Overview of Sketch Engine developments

finished, in progress, planned

Pavel Rychlý

March 21, 2012

Outline

Development priorities

- simplicity
  on all levels (user interface, API, core system)
- small size
  of data structures, programs/libraries

Development priorities

- simplicity
  on all levels (user interface, API, core system)
- small size
  of data structures, programs/libraries
- it leads to faster processing, faster development
- scalability, language independence

Results of development priorities

- simple user interface
  features/options available in the core system are not directly accessible from the user interface

Results of development priorities

- simple user interface
  features/options available in the core system are not directly accessible from the user interface
- Examples:
  - context size of a concordance (kwic, sentence)

Results of development priorities

- simple user interface
  features/options available in the core system are not directly accessible from the user interface
- Examples:
  - context size of a concordance (kwic, sentence)
  - attributes in frequency distribution (structure attributes):
    query: AJ + score
    frequecies: lemma 1L, bncdoc.genre

Bigger and Faster Corpora

- 4-byte number = max 4 billion numbers
- more than 4 bil. items (words, hits, ...)
  8-byte number
  2 times bigger, 2 times slower
- optimal storage, compression
  – smaller data, faster access

Bigger and Faster Corpora

Planned changes

- sorting of big concordances via sampling
- automated sampling for big queries
- word lists in caches

Word sketch data format

- 18–23 data files
- three parts
  - gramrel names – very small
  - $< word1, gramrel, word2 >$ lists – big
  - lists of occurrences – very big

New data format of Word Sketches

New data format of the $< w1, gr, w2 >$ lists

- unlimited number of hits
- unlimited number of $< w1, gr, w2 >$
- including commonest match
- using compression (– less files, smaller, faster)