# CS786 Assignment-2

**Amrit Singhal (150092)**
Department of Computer Science
IIT Kanpur
Kanpur, India
amrits@iitk.ac.in

**Aarsh Prakash Agarwal (150004)**
Department of Electrical Engineering
IIT Kanpur
Kanpur, India
aarshp@iitk.ac.in

## 1 Problem Statement

The aim of this assignment was to reproduce the results of word similarity judgements done by the machine as compared to human word similarity judgements. Two separate models of machine word similarity judgements were to be implemented and tested against the human scores:

- Normalized Google Distance
- Word2Vec similarity

Let use first describe each of the models being used.

## 2 Model descriptions

### 2.1 Normalized Google Distance (NGD)

The Normalized Google Distance (NGD) is a semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords. [2] The keywords that are close in the semantic sense in the natural language haev a low NGD value, whereas those that differ significantly in their semantic sense have a higher NGD value between them. The NGD between two terms $a$ and $b$ is given as:

$$NGD(a,b) = \frac{\max(\log(hits(a)), \log(hits(b))) - \log(hits(a,b))}{\log(N) - min(\log(hits(a)), \log(hits(b)))}$$

Here,

$$\begin{aligned} hits(a) &= \text{Number of pages returned by Google search that contain the word } a \\ hits(a,b) &= \text{Number of pages returned by Google search that contain both the words } a \text{ and } b \\ N &= \text{Total number of pages searched by Google multiplied by the average number of singleton} \\ &\quad \text{search terms occurring on a page (which is 1000)[2]} \end{aligned}$$

An NGD value of $0$ corresponds to very good semantic similarity, whereas a higher NGD value means a worse and worse semantic match for the terms.

### 2.2 Word2Vec

Word2Vec[5] provides us with a vector representations for words, such that it captures their linguistic context. The vectors are obtained using machine learning techniques, by training a shallow two-layer neural network. We use these representations to get the semantic similarity between words. For two given term $a$ and $b$, the cosine similarity between the two corresponding word2vec vector representations is used as the similarity score between the two words. A score of $1$ corresponds to a perfect semantic match, whereas a decreasing value corresponds to lesser and lesser match. We used the word2vec-api [3] to directly obtain this similarity score for the given terms.

# 3 Dataset

This assigment used the WordSimilarity-353 Test Collection dataset [4]. It contains two sets of English word pairs along with human-assigned similarity judgements.

# 4 Solutions

## 4.1 Solution 1

The first part required us to calculate the Normalised Google Distance between the words in each pair in the dataset. This was done using the `googlesearch` library [1] in python. The library manages the making of a request to the Google search engine, for the text query the we provide to it, and returns the number of *hits* (a hit is webpage that Google fetches for this query). To get the total number of webpages indexed by Google, we query the word *'the'*, as it is highly likely to be present in almost all of the webpages.

The runtime of the code is around 45 min i.e. 7.5s/pair. This is because we had to add a sleep delay of 2 seconds between every two queries being sent to Google, to prevent bombarding the servers with queries and the IP address being blocked consequently.

The NGD scores obtained for each pair were stored in a file (named `ngd.txt`) for future use.

## 4.2 Solution 2

This part required us to compare the NGD similarity scores obtained from part 1, against the human similarity scores from the dataset.

The human similarity scores were between 1 and 10, with 10 being for the best similarity. The NGD scores were between 0 and 1, with 0 being the best similarity. Therefore, NGD scores had to first be scaled to the same space. The scores were scaled between 0 and 10, and then every NGD score was subtracted from 10 to obtain the new NGD scores that lied in the same space and in the same order as of the human ratings.

These scores were then plotted against each other, to obtain the plot shown in figure 1.

