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 20/02
    - Discussion by email
    - Oral presentations (of design, current status)
    - => Deadline 12/03 and 13/03
    - 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
  - Word embeddings

Evaluation

- Applications of Information Access and Text Mining:
  - Student/teacher topic
  - Short presentation
    - 10 minutes sharp, ~ 10 slides
    - Presentation: **12/03 and 13/13**
  - Written report:
    - Format: http://www.acl2013.org/site/call.html
    - Deadline Report: **??/??**
    - Short paper describing an experimental work
      - < 3000 words

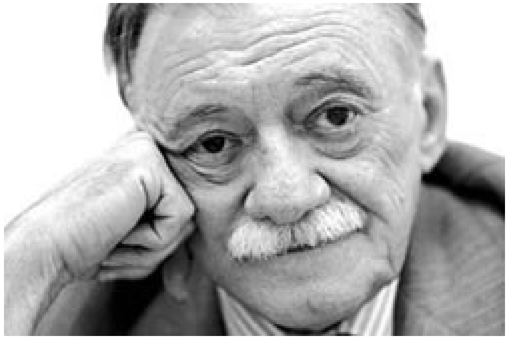
Foreword

*“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

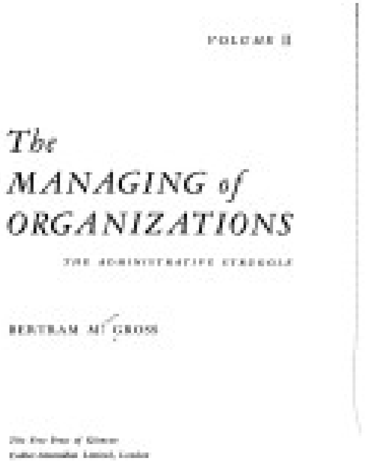
- Where are the **answers** to the new (and old) questions?
  - Introspection? Experts?…
  - From many people? … “Wisdom of the Crowd”?
  - Books, News, Tweets, … Textual Sources?
  - Multimedia sources? Images, Radio, TV ...
  - Sensors? IoT? ...
  - Anything? Everything?
- Information **overload**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

Foreword

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

Foreword

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



VOLUME II

The
MANAGING of
ORGANIZATIONS

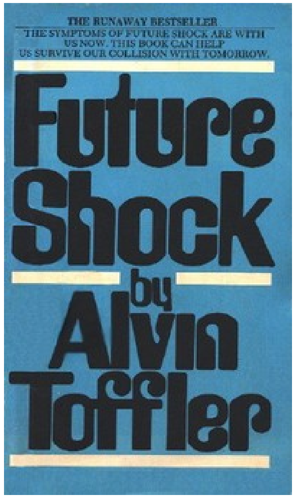
THE ADMINISTRATIVE STRUGGLE

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!

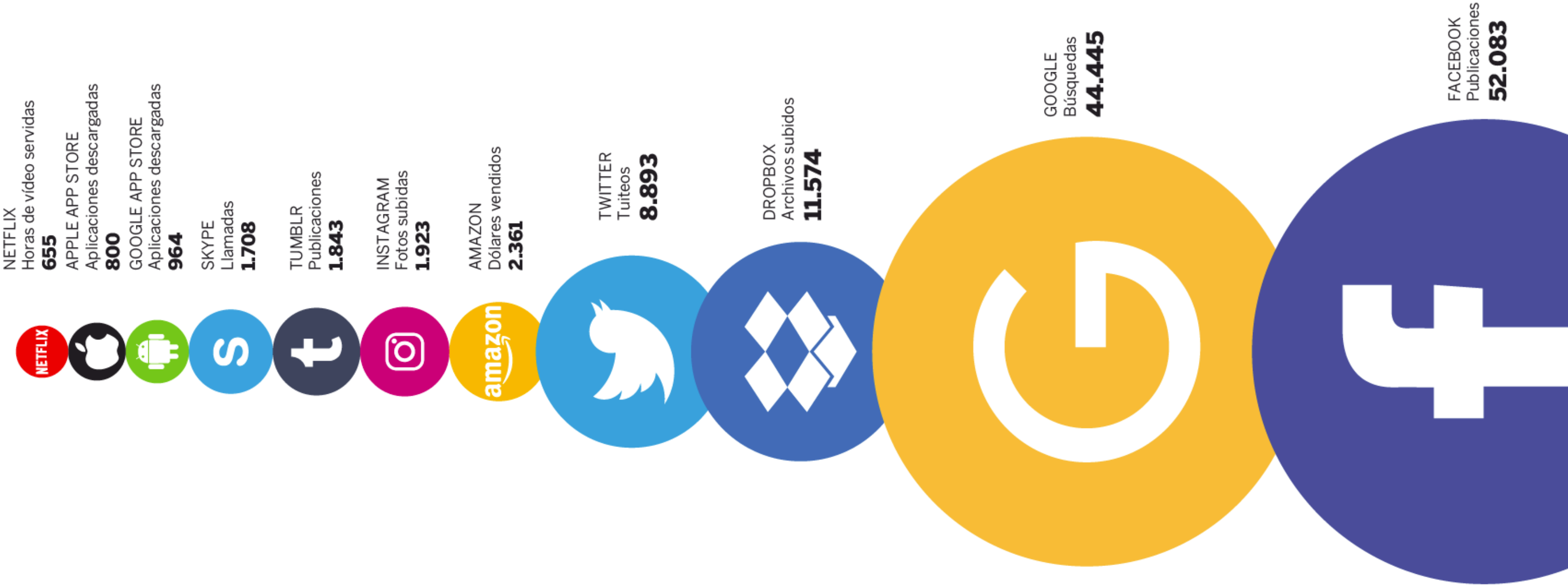
Foreword

- **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.

