Final Report - Science and religion - Network Tour of Data Science

Lucas Eckes, Lilia Ellouz, André Ghattas, Frédéric Bischoff

***Abstract*—Richard P. Feynman once claimed that "religion is a culture of faith; science is a culture of doubt"[1]. In Feynman's view, science and religion represent two distinct and opposing cultures. This is a very modern take on the relationship between science and religion. It is in fact a very mainstream opinion to have in the West [2]. However, some scientists seem to disagree. 51% of US scientists reported that they believe in at least some form of higher power in 2009 [3]. The relationship between science and religion does not appear to be as straightforward as modern society seems to think. This project has a few aims. The first one is to examine both the intra and inter relationships that we can find between science and religion, and to infer the reasons behind the relationships found such as common origins or ideas. The second one is to try to classify new documents according to the clustering that we obtain. The ultimate goal is to try to come up with a new categorization of scientific and religious articles which would reflect the ties between them more accurately and which would not take into account the typically strong opposition between them.**

I. Acquisition

The data acquisition relies on Wikipedia pages. Using the `wikipediaapi` package, we retrieve the articles under the Science and Religion categories. In order to limit the number of acquired articles, we only choose the articles belonging to the first subcategories of these two categories.

After dropping disambiguation pages, we obtain 1579 articles in the Science category and only 751 articles in the Religion category. These numbers indicate that our dataset is heavily imbalanced, which might affect the machine learning part and incur difficulties to learn about the smaller category (Religion)[4]. However, this has not impacted the accuracy of our models and we hence chose to work with this imbalanced dataset as it offered the advantage of having the maximum amount of data.

II. Exploration

A. Text processing

We construct a list of stopwords that consists of the common English stopwords compiled in the `nltk` package, to which we add some expressions that appear specifically due to the `wikipediaapi`. We then filter out the stopwords from the tokenized and lower-cased articles.

B. Constructing feature vectors

Since we would like to measure the semantic similarity of articles, we need to focus on the words. The best way to do this is to use TF-IDF, which computes a score for each word based on its frequency as well as its inverse document frequency. This means that words that are too common across all articles do not have a high score, while the *keywords* that are characteristic of the articles have a higher rank.

To count the score of each word, we used `CountVectorizer` and `TfidfTransformer` from `sickit-learn`.

This step provides us with a score for each (word, article) pair. To obtain an absolute score for each term, we sum up all article scores for each individual word. To filter out any unwanted noise, we only keep the 50 most important words, where the importance of a word refers to its TF-IDF score. Unsurprisingly, the top word was ***science***, followed by ***religion***. The top 10 words are listed in fig.1 along with their TF-IDF scores.

	tfidf
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

Fig. 1: Top 10 important words

C. Evaluating similarity between articles

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

We could replace the score for a word by 1 if it appears in an article and by 0 if it doesn't, but this eliminates too much information about the membership of an article to its category. Therefore, we choose to adopt a weighted version of the matrix, which is the generated one with TF-IDF scores instead of simply 1s and 0s.

In order to measure similarity between pages we opted for cosine similarity, which is most commonly used in higher dimensions. The magnitude of TD-IDF scores can vary according to the article length and using Euclidean distance would create a gap between a small article compared to a long one, even if they are about the same topic. The advantage of cosine similarity is that it measures similarity according to the direction of a vector and not the magnitude.[5]

Fig.2 shows the distribution of the distances between all articles. We observe that high similarities are rare: 70% of similarities between all articles are below 0.5. For comparison, if we do the same for articles of a same category (Fig.3), we see that the results are just slightly better: 95% of the similarities are below 0.63. This shows that getting a high similarity between pages is rare, even in the case of two pages belonging to the same category.

