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

6. května 2020

# OBSAH


**1 Popis projektu** **3**

**2 Způsob řešení** **3**

2.1 Preprocesing dokumentů . . . . . 3
2.2 Výpočet vah termů . . . . . 3
2.3 Implementace LSI . . . . . 4
2.4 Vyhodnocení dotazu . . . . . 4

**3 Implementace** **4**

3.1 Identify the author of Equation 3.1 below and briefly describe it in English. . . . . 4
3.2 Try to make sense of some more equations. . . . . 4

**4 Příklad výstupu** **4**

4.1 Bullet Point List . . . . . 4
4.2 Numbered List . . . . . 5

**5 Experimentální sekce** **5**

5.1 The table above shows the nutritional consistencies of two sausage types. Explain their relative differences given what you know about daily adult nutritional recommendations. . . . . 5

**6 Diskuze** **5**

6.1 How many luftballons will be output by the Listing 1 above? . . . . . 6
6.2 Identify the regular expression in Listing 1 and explain how it relates to the anti-war sentiments found in the rest of the script. . . . . 6

**7 Závěr** **6**

# 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 doporučování nejpodobnějších článků k právě čtenému.

V experimentální části projektu jsme se dále zaměřili na:

- Porovnání průchodu pomocí LSI vektorového modelu se sekvenčním průchodem databáze
- Porovnání vlivu LSI na kvalitu výsledků vyhledávání s ohledem na výskyt synonym a homonym
- Vliv různých vnitřních parametrů na výkon algoritmu (změna počtu konceptů, změna počtu extrahovaných termů, použití lemmatizace namísto stemmingu, odstranění číslovek při preprocesingu...)
- Jak se změní výsledky při použití jiného vzorce na výpočet vah termů

Celý náš projekt je volně dostupný na: (Odkaz na gitlab?).

# 2 Způsob řešení

## 2.1 Preprocesing 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. Porovnání jejich použití v programu se dále věnujeme v experimentální části.

## 2.2 Výpočet vah termů

V aplikaci vytváříme term-by-document matici, která bude mít 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, aby úloha byla výpočetně řešitelná v rozumném čase. 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
- *keep_less_freq* - udává, zda si při výběru *max_terms* termů nechat ty nejméně či nejvíce často zastoupené v dokumentech

Zkoumání vlivu změny jednotlivých parametrů na výsledek LSI se dále podrobněji věnujeme v experimentální části. V programu vždy nastavujeme *min_df* alespoň na 1, abychom odstranili termy nacházející se pouze v 1 dokumentu, které nám do LSI nepřidávají žádné užitečné informace. (pravda?)

Z této zredukované matice poté již spočteme term-by-document matici vah termů v jednotlivých dokumentech. Pro výpočet vah jednotlivých termů používáme metodiku tf-idf.

### 2.3 Implementace LSI

### 2.4 Vyhodnocení dotazu

## 3 Implementace

### 3.1 Identify the author of Equation 3.1 below and briefly describe it in English.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.1}$$

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent porttitor arcu luctus, imperdiet urna iaculis, mattis eros. Pellentesque iaculis odio vel nisl ullamcorper, nec faucibus ipsum molestie. Sed dictum nisl non aliquet porttitor. Etiam vulputate arcu dignissim, finibus sem et, viverra nisl. Aenean luctus congue massa, ut laoreet metus ornare in. Nunc fermentum nisi imperdiet lectus tincidunt vestibulum at ac elit. Nulla mattis nisl eu malesuada suscipit.

### 3.2 Try to make sense of some more equations.

$$\begin{aligned}(x+y)^3 &= (x+y)^2(x+y)\\ &= (x^2+2xy+y^2)(x+y)\\ &= (x^3+2x^2y+xy^2)+(x^2y+2xy^2+y^3)\\ &= x^3+3x^2y+3xy^2+y^3\end{aligned} \tag{3.2}$$

Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

$$A = \begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix} \tag{3.3}$$

Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem.

## 4 Příklad výstupu

### 4.1 Bullet Point List

- First item in a list
  - First item in a list
    - First item in a list
    - Second item in a list
  - Second item in a list
- Second item in a list

### 4.2 Numbered List

1. First item in a list
2. Second item in a list
3. Third item in a list

# 5 Experimentální sekce

| *Per 50g* | **Pork** | **Soy** |
|---|---|---|
| Energy | 760kJ | 538kJ |
| Protein | 7.0g | 9.3g |
| Carbohydrate | 0.0g | 4.9g |
| Fat | 16.8g | 9.1g |
| Sodium | 0.4g | 0.4g |
| Fibre | 0.0g | 1.4g |

Tabulka 5.1: Sausage nutrition.

### 5.1 The table above shows the nutritional consistencies of two sausage types. Explain their relative differences given what you know about daily adult nutritional recommendations.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent porttitor arcu luctus, imperdiet urna iaculis, mattis eros. Pellentesque iaculis odio vel nisl ullamcorper, nec faucibus ipsum molestie. Sed dictum nisl non aliquet porttitor. Etiam vulputate arcu dignissim, finibus sem et, viverra nisl. Aenean luctus congue massa, ut laoreet metus ornare in. Nunc fermentum nisi imperdiet lectus tincidunt vestibulum at ac elit. Nulla mattis nisl eu malesuada suscipit.

# 6 Diskuze

Listing 1: Luftballons Perl Script.

```
#!/usr/bin/perl

use strict;
use warnings;

for (1..99) { print $_." Luftballons\n"; }

# This is a commented line

my $string = "Hello World!";

print $string."\n\n";

$string =~ s/Hello/Goodbye Cruel/;

```

```
16 print $string."\n\n";
17
18 finale();
19
20 exit;
21
22 sub finale { print "Fin.\n"; }
```

### 6.1 How many luftballons will be output by the Listing 1 above?

Aliquam arcu turpis, ultrices sed luctus ac, vehicula id metus. Morbi eu feugiat velit, et tempus augue. Proin ac mattis tortor. Donec tincidunt, ante rhoncus luctus semper, arcu lorem lobortis justo, nec convallis ante quam quis lectus. Aenean tincidunt sodales massa, et hendrerit tellus mattis ac. Sed non pretium nibh. Donec cursus maximus luctus. Vivamus lobortis eros et massa porta porttitor.

### 6.2 Identify the regular expression in Listing 1 and explain how it relates to the anti-war sentiments found in the rest of the script.

Fusce varius orci ac magna dapibus porttitor. In tempor leo a neque bibendum sollicitudin. Nulla pretium fermentum nisi, eget sodales magna facilisis eu. Praesent aliquet nulla ut bibendum lacinia. Donec vel mauris vulputate, commodo ligula ut, egestas orci. Suspendisse commodo odio sed hendrerit lobortis. Donec finibus eros erat, vel ornare enim mattis et.

## 7 Závěr