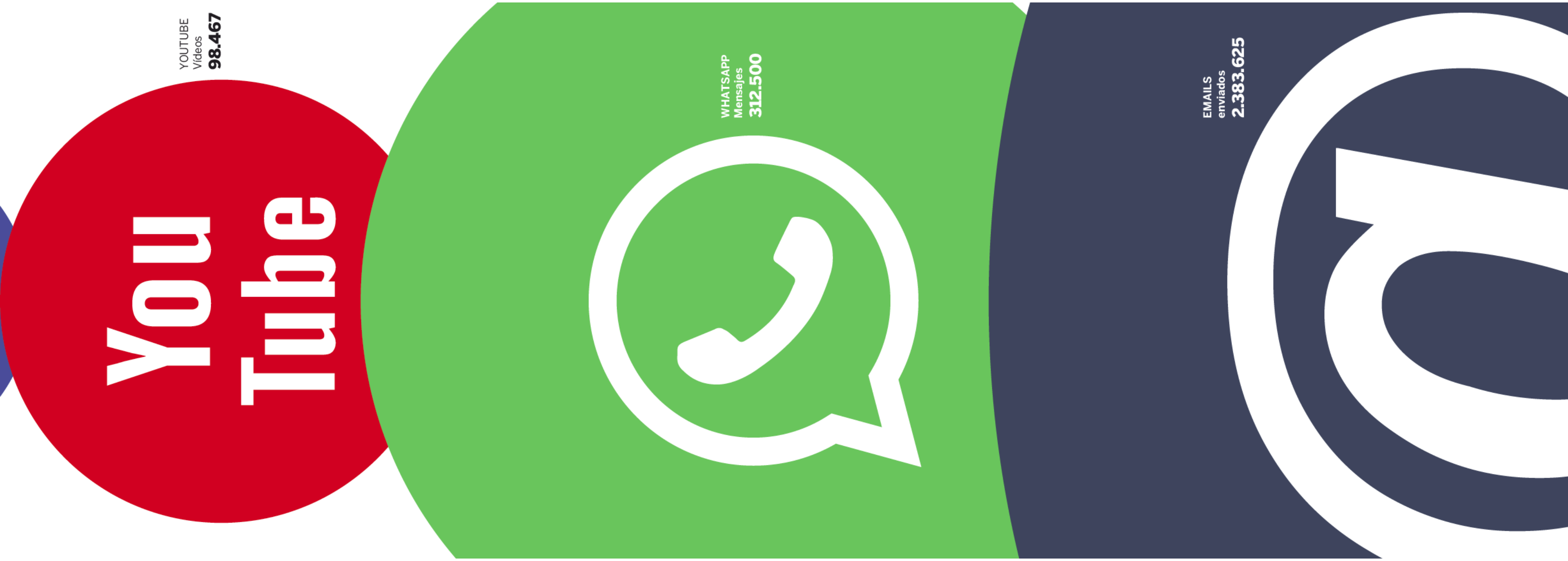
Foreword

- 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 …
- …

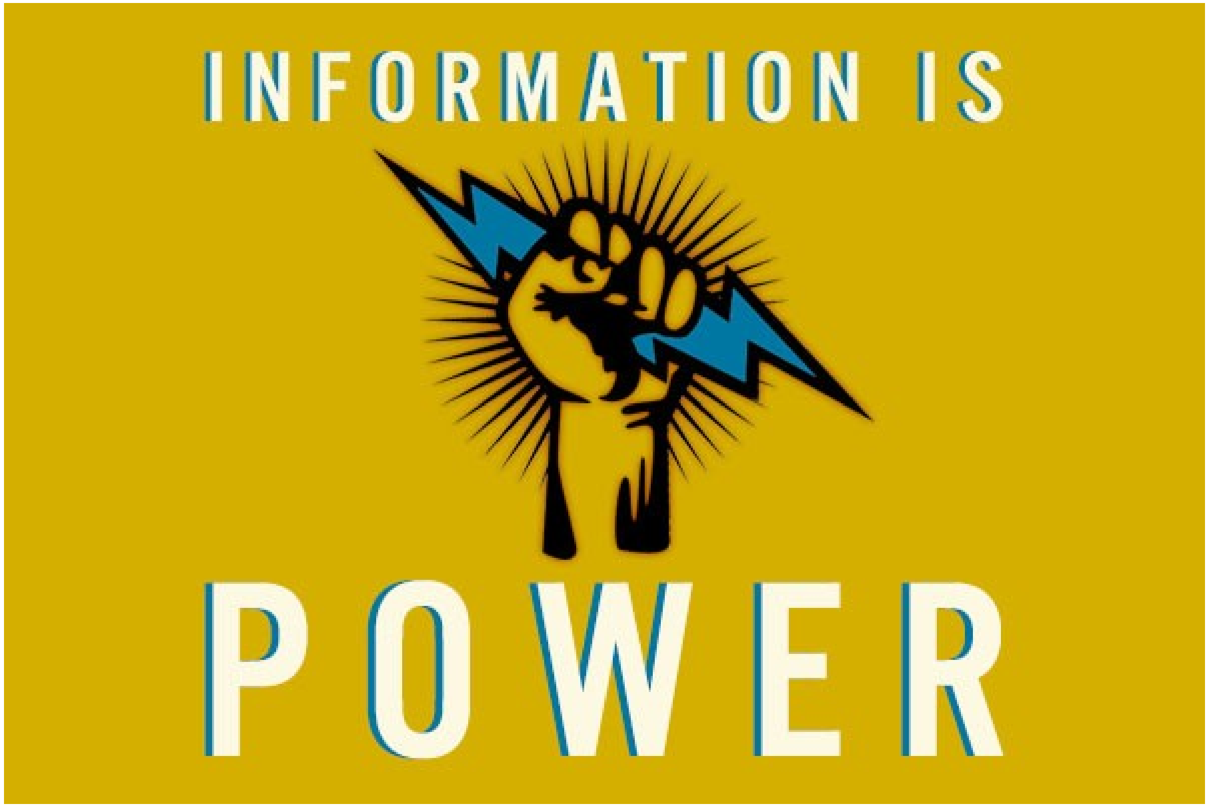
- … all kinds of e-documents …

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

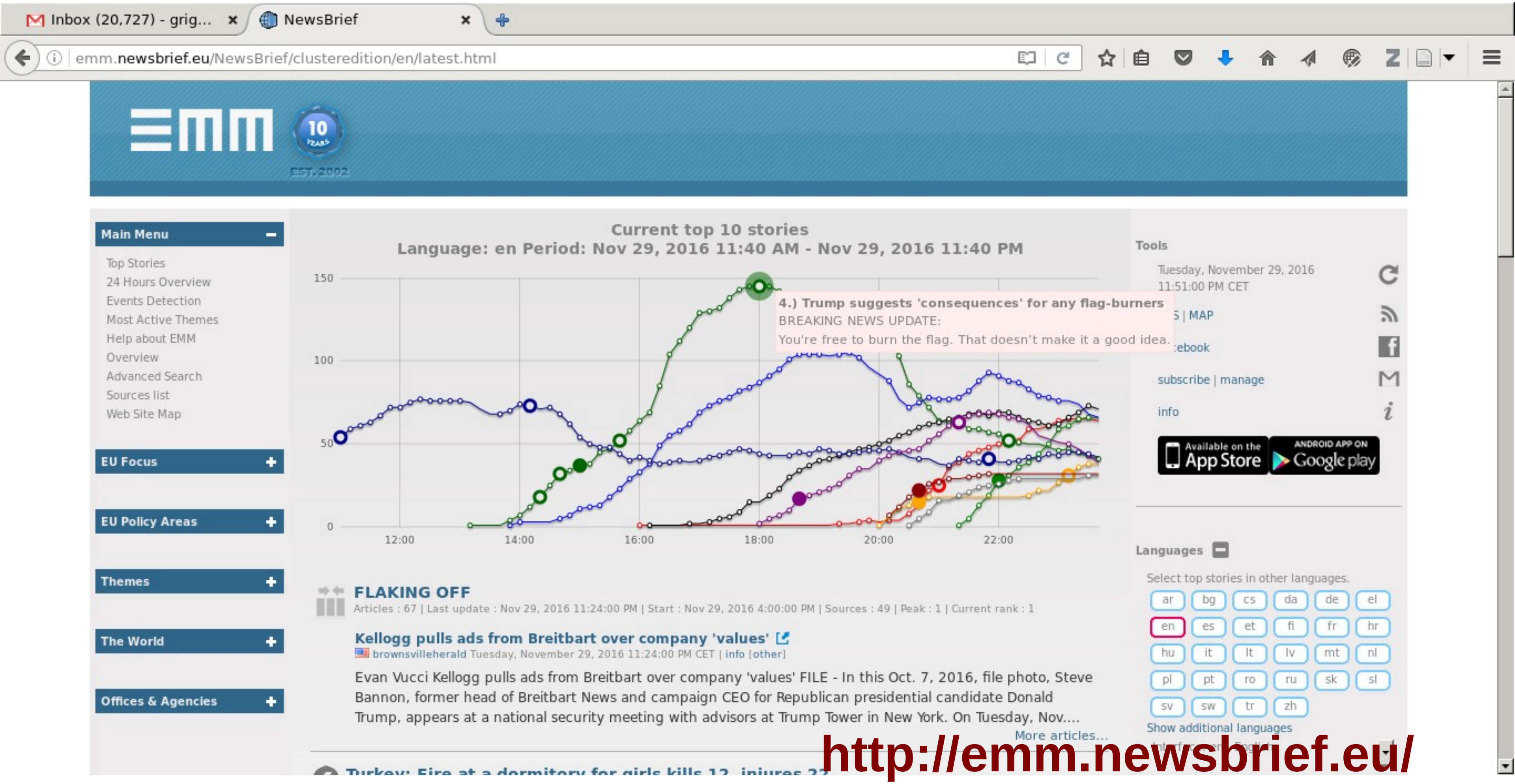
POWER

Foreword

e.g. IBM Watson ...

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

Big Data & NLP …



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.

Tuesday, November 29, 2016
11:51:00 PM CET

Select top stories in other languages.

