

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 24/02**
 - Discussion by email
 - Oral presentations (of design, current status)
 - => **Deadline 10/03**
 - 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: **10/03**
 - Written report:
 - Format: <http://www.acl2013.org/site/call.html>
 - Deadline Report: **??/??**
 - Short paper describing an experimental work
 - < 3000 words

Foreword

“Cuando creíamos que teníamos todas las respuestas, de pronto, cambiaron todas las preguntas.”

- Mario Benedetti

“When we thought we had all the answers, suddenly, they changed all the questions. ”

- Mario Benedetti



Foreword

- Where are the **answers** to the new (and old) questions?
 - Introspection? Experts?...
 - From many people? ... “Wisdom of the Crowd”?
 - Books, News, Tweets, ... Textual Sources?
 - Multimedia sources? Images, Radio, TV ...
 - Sensors? IoT? ...
 - Anything? Everything?
- Information **overload** ...

Foreword

- **Information overload ...**
 - infobesity, infoxication!

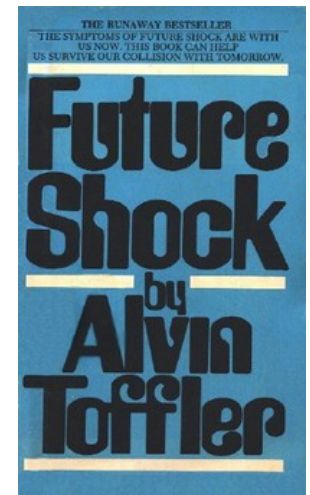
Foreword

- **Information overload ...**
 - infobesity, infoxication!
 - by Bertram Gross, The Managing of Organizations: The administrative struggle (1964)



Foreword

- **Information overload ...**
 - infobesity, infoxication!
 - by Bertram Gross, *The Managing of Organizations: The administrative struggle* (1964)
 - by Alvin Toffler, *Future Shock* (1970)



Foreword

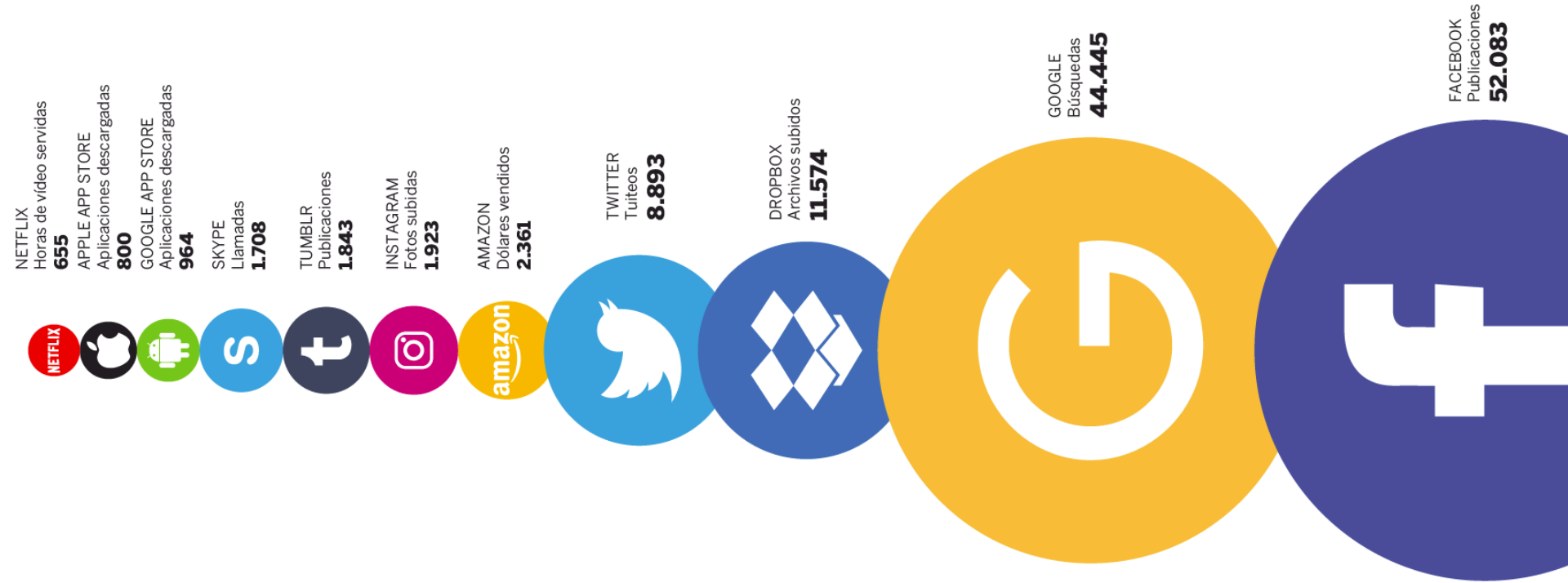
- **Information overload ...**
 - infobesity, infoxication!
 - by Bertram Gross, *The Managing of Organizations: The administrative struggle* (1964)
 - by Alvin Toffler, *Future Shock* (1970)
- Seneca complained that “*the abundance of books is distraction*” in the 1st century AD!

