Network Tour of Data Science

Science and Religion

Lucas Eckes, Lilia Ellouz, André Ghattas, Frédéric Bischoff

Final presentation
of group project

Table of Contents

I. Introduction
II. Data Acquisition
III. Exploration
IV. Exploitation
V. Conclusion

Introduction

Richard P. Feynman claimed: "*religion is a culture of faith; science is a culture of doubt*"

Is the relationship between science and religion as straightforward as modern society seems to think?

Using **Wikipedia** articles, we:

- Examined intra and inter relationships between science and religion
- Came up with a new sub-categorization of scientific and religious articles which reflects the differences between the sub-themes

Data Acquisition

Using wikipedia-api, we:

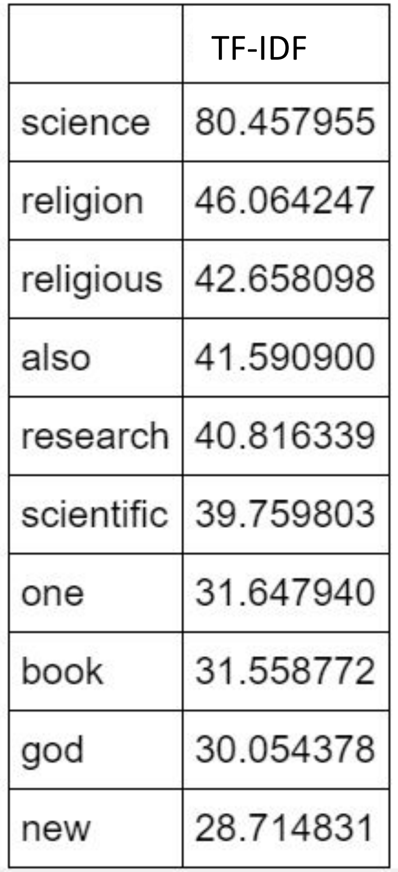
- Fetched science and religion categories
- Only chose articles in the **first** subcategories to get a **reasonable** number of articles

1579 science articles VS. **751** religion articles

Unbalanced Data
but this did not impact the machine learning phase

Exploration

Text processing and construction of feature vectors:



	TF-IDF
science	80.457955
religion	46.064247
religious	42.658098
also	41.590900
research	40.816339
scientific	39.759803
one	31.647940
book	31.558772
god	30.054378
new	28.714831

The top 10 words

- Filter out
  - ***stop words***
  - ***typical expressions*** specific to *wikipedia-api* *e.g.* ‘\displaystyle’, '<<', '>>'

- Construction of ***feature vectors*** by using ***TF-IDF***
  - ***TF-IDF*** computes a ***score*** for each word based on its **frequency** as well as its **inverse document frequency**
    - ***common*** words have ***low scores***
    - characteristic ***keywords have*** a ***high scores***

Exploration

Evaluating similarities between articles:

- For each **science** and **religion** Wikipedia article, our **feature vector** contains the TF-IDF scores for the **50 most relevant words**.

  We **keep a weighted version** of the vectors given by TF-IDF scores **instead of simply replacing weights by 1s and 0s** (to indicate the presence of a word)

  => gives a preciser information about the importance of a word

- Measure of similarity between articles with ***cosine similarity***:

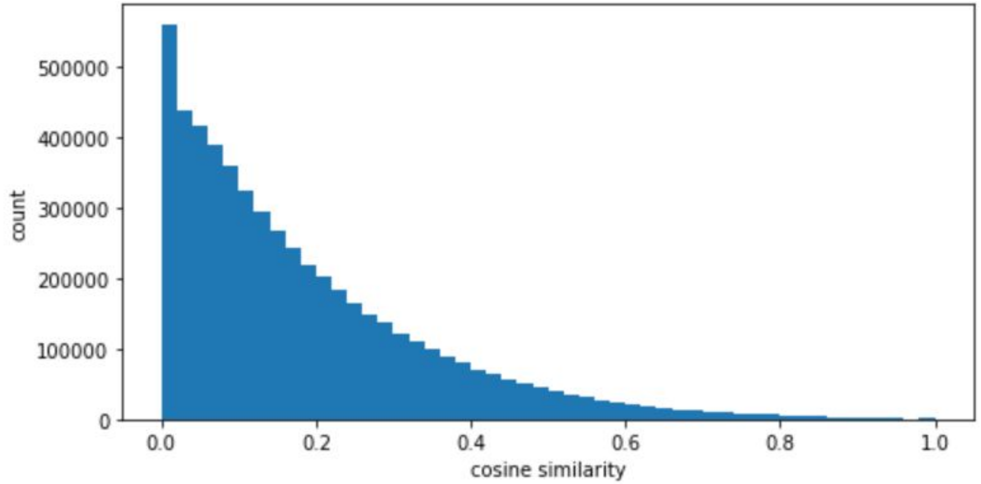
  Cosine Similarity is commonly used in **high dimensions**.
  measures similarity according to direction and not magnitude.

