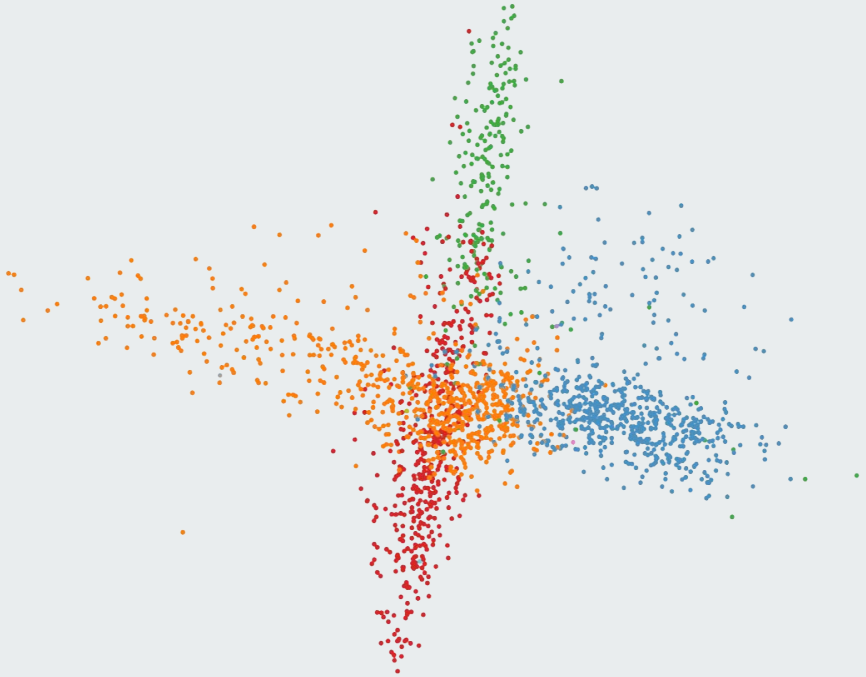
GraphLang



NTDS 2017-18 Final Project

Grégoire CLEMENT, Maxime DELISLE, Charles GALLAY and Ali HOSSEINY

What does GraphLang do?