Foreword

- **Information overload** occurs when the amount of input to a system exceeds its processing capacity.
- Decision makers have fairly **limited** cognitive processing capacity.
- Consequently, when information overload occurs, it is likely that a **reduction** in decision quality will occur.
- From (Speier et al 1999)
- Always when **advances in technology** have increased a production of information.

Foreword

- What happens in Internet every **second**? (July 2015)



Foreword

- What happens in Internet every **second**? (July 2015)



Foreword

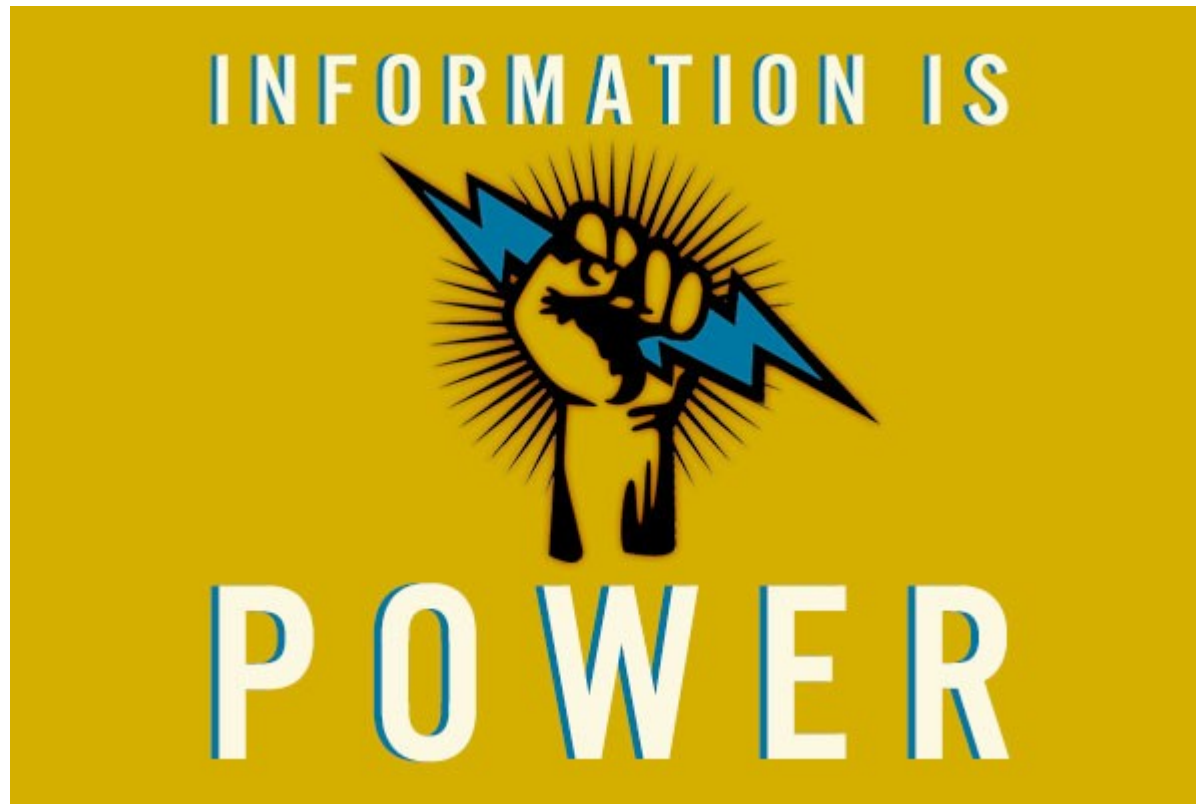
- ... not only coming from Social Media.
- LexisNexis receives **daily** 1.5M news.
- CENDOJ stores 6M judicial sentences (0.6M/year)
- 5M Electronic Health Records (EHR) ...
- 0.2M Patents ...
- ...
- ... all kinds of e-documents ...

Foreword

- **Unstructured** digital content accounts for **90%** of all information [[White paper IDC 2014](#)] ...
- Usually in the form of **texts** and documents in **multiple languages** ...
- **Only** appropriate NLP tools can access this wealth of knowledge ...
- NLP among the **top** 10 strategic technology trends for 2017 according to [Gartner](#)

Foreword

Because everybody knows that ...



