



**FAKULTA  
INFORMAČNÍCH  
TECHNOLOGIÍ  
ČVUT V PRAZE**

## **Projektová dokumentácia**

LSI vektorový model

ID projektu: I-4

Vyhľadávanie na webe a v multimediálnych databázach

BI-VWM

LS 2020/2021

Adam Makara, Matej Šutý

## Obsah:

<b>1</b>	<b>Popis projektu .....</b>	<b>3</b>
1.1	Cieľ .....	3
1.2	LSI vektorový model .....	3
1.3	Vstupy .....	3
1.4	Výstupy .....	3
<b>2</b>	<b>Spôsob riešenia .....</b>	<b>4</b>
2.1	Prístupy .....	4
2.2	Algoritmy .....	4
<b>3</b>	<b>Implementácia .....</b>	<b>4</b>
3.1	Stavba aplikácie .....	4
3.2	Použité knižnice tretích strán .....	5
3.3	Požiadavky na beh .....	6
<b>4</b>	<b>Príklad vstupu a výstupu .....</b>	<b>6</b>
<b>5</b>	<b>Experimentálna sekcia .....</b>	<b>8</b>
5.1	Experiment 1 .....	8
5.2	Experimenty 2 a 3 .....	11
5.3	Experiment 4 .....	15
5.4	Experiment 5 .....	16
<b>6</b>	<b>Diskusia .....</b>	<b>16</b>
	<b>Záver.....</b>	<b>16</b>

# **1 Popis projektu**

## **1.1 Cieľ**

Cieľom toho projektu je implementácia LSI (Latent Semantic Indexing) vektorového modelu ukladania dát (t.j. preprocessing a indexovanie) spolu s možnosťou dotazovania sa z GUI.

## **1.2 LSI vektorový model**

Vektorový model je vo všeobecnosti jeden zo spôsobov ako prehľadávať kolekciu dokumentov riešiaci nedostatky booleovského modelu. Klasický vektorový model neumožňuje zachytiť podobnosť medzi určitými slovami, teda nevie pracovať so synonymami, prípadne homonymami. Tieto nedostatky rieši práve tzv. latentné sémantické indexovanie (LSI). LSI vektorový model sa snaží zachytiť latentnú sémantiku skrytú vo vnútri dát. LSI nepracuje na štandardným priestorom termov ako klasický vektorový model, ale prevádza tzv. term-by-document maticu do priestoru konceptov. Koncepty sú tvorené lineárnymi kombináciami termov, ktoré práve zachycujú vzťahy medzi termami a dokumentami. Celý proces je založený na SVD (Singular Value Decomposition) dekompozícii na vlastné čísla matice.

## **1.3 Vstupy**

Vstupom je sada anglických článkov získaných pomocou crawleru, v ktorom sa zadala webová adresa novinového portálu a následne nám stiahol viacero článkov do textových súborov v ktorých oddelil názvy článkov od obsahu. Takéto súbory následne prejdú tzv. preprocessingom – extrakcia termov, identifikácia a odstránenie nevýznamových slov. Následne sa vypočítajú váhy termov a vykoná sa LSI.

## **1.4 Výstupy**

Výstupom je webová aplikácia bežiaca na lokálnom prostredí, ktorá umožňuje prezeranie stiahnutých novinových článkov. Každý článok je možné zobraziť si samostatne a pod jeho obsahom sa zobrazuje tabuľka so zoznamom podobných článkov, kde sa pomocou predpočítaných údajov a metódy LSI vypočíta a vypíše niekoľko najpodobnejších článkov zoradených zostupne podľa relevancie k aktuálne prezeranému článku.

## 2 Spôsob riešenia

### 2.1 Prístupy

Na začiatku projektu sme si pomocou crawleru stiahli zopár anglických článkov, aby sme mohli na nich postaviť základnú kostru aplikácie. Potom sme začali postupne pridávať a implementovať funkcionality tak, aby simulovala sekvenčné spracovanie vstupných článkov na výstupný súbor s predpočítanými hodnotami.

Vytvorili sme prvý package s názvom preprocessing, v ktorom sme pridali súbor token.py s metódami na tokenizovanie, odstránenie stop words a stematizovanie vstupných článkov. Následne sme pridali druhý package s názvom vector, kde sme vytvorili súbor vector.py, ktorý obsahuje metódy pre vytvorenie docterm vectorov, celkové vytvorenie term-by-document matice a výpočet váh pomocou algoritmu TF-IDF. Ďalej sme pridali package LSI, ktorý má na starosti výpočet vlastných čísel term-by-document matice a následnú SVD dekompozíciu s aproximáciou, pomocou ktorej získame maticu konceptov. Na základe toho sme ďalej pridali metódy na výpočet podobností dokumentov k aktuálne zadanému dokumentu. Keď nám latentné sémantické indexovanie ako tak fungovalo, vytvorili sme jednoduchú webovú aplikáciu s možnosťou prezerania článkov. Nakoniec sme pridali možnosť zobrazit' si daný článok, kde sa pri načítavaní stránky vypočíta 10 najpodobnejších článkov k aktuálne zobrazenému a tie sa vypíšu do tabuľky, viz. sekcia 4.

### 2.2 Algoritmy

Implementácia využíva 2 veľmi dôležité a zaujímavé algoritmy. Pri spracovávaní vstupných článkov sme použili algoritmus TF-IDF pre výpočet váh termov v jednotlivých dokumentoch. Pre získanie concept-by-document matice sme využili SVD dekompozíciu term-by-document matice.

## 3 Implementácia

Celý projekt je vytvorený v programovacom jazyku Python. Webová aplikácia je vytvorená pomocou high-level webového frameworku Django, ktorý je tiež napísaný v Pythone.

