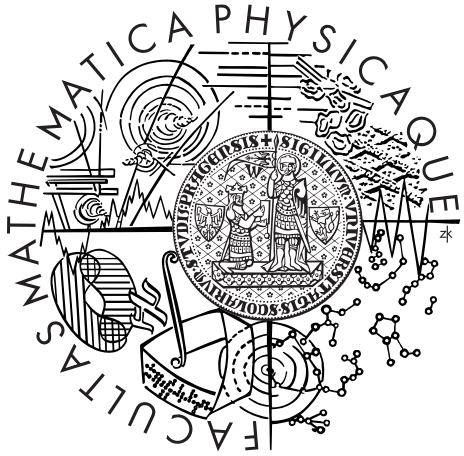
breaklines=true, breakatwhitespace=true

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Bc. Ondřej Odcházel

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 Odcházel

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 Odcházel

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. Instalace a zprovoznění

Celé anotační rozhraní je webová aplikace napsaná v jazyce Ruby a frameworku Ruby on Rails. Je k dispozici pod svobodnou licencí MIT. K jejímu spuštění potřebujete ruby verze alepoň 2.0 (nižší verze nejsou otestované), javu a ke stažení zdrojového kódu git. Program jde spustit na Linuxu a Macu.

1.1 Instalace

Zdrojový kód je volně dostupný na webu GitHubu[1]. Stáhnou tedy lze příkazem

```
git clone https://github.com/hypertornado/diplomka
```

Tento příkaz vytvoří adresář diplomka. Závislosti aplikace nainstalujete pomocí bundleru:

```
bundle install
```

Instalace může vyžadovat přístup administrátora. Dále je potřeba stáhnout knihovnu elasticsearch[2] do adresáře bin/elasticsearch. Stačí verze 1.0 a vyšší. Ve verzi 1.2.1 jsme objevili menší chybu[3], která je způsobena chybou v Javě a jde obejít nastavením delšího hostname počítače.

Nyní je možné celý projekt spustit. Nejprve se spustí elasticsearch databáze pomocí příkazu `rake es:start`, poté je možné spustit samotnou aplikaci příkazem `rails server`. Po spuštění severu je uživatelské rozhraní dostupné ve webovém prohlížeči na adrese `http://localhost:3000`. Po načtení stránky se zobrazí uživatelské rozhraní, ale veškeré AJAXové dotazy skončí chybou. V databázi nejsou importována data.

1.2 Práce s metadaty k obrázkům

Metadata k obrázkům a obrázky samotné jsou poskytovány firmou Profimedia a nejsou volně dostupné. Ke zprovoznění aplikace je nutné vložit CSV soubor `keyword-cleaned-phrase-export.csv` do adresáře data.

1.3 Překlad metadat

Soubor obsahuje metadata k obrázkům v angličtině. Jedním z úkolů této práce je poskytnout doporučování obrázků i v jiných jazycích, primárně v českém jazyce. Bylo tedy nutné metadata přeložit. Pokoušeli jsme se použít volný nástroj na překlad Moses. Ve verzi 2.1[4] nabízí volně dostupné modely pro překlad z češtiny do angličtiny. I na SSD disku trvá několik hodin, než se překladový model načte do paměti. Překlad jednoho segmentu s tímto modelem byl poměrně pomalý (překlad metadat k jednomu obrázku trval zhruba 3 sekundy) a také dosti nepřesný. Například řádek

[1] `https://github.com/hypertornado/diplomka`
[2] `http://www.elasticsearch.org/downloads/1-0-3/`
[3] `https://github.com/elasticsearch/elasticsearch/issues/6611`
[4] `http://www.statmt.org/moses/RELEASE-2.1/models/en-cs/model/`

```
"0000000003","little baby smiling","","child children
   baby babies infants kids childhood single faces
   body naked nake facial expressions smile smiling
   viewing watching laying fun amusing amusement
   amused amuse dallying frolicing playing wantoning
   open"^M
```

byl do čestiny přeložen takto:

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

Je vidět poměrně velké množství nepřeložených slov (konconky `|UNK`) a překlad je relativně nepřesný. Je možné

Nejprve příkazem `rake data:export_profimedia_words_for_translation` vyexportujeme do souboru `data/word_list.txt` seznam všech slov použitých v metadatech k obrázkům. Tento příkaz běží několik hodin i na moderním počítači s SSD diskem. Z Profimedia dat získáme seznam 352862 slov. Tento soubor je nutné přeložit z angličtiny do dalších podporovaných jazyků, v našem případě češtiny.

