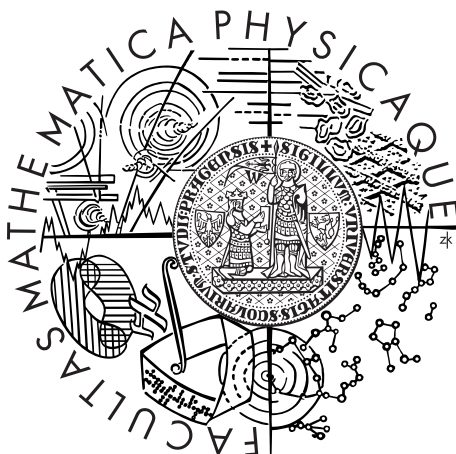


Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Bc. Ondřej Odcházal

Automatické doporučování ilustračních snímků

Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: RNDr. Pavel Pecina, Ph.D.

Studijní program: Informatika

Studijní obor: Matematická lingvistika

Praha 2014

děkuji máje, že se nebojí létat

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V dne

Podpis autora

Název práce: Automatické doporučování ilustračních snímků

Autor: Bc. Ondřej Odcházal

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: RNDr. Pavel Pecina, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt:

Klíčová slova: vyhledávání obrazových informací

Title: Automatic suggestion of illustrative images

Author: Bc. Ondřej Odcházal

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Pavel Pecina, Ph.D., Institute of Formal and Applied Linguistics

Abstract:

Keywords: information retrieval, image retrieval

Obsah

1 Úvod	2
Úvod	2
1.1 Práce s daty	2
1.2 Extrakce klíčových slov	2
1.3 Překlad do češtiny	2
1.4 Detekce jazyka	2
1.5 Webová aplikace	3
1.6 Testování	3
2 Zadání	4
3 Teorie: Jak najít vhodné obrázky	5
3.1 Popis datové sady	5
3.2 Teoretické cíle	5
3.3 Rešerše, vhodné algoritmy	5
3.3.1 TF-IDF	6
3.3.2 Extrakce bez korpusu	6
3.4 Řešení: jaké algoritmy zvoleny, získání tréninkových dat	7
3.5 Evaluace výsledků	7
4 Praktická část: Implementace moderní webové aplikace	8
4.1 Databáze: jak uložit 20M metadat obrázků	8
4.2 Backend a úprava dat: Komunikace s databází, implementace algoritmů	8
4.3 Frontend: AJAXová aplikace na zobrazování obrázků	8
4.4 Anotační rozhraní	8
4.5 Návod k použití	8
4.6 Překlad	8
4.7 Poznámky	10
5 Evaluace výsledků	11
5.1 Metodika	11
5.2 Výsledky	11
5.3 Možná zlepšení	11
Závěr	12
Seznam použité literatury	13
Seznam tabulek	14
Seznam použitých zkratk	15
Přílohy	16

1. Úvod

Cílem diplomové práce je implementovat kompletní webovou aplikaci pro doporučování a vyhledávání ilustračních obrázků v textu. Vytvořit takovou aplikaci přináší mnoho rozličných úkolů a problémů. Tato kapitola se bude snažit tyto problémy načrtnout. Další kapitola se bude jednotlivými problémy zabývat podrobně.

1.1 Práce s daty

Zadaná data obsahují 20 milionů anotací obrázků. Základním úkolem je být schopen takové množství dat vůbec nahrát do databáze a být schopný obsloužit mnoho požadavků za minutu. Bude zmíněn současný stav vývoje databázového software pro práci s velkými daty zejména s ohledem na snadnost hledání a škálovatelnost.

1.2 Extrakce klíčových slov

Extrakce klíčových slov je důležitý podobor NLP. V práci budou rozebrány algoritmy pro extrakci klíčových slov. Bude kladen zejména důraz na rychlost a nenáročnost na zdroje. Z uživatelských testování společnosti Google vychází, že rychlost načtení stránky je jedním z klíčových vlastností pro spokojenost uživatele. Klíčová slova budou mít v aplikaci dvě využití. Pokud uživatel zadá pouze text článku, extrahovaná klíčová slova se použijí na vyhledávání relevantních obrázků. Prvních několik klíčových slov bude navíc použita jako nápověda uživateli, ten pak může tato klíčová slova využít k exaktnímu omezení množiny klíčových slov.

