First Vectorized Representation of AI and Trump Administration

Northwestern University

Ali Gowani

Sunday, February 2, 2020

First Vectorized Representation of AI and Trump

Introduction & Problem Statement

Our goal in this exercise is determine which Reference Term Vector (RTV), would best suit our Data Source Items (DSI) as part of the overall corpus of the Trump Administration. Our RTVs, should represent our DSIs in a manner that they cluster well with documents of similar topic and have the appropriate level of specificity. We evaluate the results of Term Frequency – Inverse Document Frequency (TF-IDF) and Word2Vec methods to determine which terms we will selected as part of our RTV. The code from Paul Huynh was used as a baseline and starter-code for my analysis and was immensely crucial as I am new to coding, Python and NLP.

Literature Review

For me to evaluate which terms would be important, I reviewed various articles and publications. Christian Perone blog on TF-IDF, helped me understand that TF-IDF are used to convert textual representation of information into a Vector Space Model (Perone, 2015). This helps us evaluate the term and consider its importance. The publication of Efficient Estimation of Word Representations in Vector Space (Mikolov, 2013) provided a much deeper appreciation of how Natural Processing Language (NLP) methods and techniques can provide more accurate ways to high quality word vectors.

Data preparation, exploration, visualization

Our dataset included text from 123 articles relating to various topics from the Trump Administration. Articles covered a wide range of topics: trade agreements, border security,

technology and foreign relations. I am new to coding and Python so sometimes my approach is a roundabout way of conducting analysis but one way to prepare my data was to look at the corpus and extract the two articles that I had submitted: New Trump Ruling Limits AI Surveillance Exports Over China Military Fears and The Technology 202: Trump administration's CES message: We're not interested in heavy AI regulation. After having a better understanding of the data, I formulated a hypothesis that the following terms will suit my documents (or DSIs) the best as part of the overall corpus: *technology, artificial intelligence, surveillance* and *china*.

Research Design and Modeling Method(s)

After attending Paul's sync session and watching his videos couple of times, I attempted to run the code that was provided. Several attempts and few hours later, I was able to successfully execute it. However, I needed to execute the same code in a manner that would allow for me to provide the TF-IDF values for my documents. I designed my model to execute the code on the corpus and the update the data source to the file with my two documents. I did this also changed the associated variables, so I had both the corpus and my documents TF-IDF values. I also executed a Count Vectorizer library to provide term frequencies for the my selected RTVs (Brownlee, 2019). It took me many hours to get this accomplished, as well as, aggregate the results for further review and analysis. After evaluating the results and looking at the summary statistics (Figure 1), I decided to look at terms with the largest TF-IDF values in the documents to see whether anything catches my attention. It was quite apparent that perhaps I could have selected a term with a higher TF-IDF value, and it would be better associated with my documents (Figure 2). This figure shows the top 10 TF-IDF values for each of my two documents. I made a quick decision to change one of my RTV from *china* to *companies* as

companies was a top 10 value in my second document and appeared in both of my documents unlike *china*.

Figure 1:

	Doc1	Doc2
count	1416.000	1416.000
mean	0.017	0.017
std	0.020	0.021
min	0.000	0.000
25%	0.000	0.000
50%	0.030	0.000
75%	0.030	0.030
max	0.299	0.236

Summary statistics of the TF-IDF values of the two selected documents. Values rounded to 3 decimal places where appropriate.

Figure 2:

	Doc1_tf	Doc1_tfidf	Doc2_tf	Doc2_tfidf
surveillance	10	0.299	0	0.000
china	8	0.239	0	0.000
technology	10	0.212	7	0.147
restrictions	5	0.149	0	0.000
industry	5	0.106	2	0.042
export	3	0.090	0	0.000
exports	3	0.090	0	0.000
leading	3	0.090	0	0.000
military	3	0.090	0	0.000
technologies	3	0.090	0	0.000

	Doc1_tf	Doc1_tfidf	Doc2_tf	Doc2_tfidf
selfdriving	0	0.000	8	0.236
policy	0	0.000	6	0.177
ensure	0	0.000	5	0.148
federal	0	0.000	5	0.148
companies	4	0.085	7	0.147
technology	10	0.212	7	0.147
government	2	0.042	6	0.126
artificial	1	0.021	5	0.105
artificial intelligence	1	0.021	5	0.105
intelligence	1	0.021	5	0.105

Top 10 TF-IDF terms in both of my documents. Values rounded to 3 decimal places where appropriate.

