Vybrané aktuální projekty Centra ZPJ

Jan Rygl, Vít Suchomel

E-mail: xrygl@fi.muni.cz,139723@mail.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

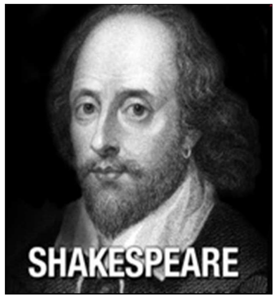
Obsah:

- Rozpoznávání autorství anonymních dokumentů na Internetu
- Budování textových korpusů z Internetu

Motivace

Proč se začalo mluvit o určování autorství?

- Může za to Shakespearovo dílo:
  - Gale Ecco, 1787:
    A Dissertation on the Three Parts of King Henry VI. Tending to Shew That Those Plays Were Not Written Originally by Shakspeare.



Motivace

Proč nás zajímá autorství dokumentů?

- Anonymní výhrůžky (omezený seznam potenciálních kandidátů)
  - Lidé dříve trestaní za šíření poplašné zprávy (bomba na letišti)
  - Všichni sousedi (příliš hlučné párty)
  - Studenti (příliš těžké zkoušky)
- Zpochybnění závěti (ověření shody autorství)
- Falešné doznání obžalovaného napsané policisty (**1. použití před soudem**)

Motivace

Proč v prostředí Internetu?

- Rasistické diskuzní příspěvky psané z internetové kavárny
- Extremistické blogy zveřejněné přes anonymní proxy servery
- Oznámení bomby na letišti pachatelem, který má čistý rejstřík
- Na Internetu se dá najít téměř vše.

Dennis Bayley, 2004:
Anonymity is the single most important enabler of criminal activity.

Praktické zadání

1. Máme dokument D a množinu dokumentů autora A. *Napsal A dokument D?*
2. Máme anonymní dokument D a množinu potenciálních autorů. *Je autorem D někdo z autorů? Kdo?*
3. Máme anonymní dokument, chceme zjistit autora.
   - Až zde potřebujeme Internet.
   - Pokud předem neomezíme množinu (autor je registrovaný na webu, bydlí v nějaké vesnici apod.), úloha je velmi "ambiciózní".
   - Předpokládá se, že skutečný **autor někdy publikoval pod svým pravým jménem** (bakalářská práce, inzerát, ...)

Jak na ověření autorství

Nejjednodušší případ, máme dokument D a texty autora A. *Jaká je pravděpodobnost, že mají shodné autorství?*

Postup:

1. Předzpracujeme texty
   - Odstranění šumu (text, který nenapsal uživatel)
   - Tokenizace
   - Značkování pro další zpracování (morfologie, syntaxe)
2. Spočítáme z textů charakteristiky autora dokumentu D a autora A.
3. Srovnáme **charakteristiky autorů** a dle jejich podobnosti určíme pravděpodobnost shody autorství.

1. Předzpracování

- Závislé na jazyce. Pro češtinu lze použít fakultní nástroje (majka, synt, set), pro angličtinu a některé další jazyky existují další kvalitní software.
- Pokud nejsme schopni předzpracování udělat, nebudeme moci použít v dalším kroku některé autorské charakteristiky.

```
<p align="justify"> <img
style="margin: 3px; float:
left" src="otaznik.jpg"
alt=""> Společnost se za
svou historii dokázala
```

```
<s>
<img src="otaznik.jpg"/>
Společnost  společnost  k1gFnSc1        N
se          sebe        k3xPyFc4        N
za          za          k7c4            N
svou        svůj        k3xOyFpXgFnSc4  Y
historii    historie    k1gFnSc4        N
dokázala    dokázat     k5eAaPmAgFnS    N
```


