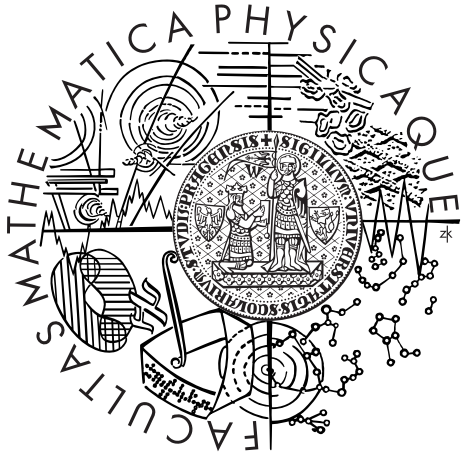
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

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 svému vedoucímu, RNDr. Pavlovi Pecinovi, Ph.D. za pomoc a cenné připomínky v průběhu celého vývoje projektu. Také bych rád poděkoval své rodině a přítelkyni, hlavně za velkou trpělivost.

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. Úvod

Cílem diplomové práce je implementovat kompletní webovou aplikaci pro doporučování a vyhledávání ilustračních obrázků v textu. Vytořit 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 testovaní 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 Translate[1]), 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. Poskytnutá data

Tato kapitola popisuje data, která slouží jako vstup projektu.

2.1 Profimedia dataset

Společnost Profimedia poskytla pro výzkumné účely korpus více než dvaceti milionů ilustračních obrázků spolu s jejich textovým popisem.

Textové popisky byly dále očištěny[4] a poskytnuty pro tento projekt ve formě souboru `profi-text-cleaned.csv`. Soubor je ve formátu CSV a obsahuje 20 014 394 řádků. Každý řádek obsahuje 4 složky:

locator
: Identifikátor obrázku v databázi Profimedia. Desetimístný řetězec číslic.

title
: Název obrázku v anglickém jazyce.

description
: Pro všechny řádky souboru prázdná položka.

keywords
: Mezerami oddělená klíčová slova obrázku.

Ukázka 2.1 Příklad obsahuje příklad jednoho řádku souboru `profi-text-cleaned.csv`.

```
1 "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
```

Ukázka 2.1: Řádek souboru profi-text-cleaned.csv

Na příkladu jsou vidět některé problémy, které data z datasetu Profimedie mají. Některá klíčová slova, jako například „czech republic“ k sobě patří a tvoří frázi. V souboru ovšem tyto víceslovné fráze nejsou vyznačené. Některým slovům chybí diakritika, například „ceska“. Některé fráze vznikly asi strojovým překladem z cizího jazyka do angličtiny. To je vidět na frázi „hradec kings“, která zřejmě původně byla názvem českého města „Hradec Králové“. Všechna slova v souboru obsahují pouze malá písmena, což například znesnadňuje detekci názvů.

Všechny popsané nedostatky dat negativně ovlivňují možnosti pro kvalitní vyhledávání v poskytnutých datech. Jedním z cílů práce je co nejvíce těchto nedostatků opravit.

2.2 Profimedia dataset s detekovanými frázemi

Některé problémy s datasetem Profimedie byly odstraněny v rámci bakalářské práce Bc. Jana Botorka[3]. Jedním z výsledků této práce je soubor `keyword-clean-phrase-expo` Ukázka 2.2 obsahuje příklad jednoho řádku tohoto souboru.

```
1 "0000000980";"Hradec kings holy ghost
    cathedral";"outdoors,nobody,urban
    scenes,architecture,houses,towers,czech,czech
    republic,Česká republika,Europe,
    buildings,build,history,historical,
    churches,church,fronts,cathedrals sv,holy ghost
    cathedral,spirit,Hradec Králové"
```

Ukázka 2.2: Řádek souboru keyword-clean-phrase-export.csv

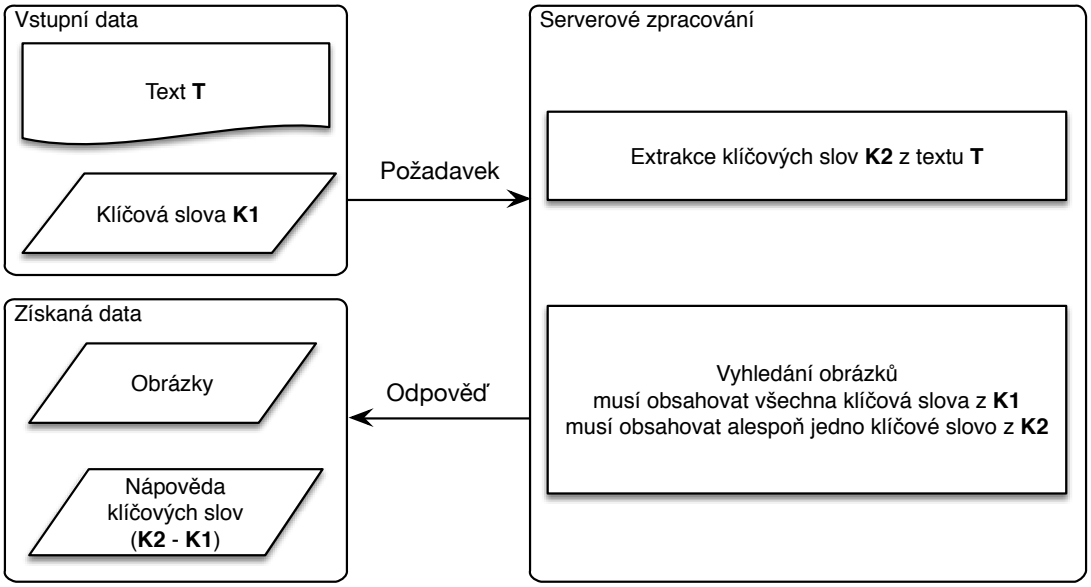
Soubor obsahuje lépe zpracovaná data z datasetu Profimedia. Nejdůležitější změnou pro tuto práci je detekce klíčových frází. Jednotlivé fráze jsou od sebe odděleny čárkou. Detekce probíhala pomocí databáze WordNet[14] a Wikipedie. Ne vždy se detekce povedla, takže v některých řádcích nejsou žádné fráze detekovány. Detekce frází je důležitá zejména pro překlad dat do jiných jazyků.

2.3 Vektor podobnosti obrázků

Kromě textových popisků jsou součástí datasetu Profimedie i samotné obrázky. Jedním z cílů práce je umožnit nad databází obrázků vyhledávání na základě vizuální podobnosti. K tomuto účelu byl poskytnut soubor `profi-neuralnet-20M.data.gz`, který je komprimovaný formátem ZIP a má velikost 129 GB.

Soubor obsahuje pro každý obrázek z datasetu Profimedie vektor 4096 reálných čísel. Pomocí vektorové vzdálenosti lze určit, jak jsou si podobné dva obrázky mezi sebou.

3. Vyhledávání relevantních obrázků



Obrázek 3.1: Základní tok dat mezi klientem a serverem.

Aplikace poskytuje uživateli dvě možnosti vyhledávání. Základní uživatelským scénářem je vložit do rozhraní text. Aplikace by v takovém případě měla poskytnout relevantní obrázky k danému textu. V dalším uživatelským scénářem je přímý požadavek na klíčová slova obrázku. Uživatel by měl mít možnost zadat přímo klíčová slova, které nalezené obrázky musí obsahovat. Oba scénáře by mělo navím být možné propojit pomocí nápovědy klíčových slov – uživatel zadá do rozhraní text, dostane výsledné obrázky a nápovědu klíčových slov, kterými může množinu nalezených obrázků více omezit. Klíčová slova ze vstupního textu se dále používají jako nápověda uživateli, který jimi může množinu hledaných obrázků dále omezit. Celý proces hledání vhodných obrázků popisuje diagram 3.1.

3.1 Extrakce klíčových slov

Extrakce klíčových slov je velmi důležitou složkou celého vyhledávání. Článek [13] shrnuje základní techniky extrakce klíčových slov z textu. Algoritmy na extrakci klíčových slov lze v zásadě rozdělit do dvou kategorií - „s korpusem“ a „bez korpusu“. Metody pracující bez korpusu jsou zajímavé a mohou dosahovat podobných výsledků jako metody s korpusem. My však máme k dispozici dataset Profimedie, takže o metody pracující bez korpusu se tato práce dále nezajímá.

3.2 TF-IDF

TF-IDF je jeden ze základních vyhledávacích algoritmů. Algoritmus využívá korpusu dokumentů D a dvou složek TF a IDF, lze ho vyjádřit jako rovnost:

$$TFIDF(t, d, n, N) = TF(t, d) \times IDF(n, N) \tag{3.1}$$

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

$$TF(t, d) = \sum_{slovo \in d} \begin{matrix} 1 & \text{pokud } slovo = t \\ 0 & \text{jinak} \end{matrix} \tag{3.2}$$

Jedná se tedy o frekvenci slova 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

$$IDF(n, N) = \log\left(\frac{N}{n}\right) \tag{3.3}$$

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

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

Výsledný vzorec pak jde shrnout jako:

$$TFIDF(t, d, n, N) = \left(\sum_{slovo \in d} \begin{matrix} 1 & \text{pokud } slovo = t \\ 0 & \text{jinak} \end{matrix} \right) \times \log\left(\frac{N - n}{n}\right) \tag{3.5}$$

3.3 TF-IDF pro extrakci klíčových slov

Algortimus TF-IDF můžeme použít pro extrakci klíčových slov z textu. Všechna znaky vstupního textu převedeme na malá písmena a text rozdělíme na slova. Dále nás nezajímá diakritika a různé speciální znaky, které může text obsahovat. Všechna slova převedeme na stemy. O převodu slov na stemmy se podrobněji píše v Kapitole 5.

Algoritmus TF-IDF pak použijeme na každé slovo vstupního textu a získáme tak skóre jeho významnosti v textu. Slova s nejvyšším skóre pak označíme za klíčová. Algoritmus musíme upravit pro naše účely. Zaprvé nás zajímají pouze taková slova, která existují v datasetu Profimedie. Slova která se v korpusu nenachází dostanou skóre 0. Část TF v algoritmu znamená četnost slova v textu.

Složitější je situace s IDF. Nabízí se použít frekvenci slov z datasetu Profimedie. Ukázalo se, že klíčová slova obrázků z datasetu nejsou pro tento účel vhodná. Klíčová slova totiž obsahují mnoho názvů a obecně méně běžných slov. Naopak obsahují velmi málo běžných slov. To pak způsobuje, že algoritmus nad

tímto datasetem dává vysoké skóre běžným slovům. Tato slova ale typicky nejsou vhodnými klíčovými slovy daného textu. Je tedy potřeba použít pro vzorec DF jiný korpus. V naší aplikaci jsme použili korpus Wikipedie, jejíž data jsou volně dostupná pro mnoho jazyků pod otevřenou licencí.

Pokud je w slovo vstupního textu, $Freq(w)$ je četnost slova ve vstupním textu a $Wiki(w)$ je četnost slova v korpusu Wikipedie, můžeme každému slovu vstupního slova přiřadit $Score$:

$$Score(w) = \begin{cases} Freq(w) \times (\frac{C}{Wiki(w)}) & w \text{ je v datasetu Profimedie} \\ 0 & w \text{ není v datasetu Profimedie} \end{cases} \tag{3.6}$$

C je experimentálně zjištěná konstanta. V našem případě je její hodnota 10000000.

