	Změna počtu konceptů . . . . .	9
5.4.2	Změna počtu extrahovaných termů . . . . .	10
5.4.3	Použití lemmatizace namísto stemmingu . . . . .	10
5.4.4	Odstranění číslovek při preprocessingu . . . . .	10
5.4.5	Použití jiného vzorce pro výpočet vah termů . . . . .	10
<b>6</b>	<b>Diskuze</b>	<b>10</b>
<b>7</b>	<b>Závěr</b>	<b>10</b>

# 1 POPIS PROJEKTU

V tomto projektu implementujeme *LSI vektorový model* sloužící k podobnostnímu vyhledávání v databázi anglických textových dokumentů. Tuto funkcionalitu následně vizualizujeme pomocí webového interface, který uživateli umožňuje procházet databázi článků na základě doporučení nejpodobnějších článků k právě čtenému.

V experimentální části projektu jsme se dále zaměřili na:

- Určení optimálního počtu konceptů
- Porovnání vlivu LSI na kvalitu výsledků vyhledávání s ohledem na výskyt synonym a homonym
- Porovnání průchodu pomocí LSI vektorového modelu se sekvenčním průchodem databáze s ohledem na čas vykonání dotazu
- Vliv různých vnitřních parametrů na výkon algoritmu (změna počtu konceptů, změna počtu extrahovaných termů, použití lemmatizace namísto stemmingu, odstranění číslovky při preprocesingu, použití jiného vzorce pro výpočet vah termů)

Celý náš projekt je volně dostupný k vyzkoušení na: <https://bi-vwm-lsi-demo.herokuapp.com/>.

## 2 ZPŮSOB ŘEŠENÍ

### 2.1 Preprocessing dokumentů

Jako první v naší aplikaci začínáme s preprocesingem dokumentů. Slova z jednotlivých dokumentů převedeme na malá písmena a odstraníme z nich nevýznamová slova a interpunkci. K identifikaci nevýznamových slov používáme seznam anglických nevýznamových slov. Jako parametr programu posíláme také, zda má z dokumentů odstranit i číslovky. Následně na zbylé termy aplikujeme *stemming* či *lemmatizaci*. Tím se snažíme slova, která mají stejný slovní základ, vyjádřit pouze jedním termem. Stemming to dělá pomocí algoritmu, kterým odsekává přípony a koncovky slova. Lemmatizace na to jde o něco chytřeji, podle kontextu slova se pokusí určit, o jaký slovní druh se jedná, a podle toho ho zkrátit.<sup>1</sup> Porovnání jejich použití v programu se dále věnujeme v experimentální části.

### 2.2 Výpočet vah termů

V aplikaci vytváříme matici  $M_w$ , která má v řádcích jednotlivé termy a ve sloupcích jejich váhy v jednotlivých dokumentech.

Začneme tím, že si vytvoříme matici počtu výskytů jednotlivých termů v jednotlivých dokumentech. Počet termů v této matici poté dále zredukujeme, abychom pracovali jen s těmi nejdůležitějšími. Funkci pro redukci termů posíláme následující parametry:

- *max\_df* - termy nacházející se ve více % dokumentů, než udává číslo  $100 * max\_df$ , z matice odstraníme
- *min\_df* - termy nacházející se v méně nebo stejně dokumentech, než udává číslo *min\_df*, z matice odstraníme
- *max\_terms* - maximální počet termů, které si v aplikaci necháme

---

<sup>1</sup>[nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html](http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html)

- *keep\_less\_freq* - udává, zda si při výběru *max\_terms* termů nechat ty nejméně či nejvíce často zastoupené v dokumentech

Z této zredukované matice poté již spočteme matici  $M_w$ . Pro výpočet vah jednotlivých termů používáme metodiku *tf-idf*. Váhu termu  $t_i$  v dokumentu  $d_j$  spočítáme podle vzorce:

$$w_{ij} = tf_{ij} \cdot idf_{ij} \quad (2.1)$$

kde  $tf_{ij}$  reprezentuje normalizovanou četnost termu  $t_i$  v dokumentu  $d_j$  a spočítáme ji podle vzorce<sup>2</sup>:

