Advanced Methods for Corpus Processing

Lluís Padró <padro@lsi.upc.edu>
Lluís Màrquez <lluismv@amazon.com>
German Rigau <german.rigau@ehu.es>

Centre de Tecnologies i Aplicacions
del Llenguatge i la Parla

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-...-2019/2020
        - Lluís Padró, Lluís Màrquez, German Rigau

Content

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

Content

- Knowledge Based methods for NLP (German Rigau)
  - 27 May: 15:00h – 18:00h
  - 28 May: 15:00h – 17:30h
  - 29 May: 15:00h – 17:30h
  - 01 June: 15:00h – 17:30h

- Statistical methods for NLP (Lluís Padró)
  - 2-4 June: 15:00h – 19:00h

- Machine Learning for NLP (Lluís Màrquez)
  - 8-10 June: 15:00h – 19:00h

- Presentations and concluding remarks (German Rigau)
  - ??: 15:00h – 18:00h

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

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

Evaluation

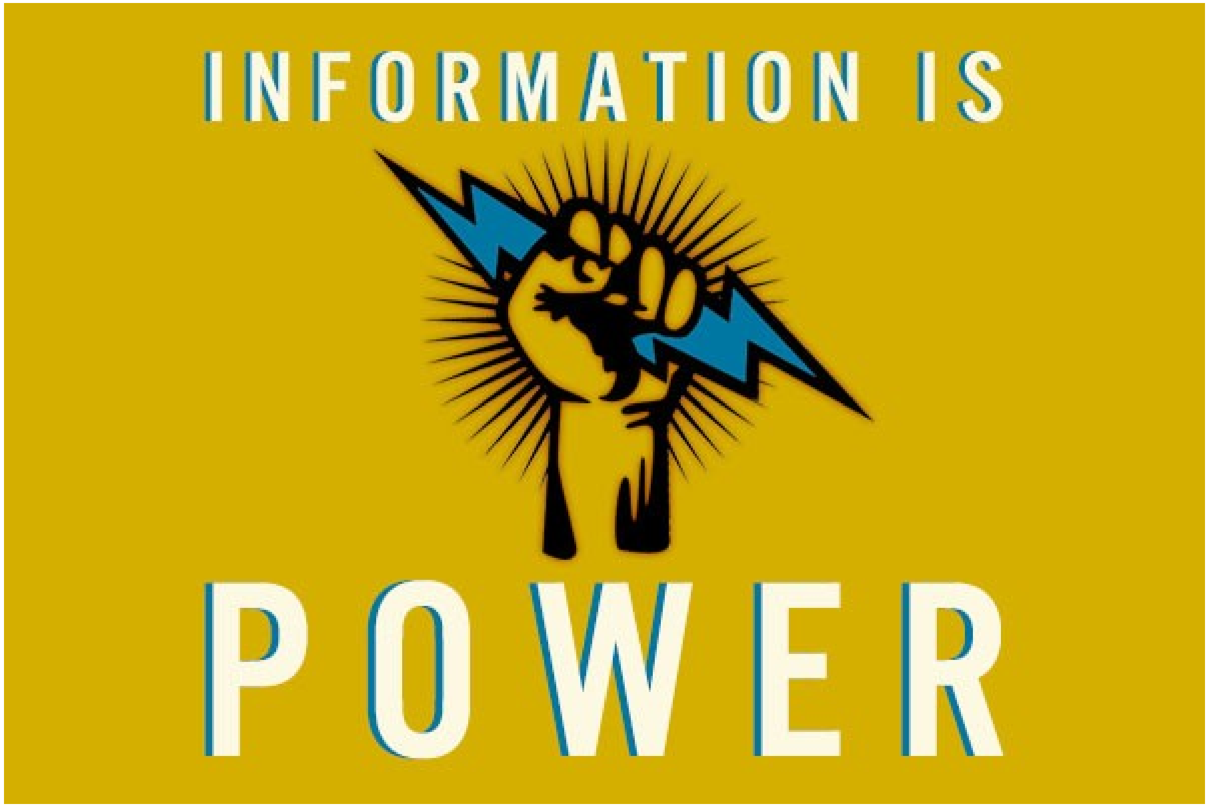
- Application of empirical methods for NLP:
  - Teacher exercises (30%)

  - Teacher/student topic
  - Short presentation (20%)
    - 15 minutes sharp, ~ 10 slides
    - Presentation: 20/06
  - Written report (50%)
    - Format: http://www.acl2019.org
    - Deadline Report: 20/07
    - Short paper describing an <u>experimental work</u>
      - ~ 3000 words

Short Motivation

Information is power!