  Magnitude of TD-IDF scores vary with article length.
  Euclidean distance would be **less relevant.**

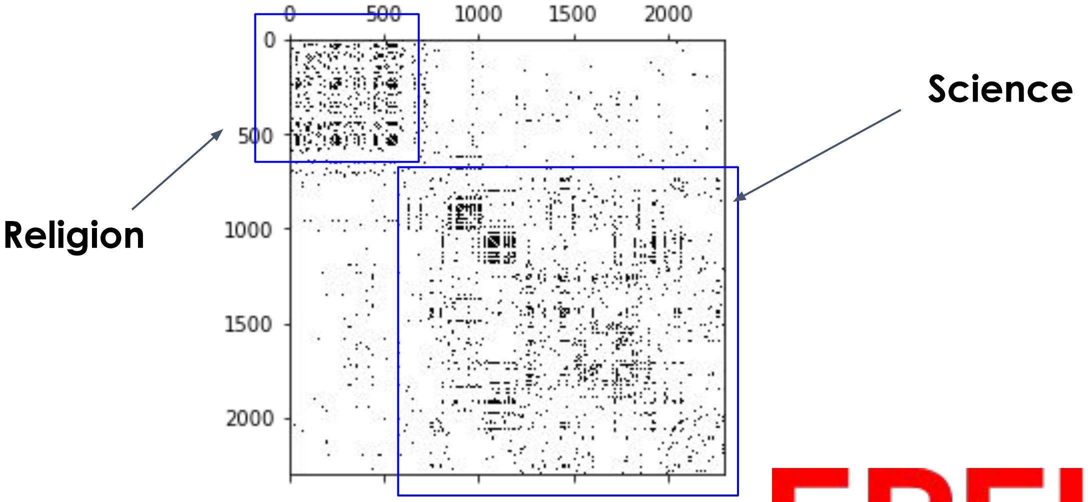
Exploration

- **High similarities** are **rare: 90% of similarities** between all articles are **below 0.6**

Histogram of cosine similarities between **articles**



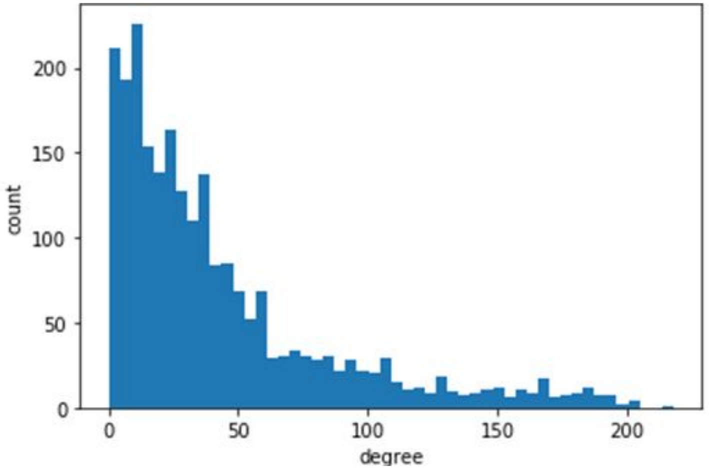
Adjacency matrix: Threshold **0.6**; 76142 links



