Unsupervised Opinion Mining

Name:	**Ankit Anil Lade**
Registration No./Roll No.:	18038
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	October 04, 2021
Date of Submission:	October 24, 2021

1 Introduction

In this dataset of assignment we find the most popular opinions for the question: "What qualities do you think are necessary to be the prime minister of India?" This dataset has 38 rows of people's opinions about the question. We find the most popular opinions using wordnet and word2vec techniques. Word2Vec models used are CBOW and Skip-gram models which were trained on Google news data sets to get the opinion from the dataset. For the wordnet techniques Wu Palmer Similarity score was used. This problem deals with opinion mining problem. After preprocessing the data we get 111 words.

2 Methods

2.1 Wordnet

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. Wu Palmer Similarity was used for the wordnet technique. It calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer). [1]

1. Removing punctuations and stopwords.
2. Creating the temporary vocabulary of nouns and adjectives using POS tagging.
3. All terms except noun and adjective should be removed.
4. Find the synsets and derivationally related forms containing nouns and adjectives and create the main vocabulary.
5. Calculate the Wu-Palmer similarities between the synsets.
6. Assign clusters to each word in the vocabulary.

7. Clustering words and extracting the names for each clusters on the basis of their frequency.

2.2 Word2Vec

Word2Vec is not a singular algorithm, rather, it is a family of model architectures and optimizations that can be used to learn word embeddings from large datasets. CBOW and Skip-gram models are used. In the CBOW model, the distributed representations of contexts are combined to predict the word in the middle. While in the Skip-gram model, the distributed representation of the input word is used to predict the context.

1. Removing punctuations and stopwords.
2. Remove everything except noun and adjective.
3. Initialize the CBOW and Skip-gram models.
4. Build vocabulary and then train the dataset.
5. Cluster the words using Kmeans Algorithm and find the similarity vector and predict the clusters formed.

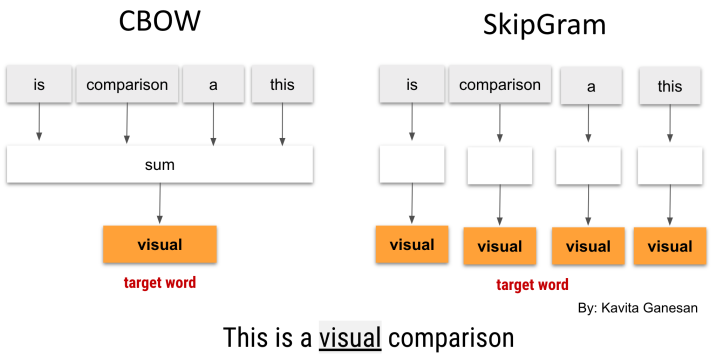

Figure 1: Difference between CBOW and skipgram techniques.

2.3 K-means

K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

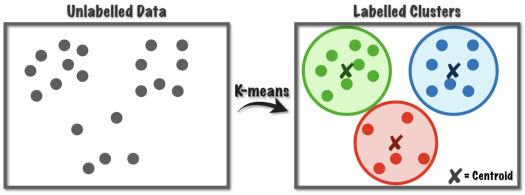

Figure 2: Example of K-means clustering

Table 1: Comparison of Words obtained from different model and the given golden labels

CBOW	**Wordnet**	**Skip-gram**	**Golden Labels**
intelligence	ability	politics	confident
honest	leadership	bravery	determination
common	integrity	problem	diplomacy
prejudices	equality	achievement	honesty
county	morality	capable height	intelligence/knowledge
term	compromise	wise	leadership
whole	basic	sure	long term vision
underprivileged	negotiator	people	political skills
eloquent	county	common	relate to diverse groups
proud	good	strong	humble height

3 Experimental Analysis

It is observed that while using the Word2Vec model with K-means clustering the extracted words change as we change the number of iterations we see change in clustering. This may be due to the fact that the clusters formed are random. On the other side it is observed that the Wordnet technique gives a same result in every case. It may be because we use Wu Palmer Similarity score to find the distance between two words which may not change. The Wu-Palmer similarity score is normalized to get a value between 0 and 1. We choose only the top-10 clusters in both the models. It is also observed that 84 clusters are formed during Wordnet technique. It is also observed that the words like county are clustered from words like country.

4 Discussion

The methods discussed are fairly accurate and work on the small dataset. WordNet and Word2Vec comparison is a narrow topic that has been thoroughly explored here. The majority of future work on Word2Vec will focus on developing tools and algorithms that use word vectors. This research establishes a solid foundation for such attempts. Here we found that the wordnet performed better than word2vec techniques. Here we also compare the results in the table form for the different methods with the given golden labels given by human experts and we find that many words are generated in these models.
A drawback that I found in this method was that it only returns a single word rather than a group of words and phrases. They represent a word as a single vector.
Various types of clustering can be explored. Here we used the centroid based clustering using Kmeans method.

References

[1] Abram Handler. An empirical study of semantic similarity in wordnet and word2vec. 2014.