Short Motivation



Short Motivation

Knowledge is power!

Short Motivation

KNOWLEDGE

IS

POWER

Short Motivation

Knowledge is power!

... and the knowledge to use ...

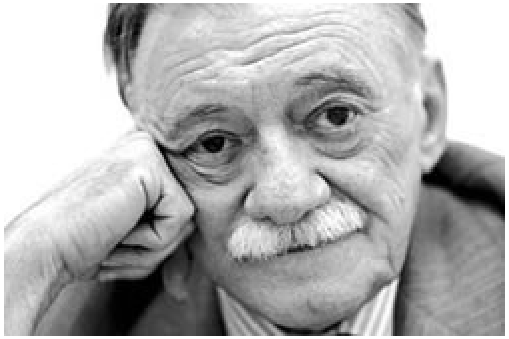
Foreword

“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

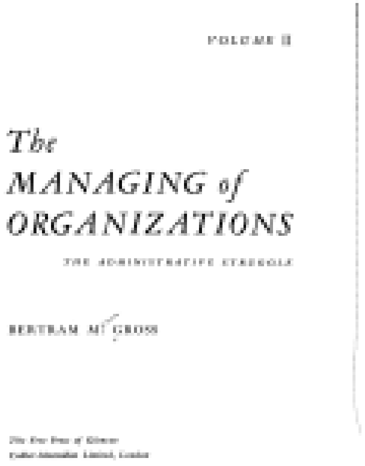
- Where are the **answers** to the new (and old) questions?
  - Introspection? Experts?…
  - From many people? … “Wisdom of the Crowd”?
  - Books, News, Tweets, … Textual Sources?
  - Multimedia sources? Images, Radio, TV ...
  - Sensors? IoT? ...
  - Anything? Everything?
- Information **overload**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

Foreword

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

Foreword

- **Information overload** …
  - infobesity, infoxication!
  - by Bertram Gross, <u>The Managing of Organizations: The administrative struggle</u> (1964)



VOLUME II

The
MANAGING of
ORGANIZATIONS

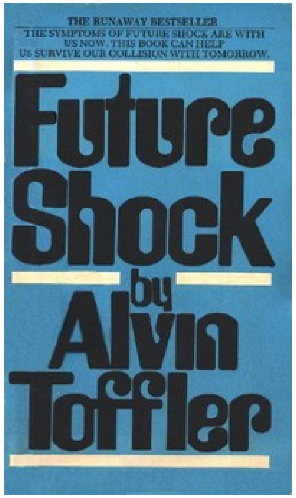
THE ADMINISTRATIVE STRUGGLE

BERTRAM M. GROSS

The Free Press of Glencoe
Collier-Macmillan Limited, London

Foreword

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


THE RUNAWAY BESTSELLER
THE SYMPTOMS OF FUTURE SHOCK ARE WITH US NOW. THIS BOOK CAN HELP US SURVIVE OUR COLLISION WITH TOMORROW.
Future Shock
by
Alvin Toffler

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!