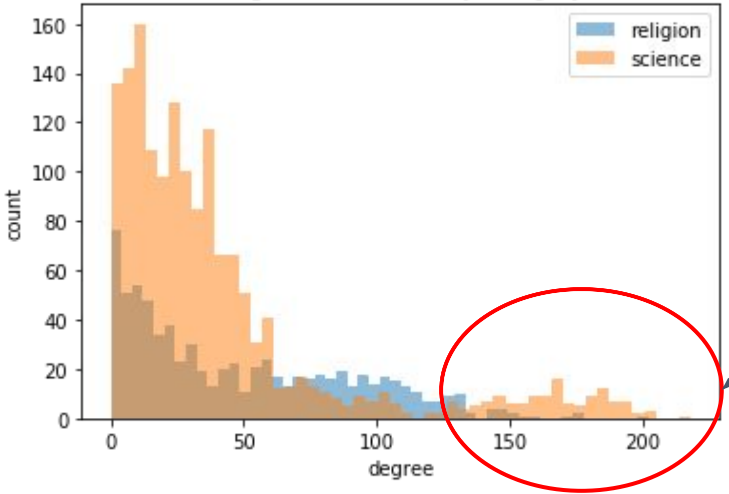
Exploration

Graph description:

Degree distribution

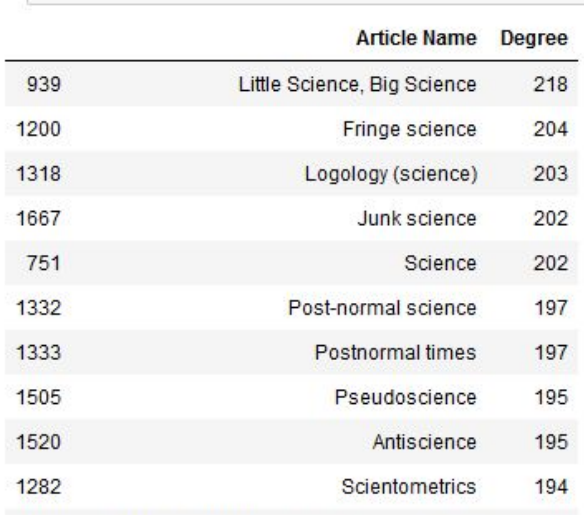


Degree distribution by category



- **Sixteen connected components**
  - **Giant** component = **2298 nodes** + **50117 edges**
  - A **negligible** second small with **2 nodes**
  - **Fourteen components** made up of **one** article each
- **Clustering coefficients**
  - **0.53 => so possibly strong links between science and religion** (not same nb of articles in each cat)
- The distribution of the degrees corresponds to a **power law** => graph is a **scale free network**
  - **Not surprising**: on Wikipedia, there is an **important number** of pages that are relatively **specific** and a **few** pages that are **very general** and have relatively many connections.

Biggest nodes are **science** pages => more articles about science than religion so **easier** to create links and find similarities

