ANLP Assignment2 Report

s2124897, s2020314

1 Introduction

There exist multiple ways to compute context vectors and similarities. It is important to study their difference and effectiveness. In our report, we use Positive Point-wise Mutual Information(PPMI) to create vectors. Vectors can be represented by keeping all the related word features or be transformed into smaller vectors with key features using Principle Component Analysis(PCA). We implement four basic similarity measures — Euclidean distance, Cosine measure, Jaccard measure and Dice measure. We mainly investigate the influence of word frequency on different similarity measures and the improvement PCA can make. We choose word pairs with different degrees of similarity — highly, moderately or not very similar. For each case, we present two sets of words — one with higher overall word frequency and the other lower. To ensure consistency, the second set is generated by substituting each word in the first set with lower-frequency similar word. We calculate the Spearman rank-order correlation[1] between the similarity score and the minimum frequency of the word in word pairs to reflect the effect of word frequency towards each method.

2 Question Identification and Word Selection

Our report is based on the assumption that word frequency may affect the measuring result of word vector similarity. In order to calculate the similarity for each pair of words, we need to fix the vector representation of word and choose the similarity measure. We define word frequency in each pair as the smaller number of occurrence for the words in the word pair. We divide the question into three main scenarios, providing word pairs (see Appendix B for example word sets) that are intuitively highly, moderately or not very similar. For scenario 1, we consider verbs that are used for expressing preference. We use emphasis spelling, antonym and synonyms as low-frequency replacement. For scenario 2, we consider most-followed female singers on twitter to be intuitively moderately related. We find out that usernames (@) are always more frequent than hashtags (#). For scenario 3, we look at alternate spellings[2]. British spelling is always less frequent than its American counterpart in Twitter.

We make the discussion by analysing word frequency effects under different scenarios. We compare the correlation of high and low word frequency word set to show whether word frequency influence each similarity method and in what degree, whether some methods are more vulnerable than others, whether different scenario change the vulnerability of each method. We try to implement PCA for vector dimension reduction to see whether it will mitigate the influence.

3 Methods Compared

3.1 Different Similarity Measures

We use Euclidean distance, cosine measure, Jaccard measure[3] and Dice measure[4] for similarity measuring, where $sim_{Euclidean}(\vec{v},\vec{w}) = \sqrt{\sum_{i=1}^{N}(v_i - w_i)^2}$, $sim_{Jaccard}(\vec{v},\vec{w}) = \frac{\sum_{i=1}^{N} min(v_i,w_i)}{\sum_{i=1}^{N} max(v_i,w_i)}$, $sim_{Dice}(\vec{v},\vec{w}) = \frac{2\times\sum_{i=1}^{N} min(v_i,w_i)}{\sum_{i=1}^{N}(v_i+w_i)}$.

3.2 Principal Component Analysis

Principal Component Analysis[5] is a method to reduce the dimensionality of our n dimensional data to k dimensions by selecting new dimensions(principal components) with the greatest variances. The first principal component points to the direction of the greatest variance. The i^{th} principal component is pointing to the greatest variance that is perpendicular to the $i-1^{th}$ principal component. The first step to compute PCA is to subtract the mean from each attribute. Secondly, we need to compute the covariance matrix that is an nxn matrix shows the covariance of two features, where n is the number of attributes and an element of the matrix is:

$$cov(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2}$$

The third step is to find the eigenvectors and eigenvalues, because the eigenvector is equivalent to the variance along the eigenvector and the eigenvector points the direction to the greatest variance. Next, we choose the highest k eigenvectors with the highest eigenvalues. Finally, we project our data to the chosen eigenvectors.

4 Analysis and Conclusion

4.1 Quantitative Analysis

The correlation between the similarity score of each word pair and the lowest frequency of the word in that pair can be used to reflect how much the corresponding word frequency and the similarity score is related. The higher the correlation, the more influential the word frequency is towards the similarity method. For each similarity method, we calculate two cases of correlation under each scenario(table 1).