Results

See the results in Figure 3 and plotting Figure 4, it showed that our document would cluster well with other documents relating to technology. Our RTVs also had specificity with terms such

as, Artificial Intelligence and Regulation to perhaps provide context and meaning to our documents.

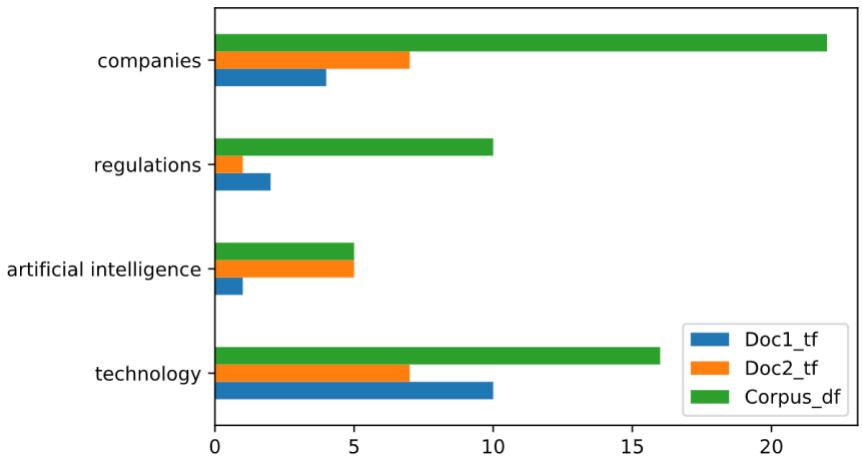
Figure 3:

	Doc1_tf	Doc1_tfidf	Doc2_tf	Doc2_tfidf	Corpus_tf	Corpus_df	Avg_TFIDF
technology	10	0.212	7	0.147	53	16	0.180
artificial intelligence	1	0.021	5	0.105	13	5	0.063
regulations	2	0.042	1	0.021	19	10	0.032
companies	4	0.085	7	0.147	54	22	0.116

Term frequency and TF-IDF values of selected terms in Doc1, Doc2 and Corpus. Values rounded to 3 decimal places where appropriate.

Figure 3 below shows a pictorial view of the four terms we selected and how it compares from the term frequency of each of our two documents and the document frequency from the corpus. It is interesting to note that only 5 documents from the entire corpus had *Artificial Intelligence* as part of its terms, so we can see that our term may be niche and may not have the same level of importance as other RTVs and documents. This would mean that the *Artificial Intelligence* reference term vector would be part of a larger cluster that may have a broader topic, like *technology*, and have RTVs with more significant TF-IDF values.

Figure 4:



Analysis and Interpretation

Further analysis shows that our two DSIs were clustered in Cluster 3 and Cluster 6. The top 10 terms for Cluster 3 are *trade, china, tariffs, chinese, trump, huawei, manufacturing, deficit, agreement* and *goods* and for Cluster 6 are *bulbs, administration, environmental, schools, climate, technology, standards, construction, projects* and *energy*. It is interesting to note that we originally had *china* as our RTV but then decided to switch it to *companies*. Even when we switched our RTV to *companies*, the document still got clustered with other documents pertaining to *china*.

Conclusions

Overall, I found this assignment to be quite interesting. What I thought were ideal RTVs did not turn out that way and after seeing the results, I am not sure whether I like any of the terms to select as ideal candidates given the low TF-IDF values. It seems that much of the effort in this assignment was to see how we think and get comfortable with trying something, knowing that we may have to go back, adjust and try again. Hopefully, the Hero Team will do a better job of selecting the terms than I did in this round!

References

Blog.christianperone.com. (2011). [online] Available at: http://blog.christianperone.com/2011/09/machine-learning-text-feature-extraction-tf-idf-part-i/ [Accessed 30 Jan. 2020].

Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

Brownlee, J. (2019, August 7). How to Prepare Text Data for Machine Learning with scikit-learn. Retrieved January 30, 2020, from https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/