Science and Technology in digital newspapers

Carlos G. Figuerola, Tamar Groves, Miguel Angel Quintanilla - ECyT Institute
University of Salamanca
II Seminar on Indicators of Scientifc and Technological Culture - 25/11/2014

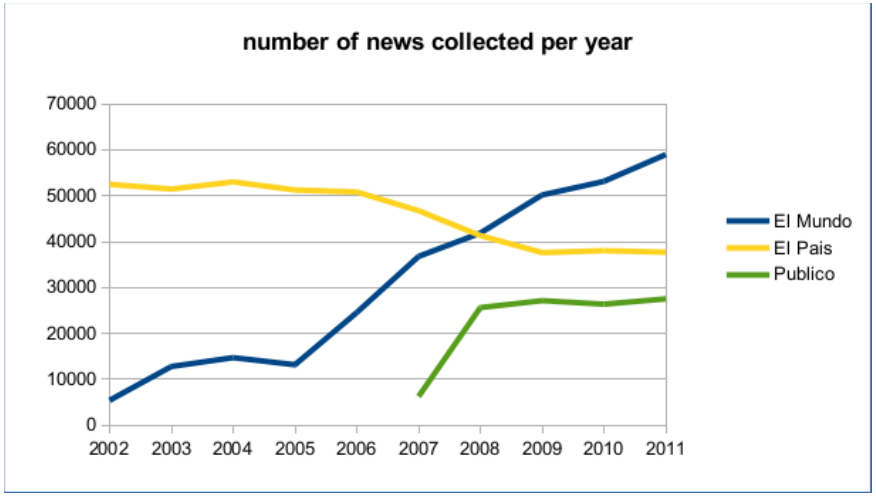
Digital Newspapers

- not as physical newspapers
- heterogeneous formats
- heterogeneous web site structures
- concerns with digital preservation

EL MUNDO
EL PAÍS
Público.es

Digital Newspapers

- Three newspapers: El Mundo, El País, Público
- Time period: 2002-2011 (except Público, only since 2007)
- More than 900.000 news



Automatic Categorization

We are only interested on news about Science & Technology

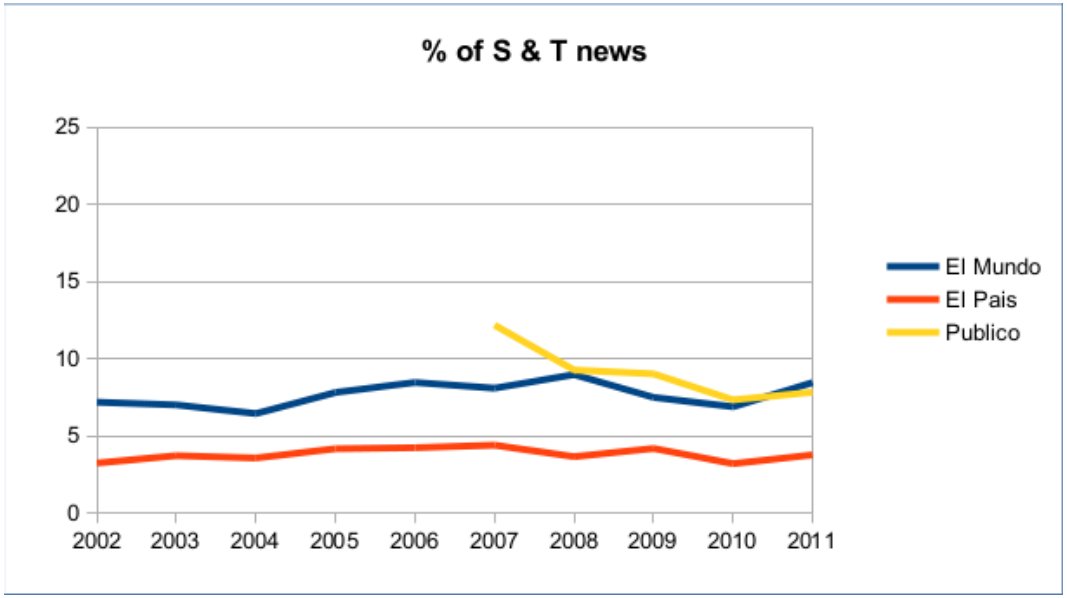
- we can use an automatic supervised classifier
- SVM is a good choice
- we can try also SVM to classify news in the categories of our theorethic model