Nyní tedy máme skóre udávající význam slova pro každé slovo vstupního textu. Slova s nejvyšším skóre vrátíme uživateli jako nápovědu pro explicitní požadavek obrázků s klíčovými slovy.

3.4 Vyhledávání obrázků

Samotné vyhledávání obrázků má dva druhy vstupních dat. Prvním typem je text článku, který uživatel vloží do uživatelského rozhraní. Máme tedy k dispozici řetězec s textem článku. Dále může uživatel zadat explicitní klíčová slova, které má hledaný obrázek obsahovat. Tato vstupní data získáváme jako pole řetězců. Úkolem algoritmu na vyhledávání obrázků je vrátit uživateli všechny relevantní obrázky v pořadí podle relevance.

Text si nejprve zpracujeme pomocí algoritmu uvedeném v sekci 3.3. Získáme skóre významnosti pro všechna slova ve vstupním textu. Pro vyhledávání použijeme pouze několik slov s nejvyšším skóre.

Nyní můžeme jako množinu relevantních obrázků označit obrázky, které ve svých klíčových slovech mají všechna uživatelem zadaná explicitní klíčová slova a alespoň jedno klíčové slovo získané extrakcí klíčových slov z textu.

Množina relevantních obrázků může být velká a obsahovat obrázky, které jsou relevantní jen velmi málo. Je proto důležité množinu relevantních obrázků správně seřadit. Máme množinu K_{text} klíčových slov získaných z textu. Každé $w \in K_{text}$ má skóre $Score(w)$. Relevantní obrázek má množinu klíčových slov K_{img}. Relevanci obrázku vůči uživatelskému dotazu můžeme ohodnotit funkcí $Rank$:

$$Rank(K_{text}, K_{img}) = \sum_{w \in K_{text}} \begin{cases} Score(w) \times \frac{1}{|K_{img}|} & w \in K_{img} \\ 0 & w \notin K_{img} \end{cases} \tag{3.7}$$

Vzorec bere v úvahu počet klíčových slov, které obrázek obsahuje. Snižuje relevanci obrázků, které obsahují mnoho klíčových slov a zvýhodňuje tím ve vyhledávání obrázky, které mají menší počet přesnějších klíčových slov.

4. Překlad obrázkových popisků

Popisky obrázků v datasetu Profimedie jsou v angličtině. Jedním z cílů aplikace je poskytnou kromě angličtiny vyhledávání i v jiném jazyce. Zaměřujeme se na češtinu, ale podobné úvahy a postupy platí většinou i pro jiné jazyky. Jsou v podstatě dvě možnosti, jak implementovat vyhledávání v češtině. Můžeme buď překládat do angličtiny hledané texty a klíčová slova, nebo předpřeložit dataset Profimedie.

My jsme se rozhodli pro druhou možnost. Musíme tedy přeložit všechny texty v datasetu Profimedie. Vzhledem k tomu, že obrázků je více než 20 milionů, nepřipadá lidský překlad v úvahu z časových i finančních důvodů. Jedinou reálnou možností je použít překlad strojový.

4.1 Strojový překlad

Strojový překlad jako obor zažívá velký rozvoj. Potřeba překladu stále celosvětově prudce stoupá. To jednak zvyšuje poptávku po kvalitním strojovém překladu a druhak to strojovému překladu dává velké množství lidských překladů, které jsou pro kvalitní strojový překlad k dispozici. Lze brát v úvahu i pravidlové překladové systémy, které ke své práci tolik dat většinou nepotřebují. Obecnější a pro většinu jazykových párů nejlepší řešení však nabízí frázový statistický strojový překlad.

S tím se zvyšuje poptávka po lepším strojovém překladu a zároveň také vzniká více lidských překladů, které pak jde použít jako zdrojová data k překladům strojovým. Existují různé metody strojového překladu. Na jednu stranu existují pravidlové systémy, které podle pravidel zapsaných překladately převádí text ze zdrojového do cílového jazyka. Na druhou stranu jedním z nejobecnějších řešení je frázový překlad.

4.2 Frázový statistický strojový překlad

Frázový statistický strojový překlad[11] ke své práci potřebuje databázi lidsky přeložených frází. Fráze jsou několikaslovné kusy přeloženého textu. Typicky se získávají z paralelního korpusu textů ve zdrojovém a cílovém jazyce. Soubor s frázovými daty obsahuje položku s textem fráze ve zdrojovém a cílovém jazyce spolu s pravděpodobností, že je takový překlad fráze správný.

Druhým důležitým vstupem strojového frázového systému je korpus cílového jazyka, ze kterého se vytvoří jazykový model. Slouží zejména k tomu, aby k sobě poskládané fráze v cílovém jazyce dobře „seděli“. Pokud máme překladový i jazykový model, můžeme vyjádřit pravděpodobnost překladu pomocí základního vzorce statistického strojového překladu. Překládáme řetězec f ve zdrojovém jazyce. K dispozici máme překladový model $p(f|e)$, který udává pravděpodobnost toho, že řetězec f ve zdrojovém jazyce je překladem řetězce e v cílovém jazyce. Jazykový model $p(e)$ nám vrací pravděpodobnost řetězce e v cílovém jazyce. Chceme získat takový řetězec $\tilde{e}$, pro který je pravděpodobnost $p(e|f)$ nevyšší. Pomocí Bayesova pravidla můžeme k hledání takové fráze použít překladový a jazykový model:

$$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e) \tag{4.1}$$

V praxi je potřeba u kvalitního strojového překladu vyřešit mnoho dalších problémů. K lepším výsledkům překladu potřebujeme reordering model, který umožňuje přesouvat pozice frází mezi zdrojovým a cílovým textem.

Dobrý frázový překlad potřebuje ke svému chodu miliony frází. Vyhledávání nejpravděpodobnějšího překladu v takovém množství dat je náročná úloha. Algoritmy, které umožňují vyhledávat ve velkém množství překladových dat jsou implementována v knihovně Moses[12], která je dostupná pod volnou licencí včetně základních překladových modelů pro některé jazykové páry.

4.3 Charakteristika dat pro překlad

Pro správné fungování vyhledávání v českém jazyce je potřeba přeložit klíčová slova v Profimedia datech do češtiny. Slova v korpusu jsou oddělena mezerou. Je ale zřejmé, že některá slova vedle sebe k sobě patří — jsou to fráze — zatímco některá nikoliv. Naskytují se tedy zhruba tři možnosti, jak takový text přeložit.

4.4 Překlad vět

Zdrojový dokument

"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 žena čištění teeth","","single čelí mladí lidé , lidé mládež rukou uvnitř interiors žena žen , žen , blonďák 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 péče let čištění

Překladač Google

"0000000102", "Mladá žena čištění zubů", "", "jednotlivé tváře lidí, lidé younge mládeží ruce interiéry ženě 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

Obrázek 4.1: Ukázka překladu metadat k obrázku pomocí Mosese a Překladače Google.

První možností je přistupovat k souboru klíčových slov u každého obrázku jako k větě a použít frázový strojový překlad — buď Moses, nebo Překladač Google[9] — k překladu z angličtiny do češtiny. Tento přístup má několik problémů. Frázový překlad se snaží aplikovat fráze z překladového modelu. V našem souboru klíčových slov ale mohou být vedle sebe slova, která tvoří frázi pouze zdánlivě. Například můžeme mít fotku dítěte, které stojí před automobilem značky Seat se dvěma klíčovými slovy vedle sebe — „child“ a „seat“. Frázový překlad

z angličtiny do češtiny pochopí tato dvě slova jako fráze, které do češtiny přeloží frází „dětské sedadlo“, která ovšem neodpovídá popisku obrázku.

Dalším problémem tohoto přístupu je časová a finanční náročnost takového řešení. Frázový překlad je dosti náročný algoritmus a překlad dvaceti milionů vět může být dosti obtížný. Překlad pomocí Mosese nás omezuje výpočetní složitostí. Na jednom stroji by takový překlad trval řádově několik dní. Pokud bychom k překladu 20 milionů vět použili Překladač Google, jsme zase omezení cenou za přístup k překladovému API společnosti Google.

Pro práci s Mosesem jsme zvolili překladový a jazykový model pro překlad z angličtiny do češtiny, které jsou poskytnuty přímo s Mosesem ve verzi XX. Obrázek 4.1 porovnává překlad metadat k jednomu z obrázků datasetu Profimedie v Mosesovi s distribuovanými překladovými daty s překladem pomocí Překladače Google. Je vidět, že Překladač Google poskytuje kvalitnější překlad.

4.5 Překlad slov

Jednodušším přístupem k překladu klíčových slov je přístup slovníkový, teda překlad každého slova zvlášť. Nejprve je potřeba ze souboru klíčových slov u všech obrázků vyextrahovat slova. Ty pak lze přeložit přímo s použitím slovníku, nebo pomocí frázového strojového překlad (ten použije jednoslovné fráze) Mosesem, či Překladačem Google. Výhodou oproti předchozímu přístupu je menší množství dat a tedy i nižší finanční a časová náročnost takového překladu. Takový systém překladu ale nedokáže detekovat fráze a kvalita překladu slovo od slova je typicky horší než překlad delších frází. Vezměmě si například anglická slova „weather“ a „vane“. Slovníkový překlad nám slova přeloží jako „počasí“ a „lopatka“. Pokud se ale tato slova nachází vedle sebe, je pravděpodobnějším překladem slovo „korouhvička“.

4.6 Překlad frází

Poslední možností je oba předchozí principy zkombinovat — nejprve detekovat v souboru klíčových slov fráze a ty pak přeložit statistickým strojovým překladem. Detekci frází z datasetu Profimedia provedl ve své bakalářské práci[3] Jan Botorek. Zkoušel detekovat N-gramy v databázi WordNet[14] a Wikipedii. Výsledky této detekce frází lze použít právě ke zlepšení překladu z angličtiny do cizích jazyků. Na slova která nejsou detekována ve frázi se použije slovníková metoda. Na překlad detekovaných frází lze použít přímo statistický strojový překlad.

4.7 Řešení

Řešením bylo nakonec použít poslední možnost. Bližší informace jsou v Kapitole [**?**]. Slova byly přeloženy pomocí Překladače Google, který překládá mnohem lépe než dostupný model pro překladový nástroj Moses.

Překlad frází by ovšem pomocí Překladače Google byl příliš drahý a pomocí Mosese zase příliš pomalý. Použili jsme tedy jednoduchou metodu, která používá pouze překladový model. Fráze z datasetu Profimedia byla přeložena pouze

tehdy, když se její překlad nacházel v překladovém modelu. Fráze, které se v překladovém modelu nenacházely, byly přeloženy slovo od slova. Metodou přesné shody s překladovým modelem bylo přeloženo 18006 z celkového počtu 899244 detekovaných frází.