1.3 Překlad do češtiny

Popisky klíčových slov jsou v angličtině. Tato práce řeší překlad množiny klíčových slov do češtiny. Kromě překladu je pro hledání také nutno implementovat algoritmus na stemming. Celá aplikace je navržena tak, aby případný další jazyk mohl být přidán co nejjednodušeji.

1.4 Detekce jazyka

Jednou z drobností, kterou ocení uživatel aplikace je detekce jazyka. Uživatel bude mít možnost zadat jazyk vstupního článku exaktně, ale aplikace bude také jazyk vstupního textu sama detekovat. Budou prozkoumány možnosti detekce jazyka. Opět se nejedná o nějakou klíčovou funkci aplikace. Výstup detekce bude moci být uživatelem měněn (podobně jako funguje Google Translate¹), důraz bude tedy kladen na rychlost a jednoduchost.

¹<https://translate.google.com/>

1.5 Webová aplikace

Všechny předchozí komponenty se spojí v jedné webové aplikaci. Webový vývoj zažívá bouřlivý rozvoj. Na backendu jsou nové zejména způsoby práce s velkým množstvím dat v distribuovaném prostředí. Ve frontendové části probíhá rozvoj pomocí implementace nových technologií, známých pod hlavičkou HTML5, do moderních prohlížečů. Práce bude rozebírat všechny možnosti tvorby moderních webových aplikací.

1.6 Testování

Aplikace bude otestována na několika úrovních. Extrakce klíčových slov bude otestována pomocí korpusu článků a klíčových slov. Bude vytvořena komplexní webová aplikace pro testování doporučených obrázků. Tato aplikace bude vydělena ze samotné webové aplikace a bude používána i nezávisle.

2. Zadání

Většina zpravodajských serverů často opatřuje publikované články tzv. ilustračními snímky, jejichž úkolem je vizuálně dokreslovat obsah článku a upoutat na něj čtenářovu pozornost. Ilustrační snímky většinou pocházejí z rozsáhlých fotografických databází, jsou vybírány autory článku a s obsahem článku souvisejí jen relativně volně. Výběr ilustračních snímků probíhá nejčastěji na základě porovnávání klíčových slov specifikovaných autorem textu a popisků, kterými jsou obrázky v databázi opatřeny (typicky svými autory).

Proces výběru ilustračních snímků (dotazování ve fotografické databázi) je obtížný jednak pro samotný vyhledávací systém (hledání relevantních fotografií na základě uživatelských dotazů), jednak pro autory, kteří musí dotazy vytvářet. Konstrukce dotazů spočívá v několika krocích: uživatel nejdříve musí identifikovat ústřední téma (či témata) článku, které chce ilustrovat vhodnou fotografií, a ta potom popsat vhodnými klíčovými slovy, zvolit a zkombinovat je tak, aby vedla k nalezení vhodného obrázku. Tento proces by mohl být zjednodušen tím, že konstrukce dotazů pro vyhledávání bude prováděna automaticky pouze na základě textu článku.

Cílem diplomové práce je navržení a implementace komfortní webové aplikace pro automatické navrhování ilustračních snímků na základě textu článku, bez nutnosti explicitně konstruovat vyhledávací dotazy. Součástí práce bude i uživatelská evaluace celého systému. Pro experimenty bude použita kolekce ilustračních snímků od společnosti Profimedia.

3. Teorie: Jak najít vhodné obrázky

3.1 Popis datové sady

Datová sada poskytnutá firmou Profimedia obsahuje 20 014 394 oannotovaných obrázků ve formě CSV souboru "profi-text-cleaned.csv". CSV obsahuje sloupce "locator", "title", "description", "keywords". Sloupec locator obsahuje ID obrázku v databázi Profimedia. Sloupec description je prázdný. Sloupce title a keywords obsahují řetězce anglických slov popi sujících obrázků a oddělených mezerou.

Příklad jednoho řádku souboru profi-text-cleaned.csv:

```
"0000000980","hradec kings holy ghost cathedral","",  
"outdoors nobody urban scenes architecture houses  
towers czech czech republic europe buildings build  
history historical churches church fronts holy ghost  
cathedral spirit ceska republika cathedrals sv hradec  
kralove" ^M
```

Na příkladu je vidět, že data obsahují fráze jako "holy ghost cathedral", tyto fráze však nejsou strojově čitelně vyznačené. Dalším problémem je špatný překlad dat do angličtiny. Fráze "hradec kings" vznikla evidentně doslovným překladem názvu "hradec králové".