- Extracts main concepts from text document
- Label text documents

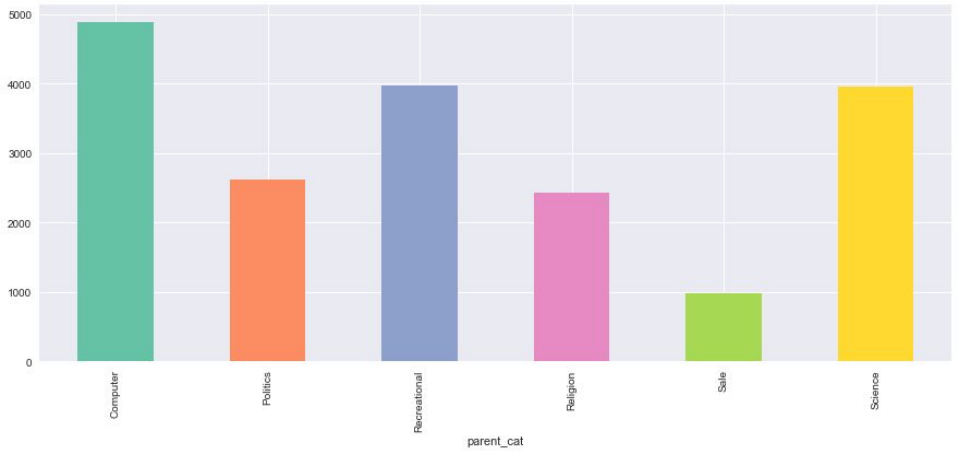
Structure of presentation

- Datasets used
- GraphLang v.1
- GraphLang v.2
- Results and shortcomings

Datasets used

20 Newsgroups:

- Around 20,000 newsgroup documents
- 6 categories
- 20 subcategories



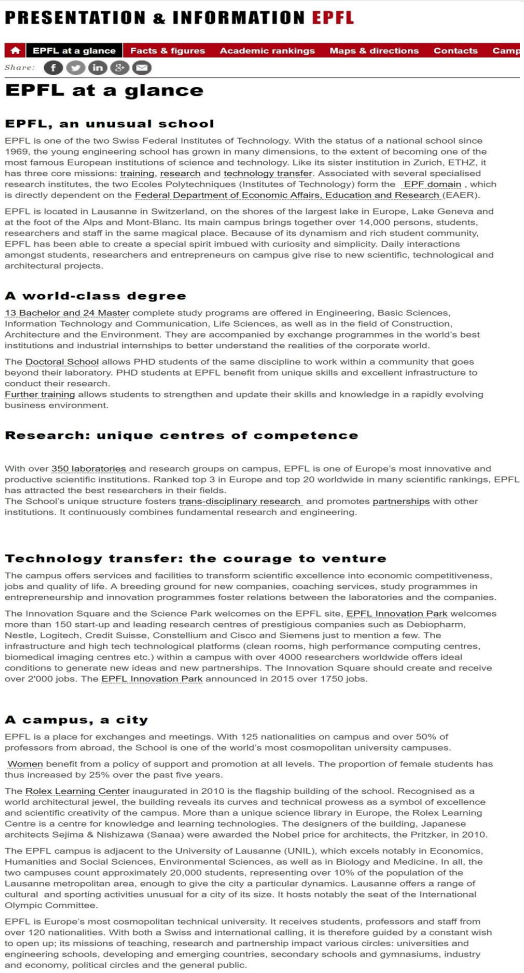
GraphLang v.1

Goal

- Leverage the power of graphs on texts
- Getting insights about text structure
- Automatic and unsupervised extraction of the most important concepts
- Interest is in on analyzing, not comparing
- Should be simple yet powerful

Does it work ?

- Someone wants to discover what that EPFL is
- In a time rush, just interested in the core ideas
- Sees description, seems nice but lengthy
- Stumbles upon GraphLang
- “Let me summarize that for you”



PRESENTATION & INFORMATION EPFL

EPFL at a glance | Facts & figures | Academic rankings | Maps & directions | Contacts | Camp

Share:

EPFL at a glance

EPFL, an unusual school

EPFL is one of the two Swiss Federal Institutes of Technology. With the status of a national school since 1969, the young engineering school has grown in many dimensions, to the extent of becoming one of the most famous European institutions of science and technology. Like its sister institution in Zurich, ETHZ, it has three core missions: training, research and technology transfer. Associated with several specialised research institutes, the two Ecoles Polytechniques (Institutes of Technology) form the EPF domain , which is directly dependent on the Federal Department of Economic Affairs, Education and Research (EAER).

EPFL is located in Lausanne in Switzerland, on the shores of the largest lake in Europe, Lake Geneva and at the foot of the Alps and Mont-Blanc. Its main campus brings together over 14,000 persons, students, researchers and staff in the same magical place. Because of its dynamism and rich student community, EPFL has been able to create a special spirit imbued with curiosity and simplicity. Daily interactions amongst students, researchers and entrepreneurs on campus give rise to new scientific, technological and architectural projects.

A world-class degree

13 Bachelor and 24 Master complete study programs are offered in Engineering, Basic Sciences, Information Technology and Communication, Life Sciences, as well as in the field of Construction, Architecture and the Environment. They are accompanied by exchange programmes in the world's best institutions and industrial internships to better understand the realities of the corporate world.

The Doctoral School allows PHD students of the same discipline to work within a community that goes beyond their laboratory. PHD students at EPFL benefit from unique skills and excellent infrastructure to conduct their research.
Further training allows students to strengthen and update their skills and knowledge in a rapidly evolving business environment.

Research: unique centres of competence

With over 350 laboratories and research groups on campus, EPFL is one of Europe's most innovative and productive scientific institutions. Ranked top 3 in Europe and top 20 worldwide in many scientific rankings, EPFL has attracted the best researchers in their fields.
The School's unique structure fosters trans-disciplinary research and promotes partnerships with other institutions. It continuously combines fundamental research and engineering.

Technology transfer: the courage to venture

The campus offers services and facilities to transform scientific excellence into economic competitiveness, jobs and quality of life. A breeding ground for new companies, coaching services, study programmes in entrepreneurship and innovation programmes foster relations between the laboratories and the companies.

The Innovation Square and the Science Park welcomes on the EPFL site, EPFL Innovation Park welcomes more than 150 start-up and leading research centres of prestigious companies such as Debiopharm, Nestle, Logitech, Credit Suisse, Constellium and Cisco and Siemens just to mention a few. The infrastructure and high tech technological platforms (clean rooms, high performance computing centres, biomedical imaging centres etc.) within a campus with over 4000 researchers worldwide offers ideal conditions to generate new ideas and new partnerships. The Innovation Square should create and receive over 2'000 jobs. The EPFL Innovation Park announced in 2015 over 1750 jobs.

A campus, a city

EPFL is a place for exchanges and meetings. With 125 nationalities on campus and over 50% of professors from abroad, the School is one of the world's most cosmopolitan university campuses.