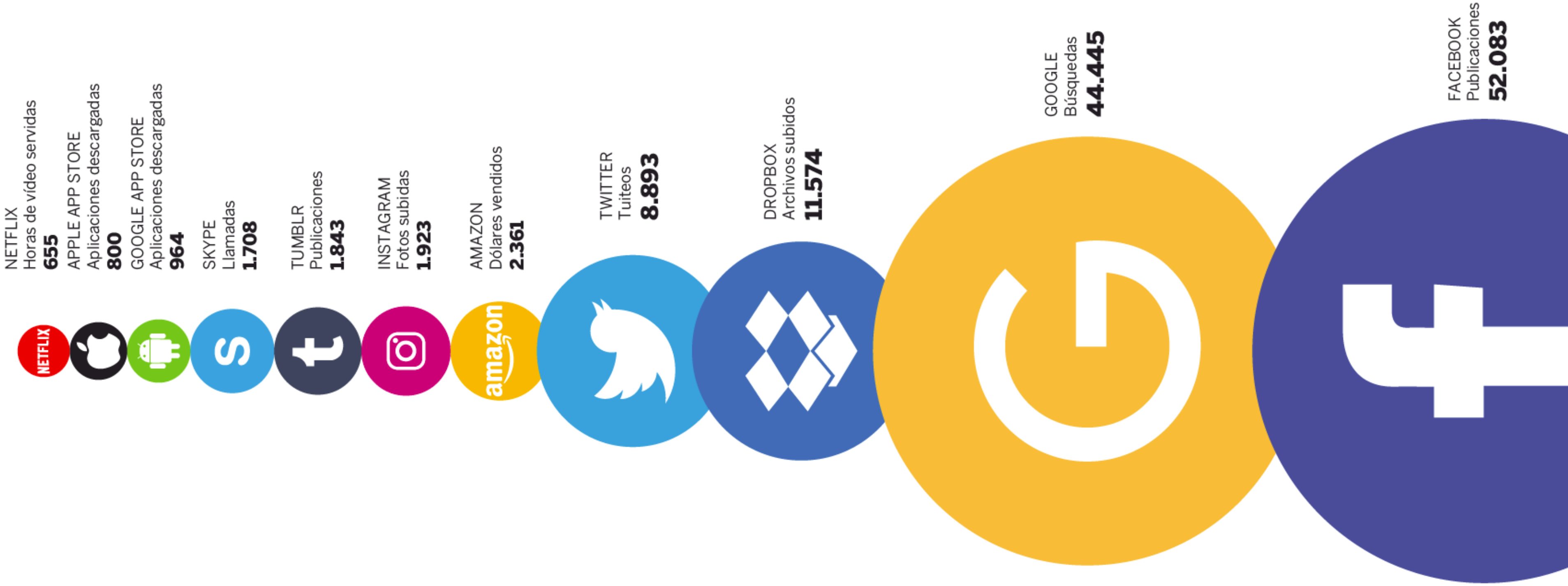
Foreword

- **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.