2. Anotační rozhraní

Vrámci této práce bylo implementováno anotační rozhraní pro vyhodnocování algoritmů, které přiřazují vhodné obrázky k textům. Anotační rozhraní je velmi univerzální. Anotátor má ve webové aplikaci v levém sloupci novinový text a v pravém sloupci galerii obrázků. Jeho úkolem je označit obrázky, které se k danému textu hodí a obrázky které se k textu nehodí. Má také možnost nechat obrázek neoznačený, pokud by se nemohl rozhodnout ani pro jednu variantu.

2.1 Instalace rozhraní

Celé rozhraní je aplikace napsaná v jazyce Ruby a frameworku Ruby on Rails. Aplikace je volně šiřitelná pod licencí MIT. Pro zprovoznění anotační aplikace je potřeba UNIXový systém (Linux, Mac). Zdrojový kód aplikace je uložen na severu GitHub[1] a nejlépe jde stáhnout pomocí gitu. Pro běh serveru je potřeba verze ruby 2.0 a vyšší. Celá aplikace se zprovozní následujícím pořadím BASH příkazů:

```
git clone https://github.com/hypertornado/cemi_anotace
cd cemi_anotace
bundle install #nainstaluje vsechny ruby zavislosti
rake db:migrate #vytvori sqlite databazi s tabulkami
rails server #spusti anotacni server na portu :3000
```

2.2 Přidání uživatelů

Po spuštění serveru je možné přidat anotátory v administračním rozhraní. Přístup je zaheslován HTTP autentifikací. Defaultní uživatelské jméno je cfo a heslo cfo85. Administrátorské přístupové údaje lze změnit v souboru `ROOT_APLIKACE/app/controlle`
Uživatelé mají pouze dvě datové položky, uživatelské jméno (Name) a heslo (Password). Uživatele jde přidávat, mazat a upravovat. Nepředpokládá se, že by anotovaná data byla vysoce citlivá, heslo je proto v databázi uloženo v plaintextu.

2.3 Import anotačních dat

Data pro anotaci lze nahrát pomocí příkazu

```
rake data:import
```

Příkaz očeká existenci souboru `ROOT_APLIKACE/public/annotation_inputs.csv`. Ten musí mít speciální formát, kdy je každý řádek rozdělen mezerami na šest sloupců s následujícími položkami:

INDEX
: unikátní číslo jedné anotace

[1] `https://github.com/hypertornado/cemi_anotace`

LABEL
interní popis pokusu

PRIORITY
priorita, celé číslo >= 0. Určuje prioritu s jakou se má anotace přiřadit. Čím vyšší číslo, tím vyšší priorita.

PREFER_USER
uživatelské jméno preferovaného anotátora. Pokud není žádný anotátor preferován, použije se pomlčka

TEXT_FILE
cesta k textovému souboru s referenčním článkem

IMAGE_FILES
seznam cest k obrázkům. Cesty nemohou obsahovat mezery a jsou oddělené středníkem.

Ukázka importovaných dat:

```
1 basics 0 - ./text/aha/aha-00263.txt.gz
   img/1.jpg;img/2.jpg;img/3.jpg
2 basics 1 - ./text/aha/aha-00006.txt.gz
   img/1.jpg;img/2.jpg
3 basics 0 - ./text/aha/aha-00009.txt.gz  img/2.jpg
```

2.4 Export anotačních dat

Hotové anotace lze exportovat příkazem

```
rake data:export
```

Tento příkaz vypíše na konzolu řádky, které mají tabulátorem oddělené položky:

INDEX
ID anotace. Stejné jako u importovaných dat.

USER
Jméno anotátora, který anotaci vytvořil.

TIME
Čas uložení hotové anotace ve formátu UNIX timestamp.

SKIPPED
Pokud uživatel anotaci přeskočil, je hodnota True, jinak False.

APPROPRIATE
Seznam obrázků které anotátor označil jako vhodné k textu ve formátu relativních cest oddělených středníkem.

NOT_APPROPRIATE
Seznam obrázků které anotátor označil jako nevhodné k textu ve formátu relativních cest oddělených středníkem.

2.5 Import obrázků a textů

Anotační texty a obrázky musí být nahrány do adresáře `ROOT_APLIKACE/public` tak, aby jejich cesty odpovídali cestám v souboru `ROOT_APLIKACE/public/annotation_inputs.csv`. Pokud tedy importovaný soubor obsahuje cestu k obrázku `img/1.jpg`, musí být nahrán odpovídající soubor do `ROOT_APLIKACE/public/img/1.jpg`.

Obrázky musí být ve formátu, který podporují webové prohlížeče, tedy hlavně JPEG a PNG. Texty musí být uloženy v textových souborech s kódováním utf-8 a komprimovaný pomocí gzip[2].

[2] `http://www.gzip.org/`

3. Závěr

Seznam použité literatury

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

Seznam tabulek

Seznam použitých zkratek

Přílohy