2. Počítání charakteristik autora

- Jazykově závislé
  - Osoba mluvčího (pohlaví, číslo)
  - Analýza gramatických značek v textu
  - Analýza počtu vět (hlavní, vedlejší, ...)
  - Chyby v textu (překlepy, hrubky, syntax)
- Jazykově nezávislé (stačí tokenizace)
  - Analýza délky vět (počet slov, znaků)
  - Analýza délky slov (porovnání histogramů)
  - Frekvence slov, bigramů, ... (ovlivněna tématem)
  - Frekvence stopslov (tematicky nezávislá)
  - δ-score (srovnání frekvencí slov v textu s běžnou frekvencí slova v korpusech)
  - Bohatost slovní zásoby

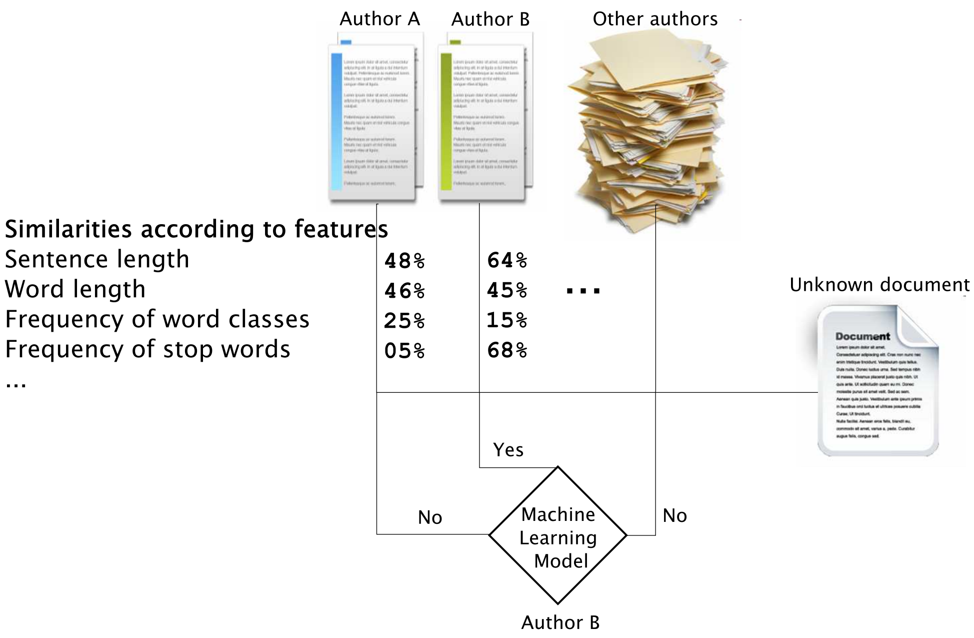
3. Porovnání autorů (1)

- Každá **charakteristika** c_i po srovnání dvou autorů vrátí **racionální číslo v intervalu** $\langle 0, 1 \rangle$
  $c_i(A, D) \sim$ **podobnost A a D**.
- Příklad (frekvence slovních druhů)
  - Pro slovní druhy 1 až 10 spočítáme relativní frekvence v dokumentu: $f_D^i, i \in \{1, \dots, 10\}$
    a ve sjednocení textů autora: $f_A^i, i \in \{1, \dots, 10\}$
  - Frekvence srovnáme pomocí druhé mocniny rozdílů odpovídajících si frekvencí:
    $$1 - \frac{\sum_{i \in \{1,\dots,10\}} {f_D^i - f_A^i}^2}{2}$$
  - Např.
    A={k1:0.5, k5:0.4, k2:0.1}
    D={k1:0.4, k5:0.3, k2:0.2, k3:0.1}
    $P = 1 - \frac{(0.5-0.4)^2+(0.4-0.3)^2+(0.1-0.2)^2+(0-0.1)^2}{2} = 1 - \frac{0.04}{2} = 0.98$