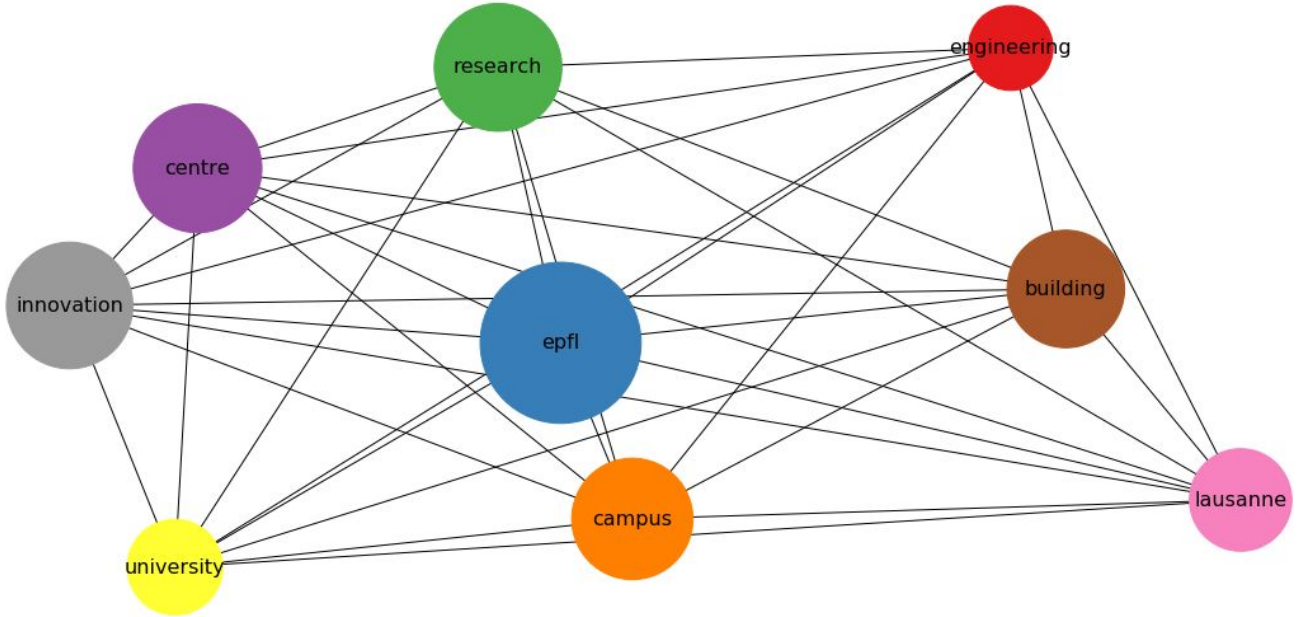
Women benefit from a policy of support and promotion at all levels. The proportion of female students has thus increased by 25% over the past five years.

The Rolex Learning Center inaugurated in 2010 is the flagship building of the school. Recognised as a world architectural jewel, the building reveals its curves and technical prowess as a symbol of excellence and scientific creativity of the campus. More than a unique science library in Europe, the Rolex Learning Centre is a centre for knowledge and learning technologies. The designers of the building, Japanese architects Sejima & Nishizawa (Sanaa) were awarded the Nobel price for architects, the Pritzker, in 2010.

The EPFL campus is adjacent to the University of Lausanne (UNIL), which excels notably in Economics, Humanities and Social Sciences, Environmental Sciences, as well as in Biology and Medicine. In all, the two campuses count approximately 20,000 students, representing over 10% of the population of the Lausanne metropolitan area, enough to give the city a particular dynamics. Lausanne offers a range of cultural and sporting activities unusual for a city of its size. It hosts notably the seat of the International Olympic Committee.

EPFL is Europe's most cosmopolitan technical university. It receives students, professors and staff from over 120 nationalities. With both a Swiss and international calling, it is therefore guided by a constant wish to open up; its missions of teaching, research and partnership impact various circles: universities and engineering schools, developing and emerging countries, secondary schools and gymnasiums, industry and economy, political circles and the general public.

Does it work ? Extracted concepts



How GraphLang V1 works

- Preprocessing
- Graph construction
- Graph analysis

Preprocessing

- First step in the pipeline
- Unwanted chunks removed
- Text normalization
  - Stemmatization / Lemmatization
  - Lowercasing