$$tf_{ij} = \frac{f_{ij}}{nt_j} \quad (2.2)$$

kde  $f_{ij}$  je četnost výskytu termu  $t_i$  v dokumentu  $d_j$ , kterou normalizujeme číslem  $nt_j$  vyjadřujícím celkový počet termů v dokumentu  $d_j$ . V přednášových slidech je použita normalizace jiná,  $tf_{ij}$  se tam počítá podle vzorce:

$$tf_{ij} = \frac{f_{ij}}{\max_i\{f_{ij}\}} \quad (2.3)$$

kde  $\max_i\{f_{ij}\}$  vrací nejvyšší četnost termu  $t_i$  přes celou kolekci dokumentů. Tento způsob normalizace nám vrací spíše horší výsledky, jejich porovnáním se zabýváme v experimentální části.  $idf_{ij}$  reprezentuje převrácenou četnost  $t_i$  ve všech dokumentech a spočítá se podle vzorce:

$$idf_{ij} = \log_2\left(\frac{n}{df_i}\right) \quad (2.4)$$

kde  $n$  je celkový počet dokumentů a  $df_i$  reprezentuje celkový počet dokumentů obsahujících term  $t_i$ .

## 2.3 Implementace LSI

Jakmile máme vytvořenou matici vah termů  $M_w$ , můžeme přistoupit k samotné implementaci LSI. Princip LSI spočívá v tom, že s pomocí *singulárního rozkladu (SVD)* seskupíme tematicky podobné články do jednotlivých  $k$  konceptů. Vlivem počtu konceptů na kvalitu výsledků se dále zabýváme v experimentální sekci.

Singulární rozklad nám matici  $M_w$  rozloží následovně:

$$M_w = U \cdot S \cdot V^T \quad (2.5)$$

kde řádky matice  $U$  jsou obrazy řádků matice  $M_w$ , sloupce matice  $V$  jsou obrazy sloupců matice  $M_w$  a matice  $S$  obsahuje na diagonále *singulární hodnoty (absolutní hodnoty vlastních čísel)* matice  $M_w$  v sestupném pořadí. Z těchto matic získáme *concept-by-document* matici  $M_{cd}$  jako:

<sup>2</sup><https://en.wikipedia.org/wiki/Tf%E2%80%93idf>

$$M_{cd} = S[k, k] \cdot V^T[k, :] \quad (2.6)$$

kde  $S[k, k]$  značí prvních  $k$  řádků a sloupců matice  $S$  a  $V^T[k, :]$  značí prvních  $k$  řádků matice  $V^T$ , kde  $k$  je počet konceptů. Nenásobíme tedy celou maticí  $M_{cd}$ , ale pouze její část podle počtu konceptů.

Matici projekce dotazu do prostoru konceptů  $M_q$  pak získáme jako:

$$M_q = U^T[k, :] \quad (2.7)$$

kde  $U^T[k, :]$  značí prvních  $k$  řádků matice  $U^T$ .

## 2.4 Vyhodnocení dotazu

Při dotazu na nejpodobnější dokumenty k dokumentu  $d_j$  převedeme dotaz do prostoru konceptů na vektor  $V_c$  pomocí vzorce:

$$V_c = M_q \cdot M_{w,j} \quad (2.8)$$

kde  $M_{w,j}$  značí  $j$ -tý sloupec matice  $M_w$ .

Vektor  $V_c$  poté pomocí *kosinové podobnosti* porovnáme se sloupcovými vektory matice  $M_{cd}$ . Indexy nejpodobnějších sloupcových vektorů matice  $M_{cd}$  pak vrátíme jako indexy nejpodobnějších dokumentů k dokumentu dotazu  $d_j$ . Spolu s indexy vrátíme i samotnou hodnotu kosinové podobnosti.

## 3 IMPLEMENTACE

Celý projekt jsme programovali v jazyce *Python*. Práci nám velmi usnadnila jeho knihovna *NLTK*<sup>3</sup> nabízející rozsáhlou funkcionalitu pro práci s přirozeným jazykem. Využili jsme například *WordNetLemmatizer* pro lemmatizaci či *SnowballStemmer* pro stemming. Dále jsme v programu hojně využívali Python knihovny *pandas*<sup>4</sup> a *numpy*<sup>5</sup>.

