MCA Assignment 3 - Text Representation and Retrieval

Shagun Uppal (2016088)

May 2020

1 Word2Vec

word2vec: Word2Vec refers to a two-layered neural network which vectorises words by taking into consideration nearby words that fall within the context window (both forward and backward). It can wither be done using continuous bag of words approach (CBOW) which basically uses context words to predict the target word or via the skip-gram model which uses the target to predict its context words. For the results reported below, the skip-gram model is used.

Vocab size: 55800Embedding size: 64Context window: 2

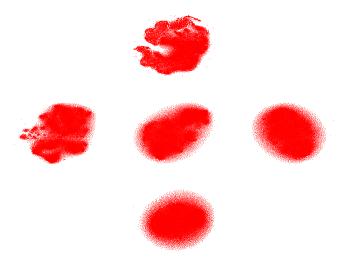


Figure 1: Latent space visualisation for the entire vocabulary for the 5 training epochs in sequential order.

Some of the similar words obtained:

• **Decision:** Limits, Results

• Beaches: Corn, Discussion

• Travelling: Communication, Mashups

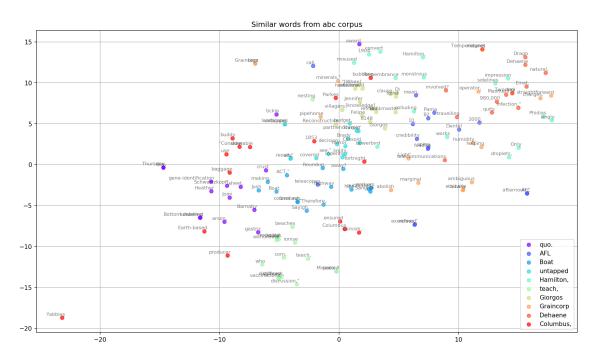


Figure 2: Latent space visualisation for some randomly chosen words and top 10 similar words for each of them represented in same color.

The loss converges well until the third epoch and after that remains almost stable.

2 Document retrieval using query expansion

Overview: Both the algorithms, relevance feedback as well as relevance feedback using query expansion uses similar technique for document retrieval wherein a user feedback is used for results on the previous query in order to determine its relevance before performing a new query. The latter also exploits similar words (synonyms) for performing each query.

Below are the results obtained over three iterations for both the algorithms.

| | mAP score | | |
|---|-------------|-------------|-------------|
| Method | Iteration 1 | Iteration 2 | Iteration 3 |
| Relevance feedback [1] | 0.5601 | 0.5915 | 0.5959 |
| Relevance feedback with query expansion [2] | 0.6090 | 0.6181 | 0.6181 |

Table 1: Mean average precision for three methods of document retrieval.

Analysis: The results obtained with relevance feedback with query expansion are better than the ones obtained with relevance feedback algorithm. This is inline with the expectations since in [1], user gives input on documents which is used to refine the query, whereas in [2], users give input on query terms or phrases. So, [2] works better since it has both local information (from query optimisation) as well global information (from automatic thesaurus generation).