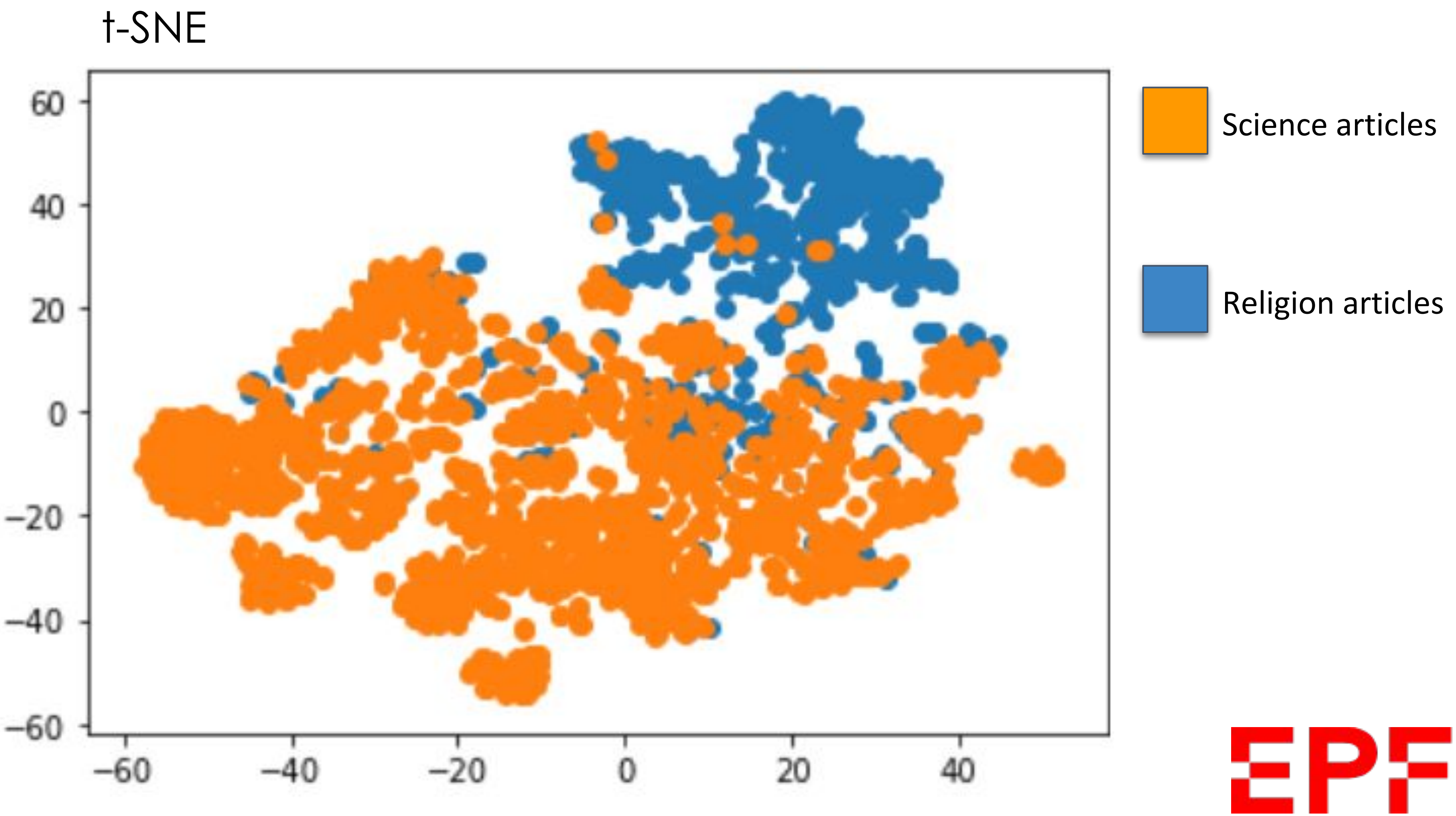


	Article Name	Degree
939	Little Science, Big Science	218
1200	Fringe science	204
1318	Logology (science)	203
1667	Junk science	202
751	Science	202
1332	Post-normal science	197
1333	Postnormal times	197
1505	Pseudoscience	195
1520	Antiscience	195
1282	Scientometrics	194

Top nodes (with highest degrees)

