Rozpoznávání české domény s Google Prediction API

Jiří Kremser

28. ledna 2011

Obsah

WebArchiv

- www.webarchiv.cz
- Projekt NKČR
- WebArchiv je digitální archiv „českých“ webových zdrojů, které jsou zde shromažďovány za účelem jejich dlouhodobého uchování.
- Inspirováno www.archive.org
- Wayback Machine
- Plošné sklizně
- Tematické sbírky
- Výběrový přístup

Výběrový přístup

Podle

- Území – všechny dokumenty (zdroje) publikované na území České republiky
- Jazyk – všechny zdroje v češtině (bez ohledu na místo vydání)
- Autorství – všechny zdroje českých autorů (bez ohledu na místo vydání)
- Předmět/obsah – všechny zdroje, jejichž obsah se týká České republiky nebo českého národa (bez ohledu na místo vydání)

WebAnalyzer

- Diplomová práce Ivana Vlčka na téma „Automatická identifikace webových stránek příslušejících k národnímu webu“
- Zásuvný modul do crawleru Heritrix $\Rightarrow$ platforma Java
- Národní web: doména cz + stránky o ČR
- Cíl sklízet národní web a uchovávat jej v LTP systému (http://ndk.cz/)

WebAnalyzer

Komponenty:

1. EmailSearcher (f_1)
2. PhoneSearcher (f_2)
3. HtmlLangSearcher (f_3)
4. GeoIPSearcher (f_4)
5. DictSearcher (f_5)

Akceptuj dokument doc pokud

$$\sum_{i=1}^{5} x_i f_i(doc) > treshold$$

$\vec{x}$ je vektor koeficientů (pevně zakódovaný)

Google Prediction API

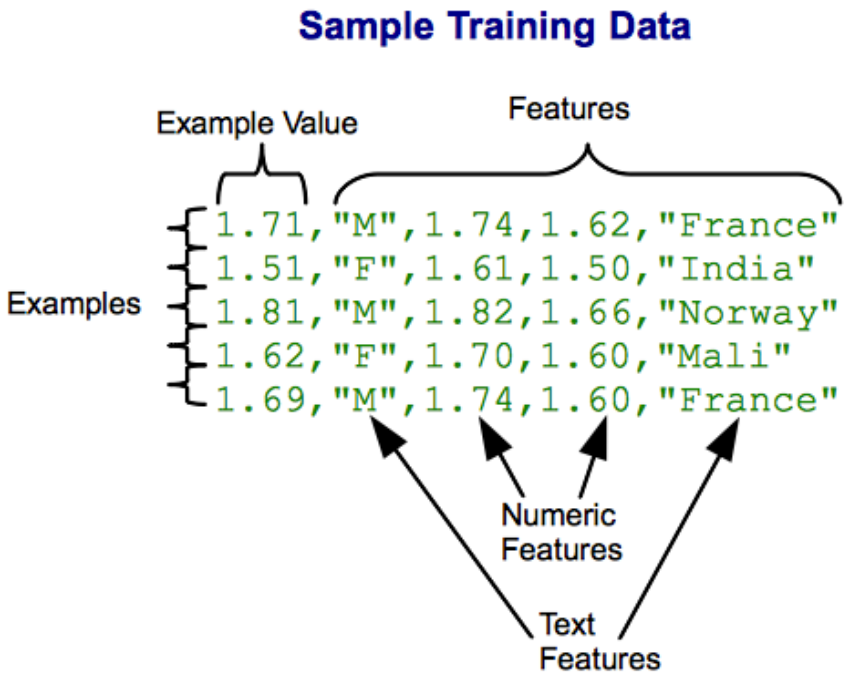
- http://code.google.com/intl/cs/apis/predict/
- XaaS, kde X=ML
- RESTful webové služby a JSON
- Supervised learning
- Produkt ve vývoji
- Black box
- Předpokladem jsou trénovací data v Google Storage

Hlavní kroky

1. Vytvoření trénovacích dat
2. Nahrání dat do Google Storage
3. Spuštění trénovacího procesu
4. Zasílání požadavků ke klasifikaci

Trénovací data

- CSV
- První sloupec je roznodovaná třída
- 2 typy dat čísla (bez uvozovek) a nečísla (s uvozovkami)



