Advanced Methods for Corpus Processing

Lluís Padró <padro@lsi.upc.edu>
Lluís Màrquez <lmarquez@qf.org.qa>
German Rigau <german.rigau@ehu.es>

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Foreword

- UPC: Artificial Inteligence
  - EMNLP: Empirical Methods for NLP
  - 2001/2002-2002/2003
    - Lluís Padró, Lluís Màrquez, German Rigau
  - 2003/2004-2004/2005
    - Lluís Padró, Lluís Màrquez, Neus Català, German Rigau
  - 2005/2006
    - L. Padró, L. Màrquez, X. Farreres, J. Daudé, G. Rigau
- EHU: NLP
  - Advanced Methods for Corpus Management
  - 2004/2005-...-2016/2017
    - Lluís Padró, Lluís Màrquez, German Rigau

Content

- Theme 1: Introduction to corpus analysis.
- Theme 2: Statistical methods.
- Theme 3: Machine learning methods.
- Theme 4: Knowledge-based methods.

Content

- Knowledge Based methods for NLP (German Rigau)
  - 5 June: 15:00h – 18:00h
  - 7 May: 18:00h – 20:00h
- Statistical methods for NLP (Lluís Padró)
  - 19-20 June: 15h – 19:30h
  - 21 June: 15h – 17:30h
- Machine Learning for NLP (Lluís Màrquez)
  - 21 June: 17:30h – 20:00h
  - 22 June: 15h – 19:00h
  - 23 June: 15h – 17:30h
- Concluding remarks (German Rigau)
  - 23 June: 17:30h - 19:30h

Content

- Statistical methods for NLP (Lluís Padró)
  - Introduction (statistical vs non-statistical NLP, what are statistical models, what is model estimation)
  - MLE Estimation
  - MaxEnt Estimation
  - Hidden Markov Models
  - Structured prediction (sequences): Log-linear models, MEMM, CRF, Perceptron
  - Generalizing structured prediction (dependency structures)

Content

- Machine Learning for NLP (Lluís Màrquez)
  - Introduction: Machine Learning and Machine Learning for NLP
  - Machine Learning: Classical Methods from AI
  - Margin-based Machine Learning Algorithms
  - Machine Learning for NLP
  - Applications

Content

- Knowledge-based NLP (German Rigau)
  - Words & Works
  - Large-scale Knowledge Bases:
    - WordNet & EuroWordNet
  - More large-scale resources
    - ConceptNet, Framenet, VerbNet, PropBank, Predicate Matrix
  - WordNet extensions:
    - SUMO ontology, eXtended WordNet, MCR
  - Ontologies:
    - AdimenSUMO
    - Reasoning, abduction
- Concluding remarks (German Rigau)
  - Combining approaches

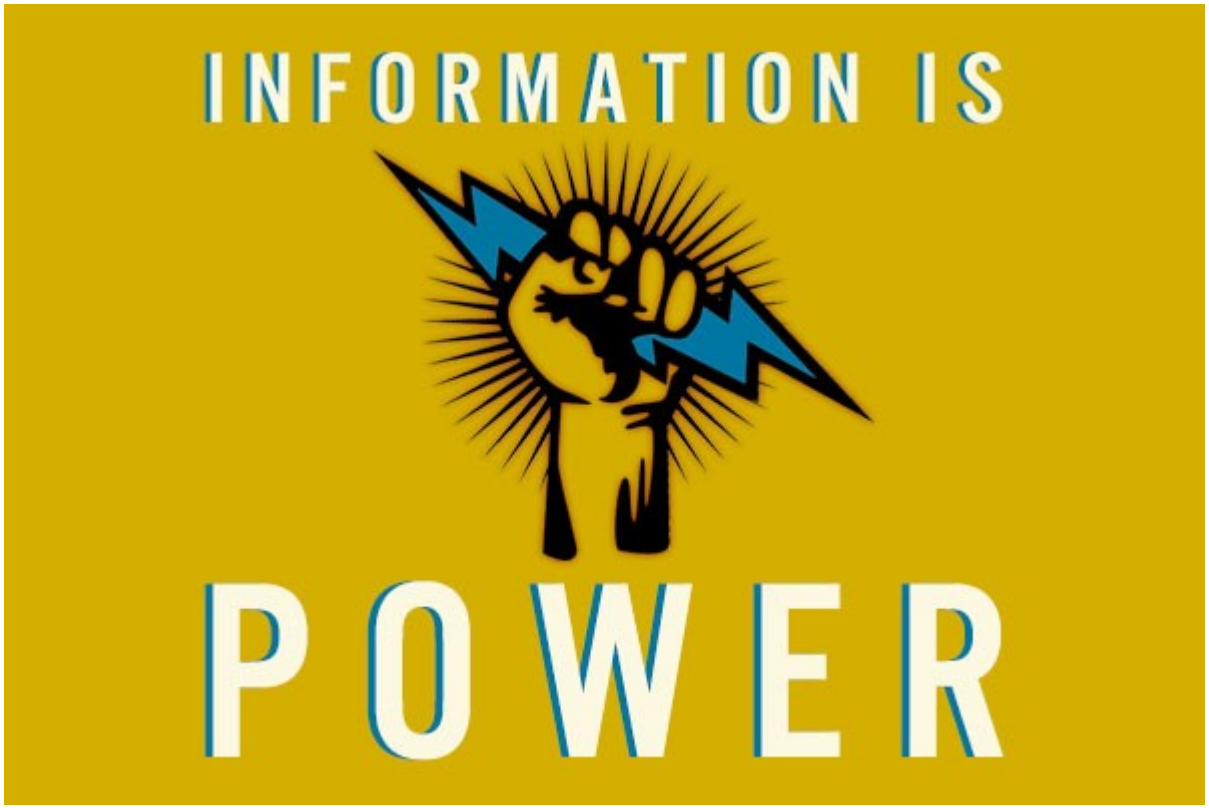
Evaluation

- Application of empirical methods for NLP:
  - Teacher exercises (30%)
  - Teacher/student topic
  - Short presentation (20%)
    - 10 minutes sharp, ~ 10 slides
    - Presentation: 23/06
  - Written report (50%)
    - Format: http://www.acl2015.org
    - Deadline Report: 23/07
    - 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 ...

Advanced Methods for Corpus Processing

Lluís Padró <padro@lsi.upc.edu>
Lluís Màrquez <lmarquez@qf.org.qa>
German Rigau <german.rigau@ehu.es>

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر