ANLP – Assignment 2

StudentIDs: s1982660, s1902539

Introduction

In this report, we focused on 2 questions considering the impact of word's frequency on the similarity scores of different methods and the implication we can make by looking at the similarity scores of some pairs. We applied 2 different methods on 2 words sets and tried to answer the questions by analysing the results.

Questions we investigated:

Question 1 (Q1): How the frequency of appearance of the words of each pair in tweets affected the similarity scores of different methods.

Question 2 (Q2): Whether the similarity scores of different methods can be used to indicate the country where some huge events (eg. natural disasters) happened in 2011.

Tested word sets:

set_A*=['apple','apricot','avocado','banana','blackberry','blackcurrant','blueberry','cherry','coconut','fig','grape','grapefruit','lemon','lime','lychee','mandarin','mango','melon','nectarine','orange','papaya','passionfruit','peach','pear','pineapple','plum','pomegranate','raspberry','strawberry','watermelon']*

set_B*=['avalanche','earthquake','hurricane','tsunami','flood','thailand','japan','egypt','greece','uk','china','australia','cameroon','austria','france','india','brazil','argentina','chile','russia','syria','georgia','spain','indonesia','kenya','algeria','portugal','ecuador','croatia','serbia']*

Both word lists contain 30 words. We considered it as a reasonable length for our word lists because, on the one hand, it was small enough to make it easier for us to focus on specific examples, but on the other hand, it was big enough to provide representative results for our statistical analysis.
For set A, we selected words in 'fruit' category from a vocabulary website[1]. Since we wanted to explore the relationship between similarity and frequency (Q1), we searched for fruits that cover a wide range of similarity, ranging from very similar, moderately similar, to not very similar. To be more specific, for 'very similar words', we selected fruits are consumed at the same seasons (e.g. watermelon and grape in summer) or belong in the same category (raspberry and strawberry both belong to the natural order Rosaceae). As for 'not very similar words', we selected fruits that are cultivated in different regions and at different seasons. We also ensured that our words covered a wide range of frequencies by checking their frequency via 'counts' document from lab8.
We wanted to use set_B to answer our second question (Q2), specifically to examine whether the similarity scores can be used to indicate the country where natural disasters happened in 2011. Thus, we decided to use a list that consists of the 5 most common natural disasters as well as 25 countries from all around the world. The countries we chose also cover a wide range of similarities and frequencies, something we ensured by applying the same selection process as for set_A.

Methods we chose:

Method A: Dice similarity measure with PPMI context vectors

Method B: Dice similarity measure with t-test context vectors

Brief descriptions

1. Dice similarity measure (method for computing similarities): Dice similarity measure is a similarity measure used to calculate the difference between two numerical vectors [2]. It is a widely used similarity measure which has been used widely in several areas, from biogeography [3] to medicine [4]. The Dice similarity measure takes values in the range [0, 1]. We used it as an alternative to the cosine similarity for Methods A and B.

To calculate the Dice similarity measure of two vectors, we divide the double of their dot product by the summation of their squared lengths.

Formula: $s_{Dice} = \frac{2*A \cdot B}{|A|^2+|B|^2}$ [2]

2. T-test (method for computing context vectors): We used t-test as an alternative for PPMI to compute the context vectors for method B. It is a measure that was widely used for collocation discovery. More specifically, given the mean and the variance of a specific sample, the t-test can be used to calculate how

probable is that the sample was drawn from a specific bigger population. The t-test assumes that the data comes from a normal distribution. [5]

Formula: $assoc_{t-test}(l,f) = \frac{P(l,f)-P(l)P(f)}{sqrt(P(l)P(f))}$ [9]

Results and analysis

Quantitative analysis

Through our quantitative analysis, we aimed to answer Q1. We used both sets of words, because they both contain words across a wide range of similarities and frequencies and because we wanted to evaluate our results on two different sets to ensure the generalization of our observations. To analyse the relationship between similarity scores and frequencies, we have used four different frequency measures; the frequency of appearances of the less and the most frequently appeared word of the pair, the total frequency of appearances for both words of each pair and the number of common appearances of both words of each pair. The Pearson Correlation Coefficients (PCCs) between the different similarity measures and the different kind of frequencies that we used in our experiments can be found in figure 5.

Looking through figure 5, we first observed that the PCCs between the similarity scores and frequencies are relatively larger for the frequency of the less appeared word and the frequency of the common appearances of the words of each pair compared to the frequency of the most appeared word and summation of frequencies of the pair's words. This observation was consistent not only across the two sets of words that we present in this report but also in many other sets of words that we used during our trials. These results can also be visualised in figures 1-4. Specifically, in figures 1,4 someone can observe that pairs with a higher frequency of the less appeared word and higher common frequency of appearance have a higher probability to get higher similarity scores too. On the contrary, the similarity values are more equally distributed across the different frequencies in figures 2,3.

