Yılmaz BAYSAL
N19139400

CMP 614 – Assignment 2 Report

In this assignment, we are expected to implement multiple dimensionality reduction algorithms with and without using a weighting function and compare the results. To be able to compare the results, we need to calculate correlation between human judgements and word similarities. Also, we need to apply statistical significance test between the correlations to prove the experimental result.

1. Implementation

I used pre-generated matrix file to load the corpus. After that, I applied the PPMI (Positive Pointwise Mutual Information) weighting to it. I used both SVD and NMF dimensionality reduction algorithms to reduce the dimensions of the matrix.

I implemented combinations of algorithms as follows:

- NMF with raw occurrence matrix
- NMF with PPMI weighted matrix
- SVD with raw occurrence matrix
- SVD with PPMI weighted matrix

2. Results

You can see the correlation results on the below figure, which clearly shows that; using PPMI weighting gives better results with dimensionality reduction on the given corpus. We cannot sure about the difference between NMF and SVD only by looking at the correlations, because they are close to each other. Thus, we have to apply a statistical significance test on the correlations to find out whether the difference between them occurred randomly or not.

```
No Weighting - SVD: 0.30582863246203584
PPMI Weighting - SVD: 0.5589331566185118
No Weighting - NMF: 0.3043505185831239
PPMI Weighting - NMF: 0.5848591858255822
```

Figure 1: Correlation value of each implementation

I used Z-Test as a statistical significance test. First, I applied z-transformation on correlations by taking inverse tangent (arctan(x)) of them. Then I applied the below formula to find z-value. After finding the z-value, I looked up the z-table to find the corresponding p-value for it.

$$Z = \frac{z_1 - z_2}{\sqrt{\left|\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}\right|}}$$

You can see the z-test results on the below figure.

```
Significance Test between SVD and PPMI-SVD   |   p-value: 0.013642988257242824
Significance Test between NMF and PPMI-NMF   |   p-value: 0.0064129957025753975
Significance Test between SVD and NMF (co-occurrence)   |   p-value: 0.49544867807536647
Significance Test between SVD and NMF (PPMI weighted)   |   p-value: 0.39367483886528415
```

Figure 2: Significance test results of the correlation pairs

2.1. PPMI or Co-occurrence Counts?

The first and second rows show the comparison between raw co-occurrence matrix and PPMI weighted version of it. The both p-values are less than 0.05, which is our level of significance value. Thus, we can say that the difference between the correlations are statistically significant and we cannot randomly achieve these results with using raw co-occurrence counts. In other words, there is a significant difference between raw co-occurrence counts and PPMI for our samples.

2.2. SVD or NMF?

The third and fourth rows show the comparison between SVD and NMF dimensionality reduction algorithms. The both p-values are higher 0.05. Thus, we can say that the difference between the correlations are not statistically significant and we can randomly achieve these results with using any of two algorithms. In other words, there is no significant difference between SVD and NMF for our samples.