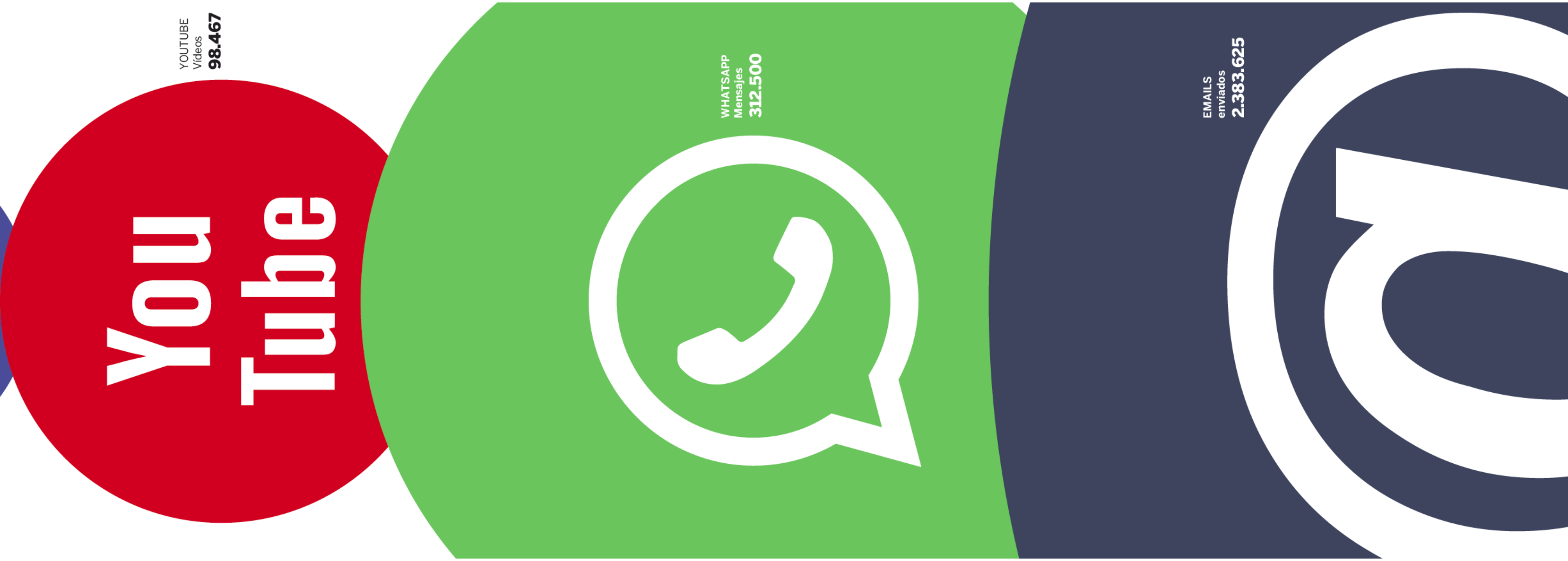
Foreword

- 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] …

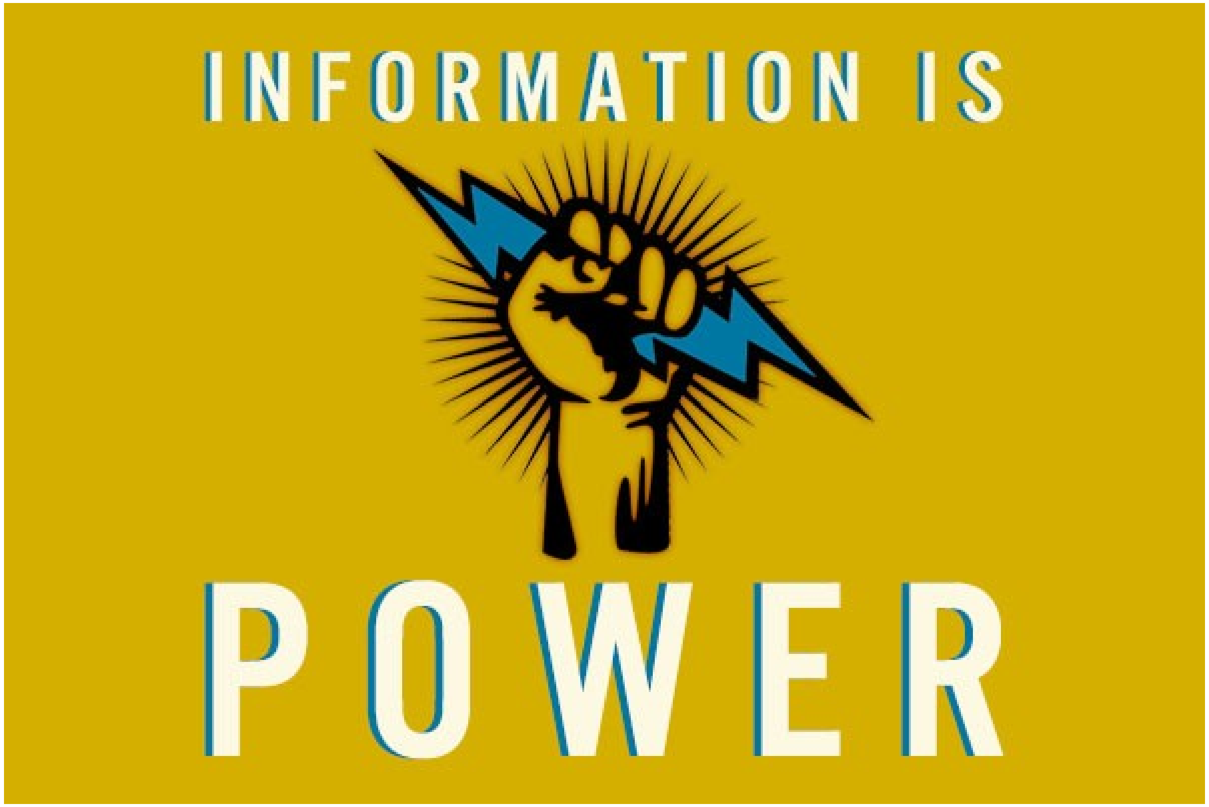
- Usually in the form of **texts** and documents in **multiple languages** …