Důležitým aspektem dat je jejich nepodobnost běžnému novinovému textu. Anotované texty neobsahují většinu nejfrekventovanějších anglických slov.

3.2 Teoretické cíle

V teoretické části je hlavním cílem práce nalézt nejvhodnější metodu extrakce klíčových slov textu. Tato klíčová slova pak budou použita při vyhledávání ilustračních obrázků v databázi Profimedia. Celá práce, pokud nebude uvedeno jinak, označuje za slova stemy vstupních slov. Jako stemmer se využívá ??? stemmer.

Vstupní text tedy nejprve rozdělíme na slova. Číslo a interpunkce nás v této úloze nezajímají, jelikož se v datech nenachází. Ze slov pak získáme stemy. Vstupem algoritmu pro nalezení klíčových slov tedy bude množina vhodných stemů. Ke každému stemu si ještě uložíme jednu jeho nestemovou variantu, kterou pak můžeme zobrazit uživateli.

Nyní můžeme použít některý z algoritmů na extrakci klíčových slov uvedených v další kapitole.

3.3 Rešerše, vhodné algoritmy

Algoritmy na extrakci klíčových slov lze rozdělit do dvou kategorií. V jedné máme k dispozici korpus podobných dokumentů, druhá kategorie tento korpus ke své práci nepotřebuje.

3.3.1 TF-IDF

Technika TF-IDF je známý algoritmus na měření významnosti slov v textu. Využívá korpusu dokumentů D a dvou složek.

$$TFIDF(t, d, n, N) = TF(t, d) \times IDF(n, N) \quad (3.1)$$

Složka TF znamená *TERM FREQUENCY* a pokud t je slovo a $d \in D$ je dokument, je TF

$$TF(t, d) = \begin{cases} 1 & \text{pokud } t \in d \\ 0 & \text{jinak} \end{cases} \quad (3.2)$$

$$TF(t, d) = \sum_{slovo \in d} \begin{cases} 1 & \text{pokud } slovo = t \\ 0 & \text{jinak} \end{cases} \quad (3.3)$$

Jedná se tedy o frekvenci slova (stemu) v dokumentu.

Složka IDF , tedy *INVERSE DOCUMENT FREQUENCY* vyjadřuje, jak moc daný termín popisuje dokument. Pokud je N počet všech dokumentů v D , tedy $N = |D|$ a n je počet dokumentů, ve kterých se vyskytuje slovo t , je IDF tohoto slova

$$IDF(n, N) = \log \left(\frac{N}{n} \right) \quad (3.4)$$

$$IDF(n, N) = \log \left(\frac{N - n}{n} \right) \quad (3.5)$$

Čím je tedy slovo v korpusu častější, tím více se s logaritmem snižuje jeho informační hodnota. Slova, která jsou velmi běžná většinou klíčovými slovy.

Výsledný vzorec pak jde shrnout jako:

$$TFIDF(t, d, n, N) = \left(\sum_{slovo \in d} \begin{cases} 1 & \text{pokud } slovo = t \\ 0 & \text{jinak} \end{cases} \right) \times \log \left(\frac{N - n}{n} \right) \quad (3.6)$$

Problémem tohoto algoritmu pro je odlišný charakter korpusu a vstupních dat. Vstupní data jsou typicky novinový článek. Pokud bychom jako korpus použili anotované obrázky, získáme špatné výsledky. Běžná anglická slova, jako "the", nebo "a" se v takovém korpusu vyskytují velmi zřídka, jejich IDF tedy bude vysoká. Naopak TF v běžném novinovém textu je vysoké. Takovýto korpus nám pak označuje jako klíčová slova běžná anglická slova.

3.3.2 Extrakce bez korpusu

Dalšími druhy algoritmů ke své práci korpus nepotřebují a pracují pouze se vstupním textem.

3.4 Řešení: jaké algoritmy zvoleny, získání tréninkových dat

Jako nejvhodnější řešení byl nakonec zvolen TF-IDF algoritmus. Jako kandidáti jsou odfiltrována slova, která se nenacházejí v datech Profimedia. Jako korpus k měření IDF byla použita data článků z Wikipedie.