Training Process

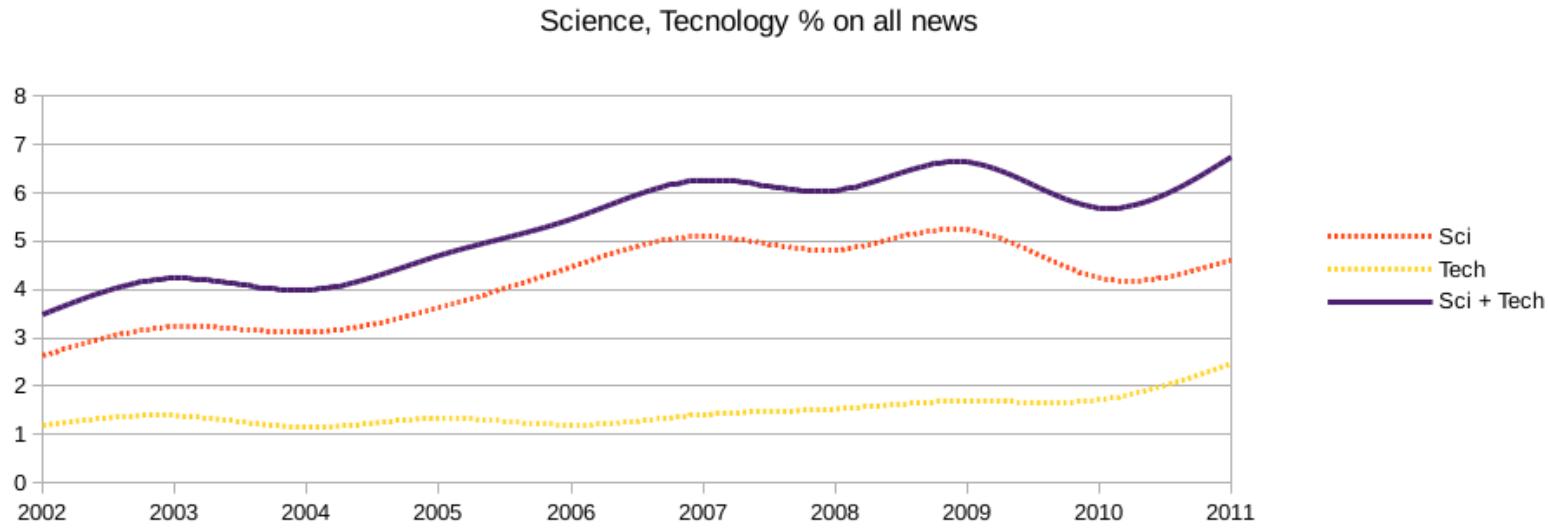
- an initial sample built by hand
- an iterative process of classify - refining sample - retraining - reclassify