Ukládání dat v projektu řešíme přes CSV soubory, ke kterým přistupujeme přes *pandas* funkce. V jednom souboru máme uložený dataset nad kterým provádíme LSI. V dalších souborech pak máme uložené matice  $M_w$ ,  $M_{cd}$  a  $M_q$ , abychom je mohli cachovat a přepočítávat jen při změně LSI parametrů, které máme uložené v souboru *server/lisa\_config.json*.

Procházení článků vizualizujeme v prohlížeči pomocí *Flask*<sup>6</sup> web serveru. Jako dataset v našem projektu používáme anglicky psané novinové články stažené z *kaggle.com*<sup>7</sup>. Dataset nepoužíváme celý, vybrali jsme z něj pouze 996 článků.

## 4 PŘÍKLAD VÝSTUPU

Na obrázku 4.1 je vidět konkrétní vstup a výstup naší aplikace. Zobrazí se název článku dotazu, jméno jeho autora a také samotný text článku. Naše aplikace dále uživateli nabídne seznam 10

<sup>3</sup><https://www.nltk.org/>

<sup>4</sup><https://pandas.pydata.org/>

<sup>5</sup><https://numpy.org/>

<sup>6</sup><https://flask.palletsprojects.com/en/1.1.x/>

<sup>7</sup><https://www.kaggle.com/snapcrack/all-the-news>

## SpaceX Launches Rocket, Its First Since Explosion on Launchpad - The New York Times

Author: Kenneth Chang, [source](#)

### Similar articles

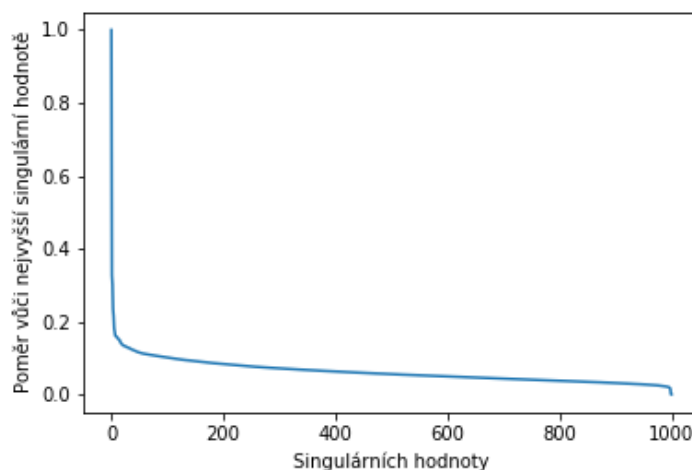
1. [WATCH: SpaceX Successfully Launches Falcon 9 Rocket, Months After Blast](#) (98.4%)
2. [Inside Elon Musk's big plan to colonize Mars](#) (84.8%)
3. [2006: a space oddity – the great Pluto debate](#) (24.4%)
4. [Scientists found bacteria inside rocks — here's what that could mean for life on Mars](#) (23.4%)
5. [Gene Cernan, Last Man To Walk On The Moon, Dies At 82](#) (21.7%)
6. [NJ warehouses decimated by hurricane get trendy makeover](#) (21.0%)
7. ['I'm not good at doing what I'm told': meet real-life Girlboss Sophia Amoruso](#) (19.5%)
8. [1 Address, 2,000 Companies, And The Ease Of Doing Business In The U.K.](#) (18.7%)
9. [Eric Holder's Airbnb Runs Controversial #WeAccept Multiculturalism Ad - Breitbart](#) (17.9%)
10. [Rosetta Crashes Into Comet, Bringing Historic Mission To End](#) (17.8%)

Obrázek 4.1: Příklad výstupu aplikace

nejpodobnějších článků i s určenou kosinovou podobností v procentech. Je vidět, že aplikace vrací víceméně přesně to, co bychom čekali. Jako nejpodobnější vrátila téměř shodný článek taktéž informující, že firma SpaceX vypustila do vesmíru raketu. Další navrhované články se také týkají firmy SpaceX, vesmíru nebo businessu.