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]

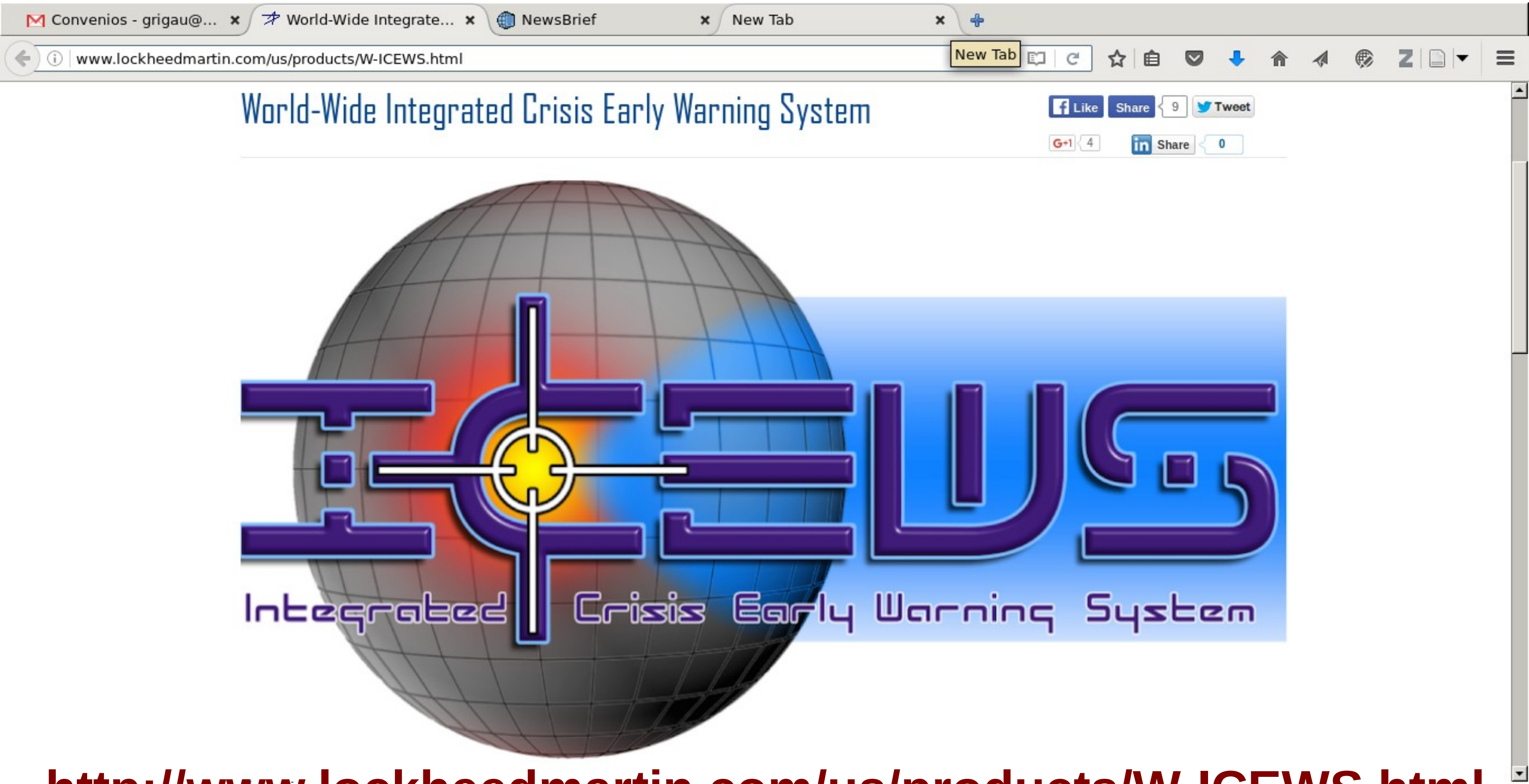
Evan Vucci Kellogg pulls ads from Breitbart over company 'values' FILE - In this Oct. 7, 2016, file photo, Steve Bannon, former head of Breitbart News and campaign CEO for Republican presidential candidate Donald Trump, appears at a national security meeting with advisors at Trump Tower in New York.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

