

LSI Vektorový model

Popis projektu

Cílem našeho projektu bylo vytvořit webovou aplikaci, která by implementovala LSI vektorový model k vyhledávání nad kolekcí textových dokumentů.

Vstupem do vyhledávacího formuláře je textový dotaz uživatele, podobně jako u klasických webových vyhledávačů, a hodnota přepínače, která určuje, zda bude vyhledáváno v kolekci sekvenčně, nebo optimalizovaně pomocí LSI vektorového modelu.

Výstupem aplikace je seřazený seznam náhledů dokumentů z kolekce, které nejpřesněji odpovídají zadanému dotazu. Daný náhled je možno rozkliknout a přečíst v plném rozsahu.

Způsob řešení

Data

Za zdroj textových dat jsme si vybrali dataset

[20 newsgroups \(http://qwone.com/~jason/20Newsgroups/\)](http://qwone.com/~jason/20Newsgroups/).

Vstupem z tohoto datasetu je tedy kolekce cca. 20 tisíc textových dokumentů rozdělených téměř rovnoměrně do 20 kategorií. Tento dataset stahujeme přímo pomocí

[knihovny funkce \(https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html\)](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html)

v scikit-learn.

Čištění dat

Textová data mohou obsahovat nežádoucí ruch, proto bylo potřeba je pročistit. S dokumenty postupně provádíme tyto úpravy:

1. všechna písmena převedeme na písmena malá
2. odstraníme e-mailové adresy
3. odstraníme non-alfabetické znaky
4. lematizujeme slova
5. odstraníme krátká slova

Vytvoření vyhledávacího modelu

Nejprve si vyrobíme term-by-document matici o rozměrech (<počet dokumentů>, <počet termů>).

Hodnoty v ní převážíme pomocí tf-idf schématu.

Tuto matici dekomponujeme pomocí singular-value-decomposition (SVD) na matice:

- **U**
 - rozměry: (<počet dokumentů>, <k = počet konceptů>)
 - concept-by-document matice
- **S**

- rozměry: ($<n = \text{počet konceptů}>$,)
- vektor konceptů
- **vt**
 - rozměry: ($<k = \text{počet konceptů}>$, $<\text{počet termů}>$)
 - koncept-by-term matice

Hledání optimálního počtu konceptů

Naším cílem bylo najít optimálního počtu k konceptů tak, aby k bylo co nejnižší (kvůli rychlosti vyhledávání) a zároveň výsledky co nejpřesnější.

Zobrazení dotazu do prostoru konceptů

Dále je zapotřebí zobrazit lematizovaný dotaz uživatele do prostoru konceptů jako vektor, následně změřit kosinovou vzdálenost v tohoto vektoru od ostatních vektorů dokumentů.

Implementace

Jazyk

K vývoji jsme použili programovací jazyk [Python \(https://www.python.org/\)](https://www.python.org/) a mikro webový framework [Flask \(https://flask.palletsprojects.com/en/1.1.x/\)](https://flask.palletsprojects.com/en/1.1.x/).

Knihovny

- [Pandas \(https://pandas.pydata.org/\)](https://pandas.pydata.org/) k analýze a zpracování dat
- [Numpy \(https://numpy.org/\)](https://numpy.org/) a [Scipy \(https://www.scipy.org/\)](https://www.scipy.org/) k práci s maticemi
- [NLTK \(https://www.nltk.org/\)](https://www.nltk.org/) k analýze přirozeného jazyka
- [scikit-learn \(https://scikit-learn.org/stable/\)](https://scikit-learn.org/stable/) k vytváření LSI modelu
- [Kneed \(https://kneed.readthedocs.io/en/stable/\)](https://kneed.readthedocs.io/en/stable/) k hledání zlomů v křivce optimálního počtu konceptů

Při testování našich nápadů jsme využili [Jupyter notebook \(../logic/logic.ipynb\)](https://mybinder.org/v2/gh/QuentinZhang/lsi-model).

Celá aplikace je kontejnerizovaná v [Dockeru \(https://www.docker.com/\)](https://www.docker.com/).

Stavba aplikace

Veškerá logika aplikace se nachází v modulu [IsiModel \(../IsiModel.py\)](https://github.com/QuentinZhang/lsi-model/blob/master/lsi_model.py) respektive ve třídě LSI uvnitř něj.