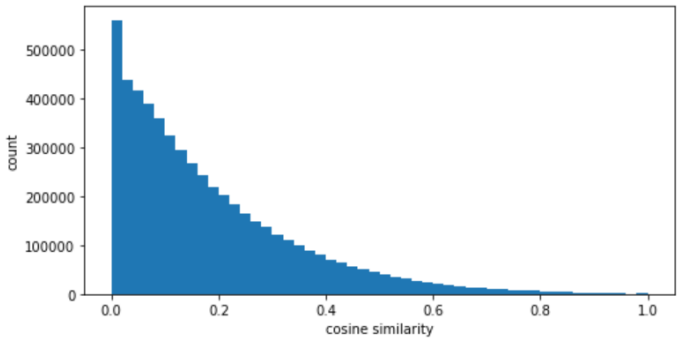

Fig. 2: Histogram of cosine similarity between pages

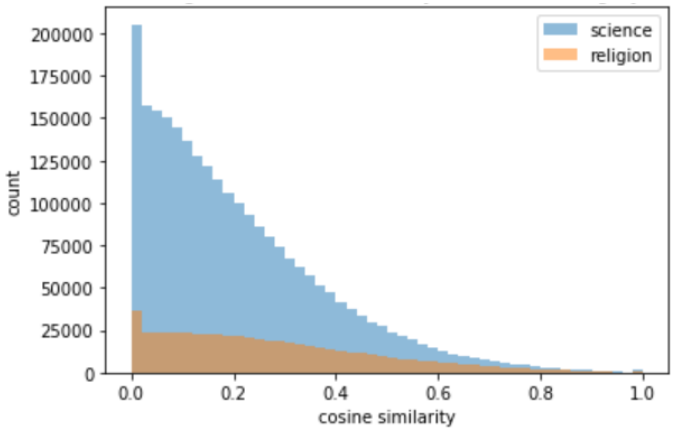

Fig. 3: Histogram of cosine similarity inside a category

These similarity measures allow us to build an adjacency matrix. We applied a threshold fixed to 0.6 in order to keep only the strongest links (i.e. the highest similarities). We expect to hence visually be able to distinguish the two main categories but to also be able to see the links that could connect nodes from the two topics. After the thresholding process, 76142 links remain. Fig.4 shows the obtained adjacency matrix: the square-shaped data points in the upper left represent the links between religion pages and the square-shaped data points in the bottom right show the links between the science pages.

D. Graph description

The graph has 16 connected components: it contains a giant component which is composed of 2298 nodes and 50117 edges, a negligible second small component made up of 2 nodes and 14 other components which are made up of one article each. The clustering coefficients of the giant component is 0.53. This result is a bit surprising as the number of religious articles we have picked is not equal to the number of scientific ones. This hence suggests that some possibly strong links have formed between scientific and religious articles.

As we can see on Fig.5, the degree distribution shows that there are a few pages with high degree (over 100), and the majority of nodes have a degree that is relatively small (above 50). Fig.7 shows to which articles nodes with highest degree

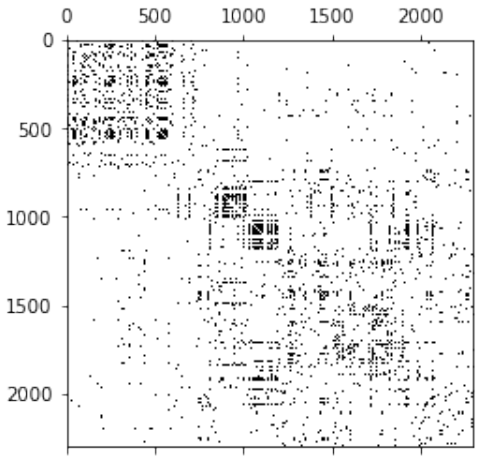

Fig. 4: Adjacency Matrix

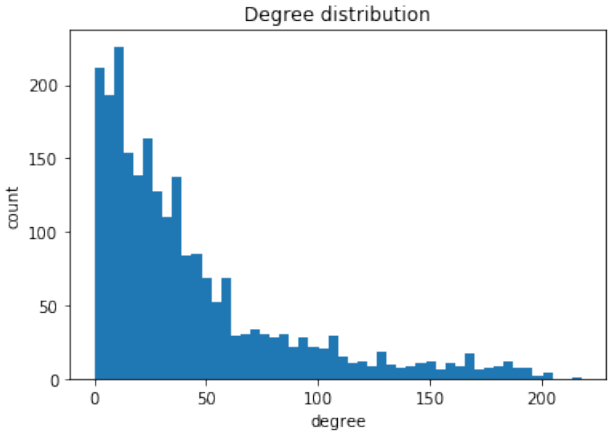

Fig. 5: Degree distribution

corresponds to. We see that the top nodes are articles that deal with broad subjects such as ”Science” and ”Pseudo-science”. Their broadness allows them to have an important number of connections. On the other hand, articles with lower degrees are more specific and hence have less links.