- **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

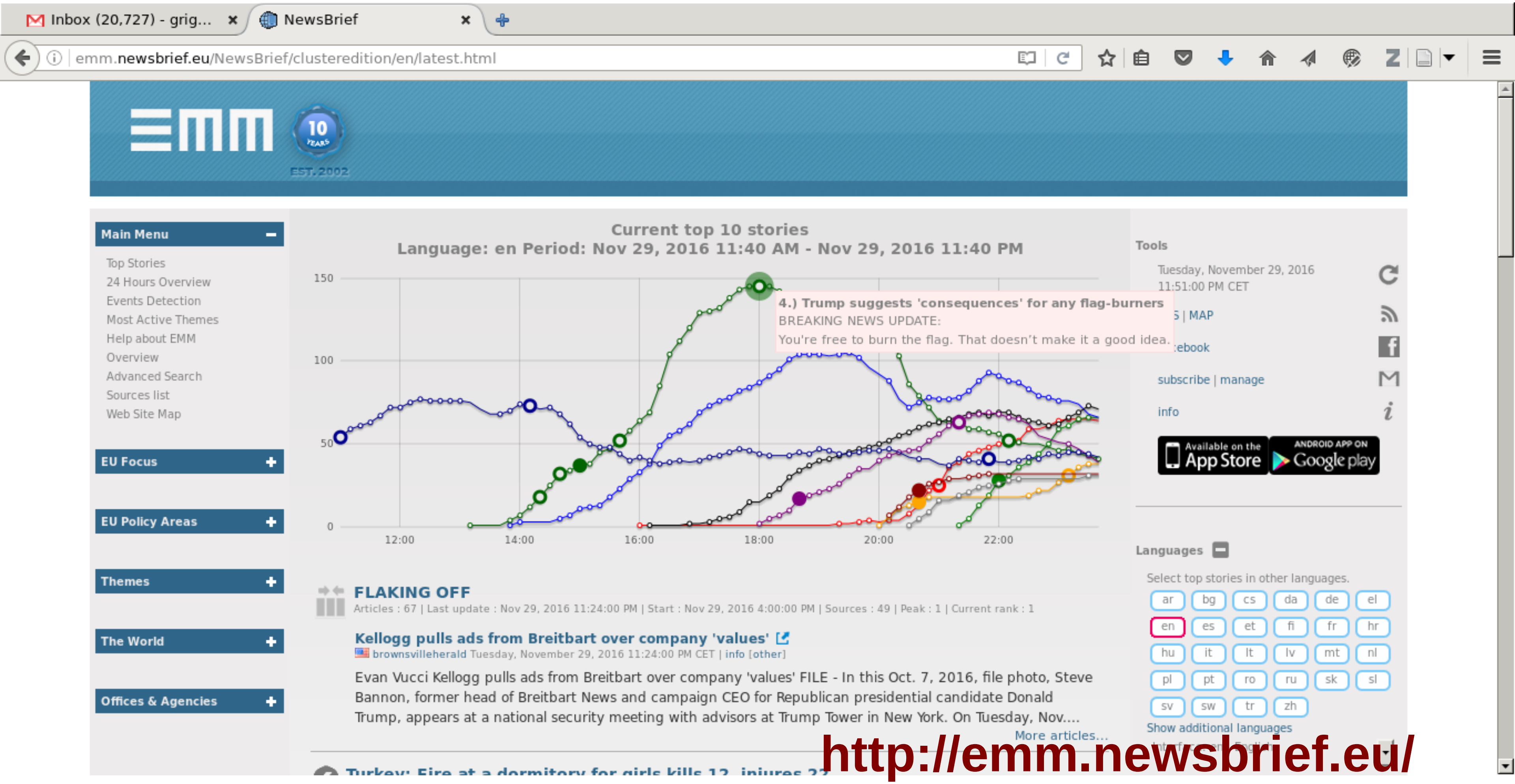
POWER

Foreword

e.g. IBM Watson ...

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

Big Data & NLP ...



Inbox (20,727) - grig... NewsBrief

emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html

Current top 10 stories
Language: en Period: Nov 29, 2016 11:40 AM - Nov 29, 2016 11:40 PM

4.) Trump suggests 'consequences' for any flag-burners
BREAKING NEWS UPDATE:
You're free to burn the flag. That doesn't make it a good idea.

FLAKING OFF

Articles : 67 | Last update : Nov 29, 2016 11:24:00 PM | Start : Nov 29, 2016 4:00:00 PM | Sources : 49 | Peak : 1 | Current rank : 1

Kellogg pulls ads from Breitbart over company 'values'
brownsvilleherald Tuesday, November 29, 2016 11:24:00 PM CET | info [other]

Evan Vucci Kellogg pulls ads from Breitbart over company 'values' FILE - In this Oct. 7, 2016, file photo, Steve Bannon, former head of Breitbart News and campaign CEO for Republican presidential candidate Donald Trump, appears at a national security meeting with advisors at Trump Tower in New York. On Tuesday, Nov....