4.8 Závěr

Překlad klíčových slov z korpusu Profimedia není typickou překladovou úlohou — nepřekládají se celé věty. Přesto je dokonalý výsledek nemožný. I lidští překladatelé s citem pro jazyk by v této překladové úloze dávali rozdílné výsledky. Strojový překlad zdaleka není na takové úrovni, aby dokázal z širšího kontextu vybrat správný překlad. Navíc lidský překladatel může u překladu klíčových slov využít přímo obrázek, ke kterému se klíčová slova vztahují. Může tak z obrázku posoudit, jestli má slovo „single“ přeložit jako „jednolůžkový“, nebo ve významu „jeden“. Lepší překlad by mohl přinést překladový model natrénovaný na specifičtější množině dat bližší korpusu Profimedie. Vytvořit takový model by ovšem bylo nad rámec této práce.

Navrhovaný mechanismus překladu poskytuje uspokojivý, i když značně nedokonalý, překlad klíčových z angličtiny do češtiny. Jednoduše jde zevšeobecnit i pro překlad do dalších jazyků pro které máme slovník a překladový model.

5. Stemmer

Důležitou vlastností systémů na vyhledávání v textu je, aby dokázali najít i v jiných tvarech slova. Pokud uživatel hledá slovo „praha“, většinou očekává, že se mu zobrazí i výsledky obsahující slovo „praze“. Je tedy potřeba mít nějaký algoritmus, který k sobě slova jako „praha“ a „praze“.

První možností je použít lemmatizér. Úkolem lemmatizéru je ke každému slovu přiřadit jeho základní tvar. U podstatných jmen je to většinou první pád jednotného čísla („praha“), u sloves infinitiv.

Alternativou lemmatizátoru může být stemmer. Ten většinou používá jednoduché heuristiky k odstranění koncovek slov. Stemmer nemusí vrátit regulérní slovo jazyka, například pro slova „praha“ a „praze“ může stemmer vrátit slovo „prah“. Výhodou stemmeru oproti lemmatizéru je, že většinou používá pouze jednoduché heuristiky. Je tedy většinou rychlejší a méně náročný na zdroje než lemmatizér. Ve vyhledávacích aplikacích je stemmer dostačující.

Pro angličtinu je nejznámějším stemmerem pro angličtinu je Porterův stemmer popsaný[?] Martinem Porterem již v roce 1980. Kromě oficiální implementace existují porty do různých jazyků včetně Ruby. Tato aplikace využívá implementaci z Ruby Gemu „stemmify“[1] v licenci MIT.

5.1 Český stemmer

Pro jazyky jako je čeština, která má bohatší morfologii než angličtina, je tvorba stemmeru náročnější. Testovali jsme několik implementací českých stemmerů. Jako nejkvalitnější byla nakonec vybrána implementace českého stemmeru v knihovně Lucene. Tuto implementaci využívá i Elasticsearch.

V rámci této práce byl portován soubor `CzechStemmer.java` ze zdrojového kódu knihovny Lucene do jazyka Ruby. Výsledkem je Ruby Gem pod licencí MIT, který lze použít nezávisle na zbytku aplikace. Instaluje se příkazem

```
1 gem install czech-stemmer
```

Knihovna obsahuje pouze jednu třídu CzechStemmer s funkcí stem, která přijímá i vrací řetězec:

```
1 require 'czech-stemmer'
2
3 CzechStemmer.stem("praha") # => "prah"
4 CzechStemmer.stem("praze") # => "prah"
5 CzechStemmer.stem("předseda") # => "předsd"
6 CzechStemmer.stem("mladými") # => "mlad"
```

[1] `https://rubygems.org/gems/stemmify`

6. Detekce jazyka

Aplikace může přijímat vstup ve více jazycích. Jedním ze způsobů, jak uživateli zpříjemnit práci s aplikací, je použít automatický detektor jazyka. Při zadání vstupu aplikace se aplikace sama pokusí detekovat jazyk zadaného textu. Tato detekce nemusí být vždy stoprocentně správná, takže by uživatel měl mít možnost volbu jazyka manuálně změnit.

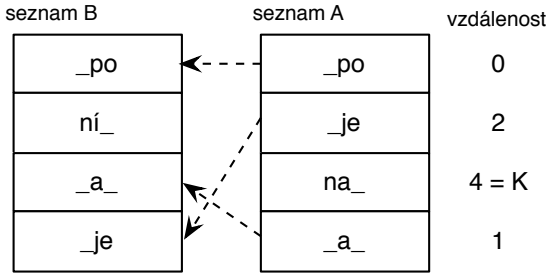
Aplikace používá algoritmus založený na statistice nejčastějších N-gramů pro daný jazyk[5]. Algoritmus nejprve použije velký korpus textů pro každý detekovaný jazyk. Pro češtinu a angličtinu lze použít například korpus Wikipedie. Z tohoto korpusu získá uspořádaný seznam K nejčastějších N-gramů. Při samotné detekci jazyka textu, pak samotný algoritmus získá stejný uspořádaný seznam nejčastějších N-gramů pro vstupní text. Tento seznam pak porovnává se seznamy nejčastějších N-gramů pro každý jazyk.

Nechť A a B jsou seznamy N-gramů s K položkami, $A[w]$ je pořadí N-gramu w v seznamu A. Potom lze vzdálenost mezi seznamy $D(A, B)$ vyjádřit vztahem:

$$D(A,B) = \sum_{w \in A} \begin{cases} |A[w] - B[w]| & \text{pokud } w \in B \\ K & \text{jinak} \end{cases} \tag{6.1}$$

Na Obrázku 6.1 je ukázka porovnání dvou seznamů pomocí funkce $D(A, B)$. Trigram *na_* se v seznamu B nenachází, proto dostane hodnotu $K = 4$.

Nyní pokud máme množinu S všech seznamů N-gramů pro jednotlivé jazyky a X je seznam N-gramů pro vstupní text, vrátí algoritmus jako jazyk takový jazyk, pro který má seznam $C \in S$ minimální hodnotu $D(X, C)$.



Obrázek 6.1: Ukázka porovnání dvou trigramových seznamů. $D(A, B) = 7$

Naše implementace používá v seznamech N-gramů trigramy. Pro texty delší než několik slov funguje velmi spolehlivě. Pro několikaslovné texty se může stát, že v seznamu nejfrekventovanějších trigramů pro vstupní text není ani jeden trigram, který by se nacházel v seznamech pro jednotlivé jazyky. Pak algoritmus může vrátit špatně detekovaný jazyk.

Vylepšení by jistě přineslo spolu s použitím trigramů použít i bigramy a unigramy. Také konstanta K by šla zvětšit z používané hodnoty 50 výš. Implementace detekce jazyka ale probíhá na klientovi, takže všechna tato vylepšení algoritmu by zvýšila množství dat, která si klient musí stáhnout. Navíc uživatel má vždy možnost detekovaný jazyk manuálně změnit a typicky pracuje spíše s delšími tex-

ty. Implementace s použitím 50 nejčastějších trigramů se tedy zdá dostačující pro daný účel.

7. Podobné obrázky

Jednou ze služeb, které výsledná aplikace poskytuje, je vyhledávání podobných obrázků. Uživatel rozhraní najde pomocí textových dotazů nějaké ilustrační obrázky a má možnost u každého z nalezených obrázků získat obrázky vizuálně podobné. Tato kapitola pojednává o tvorbě backendové služby, která vyhledávání podobných obrázků umožňuje.

Vstupními daty je soubor s vektory pro každý obrázek datasetu Profimedie. Vektor má 4096 složek s reálnými nezápornými čísly. Vektory jsou vizuální deskriptory obrázků. Tyto deskriptory byly vygenerovány pomocí software Caffe[**?**] a jsou jsou odezvami předposlední vrstvy hluboké konvoluční neuronové sítě natrénované pro klasifikaci obrázků z datasetu Profimedie do 1000 kategorii.

Mějme obrázek I_1 s deskriptorem D_1 a obrázek I_2 s deskriptorem D_2. Míru podobnosti obrázků *Similarity* pak můžeme definovat jako

$$Similarity(I_1, I_2) = \sum_{i=1}^{4096} |D_1[i] - D_2[i]| \tag{7.1}$$

V praxi se dají výsledky této míry klasifikovat zhruba do 3 kategorií. Tyto vypozorované kategorie popisuje tabulka 7. Ukázky jednotlivých kategorií podobných obrázků poskytuje obrázek 7.2. Rozdělení na 3 kategorie podobnosti podle míry *Similarity* je pouze přibližné. Nelze například zaručit, že obrázek, který by některý uživatel mohl uznačit za podobný, nebude mít míru *Similarity* vyšší než 1500.

Kategorie	*Similarity*
téměř shodné	$0 - 500$
podobné	$500 - 1500$
nepodobné	> 1500

Obrázek 7.1: Kategorie podobnosti obrázků podle *Similarity*.

Otázkou zůstává, které výsledky uživatel očekává jako výsledky vyhledávání podobných obrázků. Pravděpodobně nechce získat obrázky z kategorie „nepodobné“. Pak je otázkou, jestli uživatel chce jako výsledek získat obrázky z kategorie „téměř shodné“. V korpusu je spousta podmnožin obrázků, které byly vyfoceny v rámci jedné série. Často se liší jen malou změnou úhlu fotky, nebo jen kompresí uloženého obrázku. Pokud bychom vraceli obrázky seřezené podle *Similarity*, uživatelé u takovýchto podmnožit nedostanou příliš rozmanité obrázky. Vracet obrázky seřazené podle *Similarity* tedy nemusí být vždy vhodné, náhodné pořadí obrázků z kategorií „téměř shodné“ a „podobné“ může uživatel chápat jako lepší výsledek. Tato úvaha umožňuje použít algoritmy, které nevrací obrázky seřazené podle *Similarity*, ale jsou výrazně rychlejší.

7.1 Bootstrap implementace

První implementace nahrála všechny deskriptory do databáze Elasticsearch. V Elasticsearch lze implementovat vyhledávání pomocí vlastní porovnávací funkce.

Obrázek 7.2: Dvojice obrázků s různou kategorií podobnosti.

Jako porovnávací funkce tedy zvolíme funkci *Similarity*. Bohužel přes některé menší optimalizace algoritmu se vyhledávání nepodařilo implementovat příliš efektivně, takže fungovalo v přijatelném čase pouze pro pár tisíc deskriptorů. Naše aplikace ovšem potřebuje pracovat s více než dvaceti miliony deskriptorů. Šly by použít další optimalizace a sharding databáze na více strojů, aby algoritmus fungoval efektivněji i na větším množství dat. To by ovšem bylo příliš drahé a navíc se objevily jiné možnosti řešení.

7.2 Předgenerování výsledků

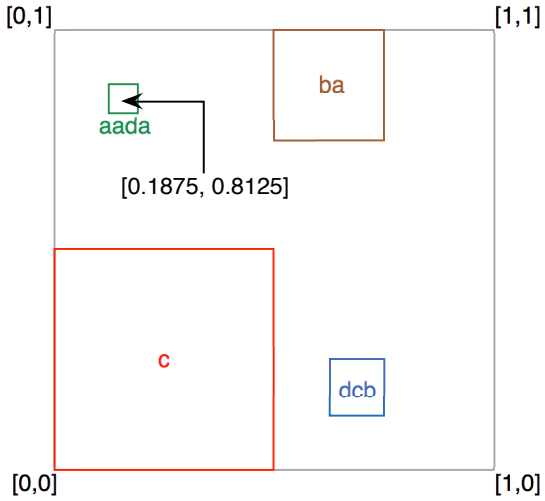
Efektivnější variantou je ke každému obrázku vygenerovat nějaké množství podobných obrázků předem. Služba která by vracela podobné obrázky by pouze vrátila položku z databáze a neprováděla žádný výpočet. Tato služba by tak byla velmi rychlá a nenáročná na zdroje. Jako problém se ovšem ukázalo právě předgenerování obrázků. Přes snahu o co nejrychlejší implementaci v jazyce go s použitím více vláken a optimalizačních heuristik, by vygenerování podobných obrázků pro každý z 20 milionů obrázků v datasetu Profimedie trvalo na běžném počítači několik týdnů.