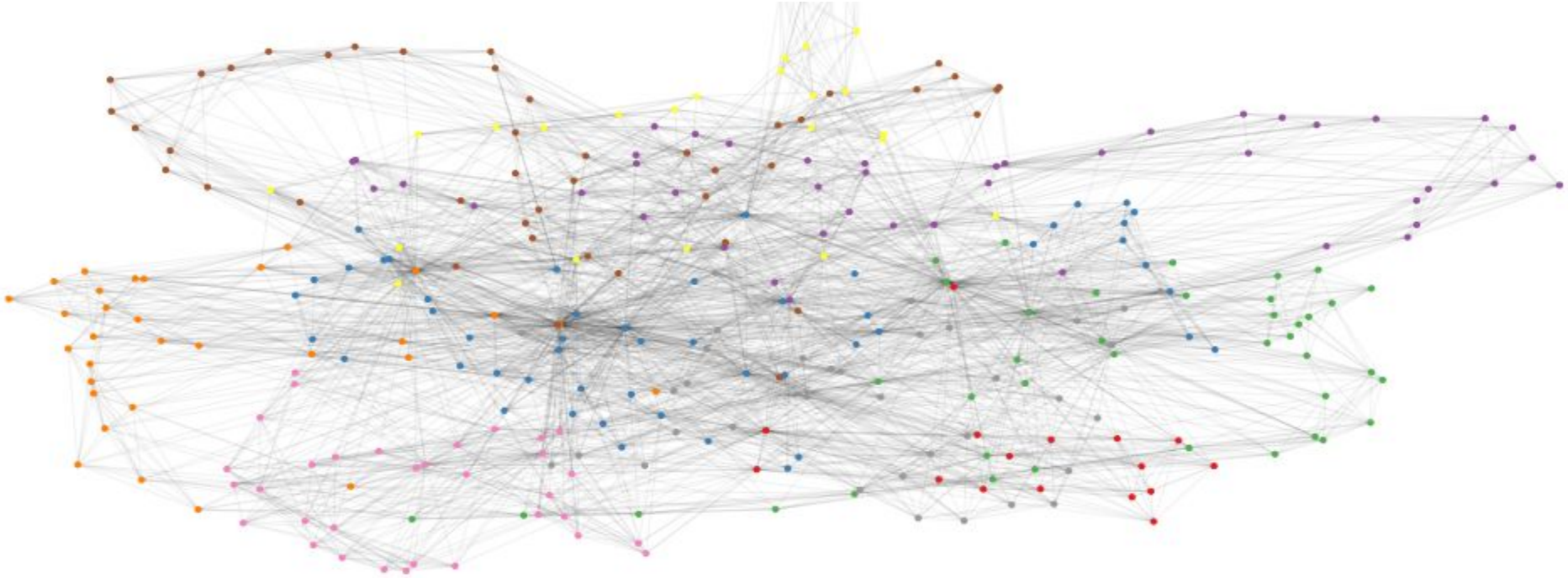
Graph Construction

- Cooccurences undirected graph
- Parametrizable window size
  - Distance taken into account for edge weights
- Stopwords considered but discarded
  - “**Banana** and **Milkshake**” cooccurence is weaker than “**Banana Milkshake**”

Graph analysis

- **Betweenness centrality** for all nodes
- **Community partition** of the graph
  - The partition of the graph nodes which maximises the modularity
  - Based on betweenness values
  - Uses Louvain heuristics