3. Porovnání autorů (2)

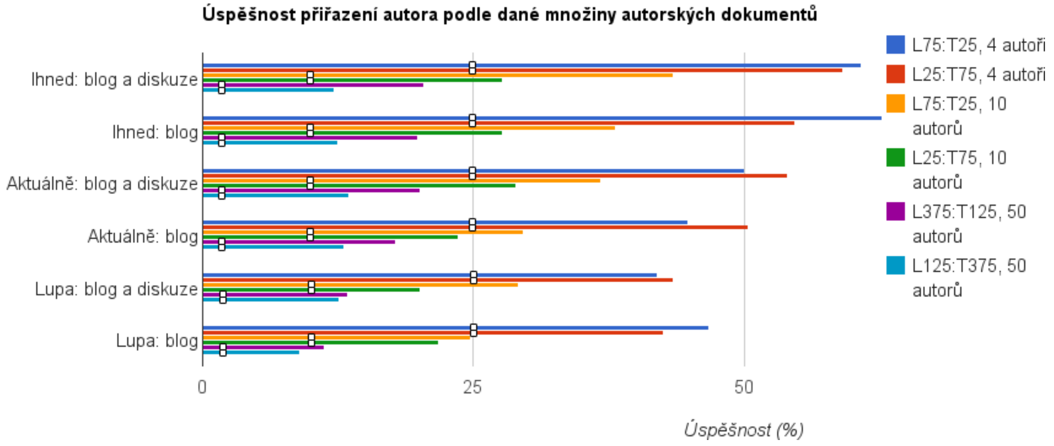
- S autory a dokumenty, u nichž známe autorství, vytvoříme velké sady dvojic autor dokument tak, abychom měli stejný případ shod i neshod (např. 10000 od každého). Každé dvojice bude reprezentována n-ticí $c_i(A, D)$.
- Použijeme **strojové učení**, aby se naučilo rozpoznávat n-tice signalizující shodu a rozdílnost autorů.
  Získáme tak model M takový, že:
  $$M\Big(\big[c_1(A, D), \dots, c_n(A, D)\big]\Big) = P\big(autor(A) == autor(D)\big)$$
- Vždy, když budeme chtít srovnávat dokument a autora, extrahujeme jejich autorské charakteristiky a předložíme je jako n-tici modelu. Ten vrátí odpověď.

Rozšíření na úlohu N autorů



Na čem závisí úspěšnost

- Počet možných autorů.
- Délka zdroje (knihy >> diskuze)
- Kvalita zdroje (je třeba detekovat citace, odkazy, ...)

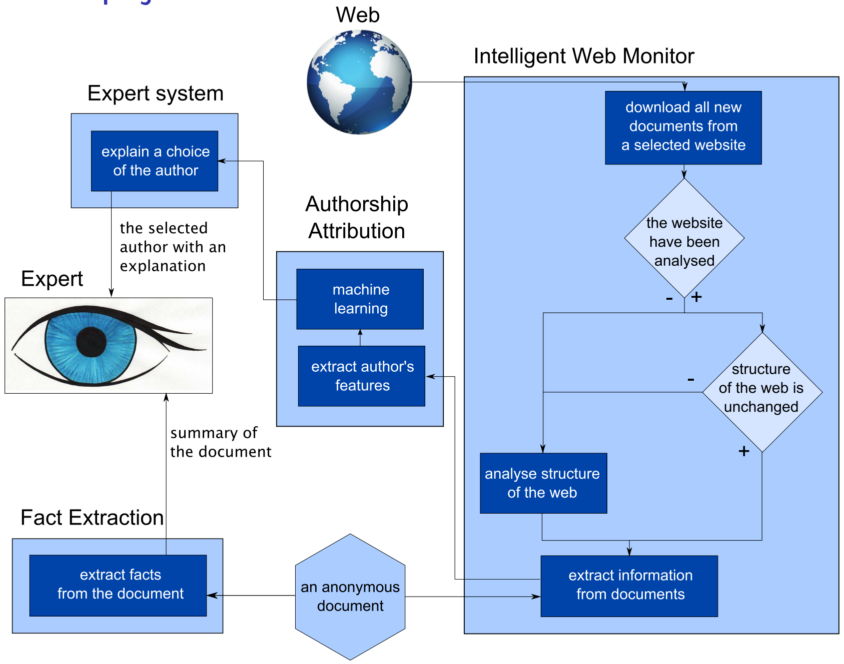


Přidání Internetu

Co potřebujeme:

- Znát weby, kde jsou texty autorů.
- Detekovat strukturu webů.
- Pravidelně stahovat nové dokumenty z webů (nutná struktura).
  - Odhalit změnu struktury webu a aktualizovat informace.
- Dokumenty spravovat v databázi.
  - Vyhledávání (stovky tisíc a více dokumentů)
  - Hledání dle času, kategorií, autora, ...
- Předzpracovávat si dokumenty.
- Ukládat si drahé mezivýsledky (nepočítat např. frekvence slovních druhů vícekrát pro jeden dokument).