Exploitation

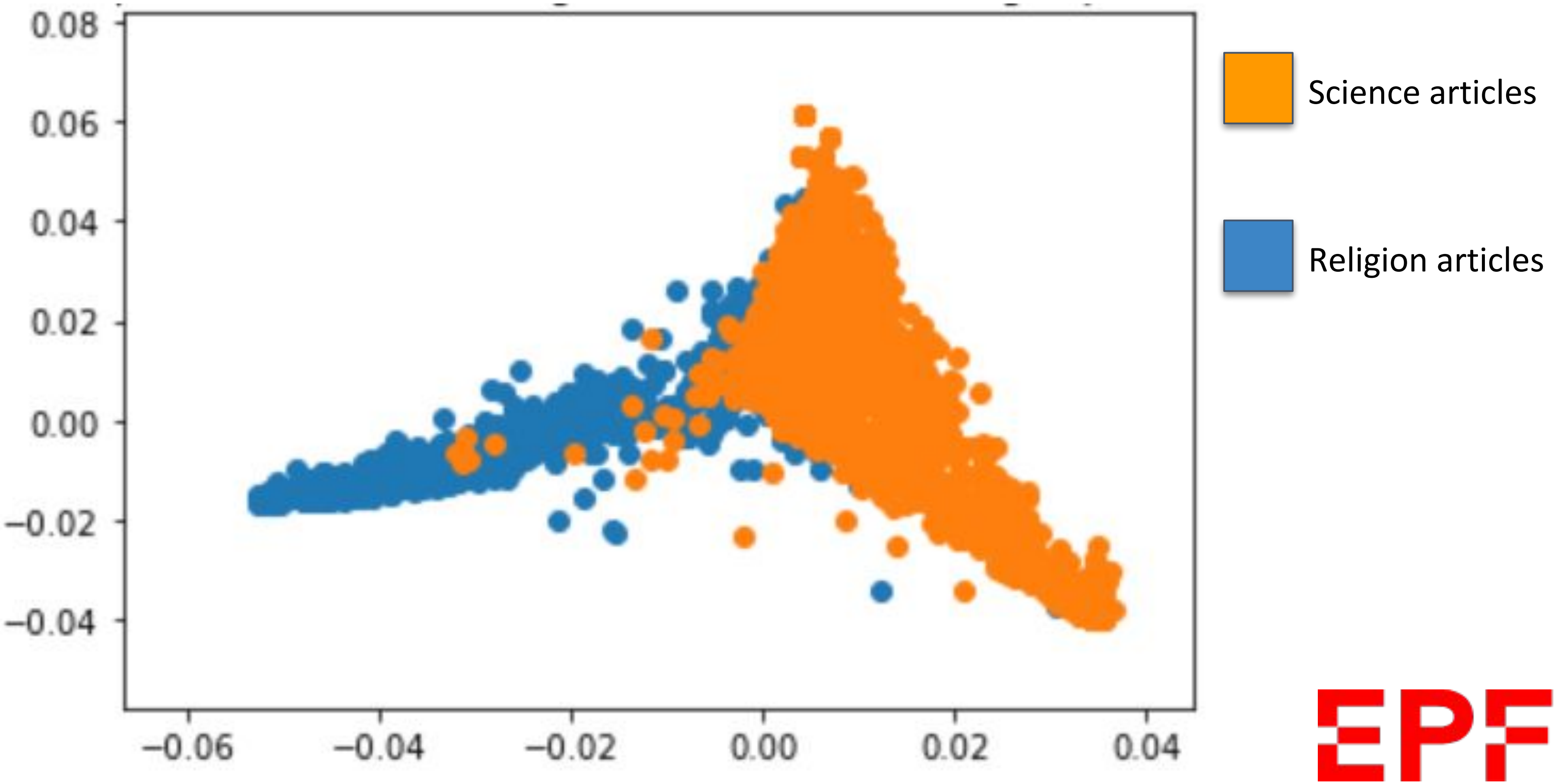
Visualization



Exploitation

Visualization

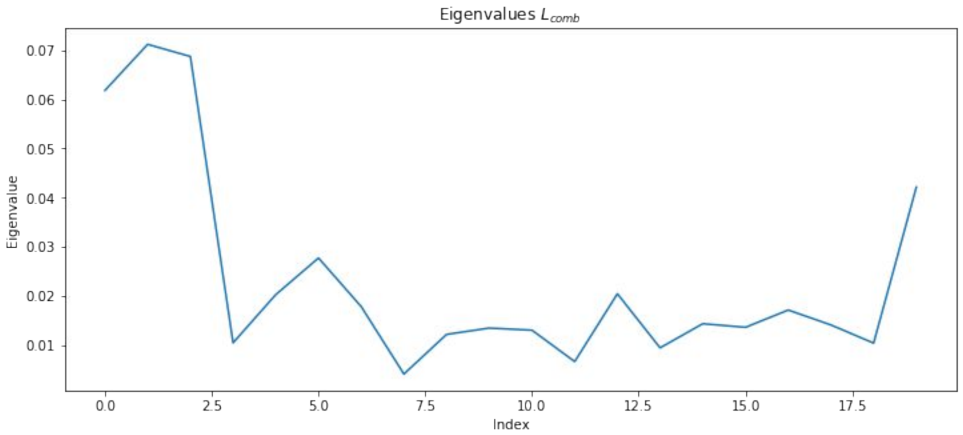
Laplacian Eigenmaps



Exploitation

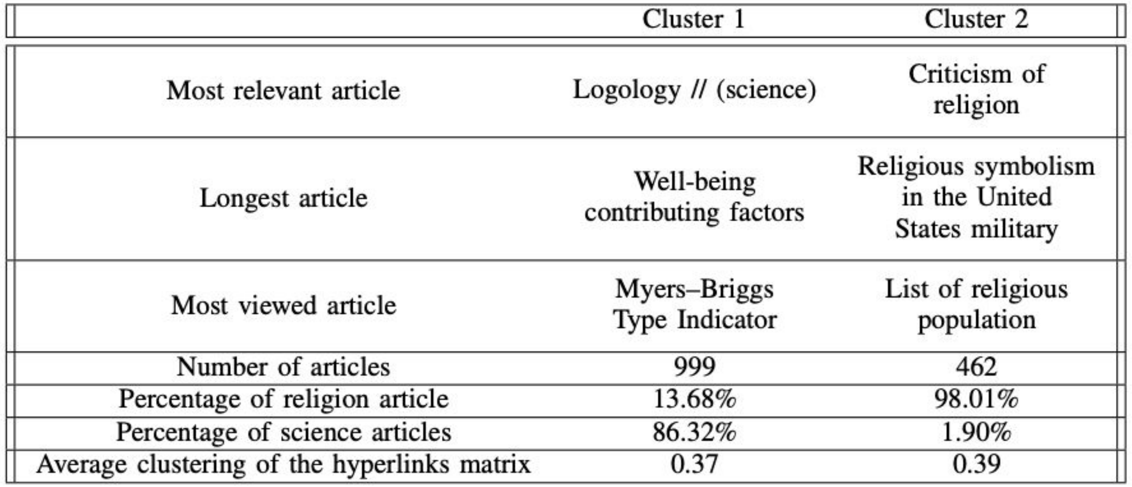
Spectral clustering of the networks

Eigengap heuristic



94.3% test accuracy

Two clusters



	Cluster 1	Cluster 2
Most relevant article	Logology // (science)	Criticism of religion
Longest article	Well-being contributing factors	Religious symbolism in the United States military
Most viewed article	Myers–Briggs Type Indicator	List of religious population
Number of articles	999	462
Percentage of religion article	13.68%	98.01%
Percentage of science articles	86.32%	1.90%