### 3.1 Stavba aplikácie

Aplikácia pozostáva z nasledujúcich komponent, ktoré medzi sebou komunikujú, jedná sa o packages:

- **Isi-web** – obsahuje celú webovú Django aplikáciu
  - Obsahuje package nazvaný **Isi**, ktorý má na starosti funkcionality na webe – prezeranie novinových článkov, zobrazenie konkrétneho článku, výpočet podobností ostatných článkov na základe zobrazeného článku a predspracovaných uložených dát. Taktiež obsahuje štýly a šablóny pre zobrazovanie webstránok
  - Ak sa náhodou nezobrazujú články alebo aplikácia hádže chybu, že nie je možné nájsť súbor, tak možno v **Isi-data/run.py** upraviť cestu k článkom a v **Isi-web/app/settings.py** úplne dole upraviť cesty.
  - Ostatné súbory sú defaultne vygenerované frameworkom Django
- **Isi-data** – obsahuje celé spracovanie uložených článkov
  - **articles** – pripravili sme 4 datasety článkov – 50,300,500 a všetky
  - **experiments** – obsahuje metódy pre testovanie experimentov a tvorbu grafov
  - **LSI** – obsahuje metódy pre výpočet SVD
  - **preprocessing** – obsahuje metódy pre tokenizovanie, odstránenie stop words a stematizovanie
  - **vector** – obsahuje metódy pre vytvorenie term-by-document matice a výpočet TF-IDF

Ďalším dôležitým súborom je súbor **run.py**, ktorého spustením sa spracujú všetky články v priečinku **articles**, vytvorí sa term-by-document matica, vykoná sa SVD dekompozícia a výsledné spracované údaje sa uložia do súboru **file.dat**.

### 3.2 Použité knižnice tretích strán

Pri implementácii sme použili nasledujúce knižnice tretích strán:

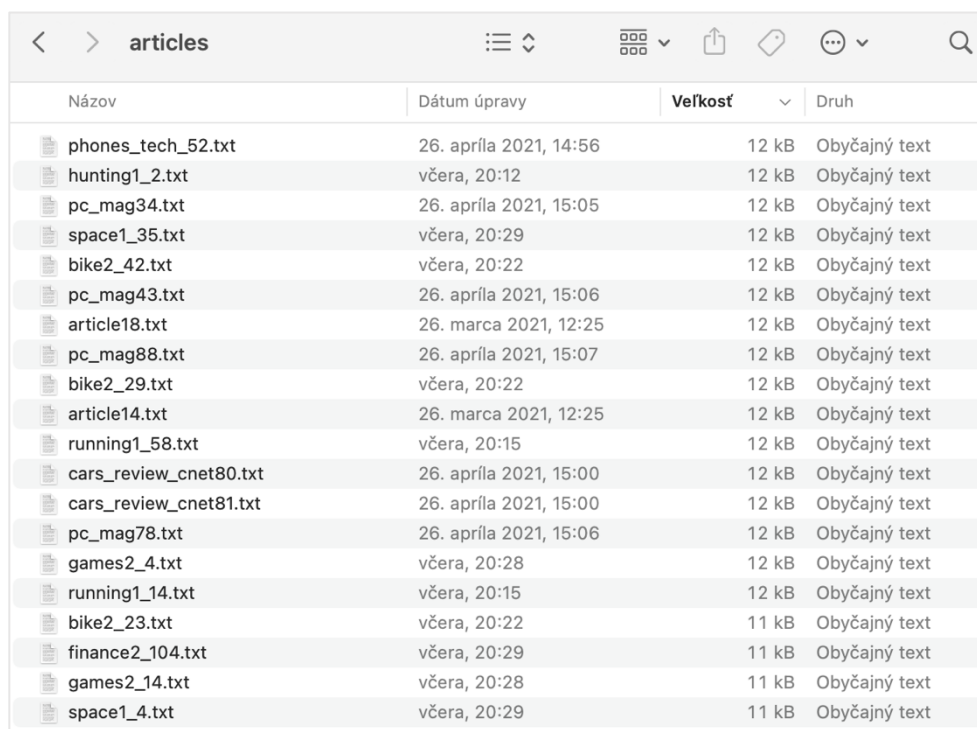
- **hashlib** – generovanie hashu pre zistenie rovnakých dokumentov
- **nltk** – použité pre tokenizovanie, stematizovanie, odstránenie stop words
- **pickle** – ukladanie a načítanie predpočítaných hodnôt z/do súboru
- **numpy** a **linalg** – práca s maticami, výpočet vlastných čísel term-by-document matice, normalizácia hodnôt, svd
- **vector** – vytvorenie docterm vektoru
- **matplotlib** – vykresľovanie grafov použitých experimentálnej sekcii

### 3.3 Požiadavky na beh

Aby aplikácia fungovala správne, vyžaduje mať nainštalované všetky vyššie spomenuté knižnice + framework Django. V priečinku **Isi-data/articles** musia byť články – textové súbory. Odporúčame pomenovať ako kategóriu o ktorej sa v článku píše, aby sa na webe správne zobrazovali kategórie. Potom stačí spustiť súbor `Isi-data/run.py`, ktorý spracuje články a v termináli sa prepnúť do package **Isi-web** a v ňom spustiť tento príkaz: **python manage.py runserver** – týmto príkazom sa spustí lokálny web server a možno prejsť na adresu `localhost:8000`.

## 4 Príklad vstupu a výstupu

Vstup je sada článkov, ktoré sa nachádzajú v priečinku `articles`. Po spustení hlavného programu `run.py` aplikácia prejde všetky články, tokenizuje, odstráni stop slová, stemmatizuje, vytvorí term-by-document maticu z ktorej vypočíta vlastné čísla a spraví SVD dekompozíciu. Všetky výsledné dáta uloží do súboru `file.dat`. Tým výpočet a spracovanie vstupu, ktorým sú články, končí.



Názov	Dátum úpravy	Veľkosť	Druh
phones_tech_52.txt	26. apríla 2021, 14:56	12 kB	Obyčajný text
hunting1_2.txt	včera, 20:12	12 kB	Obyčajný text
pc_mag34.txt	26. apríla 2021, 15:05	12 kB	Obyčajný text
space1_35.txt	včera, 20:29	12 kB	Obyčajný text
bike2_42.txt	včera, 20:22	12 kB	Obyčajný text
pc_mag43.txt	26. apríla 2021, 15:06	12 kB	Obyčajný text
article18.txt	26. marca 2021, 12:25	12 kB	Obyčajný text
pc_mag88.txt	26. apríla 2021, 15:07	12 kB	Obyčajný text
bike2_29.txt	včera, 20:22	12 kB	Obyčajný text
article14.txt	26. marca 2021, 12:25	12 kB	Obyčajný text
running1_58.txt	včera, 20:15	12 kB	Obyčajný text
cars_review_cnet80.txt	26. apríla 2021, 15:00	12 kB	Obyčajný text
cars_review_cnet81.txt	26. apríla 2021, 15:00	12 kB	Obyčajný text
pc_mag78.txt	26. apríla 2021, 15:06	12 kB	Obyčajný text
games2_4.txt	včera, 20:28	12 kB	Obyčajný text
running1_14.txt	včera, 20:15	12 kB	Obyčajný text
bike2_23.txt	včera, 20:22	11 kB	Obyčajný text
finance2_104.txt	včera, 20:29	11 kB	Obyčajný text
games2_14.txt	včera, 20:28	11 kB	Obyčajný text
space1_4.txt	včera, 20:29	11 kB	Obyčajný text

Obrázok 1: Články v priečinku `articles` pripravené na spracovanie.