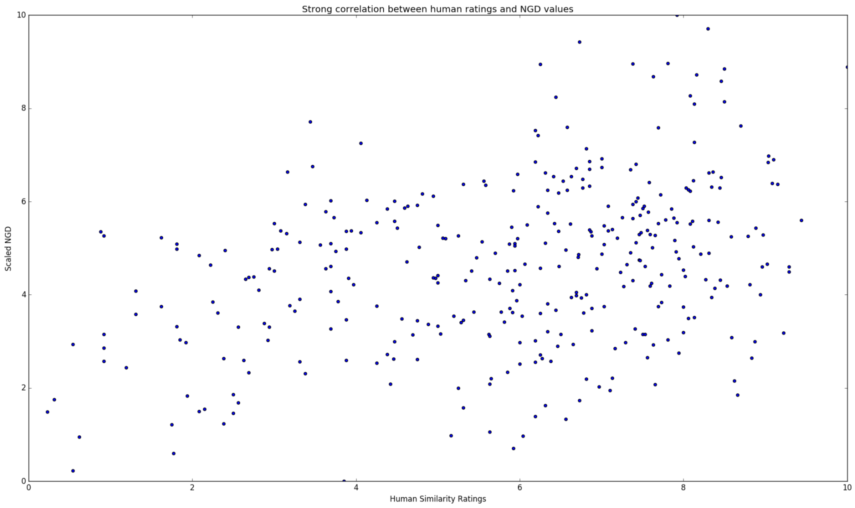

Figure 1: Scaled NGD vs Human Similarity ratings

The graph seems to indicate a very **slight positive correlation** between the scaled NGD values and the human similarity ratings.

### 4.3 Solution 3

For this part, we were required to use the Word2vec similarity and compare it to the human ratings. The word2vec-api [3] was used to get the similarity score between the word2vec embeddings of the two words in each pair of the dataset. These obtained values were scaled from $[0, 1]$ to $[0, 10]$ to match the human rating space. These scaled scores were then plotted against the human ratings, to obtain the plot as shown in Figure 2.

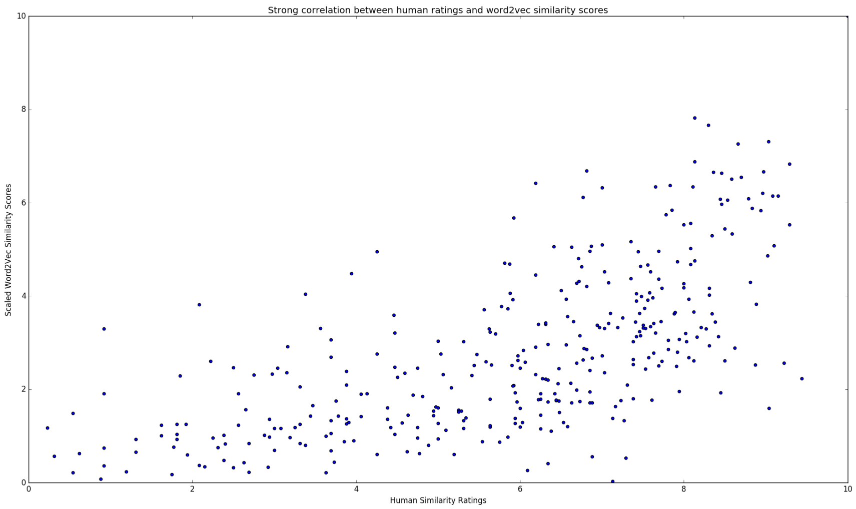

Figure 2: Scaled word2vec similarity scores vs human similarity ratings

The plot clearly indicates a **higher positive correlation** between the word2vec similarity scores and the human similarity ratings.

## References

[1] Google Search Library. https://pypi.org/project/google/.

[2] Normalised Google Distance. https://en.wikipedia.org/wiki/Normalized_Google_distance.

[3] word2vec-api. https://github.com/3Top/word2vec-api.

[4] Evgeniy Gabrilovich. The WordSimilarity-353 Test Collection. http://www.cs.technion.ac.il/ gabr/resources/data/wordsim353/.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.