

CIS530 HW4: Submission Example

5. Free-response Questions (10 points)

Results

- Task: The fourth column of `will_play_text.csv` contains the name of the character who spoke each line. Using the methods described above, **which characters are most similar? Least similar? **
- Results
 - In the play “Henry IV”, the most similar characters are FALSTAFF and PRINCE HENRY. The least similar characters are Thieves, SIR WALTER BLUNT
- Discussion

I have used 3 similarity methods (cosine similarity, jaccard similarity, dice similarity) to run the similar check for the characters in Henry IV play. For the most similar characters, the 3 methods all gave the same answer. FALSTAFF and PRINCE HENRY both represents contrasting ideas of honor, leadership, and maturity. That’s why the words they use could have most similarity. Under cosine similarity, the similarity score is high as 0.9334344176164313 (1 means the same). For the least similar pairs, the 3 methods also gave the same answer: Thieves and SIR WALTER BLUNT. SIR WALTER BLUNT is a noble knight. He represents completely different from thieves. So far, the performance between 3 similarity methods is the same. But from the score with different methods are different. The cosine similarity has highest similarity scores overall take 0.9334344176164313 as an example, the same pair has 0.6248453117563065 in dice similarity method and 0.45438183580229824 in jaccard similarity.

Based on the observation, the cosine similarity method tends to have higher score and jaccard similarity tends to have lowest score.

Extra Credit (15 points)

Results

- Explanation
- The experiment I did is to use both term to document matrix and context matrix.
- When using term to document matrix, I used cosine similarity method, Jac card similarity method and dice similarity method.
- When using context matrix, I also applied similarity method, Jac card similarity method and dice similarity

- When generate term-to-document matrix, and context matrix, I only used the words showed in Sim999.txt
- Results
 - The Sim999.txt has 1028 words.
 - The shape of term-document matrix is 1028x36
 - The shape of context document matrix is 1028x1028

For term_to_doc_matrix

Matrix Type	Similarity Method	Calculated Correlation	
TF_IDF matrix	Similarity Method	-0.042450838943562275	
TF_IDF matrix	Cosine similarity	-0.02028911744438534	
TF_IDF matrix	Jaccard similarity	-0.02028911744438534c	

For term context matrix without using PPMI

Context Window Size	Cosine Sim	Jaccard Sim	Dice Sim
1	-0.05819418	0.015818536	0.015818536
2	-0.078344083	-0.040023336	-0.040023336
3	-0.081037121	-0.049540543	-0.049540543
4	-0.072660238	-0.046024354	-0.046024354
5	-0.078471704	-0.05447142	-0.05447142
6	-0.090222029	-0.068305526	-0.068305526
7	-0.092299013	-0.073112192	-0.073112192
8	-0.092164871	-0.072227079	-0.072227079
9	-0.092025334	-0.072230931	-0.072230931
10	-0.092989052	-0.073477924	-0.073477924
11	-0.092600939	-0.07291668	-0.07291668
12	-0.092507636	-0.072859943	-0.072859943
13	-0.091782585	-0.071967962	-0.071967962
14	-0.091734707	-0.07185668	-0.07185668
15	-0.091734707	-0.071921152	-0.071921152
16	-0.09170598	-0.071849677	-0.071849677
17	-0.091744283	-0.071808798	-0.071808798
18	-0.091615011	-0.071777034	-0.071777034

19	-0.091581497	-0.071765238	-0.071765238
20	-0.091557557	-0.07178164	-0.07178164
21	-0.09159586	-0.071779243	-0.071779243

For term context matrix with using PPMI, the correlation between the similarity method and Sim999 data set

Context Window Size	Cosine Sim	Jaccard Sim	Dice Sim	
1	0.072303852	0.04705001	0.04705011	The best value
2	0.064690925	0.03533272	0.03533272	
3	0.057892787	0.02565632	0.02565632	
4	0.050686566	0.01751899	0.01751899	
5	0.05201825	0.01681684	0.01681684	
7	0.051907771	0.01387027	0.01387027	
8	0.051579796	0.01198453	0.01198453	
9	0.052379867	0.01447365	0.01447365	
10	0.05281895	0.0141957	0.0141957	
11	0.052939799	0.01478383	0.01478383	
12	0.052935771	0.01485634	0.01485634	
13	0.052754498	0.01474354	0.01474354	
14	0.052645734	0.01466298	0.01466298	
15	0.052267075	0.0143528	0.0143528	
16	0.051410499	0.01398662	0.01398662	
17	0.051539408	0.01399468	0.01399468	
18	0.051591777	0.01391411	0.01391411	
19	0.051611919	0.013898	0.013898	
20	0.051648175	0.01400273	0.01400273	
21	0.051716658	0.01412359	0.01412359	
22	0.051491067	0.01391814	0.01391814	

For term context matrix, the context window size and similarity method relationship.

