Network Features for Better Word Representation

Piyush Khushlani 15ME30057
Vineet Jain 15ME30044

Objective:

Finding out the effectiveness of some Network measures for representing better Word Similarity.

Network Structure:

A **Graph** is a pictorial representation of a set of objects where some pairs of objects are connected by links. The interconnected objects are represented by points termed as vertices, and the links that connect the vertices are called edges.

A typical Graph Structure

Subgraph for a word is obtained by considering only the directly connected nodes from the corpus with given edge weights.

Network Measures:

Pearson Correlation Coefficient:

The correlation coefficient between rows *i* and *j* is defined as

$$r_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}} = \frac{\sum_k A_{ik}A_{jk} - \frac{k_i k_j}{n}}{\sqrt{k_i - \frac{k_i^2}{n}}\sqrt{k_j - \frac{k_j^2}{n}}}$$

where n is the union of the corresponding subgraphs obtained for any two words and k_i, k_j are number of nodes in each of the subgraphs respectively. $A_{ik} = 1$ if the word i is connected to word j and similarly for A_{jk}. Therefore, $\sum_k A_{ik}A_{jk}$ is number of nodes in the intersection of 2 subgraphs obtained.

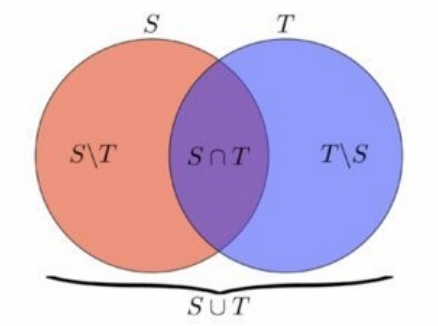
Cosine Similarity:

This similarity measure is defined as the inner product of two vectors. That is,

$$similarity(x, y) = \cos\theta = \frac{x \cdot y}{||x|| * ||y||}$$

Consider the *ith* and *jth* rows of the adjacency matrix A as vectors. Then the cosine similarity between vertices i and j is

$$\sigma_{ij} = \frac{\sum_k A_{ik}A_{jk}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$



$n_{ij} = S \cap T$, $k_i = S$ & $k_j = T$

where n_{ij} is number of nodes in the intersection of two subgraphs obtained for corresponding words and k_i and k_j are number of nodes in each of the subgraphs respectively.

Degree Centrality:

Degree centrality is defined as the ratio of the number of neighbours of a vertex with the total number of neighbours possible. Mathematically,

$$\text{Degree Centrality} = \frac{k}{N-1}$$

where k is the intersection of two subgraphs obtained for corresponding words, and N is the total number of nodes in the network.

The variance of the distribution of degree centrality in a network gives us the **centralization** of the network. One can see that a star network is an ideal centralized network,

whereas a line network is less centralized.

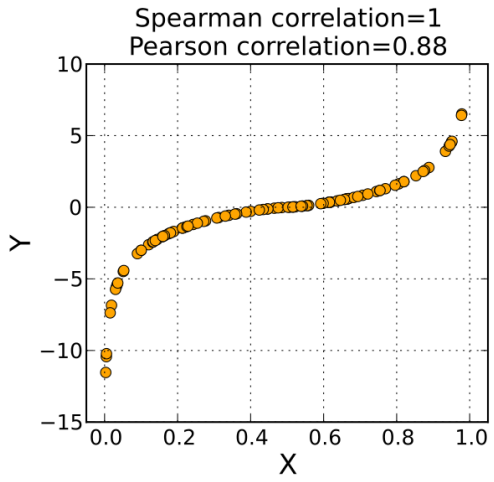
Star Network and Line Network

Experiments:

There are various types of datasets present and the experiment was conducted on the datasets namely WordSimilarity-353, ESL, MC, RG and MTURK. Here we decoded the datasets to obtain the *Word1, Word2* and *Human Mean Index.* Based upon the Network measures, we applied the idea of Pearson Correlation Coefficient, Cosine Similarity and Degree Centrality to obtain the alliance between them by the use of NetworkX package of python to create a Graph between two subgraphs obtained for *Word1* and *Word2* from corpus of Wikipedia.

Evaluation:

Spearman's rank correlation coefficient or **Spearman's rho** between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.



$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

where r_s is Spearsman's rho, d_i is the difference between the two ranks of each observation and n is total number of observations.

Therefore, the resulting measure was correlated with the actual Human Mean Index by Spearsman's correlation which evaluated the output of the network measures with Human Mean.

Results:

Accuracy on ESL dataset was observed to be 64%. For other datasets, results are as follows:

DataSet or Network Measure	Pearson Correlation	Cosine Similarity	Degree Centrality
WordSimilarity-353	rho = 0.50939 p_value=1.051 e-24	rho = 0.50939 p_value=1.051 e-24	rho = 0.50939 p_value=1.051 e-24
MC	rho = 0.85434 p_value=1.901 e-09	rho = 0.85434 p_value=1.901 e-09	rho = 0.85434 p_value=1.901 e-09
RG	rho = 0.81158 p_value=2.416 e-16	rho = 0.80792 p_value=4.172 e-16	rho = 0.80792 p_value=4.172 e-16
MTURK	rho = 0.63625 p_value=1.022 e-88	rho = 0.64038 p_value=3.297 e-90	rho = 0.64038 p_value=3.297 e-90