Výstupom je webová aplikácia, kde na úvodnej stránke je prehľad všetkých článkov a každý z nich si možno zobrazit' samostatne. Pri detailnom zobrazení daného

článku sa otvorí súbor file.dat s predpočítanými hodnotami a vypočítajú sa podobnosti články môžu byť veľmi dlhé, defaultne sme ich skryli a možno ich rozbaľiť tlačítkom.

LSI Vector Model App Home

**Does care for creation ignore the poor?**

Climate change kills millions from drought, flooding, famine, and disease and exacerbates...

[View details »](#)

**Best speakers of 2021**

Thinking about taking your home theater to the next level? Whether you're in the market for a...

[View details »](#)

**The Latest Mushroom Cultivation Technique is in "Yesterday's News"**

Here's a really quick, easy, and simple hack for growing mushrooms at home with just a few easily...

[View details »](#)

**Robin Lane Fox: all the joys of spring gardening**

You can enable subtitles (captions) in the video player Welcome to the gardens of New College...

[View details »](#)

**Nissan Micra IG-T 90 review - Lacklustre engine kills Fiesta rival's chances**

Nissan Micras haven't been a regular feature in evo-land. While perfectly suited for inner-city...

[View details »](#)

**Nissan 350Z (2003-2009): review, specs and buying guide**

It's 15 years since Nissan Launched the 350Z and writing these words makes me feel rather...

[View details »](#)

**Tears of a Dirtbag: Rapper Lil Peep Is the Future of Emo**

Pitchfork: When did you first become interested in rapping? Lil Peep: When I found out about the...

[View details »](#)

**The greatest ever Ford RS cars**

Ford is a brand synonymous with many things across the motoring spectrum, from revolutionising...

[View details »](#)

**Who decides if someone becomes a saint?**

Isabel Flores de Oliva (1586-1617) and Juan Martín de Porres Velázquez (1579-1639) were...

[View details »](#)

« < Page 1 of 41 > »

© Semestral project for BI-VWM subject, Adam Makara and Matej Šutý 2021

Obrázok 2: Úvodná stránka webovej aplikácie s prehľadom všetkých článkov.

LSI Vector Model App Home

## Nissan Micra IG-T 90 review - Lacklustre engine kills Fiesta rival's chances

[Toggle article](#)

#	Article	Category	Similarity
1	<a href="#">Nissan Micra N-Sport review - does a new engine turn Nissan's supermini...</a>	evo cars	0.7090432130976864
2	<a href="#">SEAT Ibiza updated for 2021 - debuts all-new interior</a>	evo cars	0.6953083401732246
3	<a href="#">Alpina XD3 SUV - UK specs announced for the left-field performance...</a>	evo cars	0.6449600350825673
4	<a href="#">All-new Toyota Yaris revealed - and hot Gazoo models aren't far behind</a>	evo cars	0.6270934482576684
5	<a href="#">Citroen C5 Aircross review - exactly like a Porsche 911 GT3, sort of</a>	evo cars	0.6250139968930124
6	<a href="#">Maserati Levante Trofeo review - is the most powerful series-production...</a>	evo cars	0.6148931056674175
7	<a href="#">2022 Toyota GR 86 revealed - rear-wheel drive coupe reimaged</a>	evo cars	0.5995592413255096
8	<a href="#">New Maserati Ghibli S review</a>	evo cars	0.5971702897963893
9	<a href="#">Maserati Ghibli Trofeo revealed - sober-dressed BMW M5 rival finally...</a>	evo cars	0.5675957906292345
10	<a href="#">Maserati Ghibli Diesel review - improved Italian still lags E-class</a>	evo cars	0.5426236772473421

© Semestral project for BI-VWM subject, Adam Makara and Matej Šutý 2021

Obrázok 3: Detail článku so zoznamom najpodobnejších článkov.

## 5 Experimentálna sekcia

Cieľom experimentov bolo overiť, ako vplýva aproximácia  $k$  na výber najrelevantnejších článkov.

Model LSI ponúka možnosť zvoliť aproximáciu  $k$ . Proces rozkladu na vlastné čísla SVD umožňuje redukciu matice  $A$  s veľkou dimenziou do priestoru konceptov, ktorý má oveľa menšiu dimenziu. Výberom prvých  $k$  konceptov, ktoré majú najväčšie hodnoty vlastných čísel dosiahneme, že články, ktoré sú si podobné, majú malú kosínovú vzdialenosť.

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

### 5.1 Experiment 1

Prvý experiment pozostáva z pozorovania ako reaguje LSI na zmenu aproximácie  $k$ . Sledovali sme, ktoré články sú si najviac podobné pri rozdielnych  $k$ , a aké majú hodnoty podobnosti.

#### *Nastavenie:*

LSI spracovalo 50 článkov z rozdielnych oblastí záujmu (počítače, cyklistika, financie, hudba, počítačové hry, atď.). Následne sme vybrali jeden článok s názvom *10 Ways to Boost Your Wi-Fi Signal* časopisu o počítačoch. Pre hodnoty  $k$  z  $\{1, 10, 30, 35, 45, 46, 48, 50\}$ , sme zobrazili články, ktoré LSI považuje za najpodobnejšie k článku o wifi.

#### *Pozorovanie:*

- $k=1$   
10 najpodobnejších článkov má 100% zhodu. Medzi nimi sú články z kategórií, ktoré nesúvisia s naším pozorovaným článkom, napríklad články o zdraví, financiách, cyklistike. Iba 4 články z 10 sú z podobných kategórií.