The following is the context window size VS 3 different similarity's correlation to the Sim999.txt

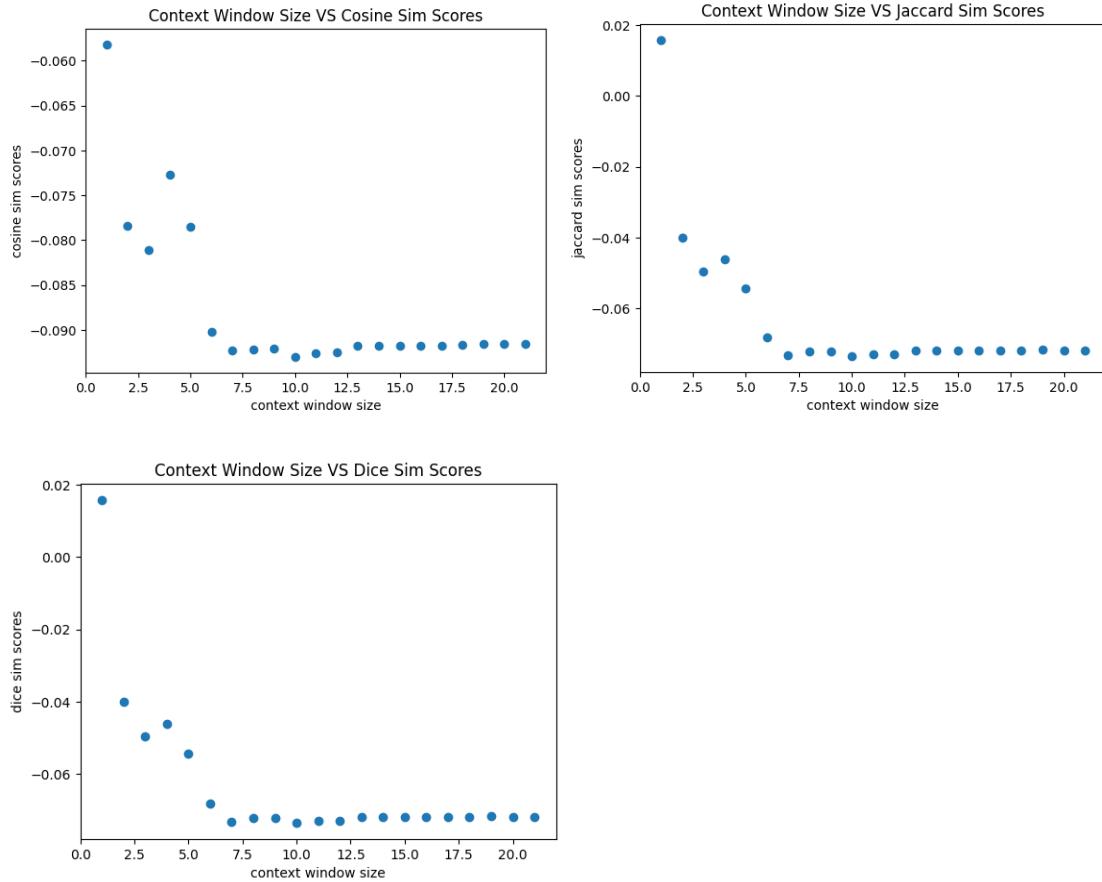
The following setup gave highest correlation:

Context Window Size = 1

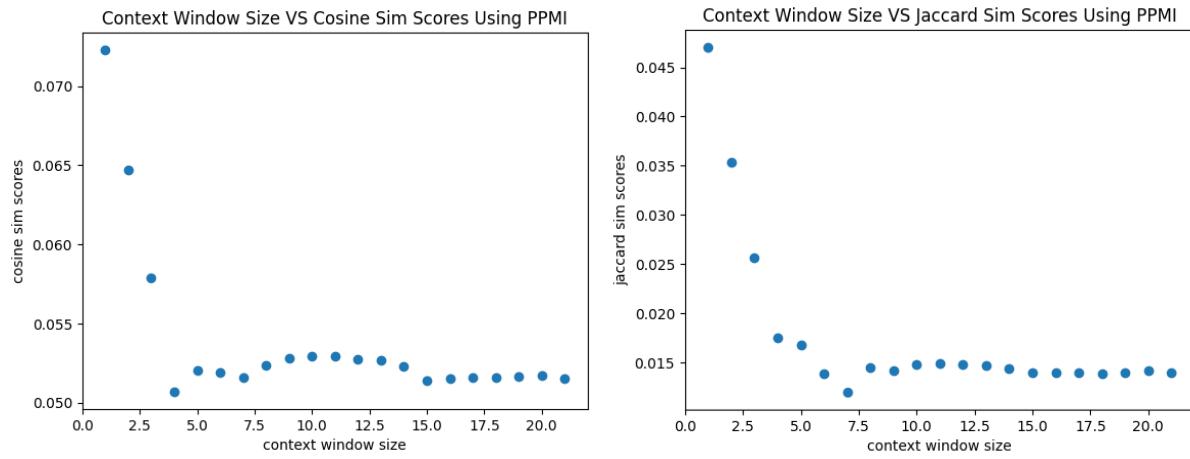
Similarity Method: Cosine Sim

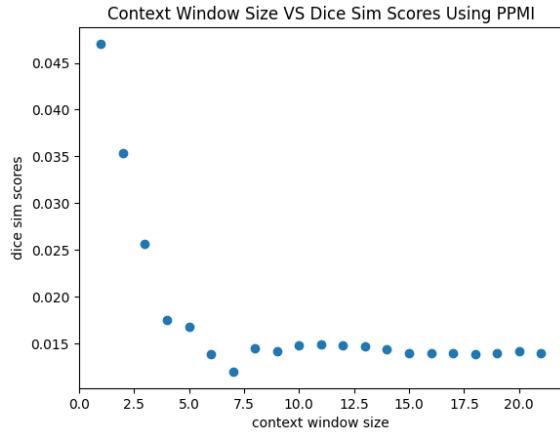
Highest correlation score: **0.072303852**

Method: Apply PPMI on the context windows



The following is the context window size VS 3 similarity method





- Discussion

Based on the experiments, I noticed that the correlation is low for tf-idf matrix and term-context matrix when there's no PPMI applied. The PPMI step is very necessary.

Why the correlation is low is because there are 224 words in Sim999 but not in Shakespeare's play. These missing words will lead to low correlation rate.