Language Research in Seznam.cz

Vladimír Kadlec

Seznam.cz a.s.

March 6, 2014

Fulltext indexing statistics (4.3.2014)

Document language	Number of documents
Czech	585 mil.
English	673 mil.
Slovak	77 mil.
German	30 mil.
Other	65 mil.
Total	1 430 mil.

Research Corpora (2013)

Language	Documents $\times 10^6$	Words $\times 10^9$
Czech	94	24
English	105	59
Slovak	6.1	2.1
German	6.7	2.6

Corpus Tools

Hadoop

- Pig, Hive
- Hadoop streaming

Query Processing

- User query
  - `úfal v praze`
- Expanded terms
  - `ústav, formální, aplikované, lingvistika, Praha, ...`

Query Analysis

- Language detection
- Entity identification
- Diacritics reconstruction

Query Expansion

- Morphology, lemmatization (`majka`), disambiguation
- Acronym expansion (queries)/extraction (documents)
- Synonyms
- Other words “related” to the query