7.3 Geohash

Další pokus o implementaci vyhledávání podobných obrázků se inspiruje algoritmem Geohash[17]. Algoritmus Geohash byl vyvinut v rámci služby geohash.org a jedná se o způsob zakódování prostorových dat. Jeho hlavním využitím je efektivní vyhledávání bodů (určených zeměpisnými souřadnicemi) v oblasti (na mapě). Algoritmus Geohash využívá i Google, nebo databáze Elasticsearch.

Zjednodušení algoritmu Geohash popíšeme na jednotkové podmnožině (jednotkovém čtverci) $\mathbb{R}^2$. Algoritmus postupuje tak, že čtvercovou plochu rozdělí na 4 čtverce a pojmenuje je písmeny „a“, „b“, „c“ a „d“. Každý ze čtverců rekurzivně rozdělí na další čtyři čtverce, kterým dá jméno pomocí suffixů „a“, „b“, „c“ a „d“ k vlastnímu jména. Rekurzi provádíme do nějaké předem určené hloubky. Bodu v jednotkovém čtverci přiřadíme jméno podle čtverce s nejdelším jménem, který bod obsahuje. Každý bod v jednotkovém čtverci tedy bude mít jméno, které má stejnou délku jako hloubka rekurze. Jako oblast pak můžeme označit jakoukoliv množinu pojmenovaných čtverců. Bod leží v oblasti právě tehdy, když je jméno nějaké ho čtverce z množiny oblasti prefixem jména bodu. K ukládání názvů bodů pak lze použít prefixový strom.

Obrázek 7.3 ukazuje některé čtverce a jejich názvy. Pokud má algoritmus hloubku rekurze 4, má bod ležící na souřadnicích $[0.1875, 0.8125]$ název *aada* a leží tedy ve čtvercích *a*, *aa*, *aad* a *aada*.

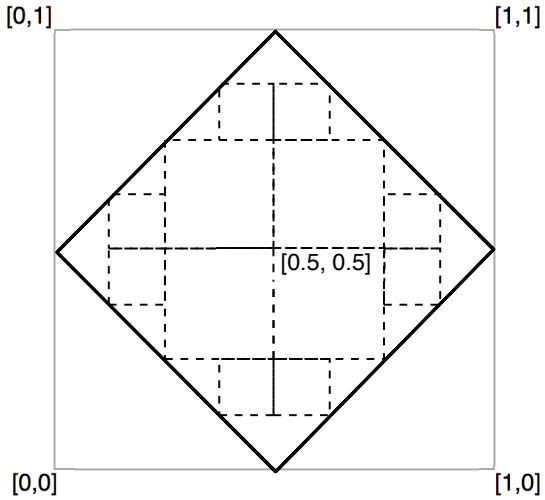


Obrázek 7.3: Ukázka čtverců algoritmu Geohash v jednotkovém čtverci.

Nyní bychom chtěli algoritmus Geohash použít pro hledání podobných obrázků v Profimedia datasetu. Každému deskriptoru bychom nejprve přiřadily název. Množina podobných obrázků by pak obsahovala obrázky, které mají *Similarity* s porovnávaným obrázkem nějak omezenou. Problémem je, že pomocí Geohash čtverců nedokážeme takovou oblat přesně definovat. Můžeme se ale pokusit co nejvíce oblast pomocí čtverců aproximovat.

Pokud bychom chtěli použít Geohash pro hledání podobných vektorů přímo, narazili bychom na několik problémů. Například oblast s ohraničenou mírou *Simiarity* nejde popsat přesně pomocí čtverců. Lze libovolně přesně aproximovat, ale s každou přesnější aproximací vzrůstá počet potřebných čtverců a zhoršuje se

efektivita algoritmu. Obrázek 7.4 ukazuje jak bychom mohli použít Geohash algoritmus pro hledání podobných deskriptorů ve dvojrozměrné dimenzi. Na obrázku je popsána situace, kdy hledáme deskriptory, které mají od bodu $[0.5, 0.5]$ vzdálenost 0.5. Tučně ohraničený čtverec vyznačuje hledanou oblast, 12 čárkovaných čtverců tvoří aproximovanou oblast.



Obrázek 7.4: Ukázka čtverců algoritmu Geohash v jednotkovém čtverci.

Ve dvou dimenzích by toto řešení fungovalo poměrně dobře. Deskriptory obrázků ovšem mají 4096 složek a počet potřebných čtverců pro stejnou míru aproximace roste s každou dimenzí exponenciálně.

7.4 Řešení

Naše řešení využívá princip algoritmu Geohash. Jde nám o to, převést hledání podobných vektorů na fulltextové vyhledávání. Vektory reálných čísel převedeme na množinu slov oddělenou mezerami — řetězec. Slova pochází z množiny $\{0, 1, \ldots 4095\}$. Každé složce vektoru deskriptoru tedy odpovídá právě jedno slovo z množiny slov. Nyní přiřadíme každému deskriptoru nějakou nějakou podmnožinu slov. Způsob přiřazení je klíčovým faktorem algoritmu, který ovlivňuje jeho efektivitu. Základním způsobem je přiřadit vektoru slova, která odpovídají nenulovým vektorům. Na deskriptorech obrázků z datasetu Profimedie to znamená, že každý vektor bude mít přiřazeno zhruba 1500 slov. To je stále příliš mnoho slov pro efektivní fulltextové vyhledávání. Elasticsearch v základní konfiguraci ani neumožňuje vyhledávat s dotazy delšími než tisíc slov. Musíme tedy množinu slov dále zmenšit. Můžeme například místo nenulovosti zvýšit požadovanou mez velikosti složky vektoru.

V našem řešení používáme jinou metodu. Seřadíme složky vektorů podle velikosti a slova přiřadíme pouze K složkám s nejvyšší hodnotou. Pro $K = 4$ pak každému obrázku přiřadíme řetězec, například "2013 432 1065 3433". Při vyhledávání podobného obrázku nejprve získáme jeho deskriptor a jeho přiřazený řetězec. Pomocí řetězce najdeme fulltextovým vyhledáváním L obrázků. Výsledky seřadíme podle podobnosti s hledaným obrázkem mírou *Similarity* a vrátíme

obrázky, které jsou v kategorii „podobné“, nebo „téměr shodné“. Pro správné fungování je nutné dobře nastavit hodnoty K a L. Jejich zvýšení vede k větší přesnosti na úkor rychlosti algoritmu.

7.5 Implementace

Celá služba je implementována nezávisle na zbytku projektu. Jedná se o program `similar_img_finder` napsaný v jazyce go. Jako databáze je použita Elasticsearch. Nejprve je nutné naimportovat data příkazem

```
1  similar_img_finder import -n 1000000 -e 9200 -k 500
     -f data.gz
```

Parametr `-n` určuje, kolik dat se má naimportovat, parametr `-e` určuje na kterém portu běží Elasticsearch, parametr `-k` odpovídá hodnotě K a parametr `-f` je cesta k souboru s importovanými daty.

Nyní můžeme spustit server na portu 8585 příkazem

```
1  similar_img_finder server -p 8585 -l 100 -e 9200
```

Parametr `-l` odpovídá hodnotě L.

Pokud chceme nyní získat id podobných obrázků k obrázku s id „0000000003“, můžeme využít JSON HTTP API a získat výsledky na adrese `localhost:8585/similar?id=00`

8. Backend

Moderní webové aplikace lze rozdělit na back a frontend. Backend je část aplikace, která běží na serveru. Pomocí svých služeb poskytuje přístup k databázi, k souborům uloženým na serveru a zpracovává uživatelské operace. Frontend těchto služeb využívá. Tato kapitola popisuje proces výběru backendových technologií pro tuto práci. Některé technologie se osvědčily, jiné se ukázaly pro daný účel nevhodné.

8.1 Databáze

Úkolem databáze je uložit data a umožnit jejich prohledávání. Důležitou vlastností naší aplikace je, že uživatelé nemají možnost databázi modifikovat. Zápis do databáze provede administrátor pouze jednou, před startem aplikace. Důležitým požadavkem je důraz na rychlost a snadnou škálovatelnost. V posledních letech vzniklo několik nových databází v kategorii vágně označené jako NoSQL[16]. Tato kategorie databází se těžko popisuje, na každou popsanou vlastnost existuje NoSQL databáze, která danou vlastnost nesplňuje. Obecně ale lze říct že NoSQL databáze nepracují s prvky v tabulkovém uspořádání a oproti standardním SQL databázím nekladou tolik omezujících požadavků na data. Jejich výhodou oproti standardním relačním databázím může být vyšší výkon a snadná škálovatelnost.

Nevýhodou je většinou obtížnější práce s daty. Většina NoSQL databází například mapodporuje databázové transakce a vůbec celý ACID. Pro práci s daty v NoSQL databázi nelze použít klasické SQL dotazy. Místo nich se používá například model MapReduce[8] vyvinutý ve formě Google.

První verze aplikace byla postavena na databázi CouchBase[7], což je právě jedna z NoSQL databází podporující MapReduce mechanismus. Výhodou CouchBase je vysoký výkon, snadná škalovatelnost, velmi dobrá dokumentace a také existence oficiálních knihoven pro nejrozšířenější programovací jazyky. Ukázalo se však, že pro implementaci vyhledávání pro naší aplikaci je model MapReduce nedostatečný a implementace vyhledávání v CouchBase by byla prakticky nemožná.

Vhodnější pro daný účel se ukázala knihovna Elasticsearch[2]. Nejedná se v pravém smyslu o databázi. Jejím hlavním cílem je poskytnout vyhledávání nad daty. Je postavená nad knihovnou Apache Lucene ke které přidává snadnou horizontální i vertikální škálovatelnost a komunikaci pomocí REST HTTP JSON API. Díky tomu, že je Elasticsearch postaven na knihovně Lucene[1], může programátor využít velkou množinu možností, které Lucene poskytuje. Například v textovém vyhledávání může využít všechny tokenizery a stemmery z knihovny Lucene. V průběhu implementace aplikace navíc vyšla stabilní verze 1.0.

8.2 Programovací jazyk

Se vzrůstající popularitou webů vzniká stále větší množství webových frameworků a dokonce programovacích jazyků zaměřených primárně na programování pro web.