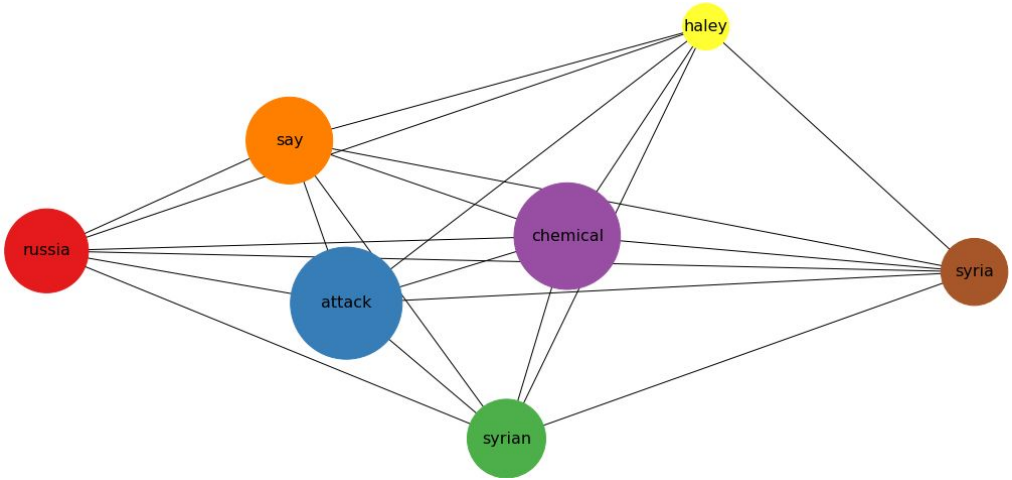
What we have so far

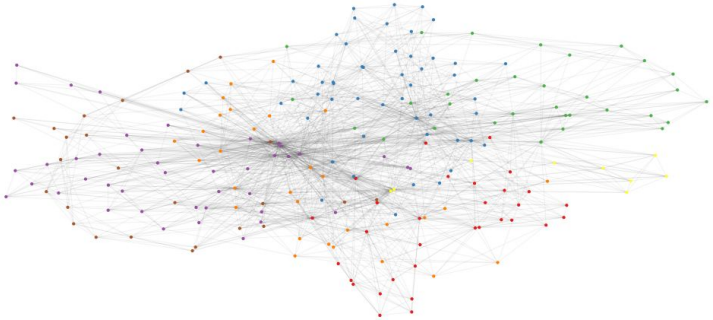


Graph analysis - Final step

- Construction of induced graph
  - Each node represents a community
- Induced graph is a “summary” of the original graph
- Nodes labelled as the most important node (word) of the corresponding community
  - Each core idea conveyed through this word
  - Importance by total degree

Example on NYT article

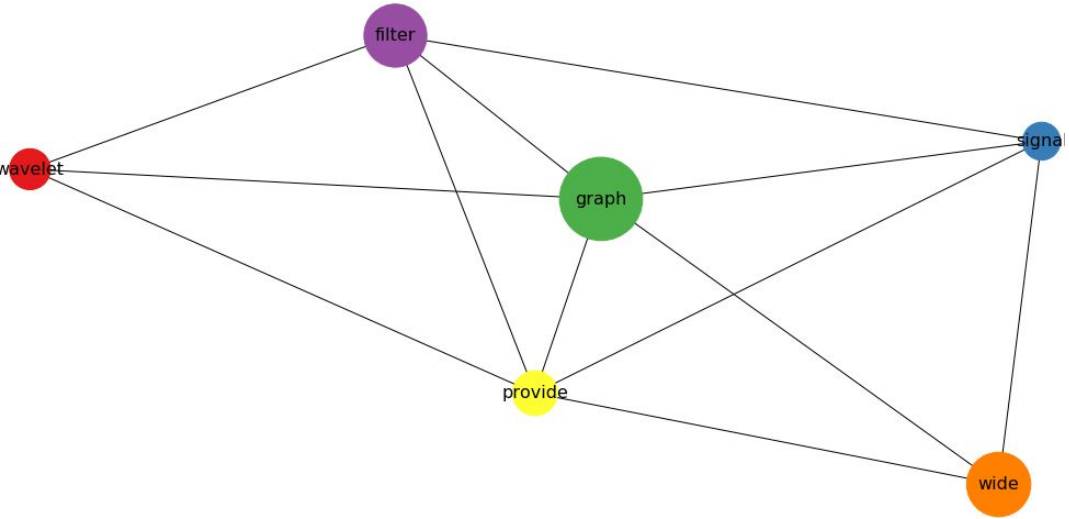


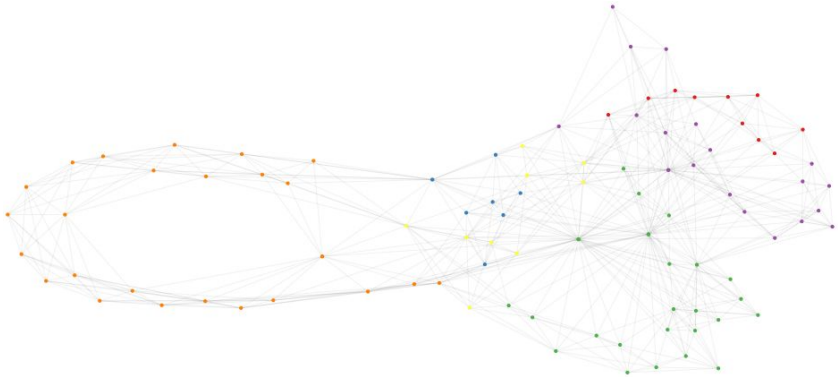


U.S. Accuses Syria of New Chemical Weapons Use

https://www.nytimes.com/2018/01/23/world/middleeast/syria-chemical-weapons-ghouta.html

Example on Python library





PyGSP: Graph Signal Processing in Python

https://pygsp.readthedocs.io/en/latest/

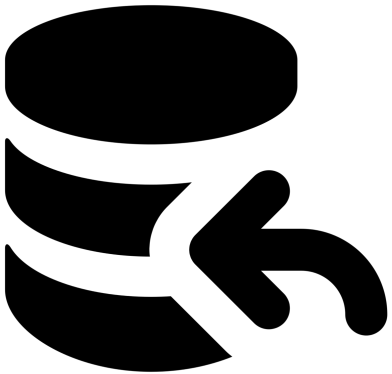
Shortcomings ?