## 10 Ways to Boost Your Wi-Fi Signal

[Toggle article](#)

#	Article	Category	Similarity
1	<a href="#">Guide to Disc Brakes</a>	bike	1.0
2	<a href="#">How this fruit became the star of Italian cooking</a>	article	1.0
3	<a href="#">Building an E-Commerce Website: 8 Technical Aspects You Need to Know</a>	pc mag	1.0
4	<a href="#">Cleaning Up the E-Waste Mess: Big Tech Needs to Do More</a>	pc mag	1.0
5	<a href="#">India is trying to build its own internet</a>	article	1.0
6	<a href="#">Antony Blinken on the global challenges facing America</a>	finance	1.0
7	<a href="#">The best at-home COVID-19 tests</a>	phones tech	1.0
8	<a href="#">The Best E-Commerce Fulfillment Services</a>	pc mag	1.0
9	<a href="#">Women of Color Die of Cancer at Higher Rates Than White Women-Here's...</a>	health	1.0
10	<a href="#">The Top 100 Tracks of 2006</a>	music	1.0

© Semestral project for BI-VWM subject, Adam Makara and Matej Šutý 2021

- $k=10$

Zhody 10 najpodobnejších článkov klesli, z prvých piatich článkov sú 4 články s podobnej kategórie (2x počítače, mobily, počítačové hry). Celkovo 7/10 článkov majú podobné kategórie. Všetky články majú veľmi vysoké zhody (viac ako 97%), napriek tomu, že ich obsah nie je natoľko podobný.

## 10 Ways to Boost Your Wi-Fi Signal

[Toggle article](#)

#	Article	Category	Similarity
1	<a href="#">Building an E-Commerce Website: 8 Technical Aspects You Need to Know</a>	pc mag	0.9886560738872333
2	<a href="#">Tips For Building a Custom Bike</a>	bike	0.9885923530519779
3	<a href="#">Now is the best time to buy a new iPhone, Galaxy S9, Note 9, Pixel 3</a>	phones tech	0.9866978265279317
4	<a href="#">21 Free Tools Your Small Business Should Be Using Today</a>	pc mag	0.9849690703754793
5	<a href="#">What can we learn from the Cyberpunk 2077 launch disaster?</a>	games	0.9831313068135854
6	<a href="#">The best at-home COVID-19 tests</a>	phones tech	0.9827967495086941
7	<a href="#">The Best E-Commerce Fulfillment Services</a>	pc mag	0.9827137795731591
8	<a href="#">Tips for Descending Hills on a Bike</a>	bike	0.980863609500321
9	<a href="#">First Impressions: Crankbrothers M20 &amp; M13 Multi Tools</a>	bike	0.9785340763384204
10	<a href="#">The Best Bluetooth and Wireless Speakers for 2021</a>	pc mag	0.9784620363268585

© Semestral project for BI-VWM subject, Adam Makara and Matej Šutý 2021

- $k=30$

Podobnosti článkov výrazne klesli (35% až 74%). 6/10 článkov majú podobné kategórie záujmu. Medzi 1. najpodobnejším a 10. článkom je veľký rozdiel.

Zaujímavé je, že 1. článok sa tematicky líši od pozorovaného článku (počítače ako niečo zlepšiť).

LSI Vector Model App Home

## 10 Ways to Boost Your Wi-Fi Signal

Toggle article

#	Article	Category	Similarity
1	<a href="#">Tips For Building a Custom Bike</a>	bike	0.7438351356578198
2	<a href="#">Now is the best time to buy a new iPhone, Galaxy S9, Note 9, Pixel 3</a>	phones tech	0.7081867950336358
3	<a href="#">The Best Bluetooth and Wireless Speakers for 2021</a>	pc mag	0.6679620331040282
4	<a href="#">Hidden Tricks Inside Windows 10</a>	pc mag	0.6458407071760343
5	<a href="#">21 Free Tools Your Small Business Should Be Using Today</a>	pc mag	0.6065401537226189
6	<a href="#">Windows 10 tips: How to take screenshots, find the secret Start menu and...</a>	phones tech	0.5216497324328794
7	<a href="#">Music is fundamentally joy, says this professor</a>	religion	0.49930099382105564
8	<a href="#">Fox Makes Big Changes to Its Forks and Shocks, Including a Whole New Model</a>	bike	0.49315169126235897
9	<a href="#">Is the music at Mass that important? The survey says yes.</a>	religion	0.4455875458195087
10	<a href="#">Building an E-Commerce Website: 8 Technical Aspects You Need to Know</a>	pc mag	0.3597631572892752

© Semestral project for BI-VWM subject, Adam Makara and Matej Šutý 2021

- $k=35$

Pri tejto hodnote už je veľmi veľký rozdiel medzi podobnosťami najpodobnejších článkov (5% až 59 %). 7/10 článkov majú podobné kategórie.

LSI Vector Model App Home

## 10 Ways to Boost Your Wi-Fi Signal

Toggle article

#	Article	Category	Similarity
1	<a href="#">The Best Bluetooth and Wireless Speakers for 2021</a>	pc mag	0.5910677775060351
2	<a href="#">21 Free Tools Your Small Business Should Be Using Today</a>	pc mag	0.48670734058975657
3	<a href="#">Hidden Tricks Inside Windows 10</a>	pc mag	0.46060790677176616
4	<a href="#">Windows 10 tips: How to take screenshots, find the secret Start menu and...</a>	phones tech	0.3957488573155937
5	<a href="#">Building an E-Commerce Website: 8 Technical Aspects You Need to Know</a>	pc mag	0.37958283875741544
6	<a href="#">The Best E-Commerce Fulfillment Services</a>	pc mag	0.27302017756247837
7	<a href="#">First Impressions: Crankbrothers M20 &amp; M13 Multi Tools</a>	bike	0.15104709820561635
8	<a href="#">Bike Fit: Here's What You Need to Know to Make Riding More Comfortable</a>	bike	0.12349050292025376
9	<a href="#">The best at-home COVID-19 tests</a>	phones tech	0.06504660007667562
10	<a href="#">Is the music at Mass that important? The survey says yes.</a>	religion	0.05919299021081982

© Semestral project for BI-VWM subject, Adam Makara and Matej Šutý 2021

- $K=45$

Pri tejto hodnote sú výsledky viac menej nepoužiteľné. Väčšina článkov sa zaoberá inými témami a podobnosti, až na najpodobnejší článok s rovnakou kategóriou, sú menšie ako 8 percent. Podobný trend je aj pri ďalších vyšších hodnotách  $k$ .