## 5 EXPERIMENTÁLNÍ SEKCE

### 5.1 Určení optimálního počtu konceptů



Obrázek 5.1: Důležitost konceptů

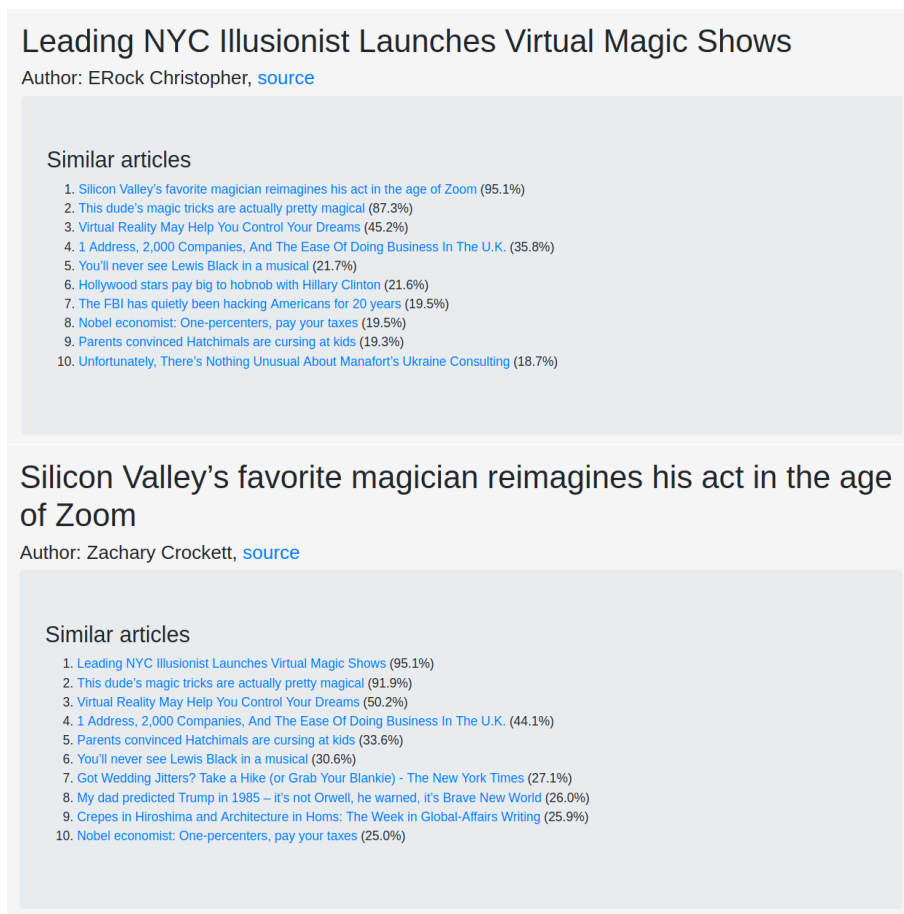
Vezmeme si singulární hodnoty, které nám vrátil singulární rozklad v rovnici 2.5. Vizualizujeme-li si v grafu 5.1, jak klesá poměr nejvyšší singulární hodnoty vůči zbylým singulárním hodnotám, vidíme z toho také, jak klesá důležitost konceptů v datasetu. Počet konceptů  $k$  v naší aplikaci tedy podle tohoto grafu určíme jako 200.

## 5.2 Porovnání vlivu LSI na kvalitu výsledků vyhledávání s ohledem na výskyt synonym a homonym

Pro zjištění kvality výsledků vyhledávání s ohledem na výskyt synonym a homonym jsme do datasetu přidali 4 články.

### 5.2.1 Synonyma

Jak si náš model poradí se synonymy jsme testovali na článcích "*Silicon Valley's favorite magician reimagines his act in the age of Zoom*" a "*Leading NYC Illusionist Launches Virtual Magic Shows*". Testovali jsme tedy anglická synonyma *illusionist* a *magician*.



Obrázek 5.2: Synonyma

Na obrázku 5.2 je vidět, že náš model synonyma v článcích identifikoval správně. V prvním článku se sice ani jednou přímo nevyskytuje slovo *magician* a v druhém ani jednou slovo *illusionist*, ale v obou se vyskytují slova se slovním základem *magic* nebo například slova *show* a *audience* a model tak články správně identifikoval jako podobné.

### 5.2.2 Homonyma

K testování jak si náš článek poradí s homonymy jsme použili články "*Little Richard, rock 'n' roll pioneer, has died at 87*" a "*Scientists found bacteria inside rocks — here's what that could mean for life on Mars*", které oba obsahují slovo *rock* ovšem jednou ve významu skály a jednou ve významu rockového muzikanta.