- **No spectral** analysis
  - But is it really that bad ? :)
- **No comparative** analysis between documents
  - “How do these texts compare ?”

GraphLang v.2

Data → Features

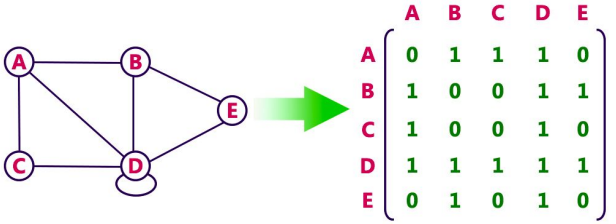
- Data acquisition (20NewsGroups)
- Features engineering (TFIDF)



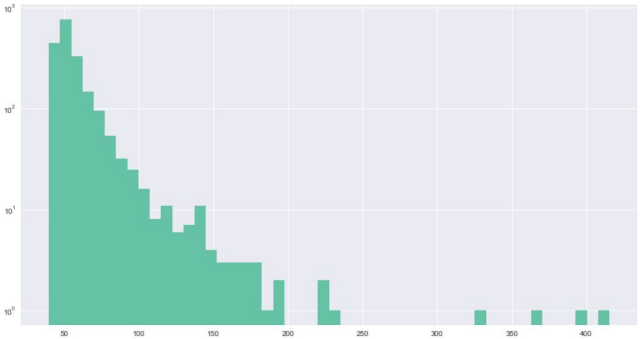
$$\text{tf-idf} = \text{tf} \times \text{idf} \qquad (1)$$

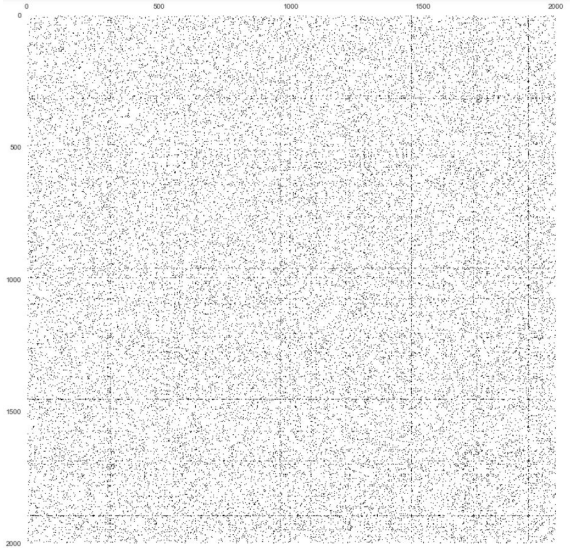
$$\text{idf}(t) = \log \frac{n+1}{\text{df}(d,t)+1} + 1 \qquad (2)$$

Graph Construction



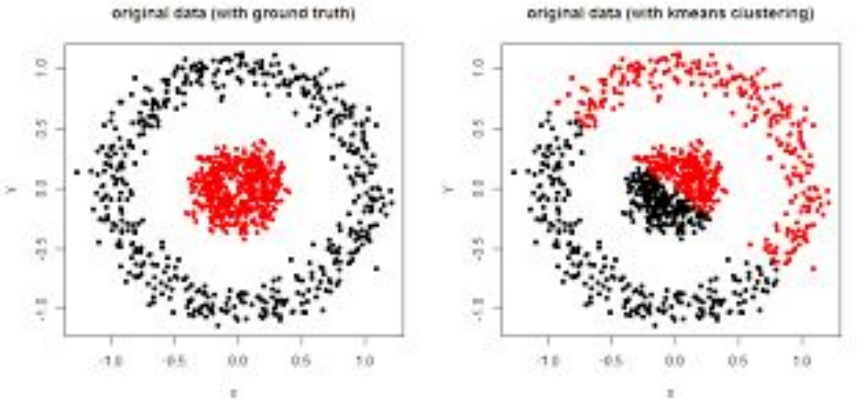
- Distance Matrix (cosine)
- Adjacency Matrix (gaussian kernel)
- Filter best neighbors (100 with highest weights)

