Building and Exploring (Web) Corpora [3]

EMLS 2008, Stuttgart

23-25 July 2008

Pavel Rychlý

pary@fi.muni.cz

NLPlab, Masaryk University, Brno

Outline

- Google as Web corpus, Unix tools
  - Why not to use a search engine as a corpus
  - Simple but powerful tools for text processing
- Exercises 3
  - Googleology
  - Unix tools exercises

Using web as corpus (local)

- pre-create
  - crawl web
  - download web pages
  - clean data
  - annotate
  - output = large ballanced web corpus (itWaC, deWac)
- advantages
  - huge corpora can be build
- disadvantages
  - time consuming
  - computer experts required

Using web as corpus (on-line)

- on-the-fly
  - input = query
  - search engine
  - download web pages/snippets
  - (annotate)
  - output = concordance lines
- disadvantages
  - limited query language
  - slow

On-line corpus - advantages

- almost no resources needed
- one can test queries using a browser
- the query language is simple
- very easy to automate
  - generate queries in any programming language
  - parse results for number of hits

On-line corpus – first problems

- the query language is very simple
  - many simple queries for simple phrase
  - post processing of results
- very easy to automate
  - huge number of request is slow

Technical problems

- 1000 queries per user per day
  - or slowing responses under a big load
- downloading pages is slow
- handling missing pages

Search engine – live black box

- black box
  - lemmatization, case sensitivity, non-words
  - automatic substitutions
    - abbreviations (GM=genetically modified, General Motors)
    - one = 1
    - no documentation
  - cannot be disabled
  - result depends on browser options
- live
  - they are improving it every day = everyday changes
  - support for different languages differs

Numbers of hits are not reliable

- result = number of pages not instances
- unknown handling of duplicates
- numbers are estimations
- changes in repeated queries
- more restrictions could get more results
  - `servercommand login` `3 hits`
  - `servercommand login name` `9 hits`

On-line corpus – moving target

- the web is constantly changing
- average life time of a URL is 6 month
- some URLs change content several times in a day
- it is hard to repeat tests with the same results

Summary: What have you learned?

Using a search engine for web as corpus processing is hard and not reliable

Exercises 3:
How a search engine works?

- Does you favourite search engine:
  - uses lemmatization or stemming?
  - limits the search on domains depending on browser language or start page (google.de, es.yahoo.com)?
  - search in links, page title
  - expands abbreviations
  - support wild-cards or regular expressions
- Can any of such features be disabled?
- Is there a (complete/any) documentation?