The distribution of the degrees corresponds to a power law. As a consequence, the graph corresponds to a scale free network. This structure of the network is not surprising. In fact, on Wikipedia, we expect there to be an important number of pages that are relatively specific to a question or topic with a relatively small number of connections, and only a few pages which are very general and which hence have relatively many connections. We also observe on Fig.7 with the top nodes and on Fig.6 that shows the degree distribution by category, that most of the biggest nodes are science pages. This is due to the fact that we have more article about science than religion and as a consequence it’s easier to create links and find similarities between science pages than for religion articles.

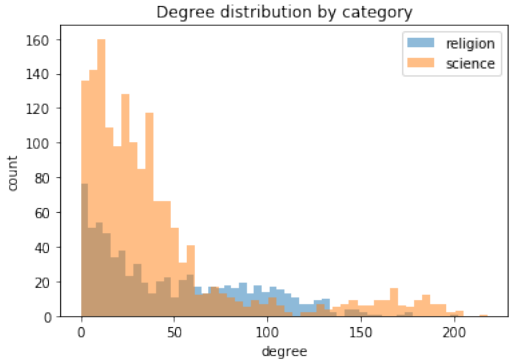

Fig. 6: Degree distribution by category

	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
1450	Amsterdam Call for Action on Open Science	193
1115	Master of Science	192

Fig. 7: Articles with highest degree.

III. Exploitation

A. Further improvements

The graph we obtained was not connected. The nodes which were not connected seemed irrelevant to our future analysis. Furthermore, we saw that the articles which were connected had a much higher importance compared to the non connected articles. Importance was measured through two scores. One was the average number of words per article, which was about three times bigger for connected articles. The other was the average number of views per month, which was 1.5 times bigger for connected articles. We hence chose to remove the nodes that were not connected in the graph.

B. Visualization

We then proceeded to visualize the graph we obtained in order to try to distinguish clusters. We did this through two methods which are commonly used for data visualization. The first was Laplacian Eigenmaps [6]. Following the usual procedure, we computed the normalized laplacian and then applied a spectral decomposition to our graph. The results of this method are displayed in Fig.8.

The second method we used was t-SNE [7]. It is a state-of-the-art algorithm used for dimensionality reduction and visualization of high-dimensional data. The results of this method are displayed in Fig.9.

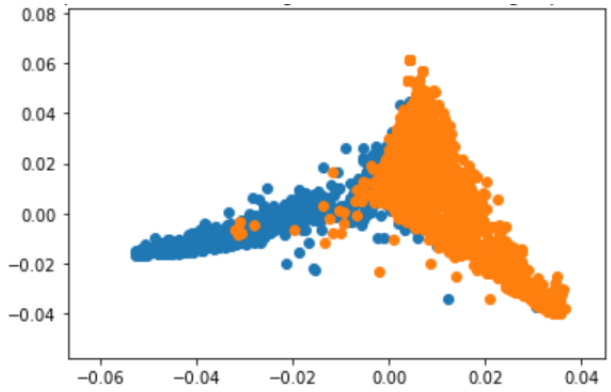

Fig. 8: Visualizing the graph using Laplacian Eigenmaps: Science articles (orange) and Religion articles (blue)

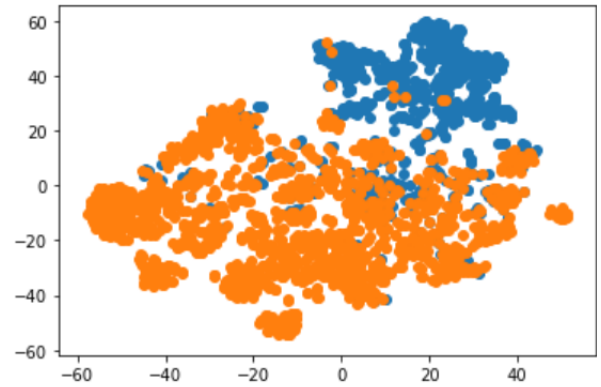

Fig. 9: Visualizing the graph using t-SNE: Science articles (orange) and Religion articles (blue

There appears to be some overlap between the two categories, but t-SNE is able to recover both the Religion and Science clusters, as the majority of the data points are well separated into two different regions.