## Scientists found bacteria inside rocks — here's what that could mean for life on Mars

Author: The Cosmic Companion, [source](#)

### Similar articles

1. 2006: a space oddity – the great Pluto debate (75.0%)
2. Epic Climate Cartoon Goes Viral, But It Has One Key Problem (67.4%)
3. Rosetta Crashes Into Comet, Bringing Historic Mission To End (64.5%)
4. Gene Cernan, Last Man To Walk On The Moon, Dies At 82 (62.2%)
5. Stephen Hawking warns against seeking out aliens in new film (61.5%)
6. Scientists just detected gravitational waves. We've entered a whole new world for astronomy. (60.1%)
7. 10 Months, 45 National Parks, 11 Rules - The New York Times (54.8%)
8. A Call From Outer Space, or a Cosmic Wrong Number? - The New York Times (53.6%)
9. Here's everything scientists know about how to avoid aging (52.6%)
10. Tiny Jumping Spiders Can See the Moon (52.5%)

## Little Richard, rock 'n' roll pioneer, has died at 87

Author: nan, [source](#)

### Similar articles

1. Chuck Berry: from enduring Jim Crow to a comeback album at age 90 (67.2%)
2. Spoon frontman Britt Daniel: 'I wanted to be a musician, not a rockstar' (57.4%)
3. Mourning Stars Through Streams, Screens And Communities (57.3%)
4. The Difficult, Adventurous, Happy Life Of Rosalie Sorrels (56.8%)
5. Ralph Stanley remembered: bluegrass's humble mountain king (56.8%)
6. Guest Dose: DJ JNETT (56.4%)
7. Review: Santigold, '99¢' (55.5%)
8. WSJ Tax Policy Reporter Richard Rubin hosting Q&A in the Hive, Fri. @1:00pm (55.4%)
9. Review: Bob Dylan, 'Triplicate' (51.2%)
10. 40 Years Ago, In A Galaxy Far, Far Away, An Iconic Film Score Was Born (49.6%)

Obrázek 5.3: Homonyma

Na obrázku 5.3 je vidět, že u prvního článku, týkajícího se vlivu objevu bakterie v podmořských skalách na možnost života na Marsu, si model poradil opravdu dobře a vrátil nám články týkající se vesmíru nebo přírody.

U druhého článku týkajícího se smrti rockového hudebníka nám model také velmi správně správně jako nejpodobnější vrátil články týkající se dalších mrtvých hudebníků. Dále vrátil také relevantní články o hudbě a celebritách.

### 5.3 Porovnání průchodu pomocí LSI vektorového modelu se sekvenčním průchodem databáze s ohledem na čas vykonání dotazu

Pro odsimulování sekvenčního průchodu databází nastavíme počet konceptů stejný jako počet dokumentů databáze. Toho docílíme tak, že v souboru `server/lsa_config.json` nastavíme parametr  $k$  na 0.

Na obrázku 5.4 je vidět, že při sekvenčním průchodu databází se čas vykonání dotazu zpomalí asi o 0.05 sekund. Při testování v prohlížeči je však rozdíl v rychlosti načítání stránky zobrazující obsah článku a 10 jemu nejpodobnějších článků mnohem výraznější, kolem 4 sekund.



```

In [21]: df_concept_by_doc, df_query_projection = transform_to_concept_space(df_tf_idf, k=0, customSVD=False)

start = time.time()
best_match = get_n_nearest(df_tf_idf, df_concept_by_doc, df_query_projection, 999, n=10)
end = time.time()
time_seq = end - start
print("Čas vykonání dotazu při sekvenčním průchodu databází: " + str(time_seq) + " sekund")

Čas vykonání dotazu při sekvenčním průchodu databází: 0.19510102272033691 sekund

In [22]: df_concept_by_doc, df_query_projection = transform_to_concept_space(df_tf_idf, k=200, customSVD=False)

start = time.time()
best_match = get_n_nearest(df_tf_idf, df_concept_by_doc, df_query_projection, 999, n=10)
end = time.time()
time_norm = end - start
print("Čas vykonání dotazu při průchodu databází pomocí LSI vektorového modelu: " + str(time_norm) + " sekund")

Čas vykonání dotazu při průchodu databází pomocí LSI vektorového modelu: 0.14609956741333008 sekund

In [23]: print("Rozdíl: " + str(time_seq - time_norm))

Rozdíl: 0.049001455307006836