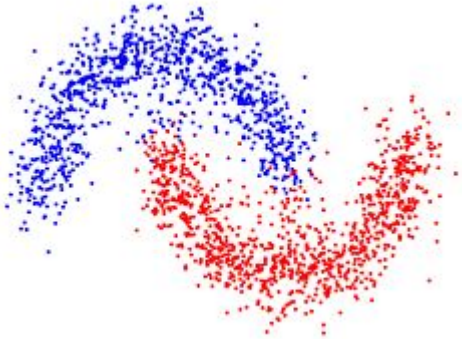




Graph Analysis

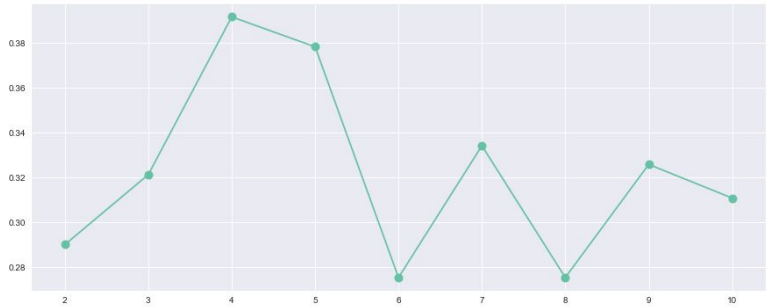
- Spectral decomposition



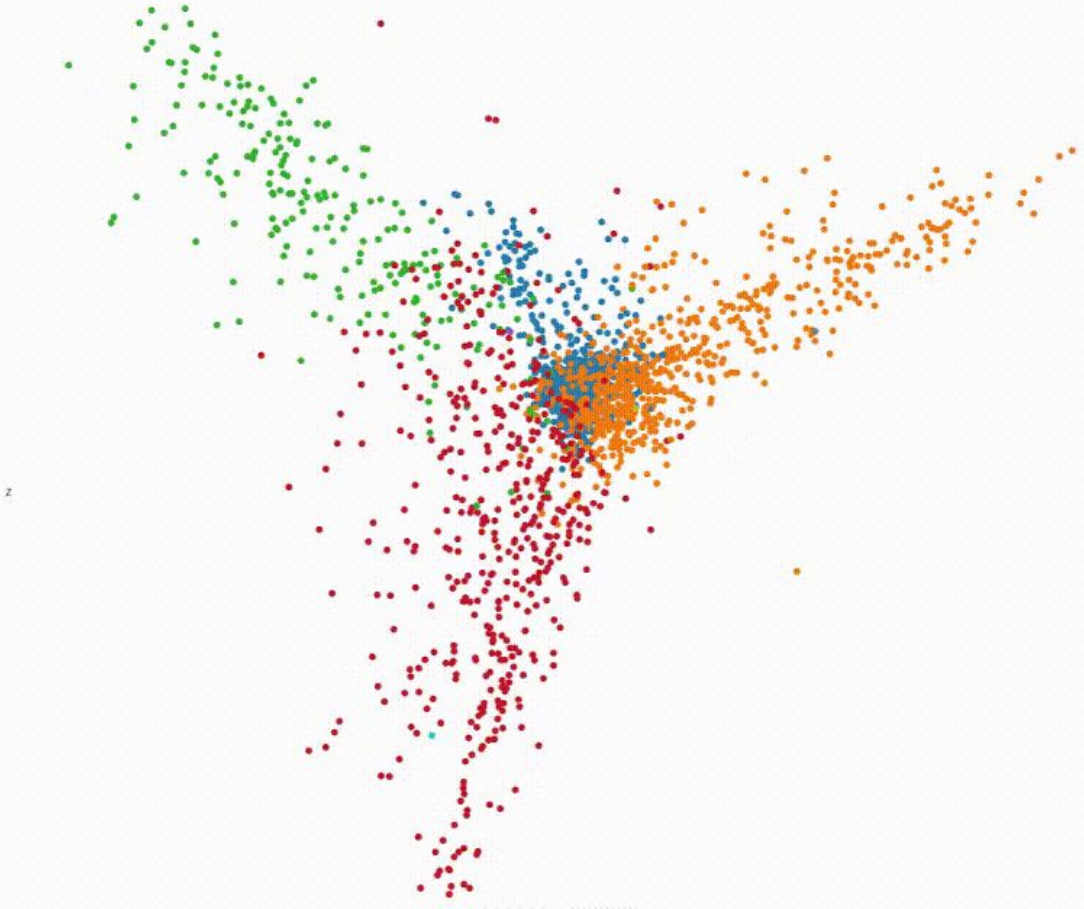


Unsupervised Clustering

- Silhouette Score
- Gaussian Mixture Model

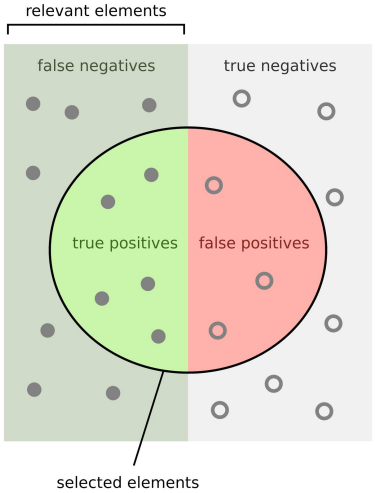


Visualization



Evaluation

- confusion matrix



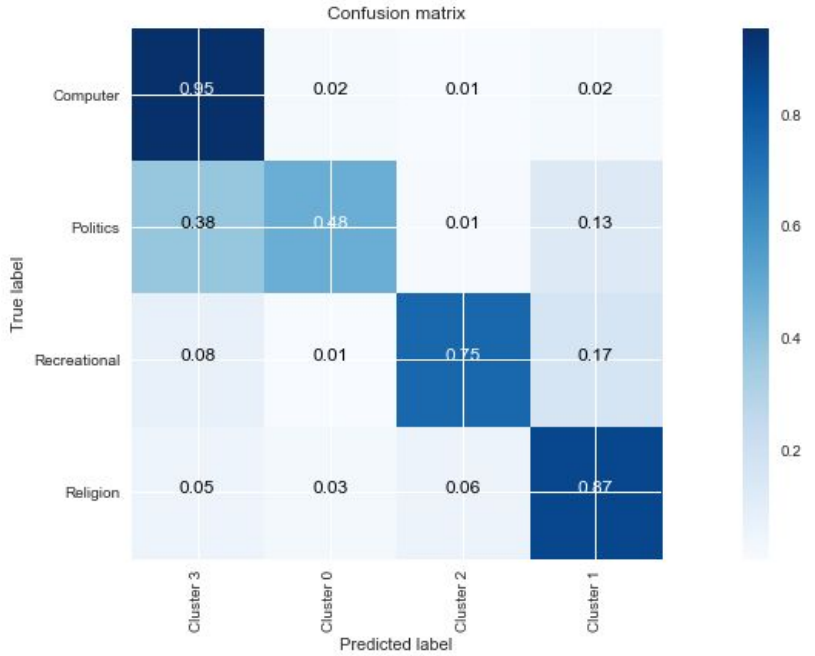
How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

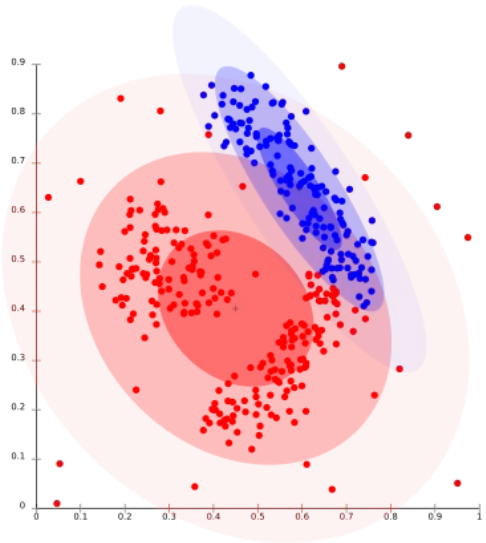