From figure 5, we can also observe that the extent of correlation between the different kind of frequencies and the different measures depends significantly on the set of words that is used in each case. For example, for Set_B there are higher absolute values of Pearson Correlation Coefficients and therefore higher linear correlation between frequency and similarity with Method B that with Method A for all types of frequencies. Nevertheless, these results cannot be generalized since for Set_A the absolute values of the Pearson Correlation Coefficients are higher with Method A, for 3 of the frequency measures. This inconsistency of results indicates that none of the methods we used could be considered better to avoid correlation between frequencies of appearance and similarity scores since this correlation also depends on the set of words that is used in each case.

Qualitative analysis

For set_B, we observed that the similarity scores between certain natural disasters (earthquake, tsunamis) and the country Japan were ranked very high in both methods (Figures 7,8). Thus, we decided to search for the keywords 'earthquake', 'tsunami' and 'Japan' on the twitter search page [6] for 2011. We found that in lots of tweets people were talking about an earthquake and a tsunami that happened in Japan in that year (Figure 6). To confirm the validity of those messages, we also found some papers [7,8] referring to the two huge events that happened in 2011. All this evidence supports the hypothesis of Q2.

Conclusions

Our results claim that there is a stronger relationship between the similarity scores and the frequency of the least appeared word of each pair and the frequency of common appearances of both words of each pair, than the relationship between the similarity scores and the frequency of the most appeared word of each pair and the summation of frequencies of both words of each pair.

Moreover, the inconsistency of the results across the two data sets we used in our quantitative analysis, indicates that none of the two methods we used could be considered superior to avoid the impact of the different types of frequency on similarity scores. That's because for each set of words the method that created a lower PCCs between the similarity scores and the different types of frequency was different.

From our qualitative analysis, we concluded that the similarity scores generated by each of the methods we used, can be indeed used to imply the country where some huge events (eg. natural disasters) happened in 2011. Thus, the short answer for Q2 is that the hypothesis is correct, although the events should be major since no other natural disaster was so obviously presented in the results to the same extent as the earthquake and the tsunami in Japan.

Output of the preliminary task

```
Sort by cosine similarity with ppmi context vectors
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
```

Figures:

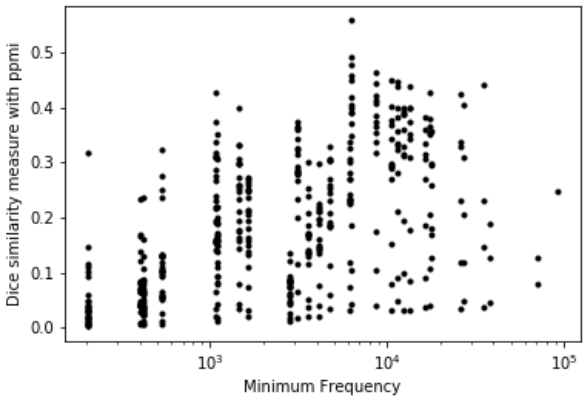

Figure 1: The number of appearances of the less frequently appeared word of each pair in tweets, by the Dice measure's value for the specific pair of words. PPMI context vectors and Set_A words are used.

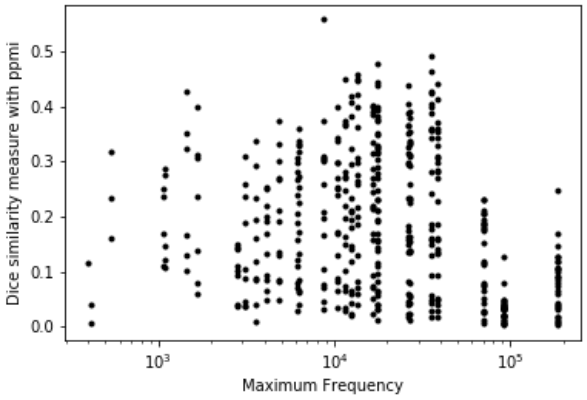

Figure 2: The number of appearances of the most frequently appeared word of each pair in tweets, by the Dice measure's value for the specific pair of words. PPMI context vectors and Set_A words are used.

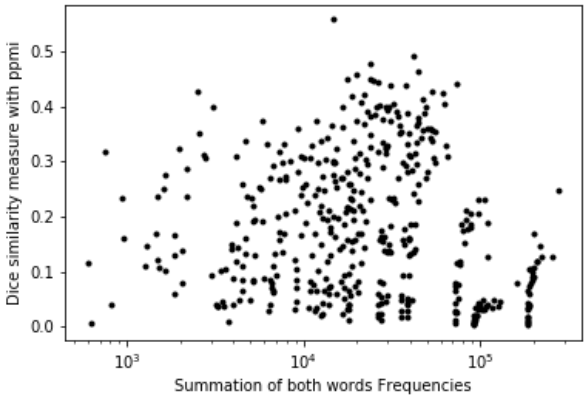

Figure 3: The summation of appearances of both words of each pair in tweets, by the Dice measure's value for the specific pair of words. PPMI context vectors and Set_A words are used.