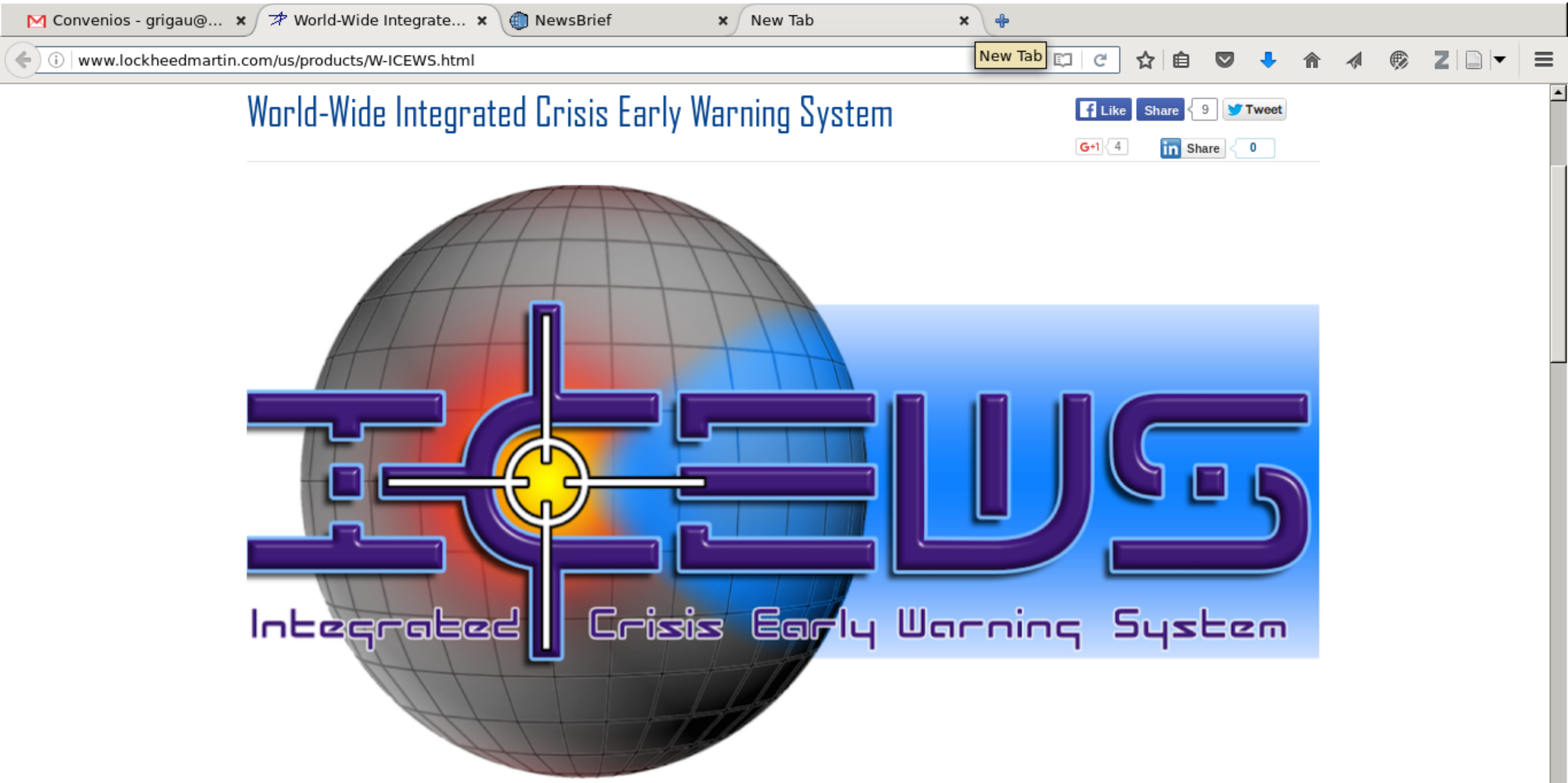
More articles...