Důležité třídní metody:

- `prepare`
 - pokud není lokálně stažen dataset s dokumenty, stáhne je
 - pročistí dokumenty
 - vytvoří model pro vyhledávání (ten se vytvoří pouze jednou při inicializaci, poté už zůstává uvnitř třídy)
- `svd_optimal_k`

- rozdělí uživatelův dotaz na slova
- pro každé slovo sekvenčně prochází term-document maticí a vrací dokumenty, ve kterých se slovo nachází

Když uživatel potvrdí dotaz ve formuláři na hlavní stránce, je tento dotaz předán LSI třídě. Ta dotaz vyhodnotí a vrátí nazpět list výsledných dokumentů. Ty jsi zobrazeny v seznamu výsledků.

[home](#)[about](#)

LSI vector model

give me seached query.

washington

LSI status

search sequentially ☐

search

This is semestral project for the BI-VWM course in semester B202 at FIT CTU

[Daniel Bukac](#) ([bukacdan](#))[Matej Latka](#) ([latkamat](#))

uživatel zadal vstup "washington" a nevybral možnost vyhledávat sekvenčně

Number of results limited to 100.
Results found in 0.15 s.

Query: washington | Lemmatized query: washington | Angle: 0.21831618861694319 | Document index: 1497 | Document category: talk.politics.misc

1 THE WHITE HOUSE Office of the Press Secretary ...

[Go to detail →](#)

Query: washington | Lemmatized query: washington | Angle: 0.3319934133122527 | Document index: 11386 | Document category: talk.politics.guns

2 National Rifle Association 1600 Rhode Island Ave. NW Washington, DC 20036-3268 1-800-368-5714 (memb...

[Go to detail →](#)

Query: washington | Lemmatized query: washington | Angle: 0.358895850377166 | Document index: 2385 | Document category: talk.politics.misc

3 From: harelb@math.cornell.edu (misc.activism.progressive co-moderator) Subject: F<O>CUS/HEALTH: How ...

[Go to detail →](#)

Query: washington | Lemmatized query: washington | Angle: 0.3597263557669946 | Document index: 11575 | Document category: talk.politics.misc

4 THE WHITE HOUSE Office of the Press Secretary ...

[Go to detail →](#)

Query: washington | Lemmatized query: washington | Angle: 0.38189183593424383 | Document index: 7229 | Document category: talk.politics.misc

5 Here is a press release from the American Federation of Teachers. HHS Secretary Shalala to Address...

[Go to detail →](#)

Query: washington | Lemmatized query: washington | Angle: 0.39407219385883335 | Document index: 8866 | Document category: sci.med

6 Here is a press release from the American Federation of State, County and Municipal Employees. Un...

[Go to detail →](#)

Query: washington | Lemmatized query: washington | Angle: 0.39666869960582735 | Document index: 4461 | Document category: talk.politics.misc

Aplikace vrátila 100 výsledků za 0.15s.

Každý výsledek má nad sebou popsany:

- původní dotaz
- lematizovaný dotaz
- úhlovou vzdálenost výsledku
- index dokumentu
- kategorii dokumentu

Result n.1

THE WHITE HOUSE Office of the Press Secretary For Immediate Release April 15, 1993 STATEMENT BY THE PRESS SECRETARY The President will travel to Pittsburgh on Saturday, April 17 to talk about his job creation plan and its impact on the state of Pennsylvania, where it would create as many as 3,818 full time jobs and up to 21,240 summer jobs. He will make a public address at Pittsburgh International Airport at 9:30 am. The President will leave Washington early Saturday morning and return that afternoon. A White House press charter will depart Andrews Air Force Base at 7:30. Filing facilities will be available in Pittsburgh.

Similar documents

Angle: 0.2194585668686967 | Document index: 1 | Document category: comp.sys.ibm.pc.hardware

1 My brother is in the market for a high-performance video card that supports VESA local bus with 1-2M...

[Go to detail →](#)

Angle: 0.25894112064439 | Document index: 8772 | Document category: comp.sys.ibm.pc.hardware

2 Hai, In a few days I'm going to buy a new motherboard with local-bus(ses). It comes with a Cirrus L...

[Go to detail →](#)

Angle: 0.26063132695642527 | Document index: 12782 | Document category: comp.graphics

3 I am using an ibm dx-50 with EISA and local bus....and I need to get a local bus video card.... The...

[Go to detail →](#)

Angle: 0.27988804755977736 | Document index: 5943 | Document category: comp.graphics

4 i am sorry, but this genoa card does nothing that the ATI ultra plus 2mb can't do, PLUS the ATI cost...