Pro rychlé testovací účely bylo oannotováno pár článků z anglických wikinews. V každém článku jsem označil pět klíčových slov. Porovnávání algoritmů na extrakci klíčových slov pak vzalo pět nejpravděpodobnějších klíčových slov podle algoritmu a porovnávalo v kolika procentech se algoritmus trefil s anotací.

Zbývá: Zkusit otestovat další metody. Překlad do češtiny.

3.5 Evaluace výsledků

Výsledkem práce by mělo být rozhraní pro anotaci obrázků využívající algoritmus na hledání klíčových slov v textu.

Tento algoritmus se dá uživatelsky testovat několika způsoby. Uživatel vidí text a několik (cca 5) vrácených obrázků algoritmem. Uživatel vybere množinu relevantních obrázků. Další možností je mezi 5 vrácených obrázků vložit jeden náhodný. Úkolem anotátora je pak vybrat ten náhodně vybraný. Přesnost algoritmu pak jde měřit pomocí toho, kolikrát se anotátor trefí do špatného obrázku (potřeba zdrojový článek)

4. Praktická část: Implementace moderní webové aplikace

Práce je z velké části implementační. Snažil jsem se tedy vytvořit moderní webovou aplikaci s využitím co nejvíce frontendových novinek v novém standardu HTML5.

4.1 Databáze: jak uložit 20M metadat obrázků

Jak uložit takové množství dat, aby se dalo rychle vyhledávat. Škálovatelnost. Dostupnost knihoven pro práci s databázema. Proč jsem si vybral nakonec ES. SQL vs NoSQL.

4.2 Backend a úprava dat: Komunikace s databází, implementace algoritmů

Proč jsem zkoušel go a proč jsem nakonec použil Ruby a Ruby on Rails. Základní popis MVC frameworku. Dostupnost knihoven pro získání stemů a práci s databází.

4.3 Frontend: AJAXová aplikace na zobrazování obrázků

Jaké jsou dnešní možnosti vývoje frontendu. Single page aplikace. Možnosti moderních prohlížečů. JavaScriptové knihovny. Proč to nedělám v jQuery, ale používám Google Closure. Google Closure Library, Templates, Compiler.

Návrh rozhraní bez jediného tlačítka. Responzivní webdesign.

4.4 Anotální rozhraní

Jak lze v Ruby on Rails vyrobit jednoduše anotační rozhraní s uživateli a s ukládáním do databáze.

4.5 Návod k použití

Popis prvků. Screenshoty aplikace.

4.6 Preklad

Dva postupy jak použít anglická data v jiných jazycích. Buď je možné přeložit vždy zadaný český dotaz do cílového jazyka. Nebo je možné přeložit všechna data

u fotek. Pak je nutné použít pro každý jazyk nějaký lemmatizer, nebo stemmer. Pro cestinu jsme nakonec zvolili druhou variantu.

Jak přeložit 20 milionu popisků? Jednou možností je překlad slovo od slova. Použít pouze slovník. Překlad pomocí google je drahý. Stalo to zhruba 1300Kč. Překlad celých frází by byl lepší (automaticky objeví fráze), ale pomocí google velmi drahý. Rozhodl jsem tedy překladový nástroj Moses s modelem přiloženým ve verzi 2.1 (<http://www.statmt.org/moses/RELEASE-2.1/models/en-cs/model/>). Po několikahodinovém nabití se model nacetl (i když mám SSD disk). Překlady v proloženém modelu jsou velmi pomale (jeden segment trvá přibližně 3s). Překlad není ideální a mám spoustu —UNK slov.

Překlad popisku prvního obrázku: "0000000003", "little baby smiling", "", "child children baby babies infants kids childhood single faces body naked naked facial expressions smile smiling viewing watching laying fun amusing amusement amused amuse dallying frolic playing wantoning open"^M

Moses: "0000000003", "little—UNK—UNK—UNK dítě smiling", "", "child—UNK—UNK—UNK děti, dětské děti kojence děti dětství jednotného čelí orgán nahé naked—UNK—UNK—UNK pořídili vyjádření usmívat usmívá odůvodněním, která zábavné sledovat zábavné zábavných i pobavena tím amuse—UNK—UNK—UNK dallying—UNK—UNK—UNK frolic—UNK—UNK—UNK hrát wantoning—UNK—UNK—UNK open"^M|UNK|UNK|UNK