Scenario(Frequency)	cosine	jaccard	dice	euclidean
Scenario1(high)	0.19**(-0.24)**	0.11**(-0.29)**	0.11**(-0.29)**	0.58**(0.58)**
Scenario1(low)	0.54**(-0.18)**	0.57**(-0.10)**	0.57**(-0.03)**	0.71**(0.32)**
Scenario2(high)	-0.17**(0.16)**	-0.03**(0.46)**	-0.03**(-0.17)**	0.36**(0.30)**
Scenario2(low)	0.50**(0.17)**	0.40**(0.27)**	0.40**(-0.31)**	0.41**(0.24)**
Scenario3(high)	0.02**(-0.03)**	-0.26**(-0.13)**	-0.26**(0.32)**	0.02**(-0.17)**
Scenario3(low)	0.36**(-0.36)**	0.55**(-0.31)**	0.55**(-0.31)**	0.55**(0.60)**

Table 1: The correlation between the similarity score of each word pair and the lowest word frequency in that pair for four similarity methods under three different scenarios. The bold values in braces indicate using PCA.

Without using PCA, for each scenario, all four similarity methods processing the word list with lower frequency returns higher correlation. Considering that two word lists under each scenario are selected to be intuitively similar but only differ in word frequency, for all scenario and all similarity method, when the word frequency in a pair is lower, the similarity will be more sensitive to the frequency. For each similarity method, the influence of word frequency are similar, except for jaccard and dice similarity method under scenario 3. The highest correlation gap of 0.81 means that jaccard and dice are more vulnerable to word frequency especially when the word pair is not very similar.

Using PCA with 2 principal components, our correlation values decrease significantly in most of the cases using Cosine, Jaccard and Dice similarities. Considering that our original number of attributes were usually above 10.000, a reduction of dimensionality can hold the risk of information loss. If we can further verify the effectiveness of PCA in preserving key information, we can deduce that the drop in correlation shows the positive effect of PCA towards decreasing the influence of word frequency to the calculation of cosine, jaccard and dice similarity. However, the correlation in Euclidean distance changes less significantly, meaning that Euclidean method bears more tolerance when we reduce the dimensionality. Also, when using PCA, it is no longer true that all the similarity methods with low frequency word lists gives higher correlations.

4.2 Qualitative Analysis

An example plot(Appendix C) shows our word vectors from Scenario 3 after using PCA. It is shown that the words with close meanings, e.g., 'color' and 'colour' or 'neighbour' and 'neighbour', are close to each other in the PCA space and each cluster is separated. It shows the power of PCA when using only 2 principal components, covering around 71% of our original dataset's variance at Scenario 1's low frequency words.

Under Scenario 1, word 'love' and 'loove' are considered the same except for different word frequency. We take another intuitively similar word 'luuv' to check the similarity between it and either of the former words. Check Appendix D figure 2. For example, using cosine method, the similarity between 'luuv' and 'loove' is around 0.125 and the one between 'luuv' and 'love' is around 0.0125. It's clearer to see the gap between these two similarities in the graph, where it looks huge proportionally. It means that frequency of words with the same meaning can change the similarity calculated. Also, the cosine similarity can show more information than dice and jaccard, presenting a similarity over zero between 'love' and 'luuv'.

Appendices

A Preliminary Task Output

Sort by cosine similarity

0.36	('cat', 'dog')	169733	287114
0.17	('comput', 'mous')	160828	22265
0.12	('cat', 'mous')	169733	22265
0.09	('mous', 'dog')	22265	287114
0.07	('cat', 'comput')	169733	160828
0.06	('comput', 'dog')	160828	287114
0.02	('@justinbieber', 'dog')	703307	287114
0.01	('cat', '@justinbieber')	169733	703307
0.01	('@justinbieber', 'comput')	703307	160828
0.01	('@justinbieber', 'mous')	703307	22265

B Word Set under Three Scenarios

Scenario 1 (Highly Similar):
High Word Frequency: love, luv, hate, like, enjoy
Low Word Frequency: loove, luuv, resent, dislike, detest
Scenario 2 (Moderately Similar):
High Word Frequency: @ladygaga, @rihanna, @nickiminaj, @selenagomez, @ddlovato
Low Word Frequency: #ladygaga, #rihanna, #nickiminaj, #selenagomez, #demilovato
Scenario 3 (Not Very Similar):
High Word Frequency: color, flavor, humor, labor, neighbor
Low Word Frequency: colour, flavour, humour, labour, neighbour