8.2.1 NodeJS

Jedním z nových trendů je tvorba webového backendu v Javascriptu pomocí knihovny NodeJS. Frontendoví vývojáři jsou prakticky nucení Javascript používat. Pokud se v Javascriptu tvoří i backendová část aplikace, může snadněji dojít ke sdílení kódu i pracovních pozic. Trend psaní všech aplikací v Javascriptu glosoval již před sedmi lety Jeff Atwood ve svém pravidle:

```
1 Any application that can be written in JavaScript,
     will eventually be written in JavaScript.
2 Vše co může být napsáno v Javascriptu, bude v
     Javascriptu napsáno.
```

Nevýhody použití Javascriptu jako backendového programovacího jazyka jsou poměrně zřejmé. Javascript byl navržen pro programování webového frontendu, jeho standartní knihovna je v porovnání s ostatními jazyky velmi chudá, podpora objektového programování je celkem nepřímočará. Obsáhlý kód v Javascriptu může být poměrně nepřehledný a jazyk svádí k vytvoření takzvaného „callback hell“, který může mít strukturu jako na obrázku (source http://strongloop.com/strongblog/node-js-callback-hell-promises-generators/). Pokud chce navíc programátor sdílet kód z backendu i na frontendu, musí být kód kompatibilní s podporovanými prohlížeči. Funkce pro jednodušší práci s polem podporuje webový prohlížeč Internet Exporer až od verze 9 [http://kangax.github.io/compat-table/es5/]. V součtu nám převažily nevýhody NodeJS frameworku na výhodami a myšlenku vývoje backendu v Javascriptu jsme opustily.

```
1 doAsync1(function () {
2   doAsync2(function () {
3     doAsync3(function () {
4       doAsync4(function () {
5     })
6   })
7 })
```

8.2.2 Go

Go (známý také jako golang)[**?**] je programovací jazyk od společnosti Google. První stabilní verze byla zveřejněna v Květnu 2012. Go je kompilovaný, staticky typovaný jazyk s garbage collectorem. Syntaxe je inspirována jazykem C a přizpůsobena pro rychlou kompilaci. Velkou výhodou je snadná práce s vlákny pomocí „go rutin“. Programátoři v Javě, nebo C++ může překvapit poněkud netypická podpora práce s objekty.

První verze aplikace byla napsána právě v jazyce Go. V praxi se ukázaly všechny výše uvedené výhody. Hlavní nevýhodou se však ukázal nedostatek kvalitních knihoven. Přestože je jazyk velmi mladý, stal se velmi populárním, o čemž svědčí například počet repozitářů na GitHubu, nebo otázek na StackOverflow. Bude to ovšem trvat ještě nějaký čas, než se knihovny pro go a jejich vývoj stabilizují. Jeden z nyní nejpopulárnějších webových frameworků pro go — Martini — v

době začátku práce na aplikaci ani neexistoval. Právě nedostatek knihoven pro go vedl k volbě jiného jazyka. Rozhodujícím pro opuštění backendu v jazyce go byla neexistující knihovna pro práci s Elasticsearch. API je sice postaveno na protokolu HTTP, takže šlo s Elasticsearch komunikovat bez použití specializované knihovny, v praxi se to ovšem ukázalo být problematické. Vývoj aplikace si žádal testování různých nastavení a rychlé prototypování v kódu.

Z interních zdrojů máme informaci o tom, že se Elasticsearch chystá vytvořit knihovnu i pro jazyk Go. Pro naše účely se ukázalo výhodnější přepsat aplikaci do dynamického jazyka s lepší podporou pro Elasticsearch.

8.2.3 Ruby a Ruby on Rails

Ruby je dynamicky typovaný jazyk, silně inspirovaný Perlem s důslednou podporou objektového programování. Popularita Ruby vzrostla zejména kvůli webovému frameworku Ruby on Rails, který je v Ruby napsaný. RoR zpopularizovaly koncept Model-View-Controller při tvorbě webových aplikací. Backend aplikace byl nakonec kompletně přepsán jako aplikace v Ruby on Rails. Elasticsearch poskytuje pro práci s Ruby vlastní knihovnu. Další výhodou se ukázala podpora knihovny Rake, což je jakási obdoba Makefile skriptů v Ruby. Pomocí Rakefilu jdou napsat přehledné úlohy pro manipulaci s daty.

Standartní implementace jazyka Ruby byla v minulosti kritizována pro svou pomalost. V průběhu práce na této aplikaci vyšla verze Ruby 2.0, která Ruby dosti zrychluje. Tato rychlost byla pro aplikaci shledána dostatečnou.

8.3 Komunikace frontend-backend

Backend poskytuje služby frontendu pomocí API. Existuje několik přístupů a technologií, jak data mezi backendem a fronendem posílat.

8.3.1 Formát dat

V současnosti jsou pro posílání dat nejběžnější dva formáty – XML a JSON. XML je klasický formát se spoustou možností a nástrojů. Moderní aplikace však stále více přechází k formátu JSON. Formát JSON je velmi úsporný datový formát, který vychází z datových typů v JavaScriptu. Právě úspornost je jednou z jeho největších výhod oproti XML – stejné informace mají v JSON typicky kratší zápis než obdoba v XML. Většina moderních browserů umí formát JSON parsovat a práce v JavaScriptu je pak vzhledem ke kompatibilitě datových typů velmi pohodlná.

Zejména kvůli poslednímu důvodu používá tato práce pro komunikaci mezi fronendem a backendem právě formát JSON.

8.3.2 REST API

REST je zkratka pro Representational state transfer. Jedná se o obecnou architekturu rozhraní. V tomto kontextu nás ale zajímá hlavně její navázání na protokol HTTP. REST využívá metody protokolu HTTP pro změnu, nebo získání stavu datových objektů. Ukázkou REST API s použitím formátu JSON může

být například knihovna Elasticsearch. Mějme například frekvenční data pro český stem „lid“ a instanci Elasticsearch na adrese `http://localhost:9200/`. Ukázky HTTP requestů XX - XX ukazují, jak taková data uložit, získat, změnit, nebo smazat.

REST API v kombinaci s JSON fromátem používá tato aplikace jak ke komunikaci mezi frontendem a backendem, tak mezi backendem a databází Elasticsearch.

8.3.3 Websocket

Alternativou k REST API a protokolu HTTP je protokol WebSocket. WebSocket je stejně jako protokol HTTP postaven nad protokolem TCP. Na rozdíl od HTTP ale poskytuje duplexní spojení. Klient je stále spojený se serverem a oba mohou posílat zprávy bez ohledu na druhou stranu. Spojení přes WebSocket má většinou menší latenci než použití protokolu HTTP [1]. Protokol je podporován všemi moderními verzemi webových prohlížečů a podpora existuje i v knihovnách pro backendové programovací jazyky.

První verze aplikace používali pro komunikaci mezi klientem a serverem právě protokol WebSocket. Nakonec však převážily nevýhody takovéhoto řešení nad výhodami. Jednou z nevýhod je nutnost udržovat spojení s klientem na serverové i klientské straně. Přináší to několik netriviálních problémů. Například v okamžik, kdy se toto spojení přeruší. Pokud se naproti tomu přeruší spojení server-klient při HTTP requestu, může klient zkusit vyslat stejný požadavek znovu.

Druhým problémem je emulace HTTP požadavků v protokolu WebSocket. Z klienta můžeme odeslat HTTP dotaz na server a dostaneme k němu přiřazenou odpověď. V protokolu WebSocket pošleme serveru zprávu a za nějaký čas můžeme dostat zprávu od serveru jako odpověď. Párování došlých zpráv z odeslanými zprávami-požadavky ale musíme implementovat vlastnoručně, například pomocí unikátních ID v těle zprávy. Navíc musíme umět řešit situaci, kdy žádná odpověď ze serveru nedojde. Například nastavením timeoutu pro čekání na odpověď.

WebSocket využijí zejména aplikace, které potřebují, aby server mohl posílat klientovi kdykoliv zprávy, nebo co nejnižší latenci. Takovými aplikacemi mohou být například různé chatovací služby, nebo online hry. Pro jiné většinou asi převáží nevýhody WebSockets oproti HTTP protokolu.

8.4 Shrnutí

Architektura je shrnuta na diagramu XX. Aplikace má backend napsaný v jazyce Ruby a frameworku Ruby on Rails. Data jsou uložená v databázi Elasticsearch. Backend komunikuje s frontendem i databází pomocí REST API, data jsou přenášena ve formátu JSON.

[1] `http://www.websocket.org/quantum.html`

9. Frontend

Možnosti tvorby webových aplikací se posledních několik let rapidně zvětšují. Prohlížeče implementují stále nové technologie, které royšiřují možnosti.

9.1 Možnosti programování frontendu

Programátor webového frontendu si dnes může vybírat z několika paradigmat tvorby webové stránky. Standardem je dnes programovací jazyk Javascript. Spíše historicky bylo výhodné místo Javascriptu používat pro frontendový vývoj různé pluginy. Nejznámější je Adobe Flash, nebo Microsoft Silverlight. Programovat pro tyto pluginy mělo velké výhody. Například snadné přehrávání videa, zobrazení stránky v celoobrazovkovém režimu, nebo přístup k uživatelově webkameře. V době, kdy byla Javascriptová API velmi chudá a rozdílně implementovaná mezi prohlížeči, nabízel Flash zejména tvůrcům webových her konzistenci mezi všemi platformami.

Velkou nevýhodou těchto pluginů byla jejich proprietálnost. Ostatní firmy se bály pustit cizí plugin do svých výrobků. Zlomový moment pro ústup Flashe ze slávy bylo uvedení telefonu iPhone, který podporu pro Flash nenabízel. Softwareoví giganti Microsoft, Apple a Google začali místo proprietálních řešení tlačit otevřenou specifikaci, která se později nazvala HTML5. HTML5 je zastřešující termín pro spoustu technologií, které snaží webová API specifikovat a vytvářet nové.

9.2 JavaScriptové frameworky

Druhým hnacím motorem moderního Javascriptu jsou frameworky. Ještě před několika lety byla práce na interaktivních webech velmi náročná. Prohlížeče se zásadně lišily v implementaci práce s eventy a DOMem. Nové frameworky do jisté míry odstínily programátora od odlišných implementací Javascriptu v prohlížečích.

9.2.1 jQuery

Nejpopulárnějším frameworkem současnosti je jQuery. Ten nabízí jednoduché rozhraní a pro velkou většinu menších webových aplikací je zcela dostačující. Čím je ale aplikace větší, tím začíná být její vývoj s pomocí jQuery náročnější. Zkusme ukázat, jak by v jQuery vypadalo přidání CSS třídy red oknu s id okno:

```
1  $("#okno").addClass("red");
```

Kód má několik problémů. Pokud neexistuje žádné okno s id okno, jQuery nevrátí žádnou chybu. Stačí malý překlep a chyba v kódu se hledá dost obtížně. jQuery nenabízí žádnou funkci typu vratObjektPodleId. Pokud by tyto funkce nabízel, velikost jeho kódu by se zvětšila. Pokud chce programátor použít knihovnu jQuery, musí si ji uživatel stránky celou stáhnout. Verze XX má XX bajtů.

9.2.2 Google Closure

Jiný přístup k vývoji frontendových aplikací přináší Google. Pro svou první webovou aplikaci Gmail vyvinul sadu nástrojů, kterou později vydal jako open source pod názvem Google Closure. Kromě Gmailu ji Google využívá v Google Vyhledávání, Google Mapách, nebo Google Dokumentech. Skládá se ze tří částí - Closure Compiler, Closure Library a Closure Templates. Closure Library je obsáhlá knihovna funkcí pro práci s DOMem, Eventy, matematickými výpočty a spoustou dalších věcí, které webový programátor může využít.