Foreword

But in fact ...

KNOWLEDGE
— IS —
POWER

Foreword

e.g. IBM Watson ...



but also Google, Facebook, Amazon, Microsoft, ...

Big Data & NLP ...



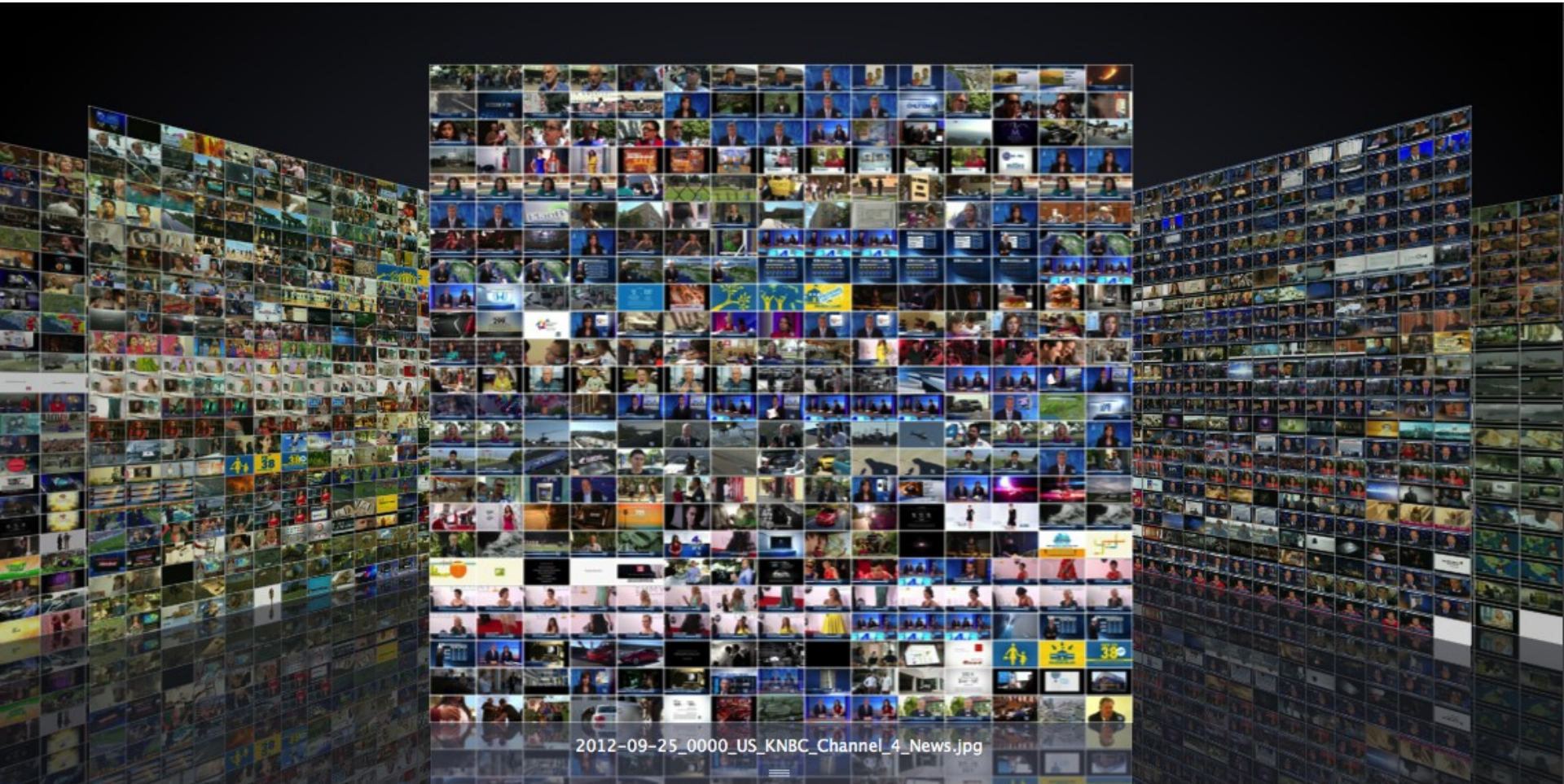
Big Data & NLP ...

World-Wide Integrated Crisis Early Warning System

Integrated Crisis Early Warning System

<http://www.lockheedmartin.com/us/products/W-ICEWS.html>

Big Data & NLP ...



<https://sites.google.com/site/distributedlittleredhen>

Information Access and Text Mining

German Rigau <german.rigau@ehu.es>