[Go to detail →](#)

Angle: 0.2872549458344135 | Document index: 1674 | Document category: comp.sys.ibm.pc.hardware

5 I UPGRADED MY OLD 386 WITH 486DX-50 LOCAL BUS MOTHERBOARD TWO MONTH AGO AND WITH IT I BOUGHT A CONT...

[Go to detail →](#)

Angle: 0.295633488375331 | Document index: 8896 | Document category: comp.sys.ibm.pc.hardware

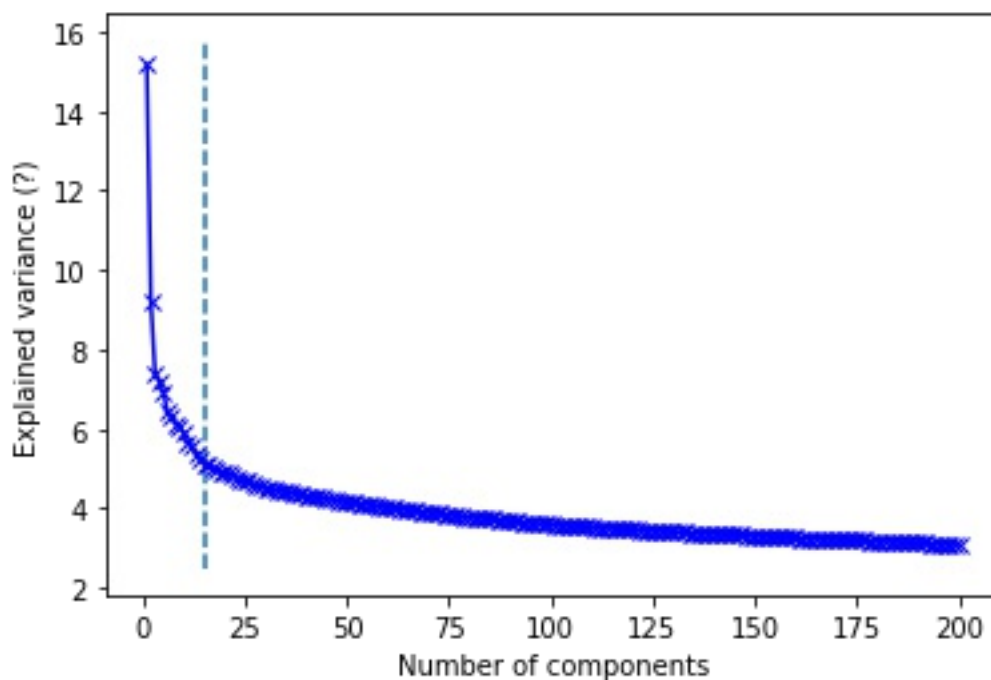
I have a WANGTFK tane controller card (Revision F) that was used with the Svtos hackin system to tak...

V horní části obrazovky je původní text (bez našich úprav) v plném rozsahu.

Pod ním jsou vylistované jemu podobné dokumenty.