C PCA Vector Graph

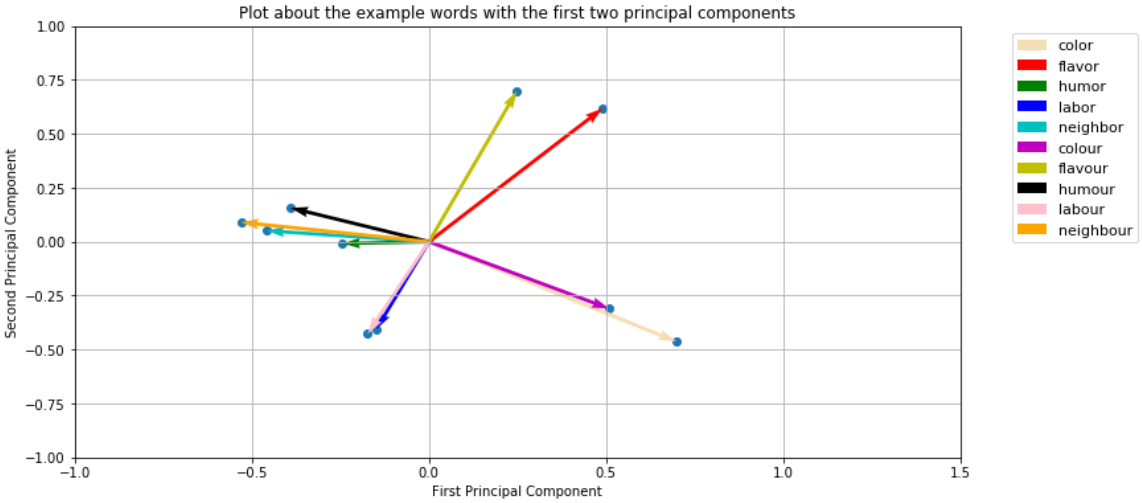
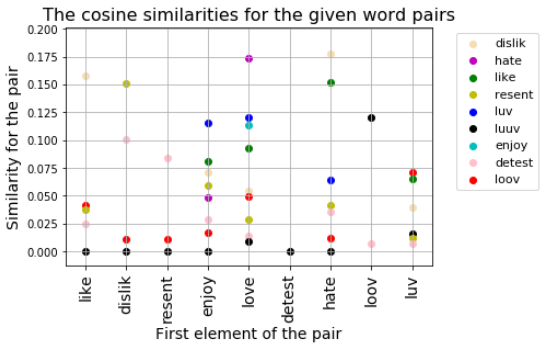
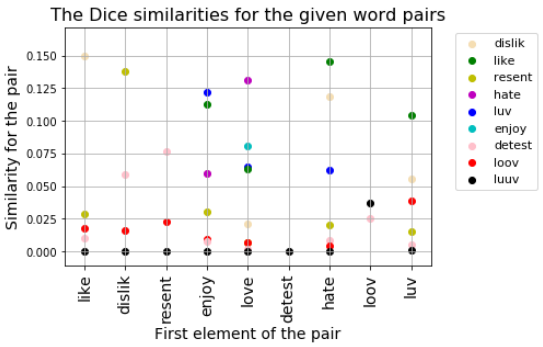