Turkey: Fire at a dormitory for girls kills 12, injures 22

Tools
Tuesday, November 29, 2016
11:51:00 PM CET
S | MAP
:ebook
subscribe | manage
info
Available on the App Store
ANDROID APP ON Google play
Languages
Select top stories in other languages.
ar
bg
cs
da
de
el
en
es
et
fi
fr
hr
hu
it
lt
lv
mt
nl
pl
pt
ro
ru
sk
sl
sv
sw
tr
zh
Show additional languages

http://emm.newsbrief.eu/

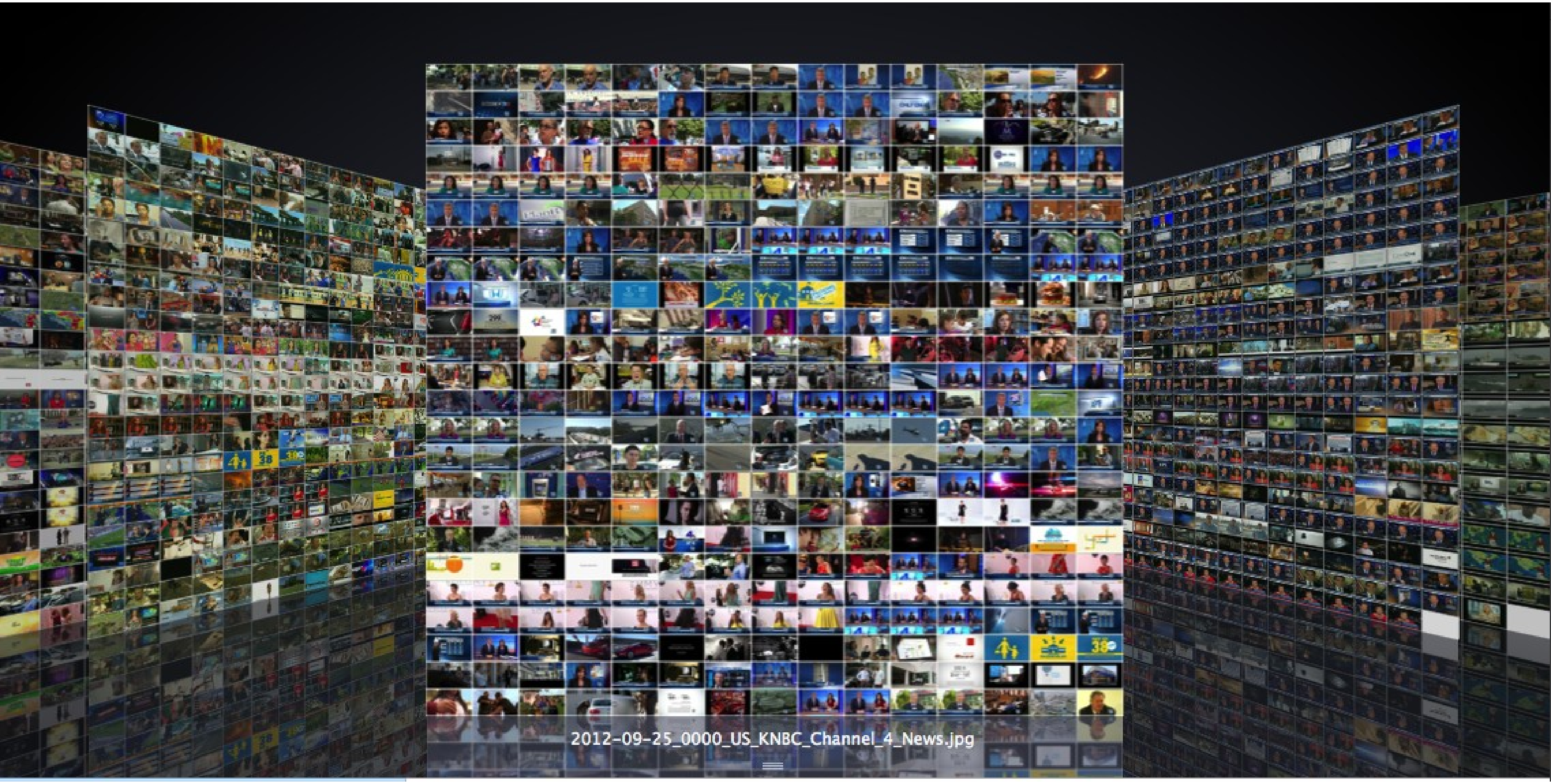
Big Data & NLP ...



World-Wide Integrated Crisis Early Warning System

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

Big Data & NLP ...



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

Advanced Methods for Corpus Processing

Lluís Padró <padro@lsi.upc.edu>
Lluís Màrquez <lluismv@amazon.com>
German Rigau <german.rigau@ehu.es>

Centre de Tecnologies i Aplicacions
del Llenguatge i la Parla