How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

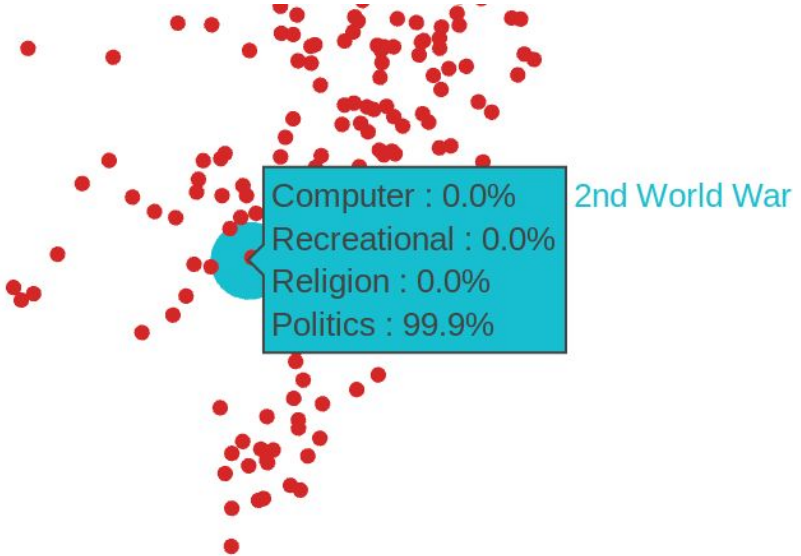
F1-Score: 76 %



Inference



- Politics



Results and shortcomings

GraphLang v.1

Impressive results

However, no ground truth

Possible improvements:

- **Addition of interactivity**
- **Better visualization of the links between each concept**

GraphLang v.2

Good results for an unsupervised approach

Does NOT scale well with the dimension of parent topics

Thanks !

Questions ?

github.com/ali-h/GraphLang