LSI Vector Model App [Home](#)

## 10 Ways to Boost Your Wi-Fi Signal

[Toggle article](#)

#	Article	Category	Similarity
1	<a href="#">The Best Bluetooth and Wireless Speakers for 2021</a>	pc mag	0.4176669564458862
2	<a href="#">Windows 10 tips: How to take screenshots, find the secret Start menu and...</a>	phones tech	0.07832432994302338
3	<a href="#">Bike Fit: Here's What You Need to Know to Make Riding More Comfortable</a>	bike	0.06024010950868161
4	<a href="#">Building an E-Commerce Website: 8 Technical Aspects You Need to Know</a>	pc mag	0.022726403702606587
5	<a href="#">The Best E-Commerce Fulfillment Services</a>	pc mag	0.011275739255702858
6	<a href="#">Tornadoes Fast Facts</a>	article	0.00020543031436471703
7	<a href="#">China's crackdown in Xinjiang has separated Uyghur children from their...</a>	article	0.0001892774439930862
8	<a href="#">How this fruit became the star of Italian cooking</a>	article	0.00003682076798082852
9	<a href="#">Writers Guild of America Awards 2021: 'Borat' and 'Promising Young...</a>	article	0.0000061153489957278415
10	<a href="#">How Apple's new iMacs color-matched today's home fashion trends</a>	phones tech	-0.000013381165731035086

© Semestral project for BI-VWM subject, Adam Makara and Matej Šutý 2021

## 5.2 Experimenty 2 a 3

### Nastavenie:

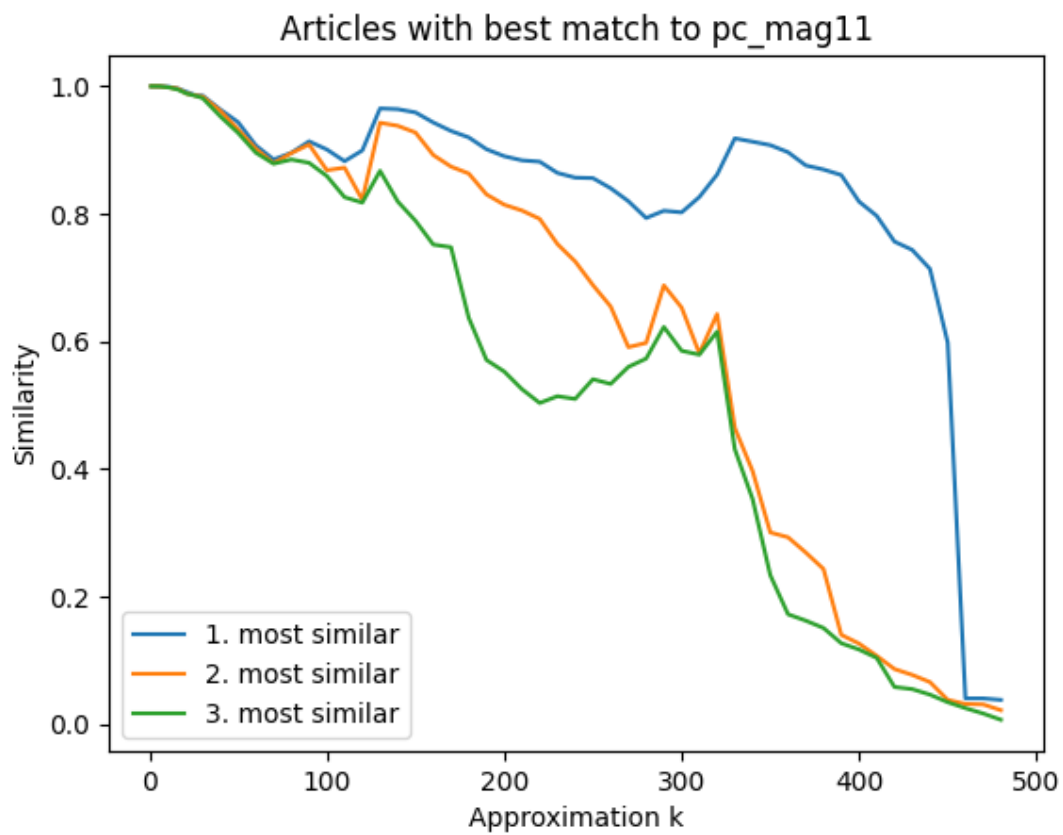
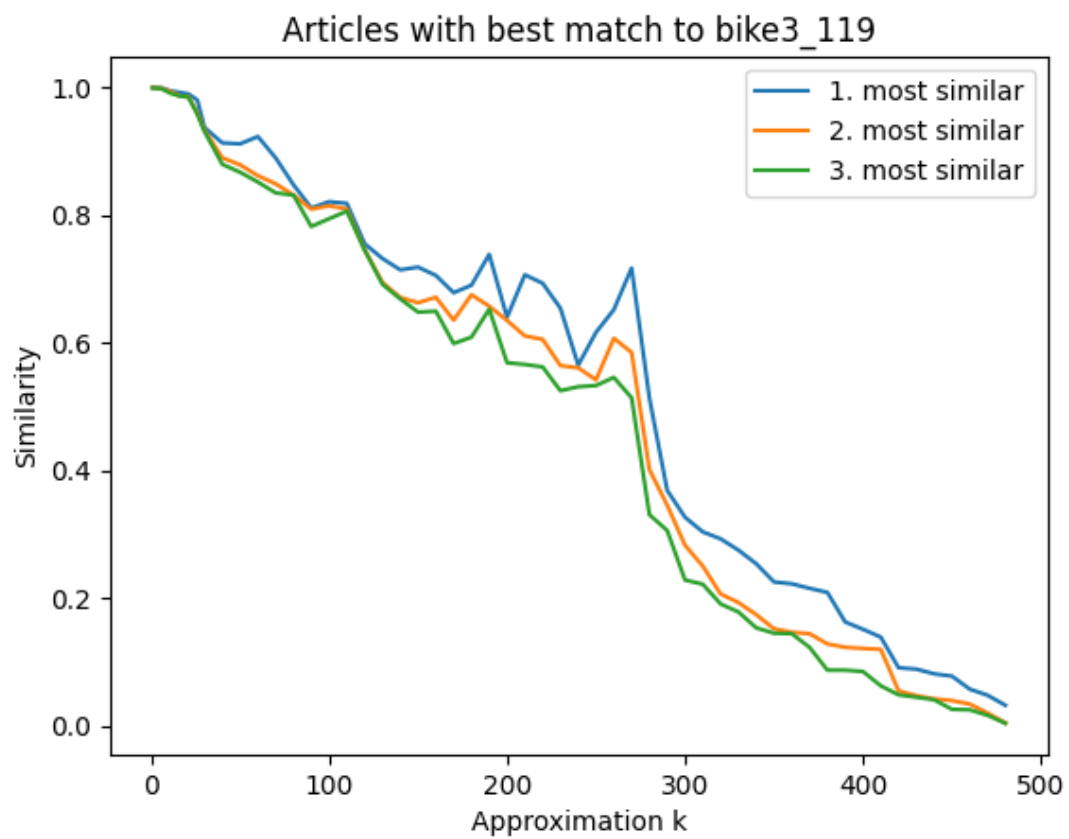
LSI spracovalo 500 článkov s rôznymi oblasťami záujmov. Z nich boli vybrané práve 3: **pc\_mag42** (počítače), **bike3\_119** (cyklistika), **pc\_mag11** (počítače).