C. Clustering

1) Finding the optimal number of clusters:
After visualizing the graph, we proceeded to apply clustering algorithms on it to see if we can detect any clusters. First, we needed to determine the optimal number of clusters K. One possible approach to doing that is through the Eigengap Heuristic [8]. The heuristic states that the number of clusters K should be equal to the eigenvalue with index K such that the difference between this eigenvalue and the eigenvalue with index $K+1$ is the biggest eigenvalue difference. We plotted the eigengaps in terms of their indices (Fig.10). We can see that the optimal K seems to be 2 according to this graph.

2) Spectral Clustering:
We then proceeded to use the optimal K in a clustering algorithm. The algorithm we chose to use was spectral clustering as compared to classic algorithms such as K-Means, it is more general and hence more powerful in detecting clusters. First, we applied spectral clustering with $K = 2$. This gave us two

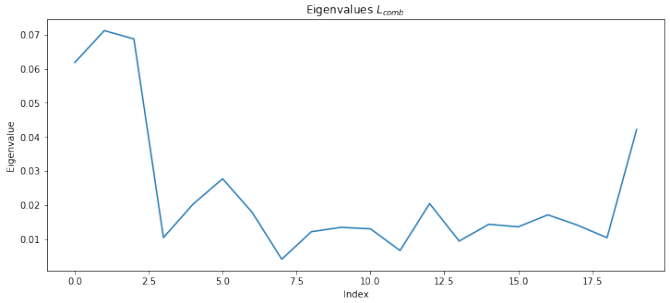

Fig. 10: Eigengaps in terms of their indices

clusters in the graph. One cluster contained mostly religious articles while the other contained mostly scientific articles. Their summary is given in table I.

TABLE I: Summary table of the clusters obtained after applying spectral clustering on the graph for $K = 2$

	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

We wanted to find out the difference between the religious articles which were grouped with the scientific articles and those which were not, so that we could get insights into the relationship between science and religion and their closeness. We hence attempted spectral clustering with $K = 5$ on the graph containing only these religious articles which were classified with scientific ones. This method gave us clusters which were more diverse and offered insights into which articles could hence be grouped together. We noticed the following:

- One cluster includes the Baháí Faith, a religion from Haifa
- One cluster contains the list of angels and topics in religion
- One cluster contains articles which mostly correspond to theosophy which is a new religion movement launched by Helena Blavatsky.
- One cluster is about the foundation of research in science and religion
- The final cluster represents philosophy and spirituality

We can clearly see that the articles are not directly about religion. They are either indirectly linked to religion, articles on more modern religions or articles that deal cover spirituality. Further details about the clusters obtained are provided in table II.

D. Predicting the category of an article

One of our goals is to predict the category to which an article belongs. For that purpose, we use the two original labels for the articles (Science and Religion) as well as the categorization we implemented in the previously described unsupervised learning process.

In order to avoid overfitting during the training process, we use 60% of our data for training and the remaining 40% is used for testing the model. We build neural networks using `PyTorch`. We use cross-entropy as a loss function and `adam` [9] as the optimizer. We constructed a Laplacian Polynomial model in DGL to perform graph filtering, as this lets `PyTorch` decide on the optimal filter coefficients. It computes the function $f(X) = \sum_{i=1}^{k} \alpha_i L^i X\theta$ where the trainable parameters are the coefficients α_i and the matrix θ. This function can be interpreted as a filtering of X by $f(X) = \sum_{i=1}^{k} \alpha_i L^i$ followed by a linear layer [10]. The model was ran on 100 epochs.