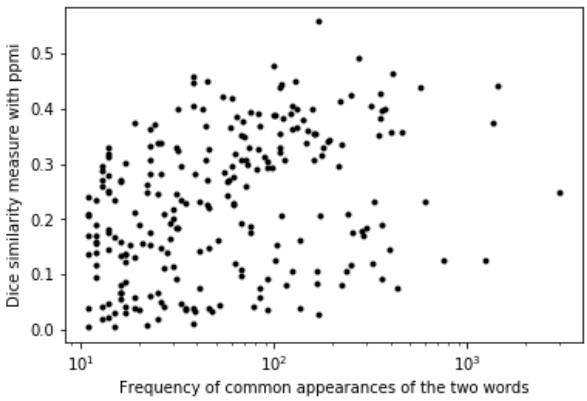

Figure 4: The number of common appearances of both words of each pair in tweets, by the Dice measure's value for the specific pair of words. PPMI context vectors and Set_A words are used.

Similarity Measure Method	Frequency Measure	Pearson Correlation Coefficient		Figure	
		Set_A	Set_B	Set_A	Set_B
Method A: Dice similarity measure with PPMI context vectors	Minimum Frequency	0.54	0.63	1	-
	Maximum Frequency	-0.23	0.10	2	-
	Summation of Frequencies	-0.11	0.27	3	-
	Common Appearance Frequency	0.37	0.57	4	-
Method B: Dice similarity measure with t-test context vectors	Minimum Frequency	0.39	0.65	-	-
	Maximum Frequency	-0.13	0.33	-	-
	Summation of Frequencies	-0.06	0.46	-	-
	Common Appearance Frequency	0.43	0.73	-	-

Figure 5: The Pearson Correlation Coefficients between frequency and similarity across different similarity and frequency measures.

Figure 6: Examples of 2011 tweets with that include the words 'tsunami', 'earthquake', 'Japan'

```
Sort by Dice similarity with ppmi context vectors (Method A)
0.49    ('earthquak', 'tsunami')        64570   28187
0.42    ('japan', 'tsunami')            145416  28187
0.42    ('japan', 'earthquak')          145416  64570
0.40    ('syria', 'egypt')              11881   20309
0.40    ('argentina', 'chile')          23509   20107
```

Figure 7: Highest Dice measure's values for set_B. Context vectors created with PPMI.

```
Sort by Dice similarity with t-test context vectors (Method B)
0.52    ('japan', 'tsunami')            145416  28187
0.47    ('japan', 'earthquak')          145416  64570
0.45    ('earthquak', 'tsunami')        64570   28187
0.31    ('argentina', 'brazil')         23509   50428
0.27    ('argentina', 'chile')          23509   20107
```

Figure 8: Highest Dice measure's values for set_B. Context vectors created with t-test.

Reference List:

[1] Fruit. (2017, March 12). Retrieved November 20, 2019, from https://www.vocabulary.cl/english/fruit.htm.

[2] Lin, D. (1998, July). An information-theoretic definition of similarity. In *Icml* (Vol. 98, No. 1998, pp. 296-304).

[3] Muldoon, K. M., & Goodman, S. M. (2010). Ecological biogeography of Malagasy non-volant mammals: Community structure is correlated with habitat. *Journal of Biogeography*, *37*(6), 1144-1159.

[4] Nguyen, T. M., & Wu, Q. J. (2011). Robust student's-t mixture model with spatial constraints and its application in medical image segmentation. *IEEE Transactions on Medical Imaging*, *31*(1), 103-116.

[5] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

[6] Twitter Search. (2011, January 1). Retrieved November 23, 2019, from https://twitter.com/search-home.

[7] Krausmann, E., & Cruz, A. M. (2013). Impact of the 11 March 2011, Great East Japan earthquake and tsunami on the chemical industry. *Natural hazards*, *67*(2), 811-828.

[8] Mimura, N., Yasuhara, K., Kawagoe, S., Yokoki, H., & Kazama, S. (2011). Damage from the Great East Japan Earthquake and Tsunami-a quick report. *Mitigation and adaptation strategies for global change*, *16*(7), 803-818.

[9] Ljubesic, N., Boras, D., Bakaric, N., & Njavro, J. (2008, June). Comparing measures of semantic similarity. In *ITI 2008-30th International Conference on Information Technology Interfaces* (pp. 675-682). IEEE.