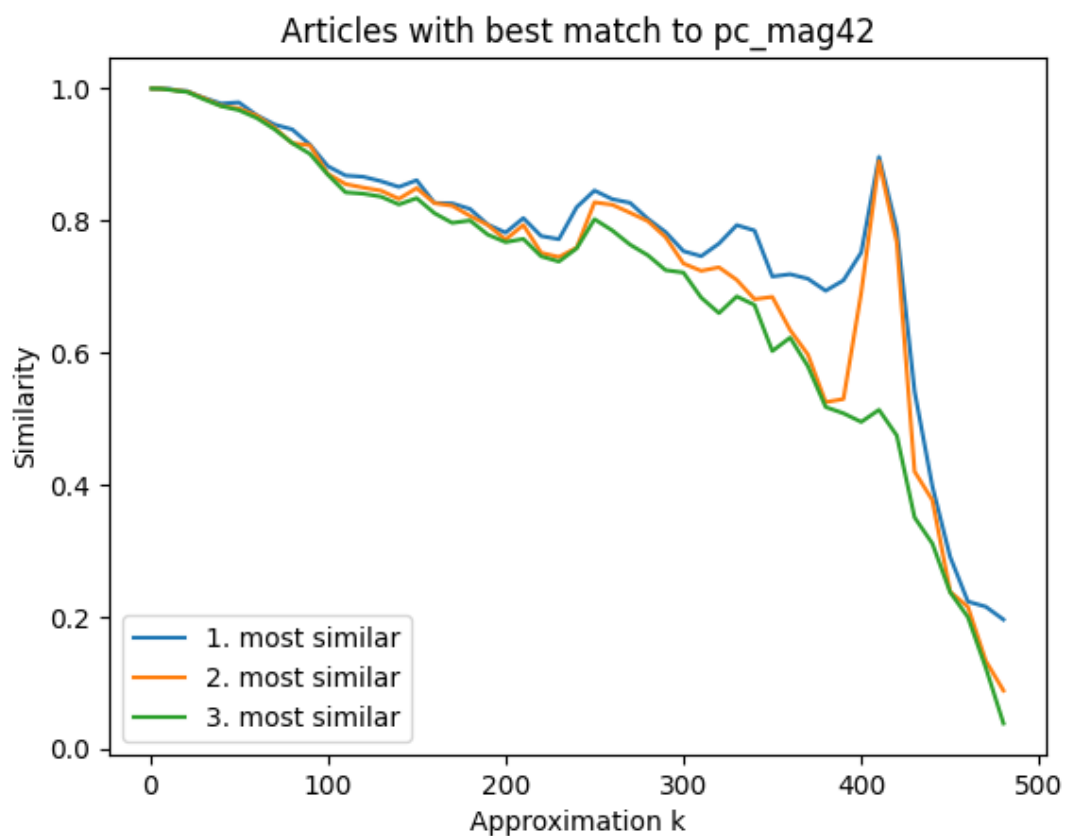
A) Pre rôzne aproximácie  $k$  sme počítali hodnoty podobnosti prvých troch najpodobnejších článkov.

B) Spočítali sme 5 článkov, ktoré sa najčastejšie vyskytujú na pozíciách 1, 2, 3.

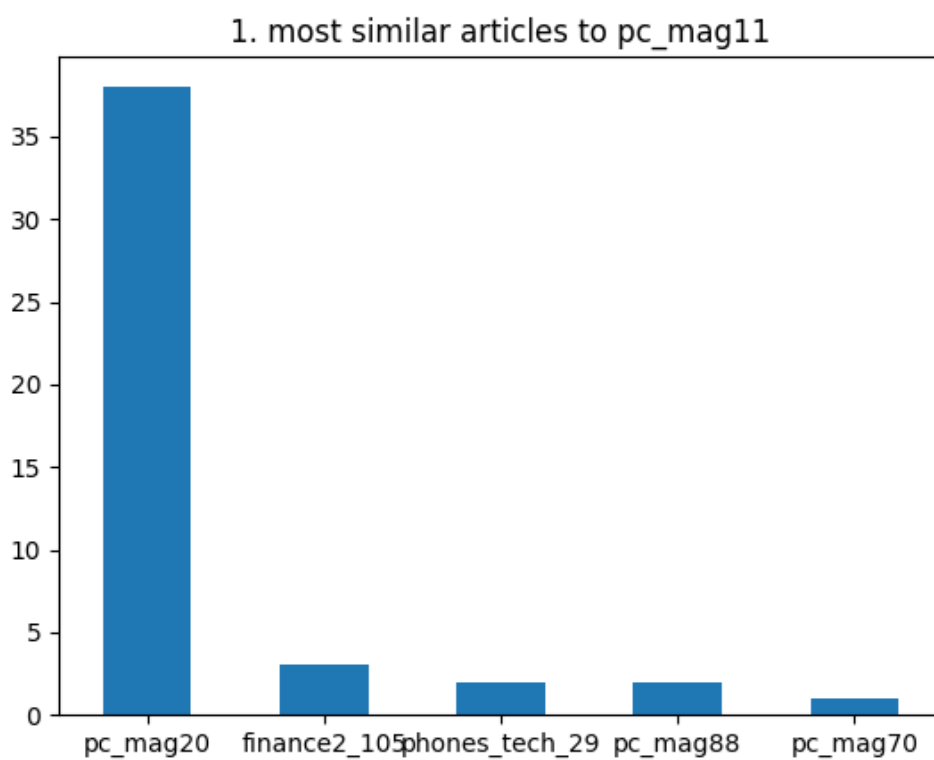
### Pozorovanie:

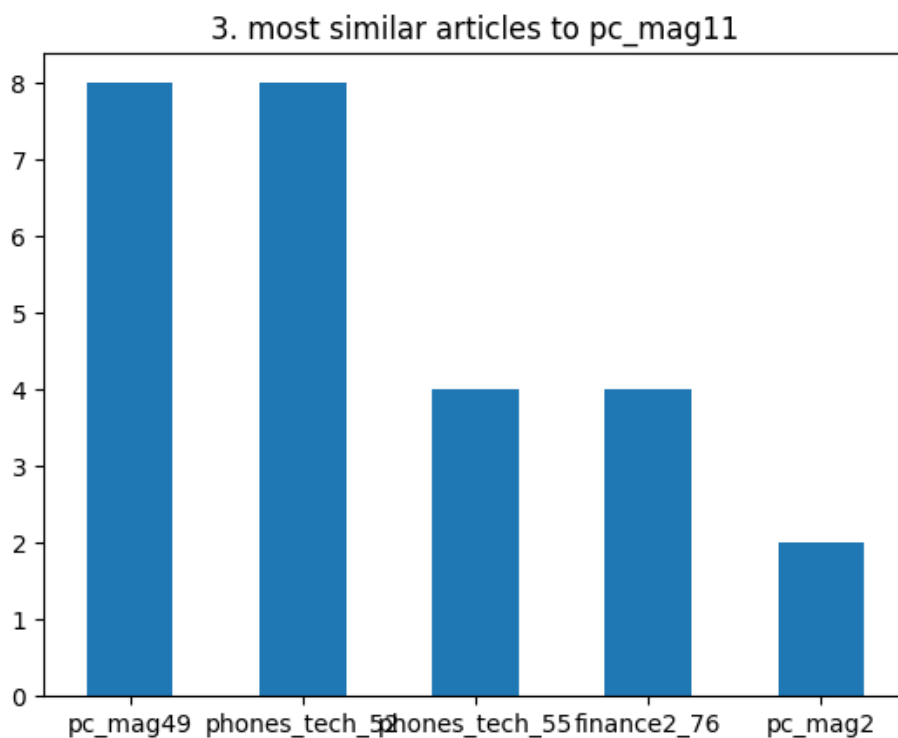
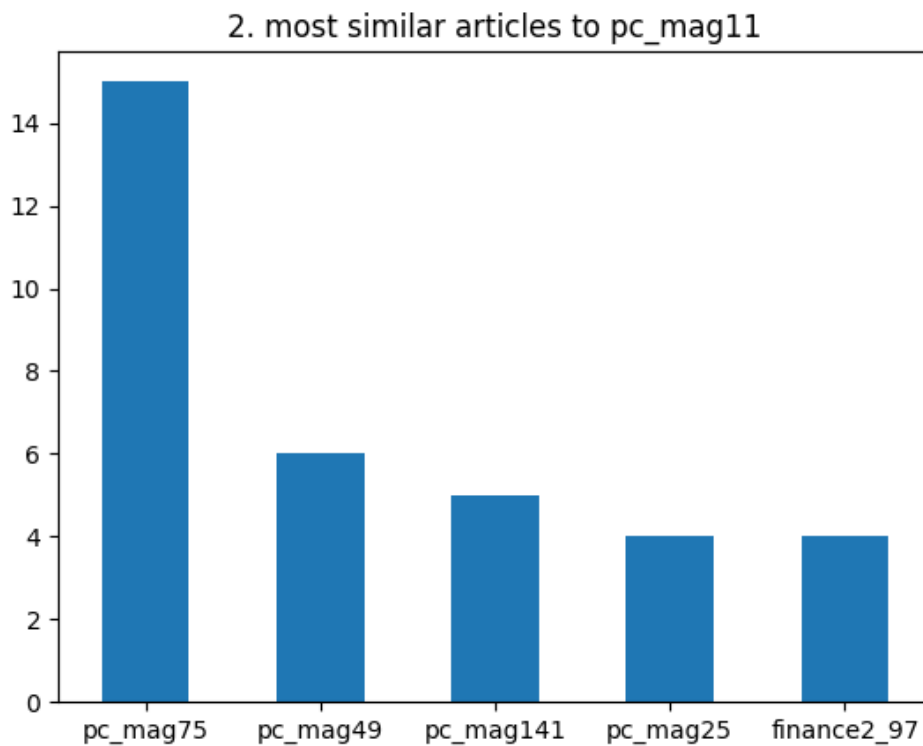
A) V grafoch nižšie môžeme vidieť, že s rastúcim  $k$ , klesá podobnosť prvých troch článkov.





B) LSI zvolilo najčastejšie články z rovnakej alebo podobnej kategórie.