1) Binary classification, Science vs. Religion:
Using the model described above and the original labels, the test accuracy was 94.3%.

2) Classifying among eight possible categories:
This model predicts whether an article has a scientific or religious theme. If it is found to be an article about religion, it can either belong to one of the 5 clusters found during the spectral clustering with $K = 5$, or it can classified as a general article about religion. This model is also capable of detecting outliers such as articles that are neither about religion nor about science (it classifies them into 'Other' in that case). The religion category and the 5 specific categories as well as the science category and the 'Other' category add up to a total of 8 possible categories.

Using the model described above, the test accuracy was 91.2%.

3) Predicting the categories of diverse articles:
The model used to classify among eight possible categories was then used to predict the category of several articles, many of which were new and did not belong to the original data. The results of this classification is displayed in Table.III. The prediction result seem to be correct in the sense that it classifies articles in the correct class between religion and science. Furthermore, the article Gleti which is a stub article about a Moon Goddess and the article about Nabeul, which is a city and is hence not connected to science nor religion, are correctly classified as not connected to either of the subjects. The article about Helena Blavatsky who is the founder of the Theosophy movement is indeed correctly classify in the Theosophy cluster. However, it would be preferable if some articles were classified in a more accurate cluster. For example, the article God in the Baháí Faith should be classified in the Baha'i cluster meaning that there are possible improvements in the sub-categorical classification.

IV. Conclusion

This project has enabled us to examine the relationship between science and religion through Wikipedia articles. It

TABLE II: Summary table of the clusters obtained after applying spectral clustering on the graph made of religion articles which were clustered with scientific articles for $K = 5$

	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

TABLE III: Predicting the categories of new articles

Article Name	Predicted Class
God	Religion
Network Science	Science
Gleti	Neither Science nor Religion
Helena Blavatsky	Theosophy
Christian angelology	Religion
Jesus	Religion
Nabeul	Neither Science nor Religion
Lectures on Faith	Religion
Principal component analysis	Science
Secular spirituality	Religion
God in the Baha'i Faith	Religion

consisted of many steps. First, we acquired the article data through the Wikipedia API. Afterwards, we explored the data we obtained. We obtained two graphs using it. We transformed it to a hyperlink graph by using the hyperlinks among articles and we got a similarity graph through the common words among articles. TF-IDF was used to keep the most important words.

We then proceeded to remove non-connected articles and to visualize the resulting graph in order to distinguish clusters. Then, we applied spectral clustering and detected seven total clusters in our graph. We used this clustering to train a Laplacian polynomial classifier and used that classifier on our test data which resulted in a good precision measure. We further tested our classifier on arbitrary articles from Wikipedia and saw that it gave decent results.

V. Further Improvements

Some further improvements are possible.

For instance, a more fine grained clustering technique could be implemented, exploring the structure of the science cluster more closely. Another improvement could be to train a better classifier as some of the results were incorrect for the arbitrary Wikipedia articles at the last step.

An aim which we initially had was to propose a better sub-categorization of Wikipedia articles relative to science and religion. The clustering we made has some inconsistency in it and it should ideally be more fine-grained. But the results we have obtained are promising as an automatic classifier to distinguish science articles, religion articles and articles belonging to neither is achievable.

References

[1] R. P. Feynman. [Online]. Available: https://www.goodreads.com/quotes/33583-religion-is-a-culture-of-faith-science-is-a-culture

[2] J. Schloss, "'faith vs. fact:' why religion and science are mutually incompatible," 2015. [Online]. Available: https://www.washingtonpost.com/opinions/science-and-theology/2015/08/03/77136504-19ca-11e5-bd7f-4611a60dd8e5_story.html

[3] "Pew research center," 2009. [Online]. Available: https://www.pewforum.org/2009/11/05/scientists-and-belief/

[4] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[5] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, vol. 4, 2008, pp. 9–56.

[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[7] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[8] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, 2007.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[10] "EPFL NTDS course: Laplacian polynomial model in DGL," 2019. [Online]. Available: https://github.com/mdeff/ntds_2019/blob/master/assignments/2_learning_with_graphs.ipynb