ASSIGNMENT-3
Computational Cognitive Science (CS786)

Sumit Kumar (18111075)

April 5, 2019

---

1) Calculate Normalized Google Distance(NGD)

We have used web scrapping using Beautiful soup to get total hits for a page. For calculating NGD between word1 and word2 ,we have followed these steps:

Step 1: For each of word1 and word2 we construct a request as https://www.google.com/search?q="+word

Step 2: Next we look for id=ResultStats in received html response.

Step 3: Corresponding to ResultsStats we have total hits for given word as we have seen on google.(About 4,55,00,00,000 results)

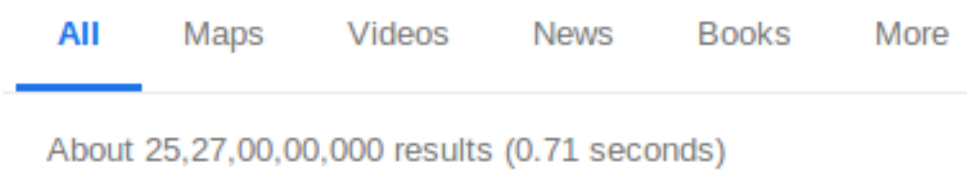

Figure 1: hits

Step 4: The number of pages indexed by Google was estimated by the number of hits of the search term "the," which was 25,270,000,000 hits. Assuming there are about 1,000 search terms on the average page this gives N = 25,270,000,000,000.

Step 5: Next we calculate NGD as follows:

$$\text{NGD}(w1, w2) = \frac{\max\{\log(hit(w1)), \log(hit(w2))\} - \log(hit(w1 + \text{"" } + w2))}{\log(N) - \min\{\log(hit(w1)), \log(hit(w2))\}}$$

2) Plot of Scaled NGD vs Human Mean

For converting distance into similarity we have used : $1/\exp(ngd)$
Then we have normalized similarity between 1 to 10:

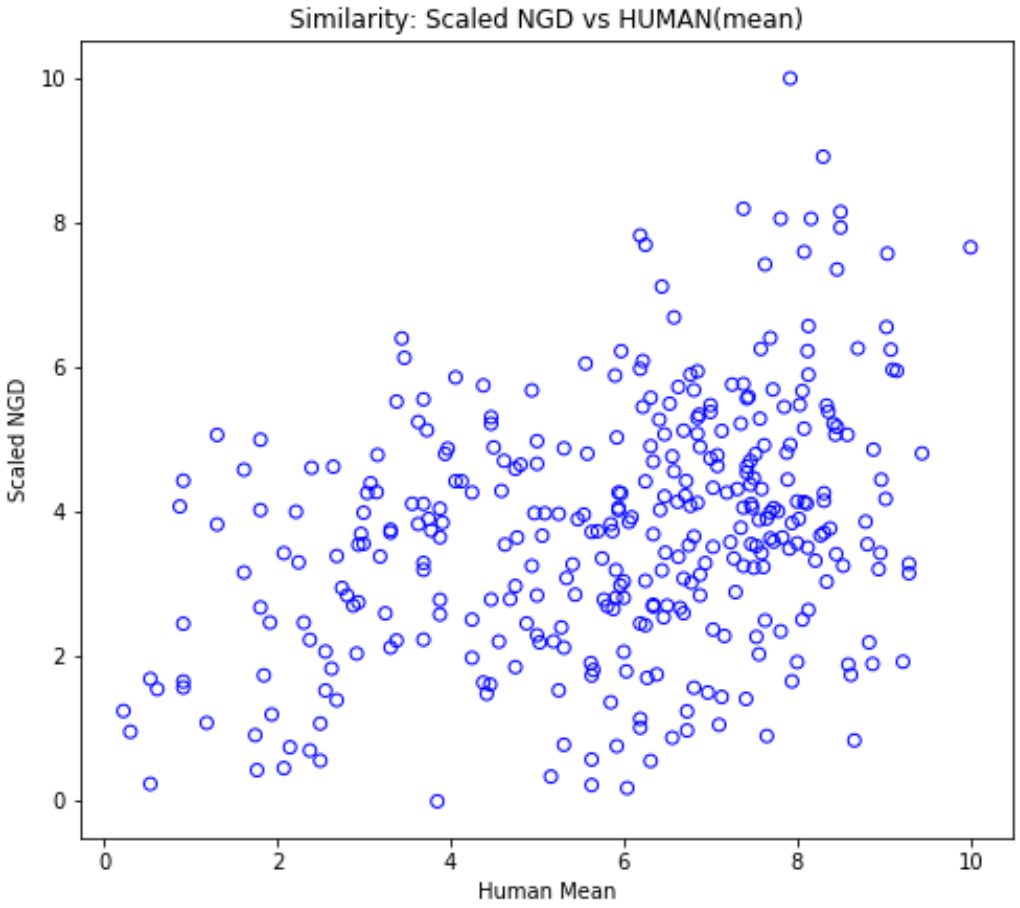

Figure 2: Similarity Human vs ScaledNGD

3) Using Word2Vec api service

We have used Google News Group pretrained model for similarity calculation. Next we normalized similarity between 0 to 10 in order for plot

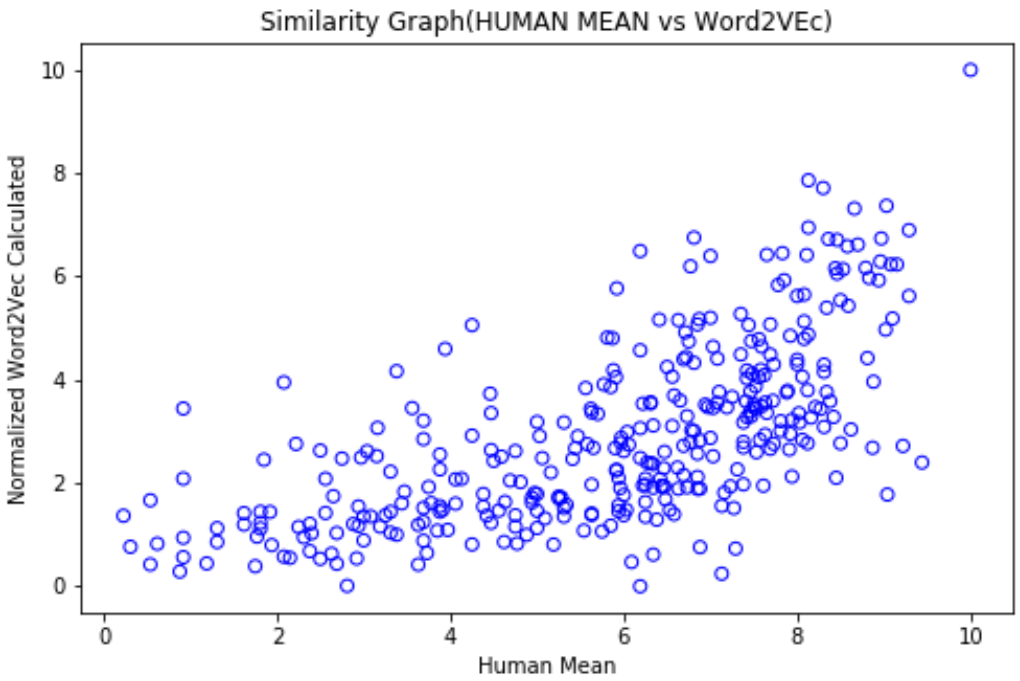

Figure 3: Similarity Human vs Word2vec