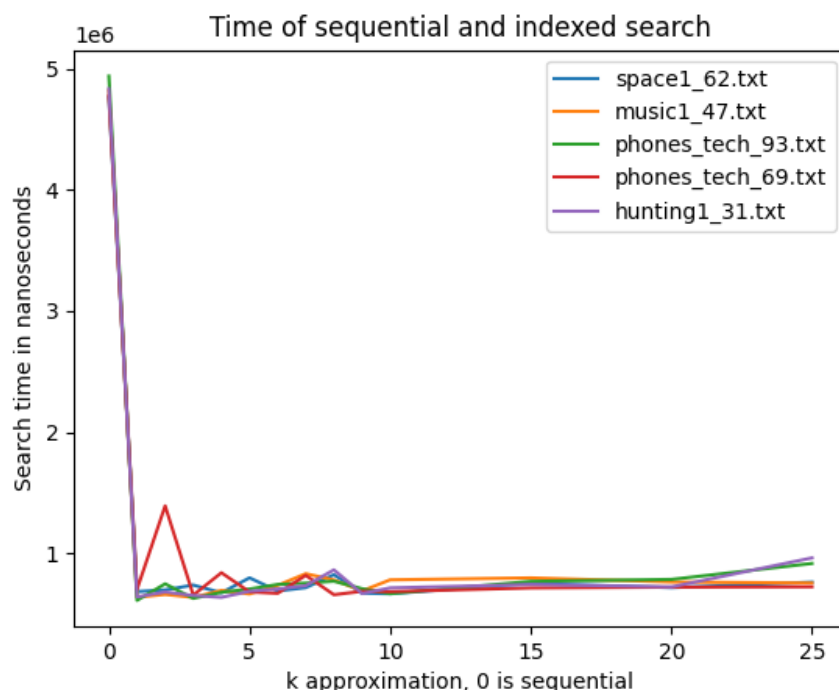


### Zhodnotenie:

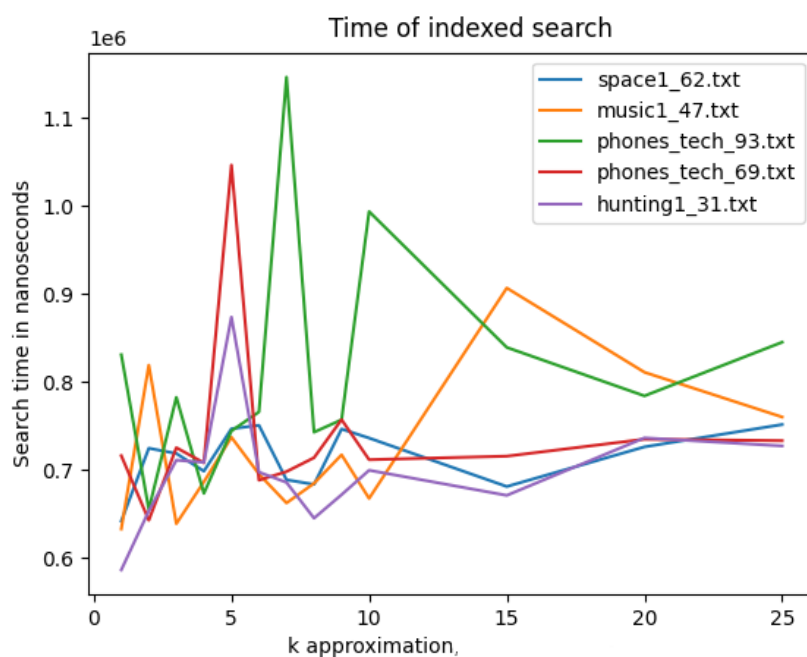
Keď zvolíme nízke  $k$ , články sa na seba budú veľmi podobieť, pretože zanedbávame vyššie dimenzie a teda priveľmi redukuje priestor v ktorom sa články nachádzajú. Pre priveľmi vysoké  $k$  je rozlíšenie LSI príliš vysoké a medzi článkami jednotlivými článkami v priestore konceptov sú také veľké vzdialenosti, že výsledky nie sú relevantné.

### 5.3 Experiment 4

V tomto experimente sme porovnali sekvenčný priechod oproti indexu vzhľadom na časovú rýchlosť výpočtu podobnosti na 50 článkoch. Toto porovnanie zachytáva graf nižšie, kde sme pre 5 článkov postupne zvyšovali aproximáciu  $k$ .  $k = 0$  znamená, že nebola použitá žiadna aproximácia a teda výpočet prebiehal sekvenčne. Je vidieť že index znížil čas výpočtu rádovo.

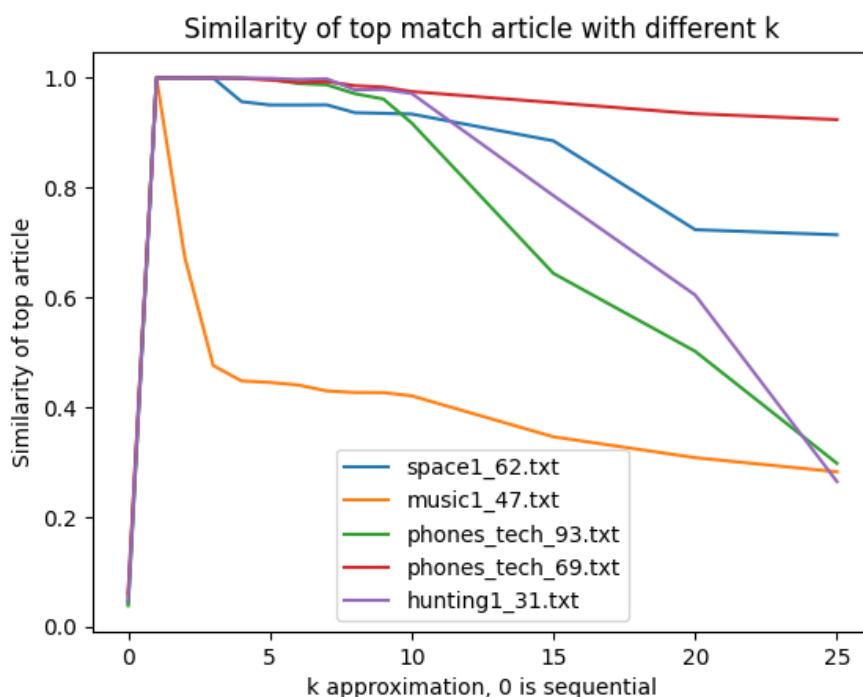


Keďže sme pre experiment použili malý počet článkov z dôvodu rýchleho počítania podobnosti, vytvorili sme ešte jeden graf, ktorý približuje rýchlosť výpočtu od  $k = 1$  a ďalej.



## 5.4 Experiment 5

V poslednom experimente sme porovnali sekvenčný priechod oproti indexu vzhľadom na podobnosť top 5 najpodobnejších článkov. Opäť sme experiment vykonávali na 50 článkoch. Toto porovnanie zachytáva graf nižšie, kde sme pre 5 článkov postupne zvyšovali aproximáciu  $k$ .  $k = 0$  znamená, že nebola použitá žiadna aproximácia a teda výpočet prebiehal sekvenčne.



## 6 Diskusia

Pri práci na projekte sme narazili na problémy hardwaru. Pri spracovávaní veľkého množstva článkov nastáva problém s využívaním RAM, pretože pracujeme s veľkými maticami, ktoré sa načítavajú práve do RAM. Pri implementácii sme dbali hlavne na jej funkčnosť, takže výpočet nie je moc optimalizovaný a na slabších počítačoch môže spôsobiť zamrznutie, prípadne pád programu. Čo sa webovej aplikácie týka, tak tá len čisto zobrazuje články. Bolo by fajn, keby vedela vyhľadávať v článkoch, prípadne pridávať nové a spúšťať celý výpočet LSI.

## Záver

Vďaka tomuto projektu sme si vyskúšali implementáciu LSI vektorového modelu. Z teoretického hľadiska lineárnej algebry si odnášame to, že je veľmi dôležitá a má široké využitie v praxi.