Report

Name: Krishna Sreenivas

Student ID: 800984436

Email ID: [ksreeniv@uncc.edu](mailto:ksreeniv@uncc.edu)

**Description of Vector embeddings:**

**Vector embedding set 1**

The dimension of 1900000 300 must be entered in the glove dataset at the start to make it compatible for word2vec.

File name: glove.42B.300d.txt

Ref Link: <https://nlp.stanford.edu/projects/glove/>

**Vector embedding set 2**

Dimension of 2000000 300 must be entered in lexvec embeddings file at the start.

File name: lexvec.commoncrawl.300d.W.pos.neg3.txt

Ref Link: <https://github.com/alexandres/lexvec>

**Problem 2**

Word embeddings of synonyms or antonyms have similar embeddings because the context in which the word appears matter. Since the vector construction requires context.

For example: Word2Vec which takes context of the words for vector construction.

Since the models take 2 different kinds of co-occurrences into consideration between words.

1. First order co-occurrence also called syntagmatic association:

This type of co-occurrence are typically found near a particular word.

For example: rode is a first order or syntagmatic associate of the word bike.

1. Second order co-occurrence also called paradigmatic association:

This type of co-occurrences take paradigms into consideration.

For example: (1) musical instrument – piano / guitar / violin / drum, and (2) vehicle – car / bus / train / plane. ‘Musical instrument’ and ‘vehicle’ are hyper-ordinates, i.e. they are names of categories which help to group together the members of the category. The arrangement is hierarchical, with a hyper-ordinate term at the top (such as ‘musical instrument’ or ‘vehicle’) and, at the next level down, a group of co-hyponyms such as ‘guitar’ and ‘violin’ or ‘bus’ and ‘train’. We can say that piano is a second order associate of guitar.

Once the vectors are created, one can find the similarity between these words by finding the cosine similarity of the 2 vectors. Since while vector construction co-occurences would be taken into consideration and hence due to this, antonyms and synonyms have similar embeddings.

**Problem 3**

Have designed 2 analogy sets which are as below

1. : cellphones

iphone apple galaxy samsung

redmi xiaomi motorola lenovo

pixel google htc-one htc

b) : national-animals

tiger india kangaroo australia

lion belgium komodo-dragon indonesia

cow nepal markhor pakistan

The set ‘a’ has brand names and manufacturers of cellphones as the analogies. And set ‘b’ has national animals as analogies.

The accuracy of both the sets was reported to be 0 against both the vector embeddings.

**Difficulties faced:**

* The parsing was computationally intensive since the vector embeddings was huge and using few inbuilt functions eased the execution process.
* Apart from the above, few trivial coding errors were encountered, which were appropriately handled.

**Performance:**

Below are the accuracy results of both vector embeddings:

Accuracy of the model for lexvec embeddings - 12.12 %

Accuracy of the model for glove embeddings - 19.08 %