Closure Compiler je inteligentní minifikátor Javascriptového kódu. Odstraňuje funkce, které nejsou volány, přejmenovává všechny názvy funkcí a proměnných na co nejkratší řetězce a v ADVANCED módu se snaží i o pokročilejší optimalizace kódu (například kód funkcí, které jsou volány pouze jednou, je vložen inline). Nejlepích výsledků dosahuje s použitím speciálních anotací, které například vynucují typ proměnné a jsou schopny udělat z Javascriptu typovaný jazyk. Closure Compiler tyto anotace vyhodnocuje a při chybném přiřazení hodnoty vrátí chybu. To umožňuje tvořit více bezpečný javascriptová kód.

Třetí částí Google Closure jsou Closure Templates, šablonovací systém pro Javascript a Javu. Pomocí Closure Templates se snadno vytváří zanořené HTML šablony. Všechny uživatelské vstupy jsou escapované což zabraňuje sniffing útokům.

Částí Templates a Compiler jdou použít odděleně v jakémkoliv Javascriptovém projektu. Používat Closure Library be Compileru nedává příliš smysl, uživatel by při návštěvě webu musel stahovat ohromné množství zbytečných dat.

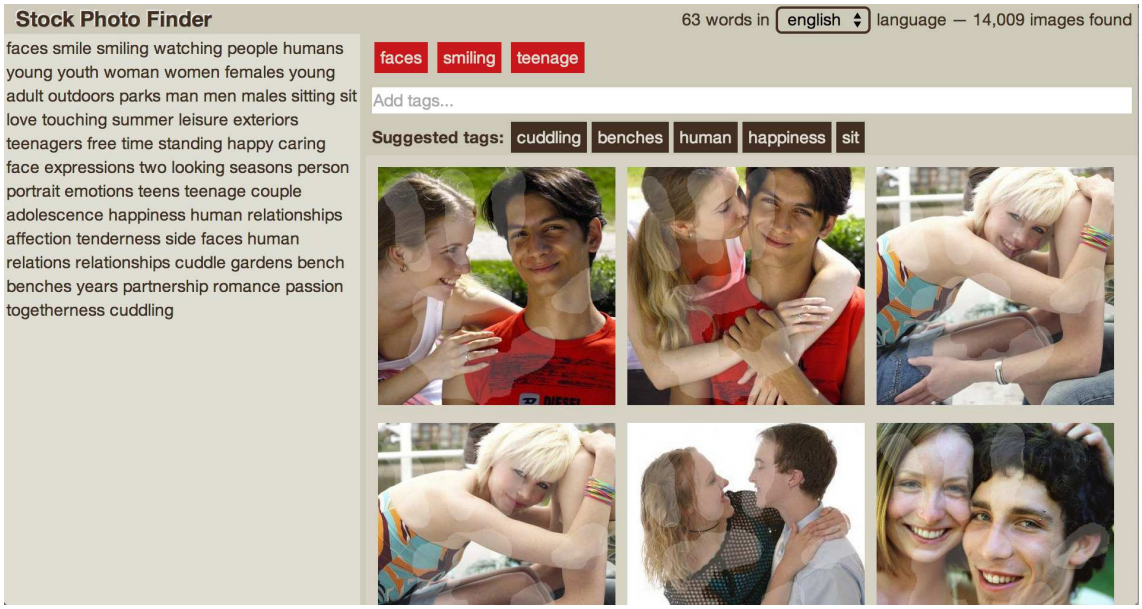
Tato práce na frontendu používá všechny části knihovny Google Closure. Díky tomu si uživatel při první návštěvě webu musí stáhnout pouze jediný soubor, který má pouze XX Kb.

9.3 Stylování uživatelského rozhraní a CSS

Specifikace HTML5 rozšiřuje i možnosti vizualizace pomocí CSS stylů. Nejviditelnějšími novinkami je podpora kulatých rohů, stínování, nebo barevných přechodů. Webový programátor nyní může ke stránce načíst i vlastní font. Všechny tyto možnosti velmi rozšířily možnosti webovým grafikům.

Dalším trendem, který v CSS světě probíhá, je používání CSS preprocesorů.

9.4 Uživatelské rozhraní



Obrázek 9.1: Ukázka čtverců algoritmu Geohash v jednotkovém čtverci.

10. 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.

10.1 Instalace

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

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

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

```
1 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.

10.2 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.

10.3 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

[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/`

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 "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:

```
1 "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 (koncovky `|UNK`) a překlad je relativně nepřesný. Je pravděpodobné, že by s lepšími daty šel natrénovat lepší překladový a jazykový model. Strojový překlad není hlavním tématem této práce, takže bylo jednodušší komerční automatický překlad Google Translate, který přeloží ukázková metadata takto:

```
1 "0000000003", "malé dítě s úsměvem", "", "dítě děti
    dítě děti kojenci děti dětství jednotlivé plochy
    těla nahá nake výrazy obličeje, úsměvu, usměvavý
    sledování sledování kterým zábava zábavné zábavní
    pobavený pobavit laškoval frolicing hrát wantoning
    otevřený" ^ M
```

I z této ukázky je zřejmé, že Google poskytuje kvalitnější překlad, než anglickočeský překladový model v releasu Mosese. Google poskytuje překlad zdarma přes webové rozhraní. Pokud se ale do překladového formuláře nahraje nespecifikované větší množství dat, přestane překlad fungovat. Google pro překlad poskytuje placené API. Platí se XX amerických dollarů za přeložené slovo. Korpus Profimedie obsahuje zhruba (wc vystup 20119222 347129204 4811998848), takze by cely překlad stál XX dollaru. Jelikož se slova v textu opakuji, lze použít překlad slovo po slovu. Ještě lepší by bylo překládat přímo klíčové fráze. Bohužel korpus Profimedie jednotlivé fráze neodděluje. Je ale možné použít učící algoritmus a klíčové fráze detekovat.

10.3.1 Export slov a frází

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.

10.4 Jazykový korpus

Tato práce potřebuje jazykové korpusy pro podporované jazyky ze dvou důvodů. První je potřeba pro určení relativní jazykové frekvence slov v algoritmu TF-IDF. Zadruhé je potřeba jazykový korpus pro rozpoznávání jazyků. Přirozeným jazykovým korpusem by mohla být samotná Profimedia data, ale pro oba účely jsou tato data nepoužitelná. Klíčová slova a názvy obrázků jsou odlišnými druhy textů, než je průměrný článek. Je tedy potřeba jazyková data získat jinde.

Wikipedia se jako korpus velmi hodí. Textový obsah je pod licencí Creative Commons a jeho struktura je velmi podobná běžnému publicistickému článku. Data se dají získat stažením přímo ze serverů wikipedie (pro češtinu zde), nebo pomocí sdílených torrentů. Práce wiki dumpy všech podporovaných jazyků v adresáři data pod názvem typu `wiki_dump_jazyk.xml`. Pro práci s daty z wikipedie je potřeba nejprve převést XML data do textového formátu. K tomu slouží python skript `lib/WikiExtractor.py` od Wikipedie. Pro převod anglických dat lze použít příkaz:

```
1 rake wiki:extract_words_from_wiki en
```

Stejný příkaz je potřeba spustit i pro ostatní podporované jazyky. Příkaz vytvoří v `data/wiki_en` adresářovou strukturu. Není potřeba převést všechna data, pouze tolik, abychom dostali reprezentativní korpus. Pro angličtinu stačí převést zhruba 10000 článků, pro podobné množství českých dat je potřeba převést zhruba 20000 článků (údaj o exportovaných článcích je průběžně vypisován na konzoli).

Nyní je možné vytvořit seznam slov v korpusu s frekvencemi. Ve skutečnosti nás zajímají pouze stemy slov, ne jednotlivé tvary. Příkaz

```
1 rake wiki:frequency_list_from_wiki
```

vytvoří seznam slov a jejich frekvencí s cestou `data/wiki_freq_list_count.en`. Ve skutečnosti nás nezajímají všechna slova z korpusu wikipedie, ale pouze ta, která se vyskytují mezi klíčovými slovy v Profimedia korpusu. Příkazem

```
1 rake data:create_tf_df_list
```

se z korpusu profimedia dat exportují stemy všech slov v profimedia korpusu spolu z hodnotami udávajícími frekvenci termínu v celém korpusu (TF) a frekvenci toho, v kolika popiskách obrázků se slovo objevilo (DF). Pro angličtinu jsou tato data uložena do souboru `data/tf_df_list_en.txt`. Nyní můžeme spárovat

informaci o stemech z wikipedie a nahrát je do databáze. Párování se provede příkazem:

```
1 rake data:pair_profimedia_and_wiki_data
```

který vytvoří soubor `data/paired_wiki_and_profimedia.txt` se statistikami všech stemů z Profimedia korpusu.

10.5 Příprava dat pro detekci jazyků

Automatická detekce jazyka probíhá ve frontendové části. Ke svému chodu ale potřebuje data z korpusu, konkrétně seznam nejčastějších trigramů pro každý podporovaný jazyk. Příkaz

```
1 rake trigrams:extract_most_frequent_trigrams
```

vytvoří pro každý jazyk seznam padesáti nejčastějších trigramů. Seznam pro angličtinu je uložen v souboru `data/most_frequent_trigrams_en.txt`. Příkaz

```
1 rake trigrams:trigrams_to_javascript_classes
```

pracuje se seznamy nejčastějších trigramů ze kterých vytvoří javascriptovou třídu `oo.diplomka.Languages.Data` v notaci Google Closure. Ta je uložena v souboru `public/js/js/diplomka/languages/data.js`.

10.6 Import dat do databáze

Veškerá data jsou importována do databáze elasticsearch. Aplikace očekává Elasticsearch připojený na portu 9200. Samotná data se importují příkazem:

```
1 rake es:import_image_metadata
```

Tento vytvoří v elasticsearch index diplomka. Poté nastaví explicitní mapování v celém indexu. Pokud se mapování explicitně nenastaví, elasticsearch se snaží data namapovat podle jednoduchých heuristik. Například text rozdělí na slova, která pak převede na malá písmena a upraví defaultním stemmerem. Aplikace se však o stemming a převedení na malá písmena stará sama, mapování tedy říká, aby nahraný text oddělil pouze pomocí mezer na slova a dále nezpracovával. Vyhledávací dotaz je pak rozdělen také pouze pomocí mezer. Nastavené mapování lze ověřit pomocí API elsaticsearch na adrese `http://localhost:9200/diplomka/_mapping/`.

Po vytvoření indexu může začít samotný import dat. Data z každého řádku Profimedia dat je převeden do formátu JSON. Data obsahují položky locator, title

10.7 Import metadat obrázků

11. 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.

11.1 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:

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

11.2 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.