Figure 1: PCA vector presentation under Scenario 3

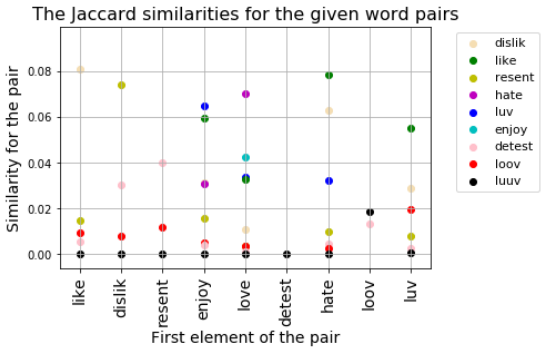
D Similarities



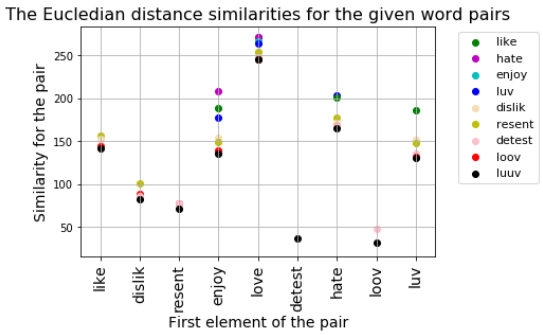
(a) cosine



(b) dice

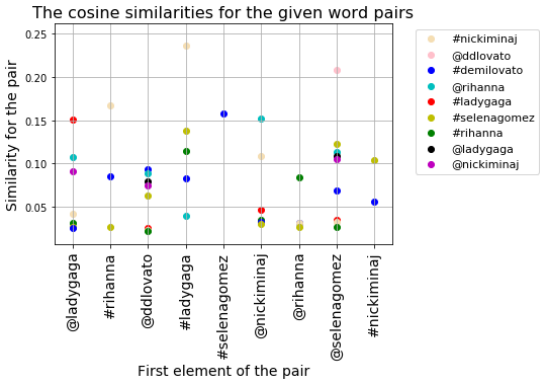


(c) jaccard

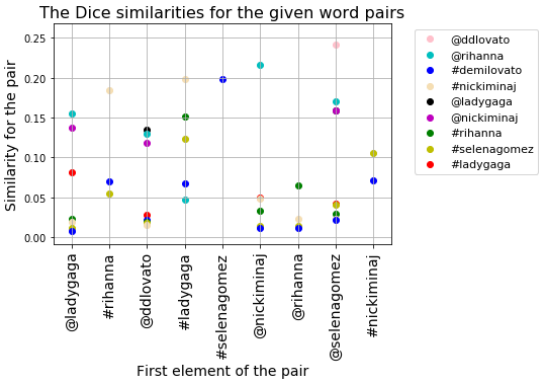


(d) Euclidean

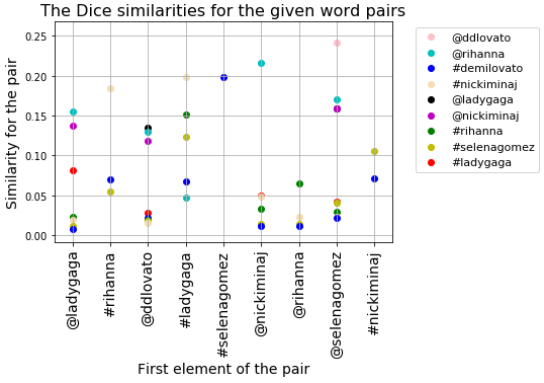
Figure 2: Scenario 1



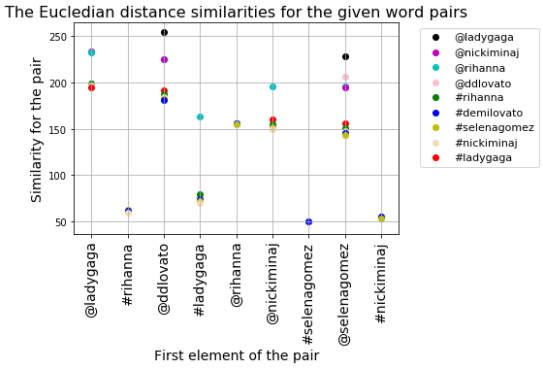
(a) cosine



(b) dice



(c) jaccard



(d) Euclidean

Figure 3: Scenario 2

References

[1] Martina Udovičić et al. "What we need to know when calculating the coefficient of correlation?" In: *Biochemia Medica* 17.1 (June 2007), pp. 10–15. ISSN: 13300962. DOI: 10.11613/bm.2007.002. URL: https://www.biochemia-medica.com/en/journal/17/10.11613/BM.2007.002.

[2] *The differences in British and American spelling — Oxford International English Schools*. URL: https://www.oxfordinternationalenglish.com/differences-in-british-and-american-spelling/.

[3] *Jaccard Index / Similarity Coefficient - Statistics How To*. URL: https://www.statisticshowto.com/jaccard-index/.

[4] Jun Ye. "Multicriteria decision-making method using the Dice similarity measure based on the reduct intuitionistic fuzzy sets of interval-valued intuitionistic fuzzy sets". In: *Applied Mathematical Modelling* 36.9 (Sept. 2012), pp. 4466–4472. ISSN: 0307904X. DOI: 10.1016/j.apm.2011.11.075.

[5] I H Witten, E Frank, and M A Hall. "Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition". In: *DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES, 3RD EDITION*. Morgan Kaufmann Series in Data Management Systems. 340 PINE STR, 6TH FLR, SAN FRANCISCO, CA 94104-3205 USA: MORGAN KAUFMANN PUB INC, 2011, pp. 1–629. ISBN: 978-0-08-089036-4; 978-0-12-374856-0.