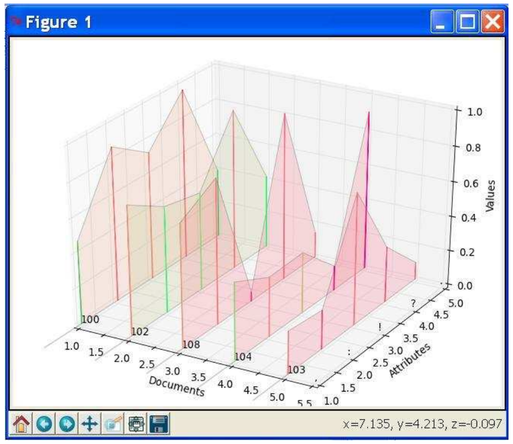
Schéma zapojení Internetu



Další výstupy

- Uživatelům často nestačí říci, který autor je nejpravděpodobnější
- Je potřeba poskytnout textovou nebo vizuální podporu výsledků

Příklad srovnání 5 autorů na autorském rysu frekvence interpunkce:



Shrnutí

- Projekt je stále ve vývoji
- Pokud vás něco zaujalo, je pravý čas se přidat
  - lingvistika nebo statistika
    Vytvářet **nové charakteristiky autora** (analýza chyb, nářečí, počet jednoduchých a složitých vět, formátování textu, či nejlépe vymyslet vlastní)
  - grafika nebo analýza dat
    Vymyslet kreativní přístupy k **vizualizaci výstupních dat**, případně k sumarizaci výsledků programu, aby jim rozuměl školený uživatel
  - programování a struktura webu
    Navrhnout nové metody pro **automatickou detekci struktury webu**, přihlašování se ke zdrojům vyžadujícím autentizaci, vyhledávání odkazů na dokumenty v doméně
  - strojové učení a analýza
    Hrát si s různými metodami **strojového učení** a frameworky
  - vše ostatní
    A mnoho dalšího, stačí se domluvit, jsou potřeba **lingvisti, programátoři, grafici, právníci,** ...

Proč potřebujeme velké korpusy?

Použití korpusů

- lexikografové: slovníky
- lingvisté: jazykové analýzy (**Word Sketches**)
- statistické nástroje ZPJ: jazykové modely pro značkovače, analyzátory, překladové systémy, prediktivní psaní,...

Přínosy velkých korpusů

- větší slovník (více různých slov)
- více/lepší příklady použití slov ve větách
- lepší pokrytí řídkých jazykových jevů
- více dat pro přesnější jazykové modely

Tradiční textové korpusy

Vznik

- obvykle na objednávku vládní instituce nebo nakladatelství
- zdroje: nakladatelství, skenování knih, přepisy rozhovorů

Výhody tradičních korpusů

- kontrolovaný obsah

Nevýhody tradičních korpusů

- nedostatenčná velikost pro některá použití
- obtížné získávání dat, vysoké náklady
- problémy s autorskými právy

Web je největší korpus

Výhody internetových korpusů

- obrovské množství dat
- dokumenty různých druhů
- aktuální podoba psané formy jazyka v populaci
- snadná dostupnost, nízké náklady

Nevýhody internetových korpusů

- neuspořádanost
- nežádoucí obsah
- duplicity
- chyby
- víme, co stahujeme?