11.3 Import anotačních dat

Data pro anotaci lze nahrát pomocí příkazu

```
1 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:

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

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

11.4 Export anotačních dat

Hotové anotace lze exportovat příkazem

```
1 rake data:export
```

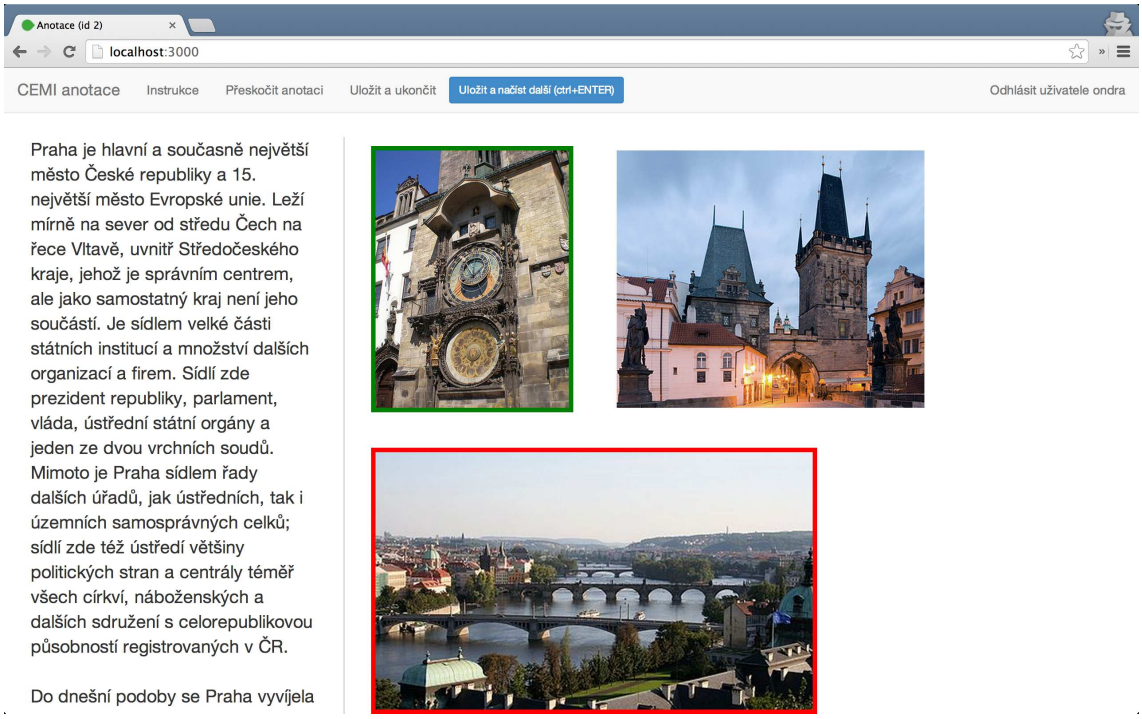
Tento příkaz vypíše na konzolu řádky, které mají tabulátorem oddělené položky:

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.



Obrázek 11.1: Anotační rozhraní. Vhodné obrázky jsou označené zeleným rámečkem, nevhodné červeným.

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.

11.5 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].

11.6 Anotační proces

Úkolem anotátora je přiřadit vhodné a nevhodné obrázky. Po přihlášení do anotačního rozhraní vidí v levé části test a v pravé obrázky. Levým tlačítkem myši může označit obrázky, které odpovídají textu, pravým tlačítkem myši označí

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

obrázky, které textu neodpovídají. Pokud si anotátor není jistý, nechá obrázek neoznačený. Uživatel může použít klávesovou zkratku ctrl+ENTER k uložení a načtení další anotace. Může také anotaci přeskočit (pak se označí jako přeskočená a nepřiřadí se jinému anotátorovi), nebo uložit a ukončit.

12. Testování výsledků

Výsledný program se skládá z několika komponent. Funkčnost každé komponenty jde testovat. Některé části jdou otestovat automaticky, například rozpoznávání jazyka. Na některé části je potřeba testování s uživateli.

Klíčovou částí celého projektu je algoritmus, který k textu přiřadí ilustrační obrázek. Úspěšnost tohoto algoritmu byla otestována na uživatelích. Rozhraní může fungovat pro více jazyků. Testování však proběhlo pro vstupní texty v češtině. Jedním z důvodů je snadnější získání anotátorů v českém jazyce. Hlavním důvodem je ovšem pravděpodobná vyšší nepřesnost algoritmu v češtině. Dataset Profimedie má data v anglickém jazyce a pro české použití musel být přeložen. Pokud se tedy ukáže, že algoritmus funguje dobře pro češtinu, měl by pro angličtinu fungovat ještě lépe. Pro testování bylo využito testovací rozhraní popsané v Kapitole 11.

12.1 Vyloučení narušitele (o_test1)

První metodou testování bylo „vyloučení nepřítele“ (anglicky „intruder detection“). Tato metoda se používá k evaluaci automaticky detekovaných shluků slov a je popsána například v [6]. V naší variantě se evaluují obrázky přiřazené k textu. Nejprve se k danému textu najdou pomocí testovaného algoritmu 4 nejvíce odpovídající obrázky. K nim se přidá jeden náhodně vybraný obrázek z datasetu a poté se náhodně zamíchá pořadím těchto obrázků. Anotátor vidí v anotačním rozhraní text a 5 obrázů. Jeho úkolem je označit obrázek, který danému textu, podle jeho názoru, odpovídá nejméně. Pokud algoritmus přiřazující obrázky textu funguje správně, měl by být uživatel schopen označit obrázek, který byl do sady vybrán náhodně a rozlišit ho od obrázků, který vybral algoritmus přiřazující obrázky.

Texty pro testování algoritmu pochází z online článků českých webových serverů. Jedná se o texty článků, které obsahují ilustrační obrázky z datasetu Profimedie. Takto omezená množina novinových textů je pro náš účel velmi výhodná. Pro různé druhy článků dataset Profimedie neobsahuje žádné vhodné ilustrační obrázky. Jedná se například o politické zpravodajství, pro které v datasetu aktuální fotky událostí, nebo například archivní fotky osobností. Oproti tomu u článků, které již nějaký ilustrační obrázek z Profimedie obsahují, jsou možnosti vhodného obrázku daleko vyšší. Jedná se většinou o různé hobby a společenské články.

Pro naše testování jsme využili články ze serveru iDnes, který obsahoval nejvíce článků s ilustračními obrázky z Profimedia. Konkrétně jich máme k dispozici 4223.

Z těchto 4223 článků jsme náhodně vybrali pro testování 60 článků. Ke každému z článků byly získány algoritmem 4 doporučené ilustrační obrázky a byl přidán jeden obrázek náhodný. Každá tato testovací sada byla otestována dvěma anotátory. Bylo k dispozici 5 anotátorů. Dohromady to znamenalo pro každého anotátora 24 a dohromady 120 anotací. Anotace byly rozděleny tak, aby každá dvojice anotátorů měla právě 6 stejných testovacích sad.

Během anotace se zjistilo, že u dvou testovacích sad se jeden z obrázků nena-

	o_test1	o_test2	celkem
anotátorů	5	5	7
textů	58	120	178
anotací	116	240	356
pozitivní shoda	45/58 = 78%	76/120 = **63**%	121/178 = 67%
negativní shoda	3/58 = 5%	15/120 = 13%	18/178 = 10%
neshoda	10/58 = 17%	29/120 = 24%	39/178 = 22%

Obrázek 12.1: Přehled výsledků uživatelského testování.

čítá. Tyto sady byly z testování vyřazeny a anotovaných testovacích sad je tedy pouze 58. Výsledky testování jsou shrnuty v tabulce 12.1 ve sloupci „o_test1“. Pobrobné výsledky testování, včetně Cohenovy kappy pro všechny dvojice anotátorů je v příloze 13.

Výsledky ukazují, že pro 78 % testovacích sad měli anotátoři pozitivní shodu. Pozitivní shoda znamená, že se oba anotátoři shodli na stejném obrázku, který je pro vstupní text nejméně vhodný. Tento obrázek byl zároveň vybrán do testovací sady náhodně. Čím je toto procento vyšší, tím lépe náš algoritmus pracuje, protože uživatelé jsou schopní detekovat náhodný obrázek. Pro 5 % testovacích sad máme negativní shodu. Oba anotátoři vybrali obrázek, který nebyl přiřazen náhodně. Znamená to tedy, že tento obrázek nebyli schopní detekovat a algoritmus tedy pro vstupní text nepracuje dobře (pokud vyloučíme možnost, že náhodně přiřazený obrázek je k textu relevantní). U poslední skupiny testovacích dat – 17 % – byl pouze jeden z anotátorů schopen vybrat náhodně vybraný obrázek.

Získaná data ukazují, že algoritmus pracuje poměrně dobře. Pokud by algoritmus přiřazoval automaticky obrázky k textům i v praxi, čtenáři by mohli by poznali rozdíl od algoritmu, který přiřazuje obrázky nahodile.

Během testování se objevil jeden zásadní problém s testovací metodou. V datasetu Profimedie jsou i obrázky, které jsou si velmi vizuálně podobné, například fotky stejné osoby z různých úhlů. Tyto fotky mají často i stejné textové popisky. Mějme tedy 4 obrázky, které jsou si vizuálně velmi podobné a mají stejné textové popisky. Pokud algoritmus označí jako nejvhodnější obrázek k textu jeden z těchto obrázků, budou i na dalších třech doporučených pozicích vizuálně podobné obrázky (pokud tedy nemáme jinou množinu obrázků, která má stejné textové popisky, ale je vizuálně odlišná). Pokud se takový text objeví v naší testovací metodě, uvidí anotátor 4 velmi podobné obrázky a k nim jeden náhodně vybraný. Nejméně vhodný obrázek pak snadno označí, aniž by vůbec četl anotační text. Ukázalo se, že v našem testování k takovému problému opravdu došlo – anotátorovi se zobrazily obrázky, z nichž ani jeden nebyl vhodným obrázkem k danému textu, přesto anotátor snadno označil nejméně vhodný obrázek. Tento problém zkresluje dosažené výsledky testování touto metodou.

12.2 Detekce správného obrázku (o_test2)

Abychom předešli problémům s metodou popsaným v předchozí sekci, provedli jsme nové testování. Úkolem uživatelů bylo nyní vybrat obrázek, který se ke vstupnímu textu hodí nejvíce. Množinu obrázků nyní tvořil jeden výstup z algorit-

mu a 4 náhodně vybrané obrázky. Zvětšili jsme i testovací sadu. Bylo testováno 120 textů náhodně vybraných z iDnes datasetu. Každý z textů byl anotován dvěma anotátory. Na každém z pěti anotátorů tedy bylo 48 anotací. Výsledky testování jsou shrnuty v tabulce 12.1 ve sloupci „o_test2“. Pobrobné výsledky testování, včetně Cohenovy kappy pro všechny dvojice anotátorů je v příloze 13.

Výsledky jsou oproti předchozí metodě o něco horší. Pro 63 % testovacích sad měli oba anotátoři pozitivní shodu, pro 13 % měli anotátoři negativní shodu a pro 24 % se anotátoři neshodli.

12.3 Shrnutí

Uživatelské testování ukazuje, že je algoritmus schopen v poměrně velkém procentu případů přiřadit automaticky „dostatečňě vhodný“ ilustrační obrázek. To, že je ilustrační obrázek dostatečně vhodný ovšem neznamená, že je nejvhodnější. Výběr nejvhodnějšího ilustračního obrázku je ovšem velmi individuální a nedá se obecně měřit.

Testování ukázalo také na slabiny algoritmu. Možná největší slabinou je fáze překladu, díky níž popisky k některým obrázkům nesprávné. Tímto problémem netrpí algoritmus pro anglické vyhledávání, který pracuje s nepřeloženými popisky. Ten však nebyl uživatelsky testován.

Jedním z problémů testování je právě omezená doména testovaných textů. Můžeme říci, že algoritmus funguje dobře na nějaké doméně textů, ale pro obecné zpravodajské články může algoritmus fungovat velmi špatně. Toto ovšem není problém testovaného algoritmu, ale dat na kterých pracuje. Rozšíření domény obrázků v korpusu Profimedie by rozšířilo i doménu textů pro které algoritmus funguje dobře.

13. Závěr

Seznam použité literatury

[1] APACHE SOFTWARE FOUNDATION. *Apache Lucene 4.9.0* [software]. 2014. Dostupné z: `http://lucene.apache.org/`.

[2] ELASTICSEARCH BV. *Elasticsearch 1.3.0* [software]. 2014. Dostupné z: `http://www.elasticsearch.org/`.

[3] BOTOREK, Jan. *Tvorba nástroje pro zpracování textových popisů multimediálních dat* [online]. 2012 [cit. 2014-07-17]. Bakalářská práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Petra Budíková. Dostupné z: `http://is.muni.cz/th/359815/fi_b/`.

[4] BUDÍKOVÁ, Petra, Michal BATKO a Pavel ZEZULA. *Evaluation Platform for Content-based Image Retrieval Systems*. In *International Conference on Theory and Practice of Digital Libraries 2011, LNCS 6966*. Berlin: Springer, 2011. s. 130-142, 12 s. ISBN 978-3-642-24468-1.

[5] CAVNAR, William B., et al. *N-gram-based text categorization. Ann Arbor MI*, 1994, 48113.2: 161-175.

[6] CHANG, Jonathan, et al. *Reading tea leaves: How humans interpret topic models*. In: *Advances in neural information processing systems*. 2009. s. 288-296.

[7] COUCHBASE. *Couchbase 2.5.1* [software]. 2014. Dostupné z: `http://www.couchbase.com/`.

[8] DEAN, Jeffrey; GHEMAWAT, Sanjay. *MapReduce: simplified data processing on large clusters*. Communications of the ACM, 2008, 51.1: 107-113.

[9] GOOGLE INC. *Překladač Google* [online]. 2014 [cit. 2014-07-25]. Dostupné z: `http://tranlslate.google.cz`.

[10] JIA, Yangqing. *Caffe: An open source convolutional architecture for fast feature embedding*. 2013 Dostupné z: `http://caffe.berkeleyvision.org`.

[11] KOEHN, Philipp. *Statistical machine translation*. Cambridge University Press, 2009.

[12] KOEHN, Philipp, et al. *Moses: Open source toolkit for statistical machine translation*. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2007. s. 177-180.

[13] LOTT, Brian. *Survey of Keyword Extraction Techniques*. UNM Education, 2012.

[14] MILLER, George A. *WordNet: A Lexical Database for English*. Communications of the ACM, 1995, roč. 38, č. 11, s. 39-41.

[15] PORTER, Martin F. *An algorithm for suffix stripping*. Program: electronic library and information systems. MCB UP Ltd, 1980, roč. 14, č. 3, s. 130-137.

[16] STRAUCH, Christof; SITES, Ultra-Large Scale; KRIHA, Walter. NoSQL databases. Lecture Notes, Stuttgart Media University, 2011.

[17] WIKIPEDIA *Geohash — Wikipedia, The Free Encyclopedia* [online]. 2014 [cit. 2014-07-24] Dostupné z: `http://en.wikipedia.org/w/index.php?title=Geohash&oldid=609521802`.

Seznam tabulek

Seznam použitých zkratek

Příloha 1