Results: the SCSC

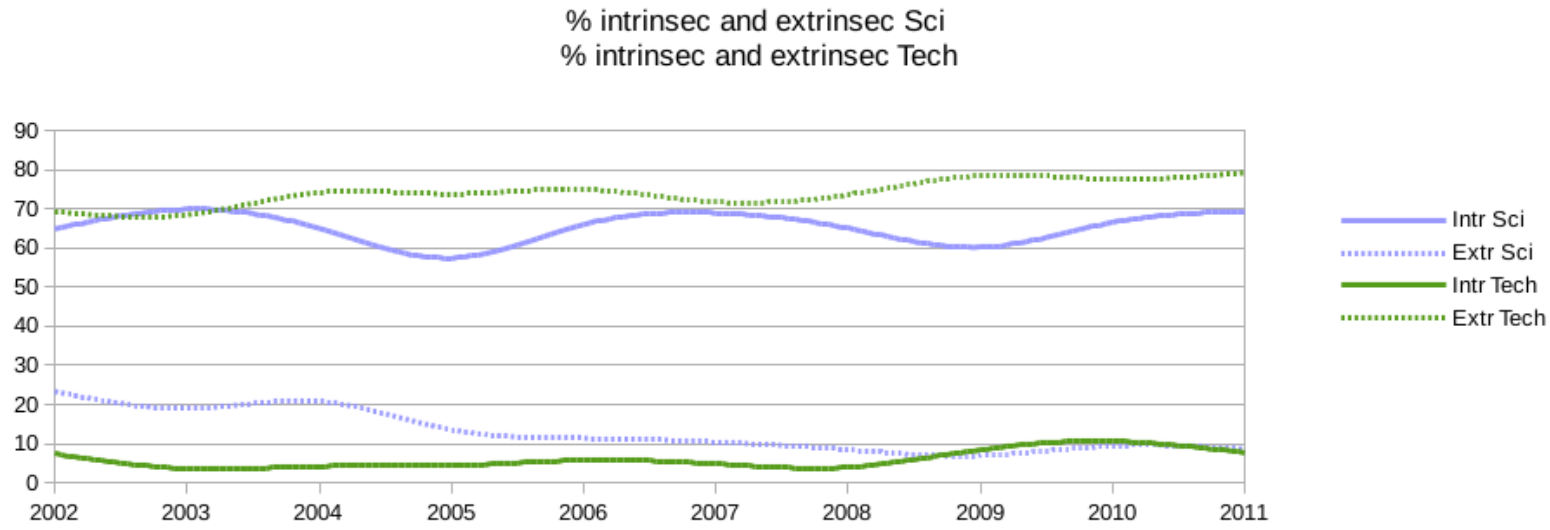
50,753 news about S & T



More Results: Science vs. Technology



Intrinsec and extrinsec features



Topics Discovering using SNA Techniques

- objects can establish relationships between them
- we can map objects and relationships towards a network or graph
    - objets are nodes
    - relationships are edges or links between nodes

Establishing relationships between news

- we can compute semantic simmilarity between documents
  - using borrowed techniques from the Information Retrieval field
  - applying the well known Vector Space Model
  - based on words and weights of each word inside each document
- news are nodes in a network
- there is an edge between two docs if they are simmilar
- the weight of this edge is the simmilarity's degree between both docs

Detecting Communities

- in a network, a community is a bunch of nodes
    - strongly linked between them
    - links weakly with nodes outside the bunch
- in our network of news, a community is a topic
- they are several algorithms to find communities in networks
- we use InfoMap: fast and efficient, accurate results

Analyzing Results

Communities listing

community	topic
1	Public Health
2	Biomedicine
3	Energy
4	Human Development
5	Natural Resources
6	Aerospace Research
7	Biodiversity
8	Astronomy & Cosmology
9	Information Technology
10	Science Policy
11	Protected Species - Spain
12	Human Evolution
13	Contamination

Analyzing Results

Subcommunity	Topic	Subcommunity	Topic
1.1	influenza	1.11	infections, E. Coli,
1.2	AIDS	1.12	cholera
1.3	mortality	1.13	Legionella
1.4	drugs	1.14	polio
1.5	vaccines	1.15	mad cow disease
1.6	malaria	1.16	foot and mouth disease
1.7	SARS	1.17	dengue
1.8	tuberculosis	1.18	insect infections
1.9	hepatitis C	1.19	Chagas
1.10	antibiotics, bacteria	1.20	bio-bac

Conclusions

- more Sci than Tech
- in Sci news more intrinsecallity
- predominance of academic model of science communication
    - journalists tend to reproduce scientific information and they don't enter into questions of its social political or moral implications
- topics:
    - predominance of biomedicine
    - progressive growing of Information technologies
    - specific events produce punctual growth in news about ecology, pollution, ...

Conclusions: big data treatment

- We tried using automated information retrieval procedures to recuperate science news and several kinds of specialized software to classify and analyze it.
- Their usage was efficient in analyzing our vast corpus and reaching some preliminary conclusions.
- However we are left with the challenge of explaining the high number of unclassified articles related to our model.
- There is a need to analyze more carefully the sub clusters and their significance.

<Thank You!>

Important contact information goes here.

e-mail

www