Web crawler

Web crawler je druh počítačového programu

- prochází internet (stránky propojené odkazy)
- stahuje dokumenty (metainformace, obsah)
- ukládá části dokumentů v různých formátech k dalšímu použití

Crawlery

- k získávání obsahu dokumentů – GoogleBot (navíc k indexování), Heritrix a mnoho dalších
- ke sbírání odkazů
- k získávání textových dokumentů pro ZPJ – SpiderLing

Základní návrh crawleru

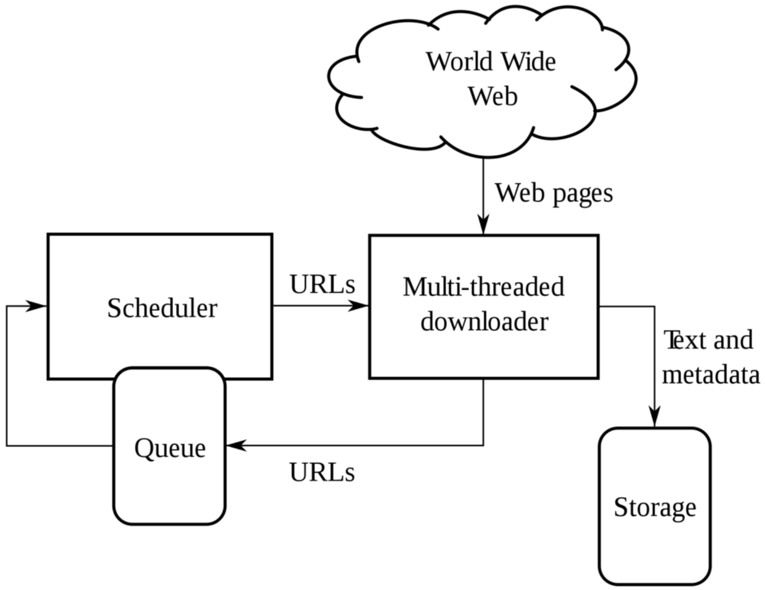


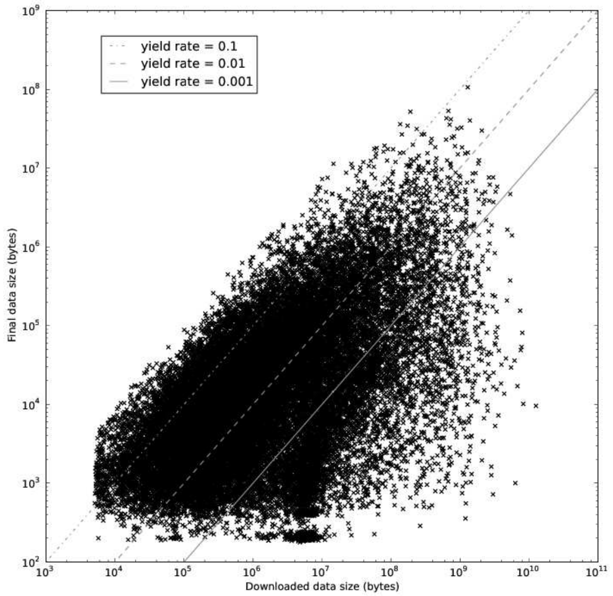
Schéma z http://en.wikipedia.org/wiki/Web_crawler

SpiderLing – crawler pro textové korpusy

- důraz kladen na efektivitu stahování
- **míra výtěžnosti** $= \dfrac{\text{velikost výsledných dat}}{\text{velikost stažených dat}}$
- crawler průběžně vyhodnocuje výtěžnost webových domén, zaměřuje se na „textově bohaté“ a odkládá stahování (nebo vůbec nestahuje) z neperspektivních webů
- cílem je sestavit korpusy velikosti $\geq 10^{10}$ slov pro všechny významné jazyky