```
1  Výsledek testu o_test1: 103 / 116 = 88.79 %
2  Výsledek testu o_test2: 196 / 240 = 81.66 %
3  Celkové výsledky dohromady: 299 / 356 = 83.98 %
4
5
6  havel
7   o_test1: 20 / 22 = 90.90 %
8   o_test2: 36 / 48 = 75.0  %
9   celkem:  56 / 70 = 80.0  %
10   odchazel_o
11    o_test1
12     3   0
13     1   1
14     Cohen's kappa = 0.5454545454545455
15   odchazel_v
16    o_test2
17     10  2
18     0   0
19     Cohen's kappa = 0.0
20   paroubkova
21    o_test1
22     6   0
23     0   0
24     Cohen's kappa = NaN
25    o_test2
26     6   1
27     3   2
28     Cohen's kappa = 0.2727272727272727
29   pavlovic
30    o_test1
31     5   0
32     0   0
33     Cohen's kappa = NaN
34   rakosova
35    o_test1
36     6   0
37     0   0
38     Cohen's kappa = NaN
39    o_test2
40     9   0
41     1   2
42     Cohen's kappa = 0.7499999999999999
43   semerad
44    o_test2
```

```
    6   2
    0   4
    Cohen's kappa = 0.6666666666666667

odchazel_o
 o_test1: 21 / 23 = 91.30 %
 celkem:  21 / 23 = 91.30 %
  havel
   o_test1
    3   1
    0   1
    Cohen's kappa = 0.5454545454545455
  paroubkova
   o_test1
    4   1
    0   1
    Cohen's kappa = 0.5714285714285715
  pavlovic
   o_test1
    6   0
    0   0
    Cohen's kappa = NaN
  rakosova
   o_test1
    6   0
    0   0
    Cohen's kappa = NaN

odchazel_v
 o_test2: 42 / 48 = 87.5  %
 celkem:  42 / 48 = 87.5  %
  havel
   o_test2
    10  0
    2   0
    Cohen's kappa = 0.0
  paroubkova
   o_test2
    10  1
    0   1
    Cohen's kappa = 0.625
  rakosova
   o_test2
    9   1
    0   2
    Cohen's kappa = 0.7499999999999999
  semerad
   o_test2
```

```
    11  0
    0   1
    Cohen's kappa = 1.0

paroubkova
 o_test1: 20 / 24 = 83.33 %
 o_test2: 43 / 48 = 89.58 %
 celkem:  63 / 72 = 87.5  %
  havel
   o_test1
    6   0
    0   0
    Cohen's kappa = NaN
   o_test2
    6   3
    1   2
    Cohen's kappa = 0.2727272727272727
  odchazel_o
   o_test1
    4   0
    1   1
    Cohen's kappa = 0.5714285714285715
  odchazel_v
   o_test2
    10  0
    1   1
    Cohen's kappa = 0.625
  pavlovic
   o_test1
    4   0
    2   0
    Cohen's kappa = 0.0
  rakosova
   o_test1
    5   1
    0   0
    Cohen's kappa = 0.0
   o_test2
    11  1
    0   0
    Cohen's kappa = 0.0
  semerad
   o_test2
    10  2
    0   0
    Cohen's kappa = 0.0

pavlovic
```

```
 o_test1: 21 / 23 = 91.30 %
 celkem:  21 / 23 = 91.30 %
  havel
   o_test1
    5   0
    0   0
    Cohen's kappa = NaN
  odchazel_o
   o_test1
    6   0
    0   0
    Cohen's kappa = NaN
  paroubkova
   o_test1
    4   2
    0   0
    Cohen's kappa = 0.0
  rakosova
   o_test1
    3   1
    1   1
    Cohen's kappa = 0.24999999999999986

rakosova
 o_test1: 21 / 24 = 87.5  %
 o_test2: 39 / 48 = 81.25 %
 celkem:  60 / 72 = 83.33 %
  havel
   o_test1
    6   0
    0   0
    Cohen's kappa = NaN
   o_test2
    9   1
    0   2
    Cohen's kappa = 0.7499999999999999
  odchazel_o
   o_test1
    6   0
    0   0
    Cohen's kappa = NaN
  odchazel_v
   o_test2
    9   0
    1   2
    Cohen's kappa = 0.7499999999999999
  paroubkova
   o_test1
```

```
    5   0
    1   0
    Cohen's kappa = 0.0
   o_test2
    11  0
    1   0
    Cohen's kappa = 0.0
  pavlovic
   o_test1
    3   1
    1   1
    Cohen's kappa = 0.24999999999999986
  semerad
   o_test2
    9   0
    0   3
    Cohen's kappa = 1.0

semerad
 o_test2: 36 / 48 = 75.0  %
 celkem:  36 / 48 = 75.0  %
  havel
   o_test2
    6   0
    2   4
    Cohen's kappa = 0.6666666666666667
  odchazel_v
   o_test2
    11  0
    0   1
    Cohen's kappa = 1.0
  paroubkova
   o_test2
    10  0
    2   0
    Cohen's kappa = 0.0
  rakosova
   o_test2
    9   0
    0   3
    Cohen's kappa = 1.0


o_test1
 nesprávné označení (text, uživatel)
  00207 o_paroubkova
  01337 o_pavlovic
  01595 o_paroubkova
```

```
  01921 o_rakosova
  03758 o_odchazel_ondrej
  03758 o_paroubkova
  03907 o_paroubkova
  04107 o_rakosova
  07604 o_havel
  07604 o_odchazel_ondrej
  09093 o_pavlovic
  09093 o_rakosova
  09551 o_havel
celkem 116 anotací 58 textů
každý text oanotován dvěma anotátory
 oba správně : 45 / 58 = 77.58 %
 oba špatně  : 3 / 58 = 5.172 %
 jeden špatně: 10 / 58 = 17.24 %

o_test2
 nesprávné označení (text, uživatel)
  00124 o_semerad
  00405 o_havel
  00405 o_paroubkova
  00623 o_paroubkova
  01907 o_rakosova
  01911 o_odchazel_vojtech
  01911 o_paroubkova
  02505 o_odchazel_vojtech
  02505 o_rakosova
  03610 o_havel
  03610 o_semerad
  04211 o_odchazel_vojtech
  04282 o_havel
  04282 o_semerad
  04368 o_odchazel_vojtech
  04368 o_semerad
  04480 o_semerad
  04530 o_havel
  05102 o_rakosova
  05102 o_semerad
  05692 o_semerad
  06379 o_paroubkova
  06777 o_rakosova
  07326 o_rakosova
  07326 o_semerad
  07587 o_odchazel_vojtech
  07587 o_rakosova
  08015 o_havel
  08015 o_semerad
  08621 o_havel
```

```
  12931 o_odchazel_vojtech
  13674 o_havel
  13674 o_semerad
  13803 o_havel
  13803 o_paroubkova
  14257 o_havel
  14257 o_rakosova
  14329 o_rakosova
  14329 o_semerad
  14353 o_havel
  14353 o_rakosova
  14988 o_semerad
  15726 o_havel
  15770 o_havel
celkem 240 anotací 120 textů
každý text oanotován dvěma anotátory
 oba správně : 76 / 120 = 63.33 %
 oba špatně  : 15 / 120 = 12.5  %
 jeden špatně: 29 / 120 = 24.16 %
```