More articles…

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

http://emm.newsbrief.eu/

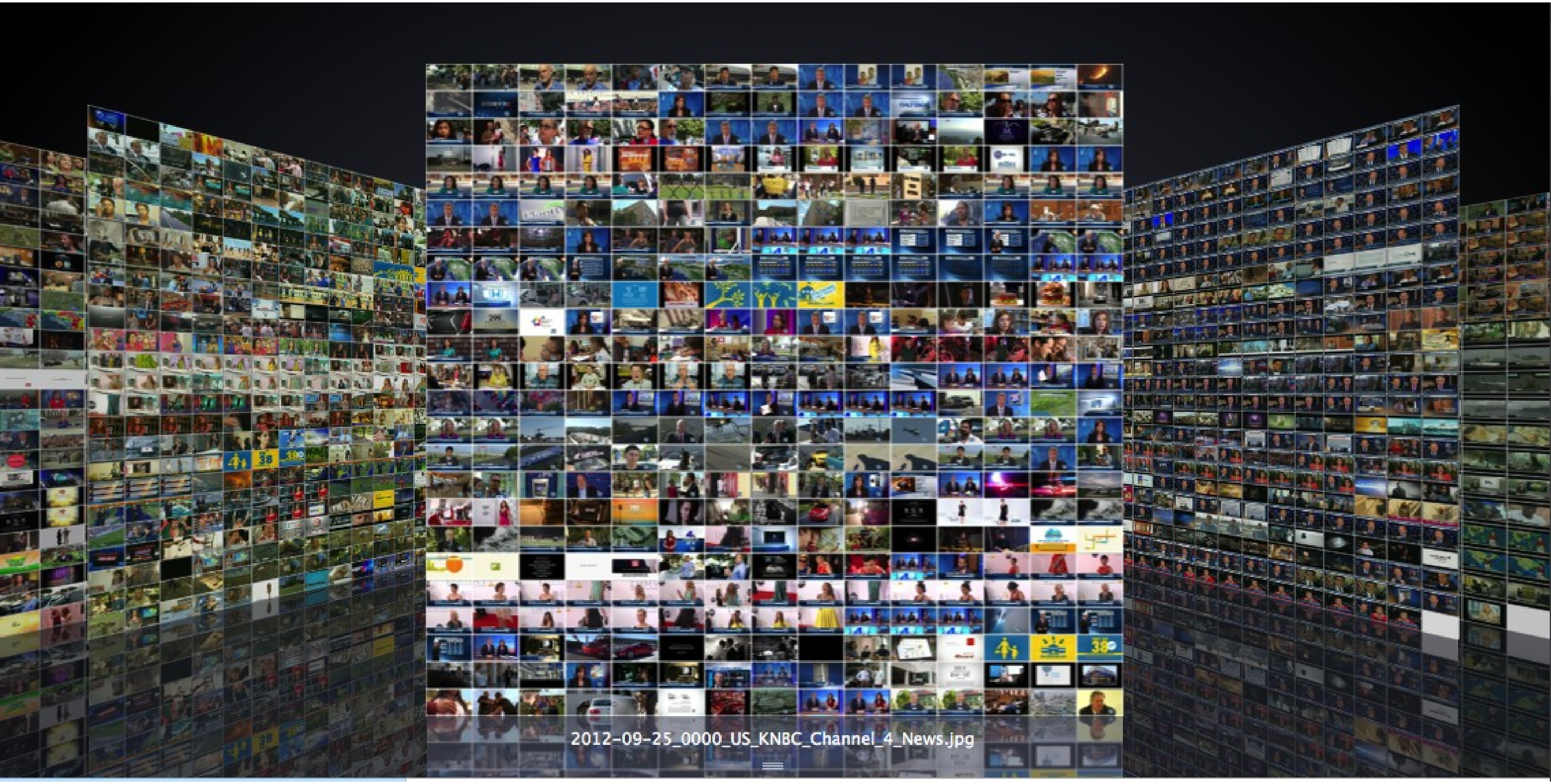
Big Data & NLP ...



Convenios - grigau@...
World-Wide Integrate...
NewsBrief
New Tab
www.lockheedmartin.com/us/products/W-ICEWS.html
New Tab
World-Wide Integrated Crisis Early Warning System
Like
Share
9
Tweet
G+1
4
Share
0
ICEWS
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>