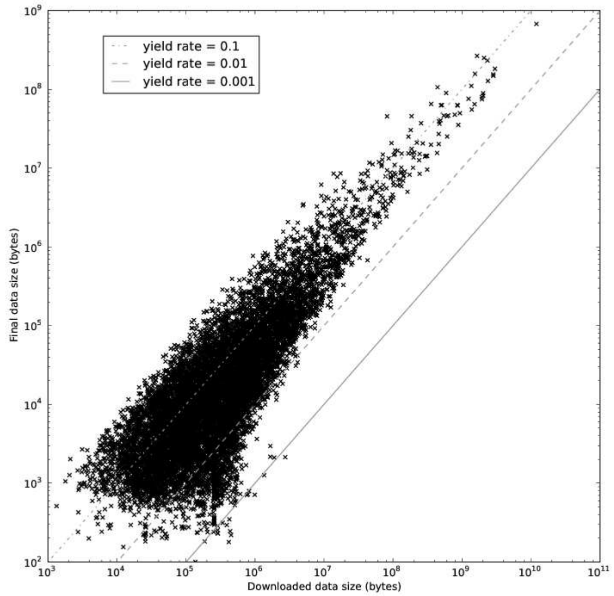
Efektivita stahování z webových domén

Heritrix, portugalský internet



Efektivita stahování z webových domén

SpiderLing, český internet



Postup získávání webových korpusů v CZPJ

- příprava jayzkově závislých modelů používaných v dalších krocích – učení na dokumentech z Wikipedie
- spuštění crawleru (`SpiderLing`)
- zpracování a vyhodnocování během běhu crawleru
  - detekce znakové sady dokumentu (`Chared`)
  - filtrování jazyka (vektor trigramů znaků)
  - odstraňování nežádoucího obsahu (`Justext`)
  - kontrola duplicitních dokumentů
  - vyhodnocování průběžné výtěžnosti webových domén
- zpracování získaných dat
  - odstranění podobných odstavců (`Onion`)
  - tokenization (`Unitok` nebo jiný nástroj)
  - značkování morfologické a syntaktické – externími nástroji, jsou-li dostupné
  - zakódování a nahrání do korpusového manažeru (`Manatee`/`Bonito`)

Více v předmětu PA154 nástroje pro korpusy

Detekce kódování znaků

Používáme nástroj Chared (`http://nlp.fi.muni.cz/projects/chared`)

- 57 podporovaných jazyků, samostatný model pro každý
- model obsahuje vektor četností trigramů bajtů pro každé z používaných kódování
- klasifikátor vybere kódování, jehož vektor má největší skalární součin s vektorem testovaného dokumentu
- demo na stránce nástroje

Odstraňování nežádoucího obsahu

Nežádoucí obsah

- html značky, styly, poznámky
- negramatické věty: navigace, reklamy, tabulky, příliš krátké úseky,. . .

Používáme nástroj jusText (`http://nlp.fi.muni.cz/projects/justext`)

- rozdělení na odstavce
- slovník častých slov v daném jazyce
- klasifikace odstavce podle délky, hustoty slov ze slovníku, hustoty odkazů, třídy okolních odstavců
- demo na stránce nástroje

Odstraňování duplicit

Duplicity

- shodné dokumenty nebo odstavce – snadné
- podobné dokumenty nebo odstavce – obtížné

Používáme nástroj onion (`http://nlp.fi.muni.cz/projects/onion`)

Velké textové korpusy získané z internetu v CZPJ

jazyk	surová data [GB]	velikost korpusu [GB]	míra výtěž-nosti	velikost korpusu [10^9 tokenů]	doba běhu [dny]
angličtina	2859	108	3.78 %	17.8	17
am. španělština	1874	44	2.36 %	8.7	14
arabština	2015	58	2.89 %	6.6	28
čeština	4000			5.8	40
francouzština	3273	72	2.19 %	12.4	15
japonština	2806	61	2.19 %	11.1	28
ruština	4142	198	4.77 %	20.2	14
turečtina	2700	26	0.97 %	4.1	14

V NLPC máme k dispozici také kolekci dat ClueWeb '09 – vyčištěná anglická část obsahuje zhruba 70 miliard tokenů.

Ukázky internetových korpusů

- srovnání korpusů BNC (tradiční) a enTenTen (internetový)
  - novější bývá lepší
  - větší bývá lepší
- získávání termínů z doménového korpusu
- korpus tádžické perštiny