Project Summary - Network Tour of Data Science

Frederic Bischoff, Lucas Eckes, Lilia Ellouz, André Ghattas

I. Story

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.

II. Acquisition

We will take the articles under the categories science and religion as well as the articles belonging to the first subcategories of these two categories. Each article will be represented by a binary vector $\mathbf{a}$ such that $\mathbf{a}_i = 1$ if word number i is in the article and $\mathbf{a}_i = 0$ otherwise. The most frequent words in all the articles will be the ones taken into consideration. This will give us a vectorial representation for each article, which we'll use to compute the distance between the articles. Our first graph will hence be created such that each article is a node and an edge will be created between the nodes if the distance is below a certain threshold.

Before constructing the graph, we need to properly treat the articles. We will tokenize the content of the Wikipedia page to have a better word segmentation. Then, we will remove the stopwords (i.e. words that are extremely common words which don't give any indication of the nature of the article), the words that are one character long, the numbers and the symbols (e.g. '...', '='). We will write every word in lowercase so we don't encounter issues (e.g. to treat 'Religion' and 'religion' similarly). In our preliminary exploration, we saw that even after this treatment, our data set still had many neutral words in the list of most used (e.g. 'also' and 'however'). We will hence implement TF-IDF so that these words will have less weight in the final model.

Another graph which we may create is a directed graph based on the hyperlinks among the articles present in the two categories. We may then compare both graphs as we expect them to have a similar structure.

III. Exploration

Using the acquired data, we will visualize the Euclidean distance between the articles. We will explore the eventual existence of a giant component of the constructed graph as well as the average clustering coefficient. By fine tuning the sigma and epsilon parameters, we will be able to compute and visualize the similarity matrix of the graph. Based on the structure of the obtained graph, we will use NetworkX's Pagerank to compute a ranking of the nodes in the graph, which will provide us with the most relevant articles. Via the eigenvalues and eigenvectors of the adjacency matrix, we will use spectral graph theory to partition the graph. We will compute the Fourier transform from the adjacency matrix, which we will use to introduce filtering later on.

IV. Exploitation

Our graph will give a measure of proximity between articles of different categories or subcategories by analyzing the most frequent words in these articles. The goal of the exploitation will be to clearly distinguish some clusters of articles in order to see later if these clusters only contain articles of a given category or gather articles that initially come from different categories. As the dimensionality of our graph is high due to the fact that it takes into account a high number of words, it will be important to decrease its dimensionality with methods like PCA or t-SNE. Once we have a clear network, we can analyze more concretely the diversity of the articles present in a given cluster and apply machine learning methods in order to realize some classifications of articles. It would be for example interesting to see if it's so easy to distinguish a page dealing with religion from another one about science. We would also like to cluster subcategories of the two categories, which could represent a better sub-categorization of articles in Wikipedia. One subcategory may be the connection between science and religion.

V. References

[1] Goodreads page with Richard P. Feynman quote
[2] *'Faith vs. Fact:' why religion and science are mutually incompatible* by Jeffrey Schloss, August 3, 2015
[3] *Scientists and Belief*, a study by the Pew Research Center, November 5, 2009