PA164

Podpora strojového učení pro WebAnalyzer a otestování Google Prediction API

autor
Bc. Jiří Kremser

WebArchiv

- www.webarchiv.cz
- Projekt NKČR
- WebArchiv je digitální archiv „českých“ webových zdrojů, které jsou zde shromažďovány za účelem jejich dlouhodobého uchování.
- 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

- Komponenty:
  - EmailSearcher (f1)
  - PhoneSearcher (f2)
  - HtmlLangSearcher (f3)
  - GeoIPSearcher (f4)
  - DictSearcher (f5)
- Systém zjednodušeně

  if $\sum_{i=1}^{5} x_i f_1(doc) > treshold$ then accept doc

Google Prediction API

- http://code.google.com/intl/cs/apis/predict/
- RESTful interface s JSONem
- Předpokladem jsou trénovací data v Google Storage
- Google Storage ceník:
  - Upload $0.10/GB
  - Download $0.15/GB for Americas and EMEA
  - PUT, POST, LIST—$0.01 per 1,000 requests
  - GET, HEAD—$0.01 per 10,000 requests

Google Prediction API demo

curl -d "text=In terms of scale and style it is, as Nolan intended, comparable to Bond’s best excursions — yet filtered through a brain-frying, subconscious-spelunking, time-dilating structure that boldly frames action sequences around each other. So we get an explosive Arctic mountain vault-storming within a zero-gravity scramble within a vehicle-crunching chase. In effect, the set-pieces are simultaneous. Which is insane, but brilliant as, while he at times boggles through the necessarily complex editing, Nolan never corrupts his multiverse’s internal logic." http://text-processing.com/api/sentiment/

Response in JSON:

{"probability": {"neg": 0.09206104904798176, "neutral": 0.26419740580479445, "pos": 0.9079389509520182}, "label": "pos"}

Shrnutí

- Preprocessing dat
- Data do Google Storage
- Spustit proces uceni nad ”bucketem” dat
- Funkční komponenta s jasně definovaným rozhraním

Konec

Děkuji za pozornost

Dotazy