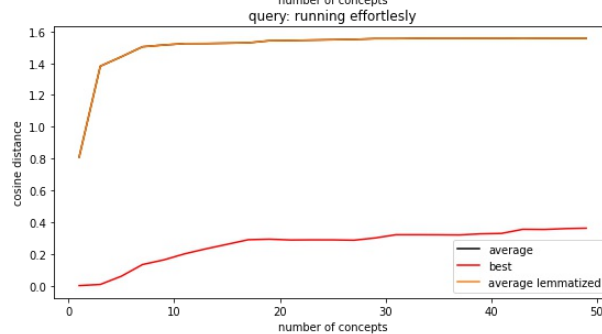
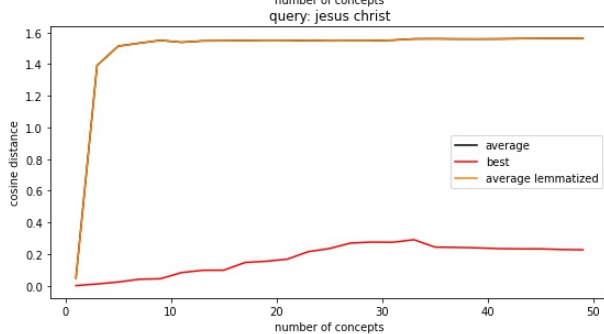
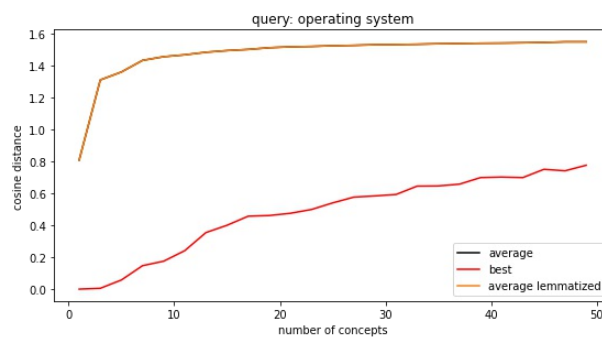
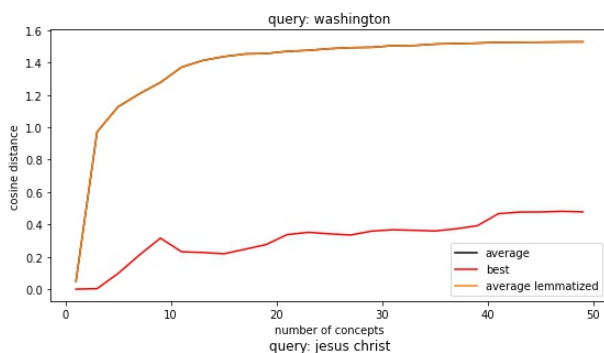
Experimentální sekce

Nejprve bylo potřeba určit optimální počet K konceptů. Vyzkoušeli jsme pro K hodnoty z intervalu $\langle 1, 200 \rangle$ a pozorovali hodnoty v matici S při singular value decomposition.



Na ose x je vynesen počet konceptů, na ose y hodnoty singular values tedy "důležitost" konceptů. Křivka se lomí v bodě $x=15$

V dalším experimentu zkoušíme hodnoty pro K z intervalu $\langle 1, 50 \rangle$ a inkrementujeme o 2 (pro více hodnot trval výpočet příliš dlouho). Zároveň pro každou hodnotu K zkoušíme dotaz zpracovat s lematizací i bez. Výsledky zkoušíme na 4 různých dotazech.



V grafech je vidět, že pro nižší hodnoty K je kosinová vzdálenost menší, nicméně při prozkoumání výsledných dokumentů se ukázalo, že nejsou příliš relevantní pro dotaz. Rostoucí funkce průměrné kosinové vzdálenosti v závislosti na K se láme zhruba okolo bodu $K=15$, což odpovídá předchozímu experimentu.

Zároveň se ukázalo, že lemmatizace dotazu nemá žádný vliv na výsledky (křivka průměrné vzdálenosti s lemmatizací kopíruje křivku bez lemmatizace).

Diskuze

Největším problémem modelu je, že pokud je mu zadán dotaz, který se neobjevuje v žádném z dokumentů a ani v žádném z konceptů, tedy vektor tohoto dotazu je nulový, všechny dokumenty v kolekci jsou stejně dobré, tedy mají stejnou kosinovou vzdálenost. Model proto vrátí jako nejlepší výsledek první dokument v kolekci (shodou náhod o Pittsburgh Penguins a Jaromíru Jágrovi).

Dalším nedostatkem je řešení sekvenčního prohledávání. V ideálním případě by mělo být realizováno pomocí nastavení počtu konceptů K na maximální hodnotu (v našem případě počet dokumentů).

Pro takto vysokou hodnotu (téměř 20000) nám však nestačila operační paměť a program zkolaboval.

Závěr

Podařilo se nám implementovat LSI vektorový model k information retrieval. Při zadání dotazu, který je možné v dokumentech najít, model vrátí relevantní výsledky.

Při řešení jsme se potýkali s menšími problémy, nejzávažnější pro logiku modelu bylo správné zobrazování dotazu do prostoru konceptu. Samozřejmě nemohly chybět ani zádrhely s webovým GUI.

Projekt hodnotíme jako zajímavý a přínosný.