Average clustering of the hyperlinks matrix	0.37	0.39

The second cluster is almost **exclusively** made of **religion** articles but how can we explain the small number of **religion** articles that were put into the first cluster which **mostly** contains **science** articles?

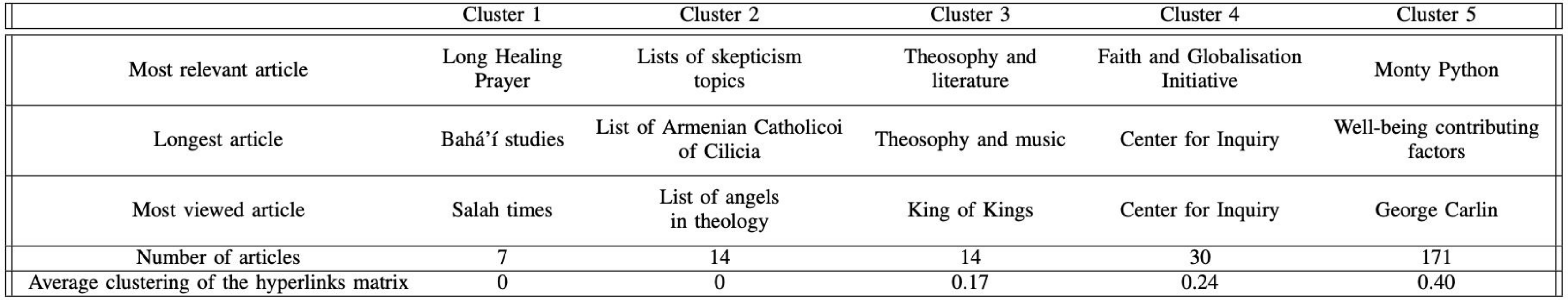
Exploitation

Run the **religion** articles found in **Cluster 1** through the same clustering pipeline:

- Compute new 50 words based on the TF-IDF from this corpus
- Create a graph
- Do spectral clustering
- **91.2%** test accuracy for all articles (7 clusters)

The **topics** of the clusters were not not easy to determine which is why we computed the **most relevant** words using **TF-IDF** scores.