Obrázek: Zdroj: http://code.google.com/apis/predict/docs

Nahrání dat

- Utilita gsutil v Pythonu
- Vytvoření nového „bucketu“: `$gsutil mb gs://my_bucket`
- Nahrání trénovacích dat: `$gsutil cp data.csv gs://my_bucket`

Učení

- Asynchronní HTTP POST volání na

  https://www.googleapis.com/prediction/v1.1/training?data=my_bucket%2Fdata.csv
- Ověření stavu učení: GET na stejné URL
- Služba je zabezpečena metodou OAuth, proto je vhodné využít připravených skriptů a oacurl

Klasifikace

- Asynchronní HTTP POST volání na

  https://www.googleapis.com/prediction/v1.1/training?data=my_bucket%2Fdata.csv/predict

- Data musí být stejného formátu jako při učení, ale první sloupec chybí
- Služba vrací i stupně konfidence příslušnosti k jednotlivým třídám {"data":{"kind":"prediction#output",

"outputLabel":"positive",

"outputMulti":[{"label":"negative","score":156.3161020576954},

{"label":"positive","score":231.29281866550446}]

}} (http://en.wikipedia.org/wiki/Czech_language)

Google Storage ceník

- **Storage** — $0.17/GB/month (3 Kč)
- **Síť**
  - Upload data to Google — $0.10/GB (1,76 Kč)
  - Download data from Google
    $0.15/GB for Americas and EMEA (2,64 Kč)
    $0.30/GB for Asia-Pacific (5,29 Kč)
- **HTTP požadavky**
  PUT, POST, LIST — $0.01 za 1000 požadavků (0,17 Kč)
  GET, HEAD — $0.01 za 10000 požadavků (0,17 Kč)

Google Prediction API ceník

- **Varianta zadarmo**
  - Pouze prvních 6 měsíců
  - Limit 100 predikcí/den a 5MB trénovacích dat/den (učení)
  - Maximálně 20 000 predikcí celkem
- **Placená varianta**
  - $10 měsíčně pokrývá dalších 10 000 predikcí (každých 1000 predikcí je pak za $0.5)
  - $0.002 za 1MB trénovacích dat (max. velikost CSV souboru je 100MB)

WebAnalyzer 2.0

- Trénovací data (URL z logu crawleru pro sklizeň domény .cz)
- Negativní příklady
- **Vektor vlastností**
  - CZHtmlLang — atribut značky HTML
  - CzechNames — množství česk/czech/tschechi ...
  - CZLinks — množství http://*.cz
  - CZEmails — množství emailů končících .cz
  - CZPhoneNum — množství českých tel. čísel
  - CZIP — true, když je ip z ČR
  - LangID — country code jazyka
  - Words — náhodně vybraných 100 slov ze stránky

Příklad

"positive","false","a_lot","few","none","none","false","en","for and general embassy east in embassy under agency edit embassy north the abroad to republic czech a community kenya embassy uruguay lebanon of status sovereign vilnius consulate czech athens vojvodina search embassy angeles embassy norway ..."

(http://en.wikipedia.org/wiki/List_of_diplomatic_missions_of_the_Czech_Republic)

Technologie použité k preprocessingu

- wget
- html2text
- Bash
- sed&grep
- curl pro volání geolokačního API a API pro detekci jazyka
- oacurl pro volání Google Prediction API

Wekka J48 (bez textu)

Precision	Recall	F-Measure	ROC Area	Class
0.967	0.968	0.968	0.982	negative
0.985	0.985	0.985	0.982	positive
0.979	0.979	0.979	0.982	

Matice záměn

a	b	
922	30	a
31	1974	b

Wekka J48 s textem (StringToWordVector filter)

Precision	Recall	F-Measure	ROC Area	Class
0.974	0.965	0.969	0.984	negative
0.984	0.988	0.986	0.984	positive
0.98	0.98	0.98	0.984	

Matice záměn

a	b	
919	33	a
25	1980	b

Google Prediction API

- Accuracy 0.98
- ```
  $./oauth-check-training.sh cz-analyzer/data_google.csv
  {"data":{"data":"cz-analyzer/data_google.csv",
  "modelinfo":"estimated accuracy: 0.98"}}
  ```
  ```

Demo

Obrázek: Mrak vyrobený z náhodně vybraných 10 000 sklizených slov

Konec

Děkuji za pozornost

Q&A