Tyler Percy CNN Data Analysis HW 1

For my data, I used 105 articles from the top CNN new articles according to google at the time. The articles were taken near the start of February and due to the recent presidential inauguration and thus were mostly political in nature, however there still a good number of articles that are not politically focused. I used 105 instead of 100 in case there were some errors in reading the text in some of the articles. I used a frequency data matrix for analyzing the similarity between the articles but I did not TFIDF to normalize the matrix. The first thing I noticed from the ordered ranking of all three similarity scores was some flaws in the data that I found due to closer inspection of the results. The top result for all 3 similarity scores was the same, but this is because the two articles were exactly the same. The results also showed there were 2 articles that the links ended up being broken so they have no text inside of them. The Jaccard and Euclidean showed these two as exactly the same and the cosine and didn't list the results for that particular pair at all. They must have been filtered out due to the way the cosine score was calculated. Another mistake I made was not realizing the at similarity score matrices were upper triangular so all of the results appear twice in the rankings as sim(i, j) and sim(j, i). Overall, the cosine score was extremely accurate according to my own judgement, while the Jaccard similarity gave articles related at a high level top rankings and the Euclidean similarity was solidly in last place, giving results that were not related in any way besides possibly poorly chosen feature words for the data set.

Cosine Similarity Results:

The next couple results for the cosine similarity are all articles that have the same long preface that was not actually part of the article. Only a few of the articles had this preface, and because of it they all received very high scores because I accidently chose words that were used multiple times in the preface without me knowing. After sifting through these mistakes, the actual highest similarity articles is actually pretty accurate for cosine. They are both articles about the failed yemen raid. One is focused on the death of the military member and the other is focused on the video that was released but the content of the articles is still very similar nonetheless. The next highest pair are both articles about the Travel ban, a very hot topic at the time I took all of the articles from CNN. One article is about the lawsuits coming from the ban, and the other is about Homeland security not enforcing the ban due to the judicial ruling. Again, these articles are indeed very similar and should be near the top. The third highest pair of articles are both about Trump's diplomacy. Both articles are talking about Trump's interactions with the Australian Prime Minister and the President of Mexico. One of the articles is more focused on the discussion with the President of Mexico but both of them discuss mostly the same material. Overall the top 3 cosine similarity rankings are all extremely accurate.

Euclidean Similarity Results:

The top actual result for Euclidean Similarity was not very accurate at all. The first article in the pair was about a sushi flavored Kit-Kat made in Japan and the other was about a boy in Texas finding a rattlesnake in a toilet. The next highest pair was the same article about the Kit-Kats paired with an article about teething products with a potentially harmful chemical on them. Again, these articles are not very similar. The third highest result was between two articles that again were not very similar at all. The first was about millions of year old tissue found on a dinosaur bone and the other was about a duet album between a contemporary R&B singer and a late rapper. The first result that the two articles were actually even slightly related on the ranking was about 6 pairs down. The first article was about a sickness in India and the other was about removing a tape worm from a sick man so even these two articles are only sort of related. Overall the Euclidean Similarity results left a lot to be desired and was easily the least effective similarity rating.

Jaccard Similarity Results:

The top meaningful result from Jaccard Similarity gave the same Kit-Kat article but with a different pair. This time, the pair was with an article about Goat Yoga. These articles aren't too similar either besides both articles focus on something new and viewed as strange by the majority of people. The next highest pair is the snake in the toilet again, paired with a story about a zoo in Iraq. These stories are both about animals so that may be the reason they came up as similar, but not even close to as accurate as the cosine similarities. The third highest similarity was two stories that were again, not very similar at all on the surface but ultimately both about leaving an old place for a new one. The first was about a mosque burning down in Texas and the other was about the venue of the Rio Olympics being abandoned in the months past the end of the games. Another pair of articles further down are both about very old things, the same dinosaur rib article and an article about a lost continent that was found recently. To me it seems like these all of these articles are at least slightly related at a higher level that was passed over in the other similarity scores. The problem with these pairs not being as accurate as the cosine might be that most of the feature words that I chose have to do with politics so it might be possible that articles that are not about politics at all may share a few common words and give a false high similarity score allowing the non political articles to float to the top.

Concusion:

I was very impressed with the accuracy of the cosine similarity score! I was worried that something had gone wrong when the results of the Euclidean and Jaccard were so poor, but the cosine results were almost spot on for the top rated articles. It is possible that the Euclidean and Jaccard may benefit more from normalization. The Euclidean is the most simple similarity score and so it was not too much of a surprise that it performed rather poorly. This was my first time reading data from websites and first data analysis task and I enjoyed it a lot! Below are a couple tables that show the top words in common for the top result for each similarity score to possibly gain some insight into why the score ranked the pair so high.

These tables contain word counts of all words, not just the 100 feature words used. This will show the similar words beyond just the metric used to find the similarity score.

Top 10 Meaningful Word Counts in Cosine Similarity Top Pair			
Article 7		Article 68	
raid	12	US	14
video	12	military	7
US	10	Yemen	6
military	10	raid	6
Al Qaeda	9	Al Qaeda	5
released	6	official	4
officials	5	killed	4
CENTCOM	5	service	4
Yemen	4	Trump	4
information	4	members	3

As we can see, these articles focused on many of the same topics, even if not all of them were contained in the feature words.

Top 10 Meaningful Word Counts in Jaccard Similarity Top Pair			
Article 42		Article 99	
KitKat	9	yoga	9
sushi	6	Morse	9
KitKats	4	goat	7
edition	4	goats	7
Chocolatory	4	people	4
Japan	3	farm	4
Nestle	3	Oregon	3

Ginza	3	edition	3
varieties	2	animals	2
shop	2	little	2

The jaccard top pair only has one similar word in the top 10. As we can see from the top words, these articles are not very similar at all and show the poor performance of the Jaccard on this particular set of features.

Top 10 Meaningful Word Counts in Euclidean Similarity Top Pair			
Article 42		Article 36	
KitKat	9	snake	5
sushi	6	toilet	3
KitKats	4	found	3
edition	4	edition	3
Chocolatory	4	Mcfadden	3
Japan	3	Isac	3
Nestle	3	Cassie	3
Ginza	3	homes	3
varieties	2	five	2
shop	2	company	2

Interestingly enough, these two articles also share the sole common word that the Jaccard top pair shared in “edition”. This word might have been the key to them being matched together for the Euclidean and Jaccard Similarity scores due to the fact they both don’t contain much political language.