Information Access
and
Text Mining

German Rigau <german.rigau@ehu.es>

Content

- NLP Tools & Resources for Web Search (German Rigau)
- Hands-on with NLP tools (Rodrigo Agerri)
- Basic Techniques for Web Search (German Rigau)
- Hands-on with search engine (Rodrigo Agerri)
- Projects (students)
  - Discuss proposals (for title and one paragraph summary) => Deadline 26/02
  - Discussion by email
  - Oral presentations (of design, current status)
  - => Deadline 18/03
  - Full report
  - => Deadline ??/??

Content

- Basic Techniques for Web Search
  - Information Retrieval (IR)
  - Information Extraction (IE)
  - Question Answering (Q&A)
  - Clustering
  - Classification
  - Summarisation
  - Multilingüism
    - Cross-lingual Information Retrieval (CLIR)
    - Machine Translation (MT)

Content

- NLP Tools
  - Basic Tools
    - Tokenization
    - Sentence Splitting
    - Language Identifiers
    - Lemmatization, POS tagging
    - Named Entity Recognizers and Categorizers (NERC)
    - Parsing
    - Word Sense Disambiguation (WSD)
    - Semantic Role Labelling (SRL)

Content

- Resources
  - Words & Works
  - Ontologies:
    - Mikrokosmos
  - Large-scale Knowledge Bases:
    - WordNet & EuroWordNet
  - More large-scale resources
    - ConceptNet, Framenet, VerbNet, PropBank, ...
  - Building Wordnets
  - WordNet extensions:
    - SUMO ontology, eXtended WordNet, Meaning project
  - Reasoning

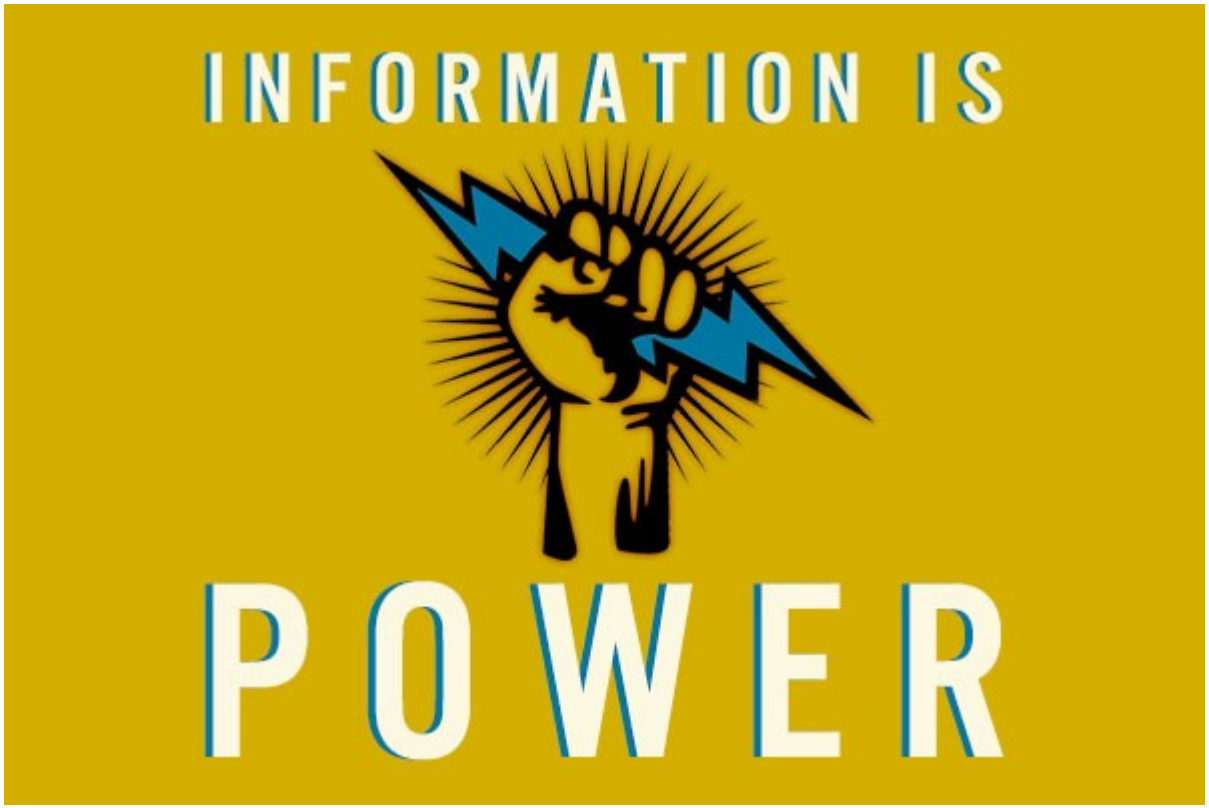
Evaluation

- Applications of Information Access and Text Mining:
  - Student/teacher topic
  - Short presentation
    - 10 minutes sharp, ~ 10 slides
    - Presentation: **18/03**
  - Written report:
    - Format: http://www.acl2013.org/site/call.html
    - Deadline Report: **??/??**
    - Short paper describing an experimental work
      - < 3000 words

Short Motivation

Information is power!

Short Motivation



Short Motivation

Knowledge is power!

Short Motivation

Short Motivation

Knowledge is power!

… and the knowledge to use …

Short Motivation

More than **90%** of digital information available is **unstructured** information in the form of texts and documents (written or spoken) in multiple languages ...

Information Access
and
Text Mining

German Rigau <german.rigau@ehu.es>