```

Obrázek 5.4: Porovnání času vykonání dotazu při sekvenčním průchodu databáze

## 5.4 Vliv různých vnitřních parametrů na výkon algoritmu (změna počtu konceptů, změna počtu extrahovaných termů, použití lemmatizace namísto stemmingu, odstranění číslovek při preprocesingu, použití jiného vzorce na výpočet vah termů)

### Little Richard, rock 'n' roll pioneer, has died at 87

Author: nan, [source](#)

#### Similar articles

1. Chuck Berry: from enduring Jim Crow to a comeback album at age 90 (67.3%)
2. Mourning Stars Through Streams, Screens And Communities (50.0%)
3. Spoon frontman Britt Daniel: 'I wanted to be a musician, not a rockstar' (49.6%)
4. Video Appears To Show Tourists Destroying Popular Oregon Rock Formation (49.1%)
5. Guest Dose: DJ JNETT (49.1%)
6. Ralph Stanley remembered: bluegrass's humble mountain king (46.3%)
7. WSJ Tax Policy Reporter Richard Rubin hosting Q&A in the Hive, Fri. @1:00pm (45.6%)
8. Chris Rock, Oscar host who really seems to hate the Oscars, now at center of storm (45.0%)
9. Review: Bob Dylan, 'Triplicate' (44.5%)
10. Review: Santigold, '99¢' (43.8%)

Obrázek 5.5: Špatné určení článku obsahujícího homonymum jako podobného

### 5.4.1 Změna počtu konceptů

Při snížení počtu konceptů na 100 vrací model stále relevantní výsledky a také určování homonym a synonym funguje správně. Na obrázku 5.5 je vidět, že při zvýšení počtu konceptů na 300 už model špatně určí jeden článek obsahující homonymum jako podobný. Tato chyba je se zvyšujícím se počtem konceptů čím dál častější.

#### 5.4.2 Změna počtu extrahovaných termů

V naší aplikaci ve funkci na redukci termů (o které píšeme v části 2.2) odstraňujeme pomocí parametru *min\_df* pouze termy vyskytující se pouze v jednom dokumentu. Parametr *max\_df* necháváme defaultně nastavený na 1, po preprocessingu nemáme v našem datasetu termy vyskytující se v hodně dokumentech. Parametr *max\_terms* také necháváme defaultně nastavený na 0. Uřednostňujeme nechat si v datasetu všechny důležité termy a mít přesnější výsledky před rychlejší dobou běhu.

#### 5.4.3 Použití lemmatizace namísto stemmingu

```
In [72]: start = time.time()
df_words_stemmed = preprocess_docs(df_data, use_lemmatizer=False, remove_numbers=True)
end = time.time()

print("Čas preprocessingu s použitím stemmingu: " + str(end-start) + " sekund")
Čas preprocessingu s použitím stemmingu: 7.0390825271606445 sekund

In [73]: start = time.time()
df_words_lemmatized = preprocess_docs(df_data, use_lemmatizer=True, remove_numbers=True)
end = time.time()

print("Čas preprocessingu s použitím lemmatizace: " + str(end-start) + " sekund")
Čas preprocessingu s použitím lemmatizace: 29.658216953277588 sekund
```

Obrázek 5.6: Čas preprocessingu při použití stemmingu versus při použití lemmatizace

Na obrázku 5.6 je vidět, že při použití lemmatizace namísto stemmingu se čas preprocessingu více než zčtyřnásobí na téměř 30 sekund. Uděláme-li si graf důležitosti konceptů jako na obrázku 5.1 vyjde při použití stemmingu i lemmatizace úplně stejný. V naší aplikaci jsme se rozhodli používat lemmatizaci a upřednostnit tak přesnější převádění slov z dokumentů na jednotlivé termy na úkor delšího času preprocessingu.

#### 5.4.4 Odstranění číslovek při preprocessingu

Číslovky z dokumentů jsme se v naší aplikaci rozhodli odstraňovat, jelikož nám pak aplikace vrací o něco lepší výstupy. Číslovky v našem datasetu novinových článků nejsou většinou pro podobnost moc relevantní spíše naopak.

#### 5.4.5 Použití jiného vzorce pro výpočet vah termů

TODO

## 6 DISKUZE

## 7 ZÁVĚR