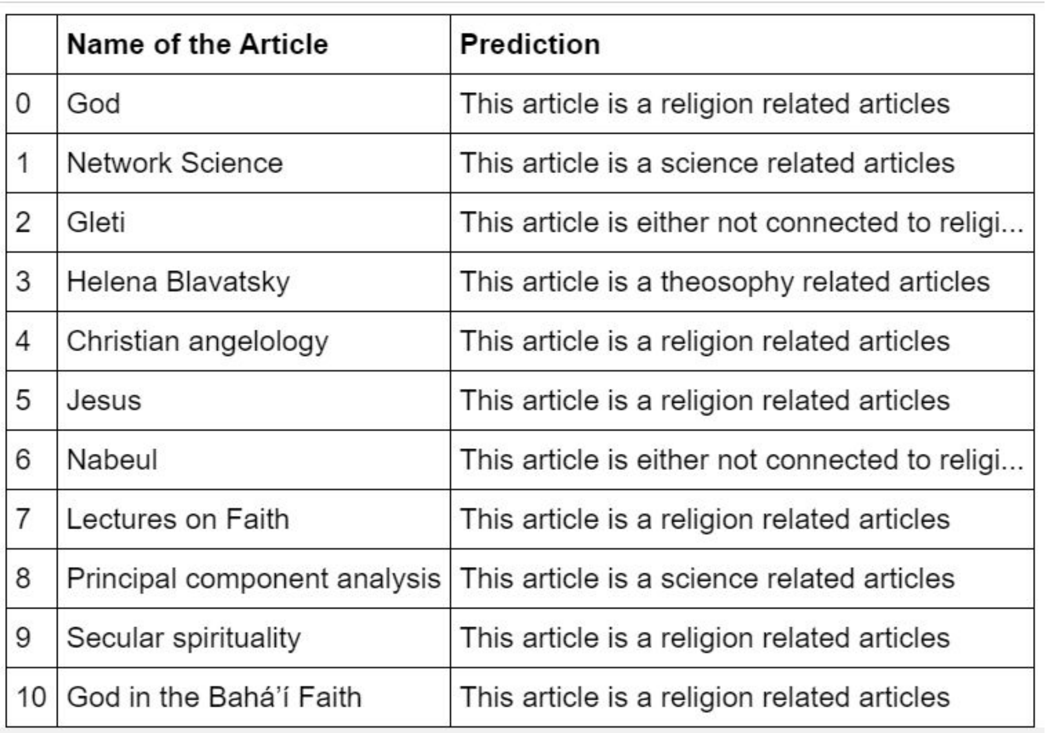
Five clusters gave coherent results



	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Most relevant article	Long Healing Prayer	Lists of skepticism topics	Theosophy and literature	Faith and Globalisation Initiative	Monty Python
Longest article	Bahá’í studies	List of Armenian Catholicoi of Cilicia	Theosophy and music	Center for Inquiry	Well-being contributing factors
Most viewed article	Salah times	List of angels in theology	King of Kings	Center for Inquiry	George Carlin
Number of articles	7	14	14	30	171
Average clustering of the hyperlinks matrix	0	0	0.17	0.24	0.40

Exploitation

Could our **model** classify articles from Wikipedia into the **7 clusters** or detect if the article did not go into **either** category?



	Name of the Article	Prediction
0	God	This article is a religion related articles
1	Network Science	This article is a science related articles
2	Gleti	This article is either not connected to religi...
3	Helena Blavatsky	This article is a theosophy related articles
4	Christian angelology	This article is a religion related articles
5	Jesus	This article is a religion related articles
6	Nabeul	This article is either not connected to religi...
7	Lectures on Faith	This article is a religion related articles
8	Principal component analysis	This article is a science related articles
9	Secular spirituality	This article is a religion related articles
10	God in the Bahá'í Faith	This article is a religion related articles

Our model might need some **refinement**, considering all the possible classes. But it proved to be **highly efficient** in classifying religion and science articles.

Is an **automated** categorization of Wikipedia pages possible?

Conclusion

- A science vs. religion article classifier is **achievable**!
- Clustering could be **more fine-grained**, especially to explore the **structure** of the religion cluster.
- A classifier that would perform **better** than the current one.

Questions? :)

I. Introduction
II. Data Acquisition
III. Exploration
IV. Exploitation
V. Conclusion