Google: "0000000003", "malé dítě s úsměvem", "", "dítě děti dítě děti kojenci děti dětství jednotlivé plochy těla nahá naked výrazy obličeje, úsměvu, usměvavý sledování sledování kterým zábava zábavné zábavní pobavený pobavit laškoval frolic hrát wantoning otevřený"^M

Překlad jiného obrázku: "0000000102", "young woman cleaning teeth", "", "single faces people humans young youth hands indoors interiors woman women females blond fair young adult s girls close view beauty home home dental bathrooms person portrait adult years half length portrait open mouth hygiene teeth dental care years cleaning toothbrush underwear bras"^M

Moses: "0000000102", "young—UNK—UNK—UNK žena čištění teeth", "", "single—UNK—UNK čelí mladí lidé, lidé mládež rukou uvnitř interiors—UNK—UNK—UNK žena žen, žen, blondák spravedlivé mladé dívky zavřít dospělé s cílem krásy vnitřní vnitřní stomatologické koupelny portrét dlouhé roky polovina dospělé osoby portrét otevřené úst hygienické zuby zubní kartáček prádlo bras"^M|UNK|UNK|UNKpeletitn

Google Translate: "0000000102", "Mladá žena čištění zubů", "", "jednotlivé tváře lidí, lidé younge mládeží ruce interiéry žen ženám ženy ženskému blond fair mladý dospělý s dívek close view krása domov domácí zubní koupelny osoba portrét dospělý let poloviční délka portrét otevřená ústa hygienické zubů zubní péče roky čistící kartáček na zuby spodní prádlo podprsenky"^M

Detekované fráze: <http://mufin.fi.muni.cz/xbatko/keyword-clean-phrase-export.csv.gz>
mail: <https://mail.google.com/mail/u/0/search/pecina+>

export frází bez omezení: `wc phrasesist.txt1143513515183137131575phrasesist.txt—takesenedaprelozitrozumnegoogle`

omezení na maximální délku 4 slova: `wc phrasesist.txt899244220461615677087phrasesist.txt`

extrakce frází: <http://stackoverflow.com/questions/1643616/algorithms-to-detect-phrases-and-keywords-from-text>

detektor jazyka: trigramy, nefunguje na frázi "hello world", chce to dlouhá data, ale pro bezné články to funguje dobře

4.7 Poznamky

Pouziji data z wiki dumpu.

Cesky dump: <http://dumps.wikimedia.org/cswiki/20140612/> a [cswiki-20140612-pages-articles-multistream.xml.bz2](http://dumps.wikimedia.org/cswiki/20140612/pages-articles-multistream.xml.bz2) /data/wiki_dump_c.s.xml rakees : *extract_words_from_wiki[cs]* es: 60MB textu cestina obsahuje nazacatku hodne kratkych vygenerovanych clanku typu > 1. leden <

Anglicky dump: torrent z piratebay <http://thepiratebay.se/torrent/8114722/Wikipedia2013>, *extract_words_from_wiki[en]* extrakcedoid10000 > 70MB textu

application.rb obsahuje konstanty aplikace

vyrobim nejcastejsi trigramy pro cestinu a anglictinu z wiki dat pomoci: rake es: *extract_most_frequent_trigrams*

Language data jsem stahnul z: <http://dumps.wikimedia.org/cswiki/latest/>

Ceska data jsou z: /net/seznamdata/profiset/profi-text-cleaned.csv

prekladova data: moses - <http://www.statmt.org/moses/RELEASE-2.1/binaries/macosx-mavericks/bin/>

prekladovy model: <http://www.statmt.org/moses/RELEASE-2.1/models/en-cs/>

5. Evaluace výsledků

5.1 Metodika

Jak budu měřit. Porovnání naivního a nejlepšího algoritmu. Nejlepší algoritmus vybrán pomocí oannotovaných klíčových slov ve Wikinews článcích. Pak anotátoři se budou snažit najít nejméně vhodný obrázek z nabízených hodnot.

Evaluace jenom pro angličtinu, nebo i pro češtinu?

5.2 Výsledky

Grafy, tabulky.

5.3 Možná zlepšení

Jak bychom mohli mít lepší data a co omezuje použitý algoritmus.

6. Závěr

Seznam použité literatury

- [1] LAMPORT, Leslie. *ΛT_EX: A Document Preparation System*. 2. vydání. Massachusetts: Addison Wesley, 1994. ISBN 0-201-52983-1.

Seznam tabulek

Seznam použitých zkratek

Přílohy