NLP: Assignment 6

- *Shreyash Arya (2015097)*

1.

The model is trained on the full ‘GoogleNews-vectors-negative300.bin’ dataset. (Results may vary if the model is trained partially)

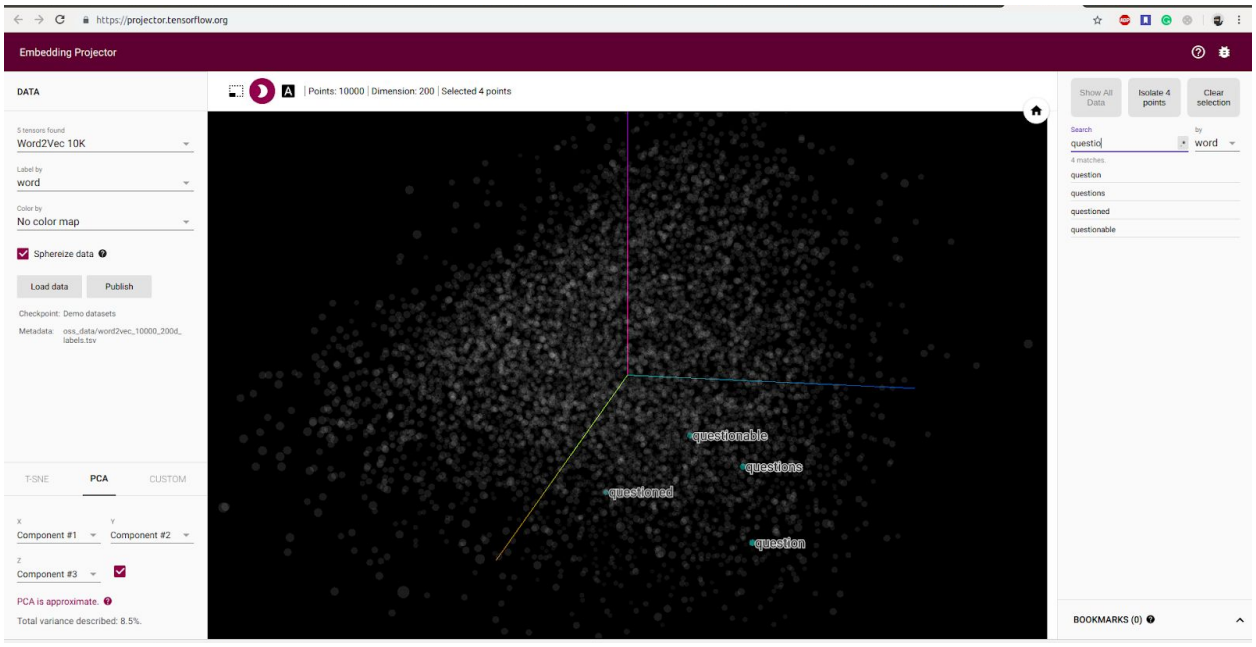
Results:

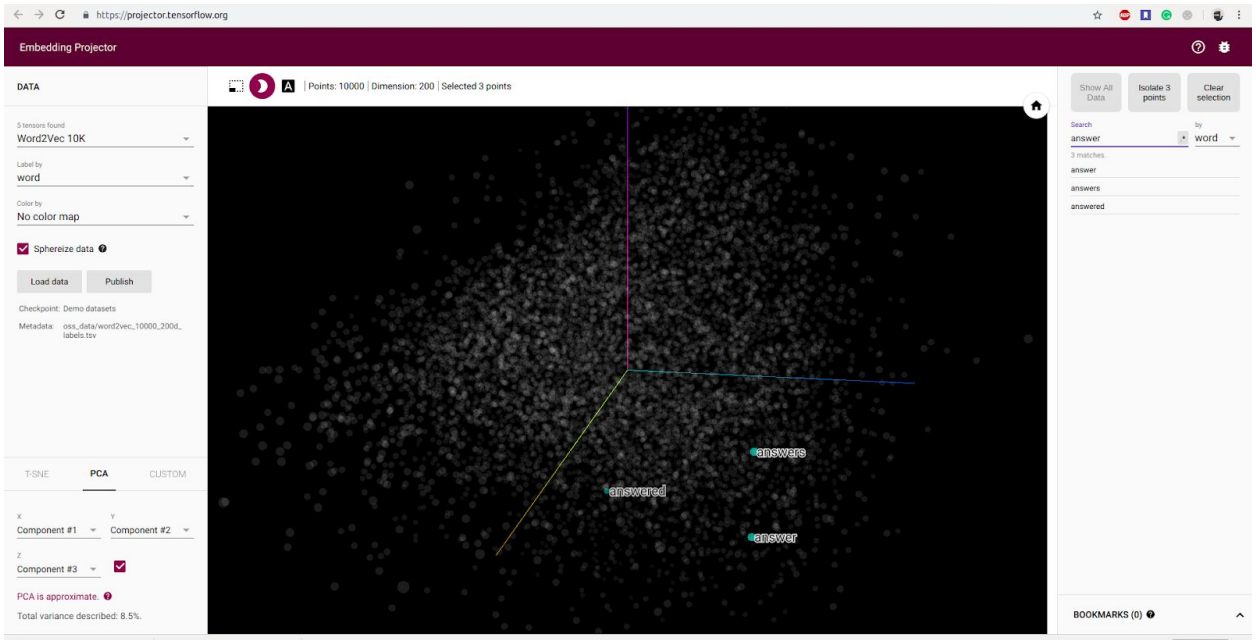
If Delhi is the capital of India then what is the capital of China?
→ Answer: Beijing with score 0.7975

If ISRO is related to India then what is related to USA?
→ Answer: STScI with score 0.4270

STScI stands for Space Telescope Science Institute.
Note: It seems quite intuitive that the result should have been ‘NASA’ but if we try to train the model partially say limit = 50000 parameter, then the result comes out to be ‘NASA’. Also, if we use ‘America’ instead of ‘USA’ then still we get ‘NASA’ as the result. This can be because of how word vectors lie in the space and how the similarity is compared by taking the vector difference.

Visualization:





Inference: As we can infer from the above visualizations that the words are vectorized in the space the words which are in past tense lie in the same direction in both the cases (question and answer) and same for the plural form which lies to the right of the word mentioned. This can be used to calculate the vector difference and score different similarity measures between the words.

2.

- Gensim doc2vec is trained on total data provided except 20 files from ‘comp.graphics’ which is used for testing.
- One single global test file is separated from the test data which is used for both testing measures.
- For different group testing, the global test file is compared from one file from all other folders and normalized score is stated.
- For same group testing, the global file is compared with the other 19 files from the testing data.
- Cosine similarity is used for the comparisons of the document vectors.

Results:

{'soc.religion.christian': 0.2604871, 'talk.politics.guns': 0.15982503, 'talk.politics.misc': 0.32984304, 'rec.motorcycles': 0.14504614, 'talk.religion.misc': 0.26163802, 'comp.windows.x': 0.20664674, 'misc.forsale': 0.277412, 'sci.electronics': 0.35208926, 'sci.med': 0.1697625, 'sci.crypt': 0.31659266, 'sci.space': 0.33144468, 'talk.politics.mideast': 0.3493191,

'rec.sport.hockey': 0.011780897, 'comp.sys.ibm.pc.hardware': 0.40116343, 'rec.sport.baseball': 0.23652647, 'alt.atheism': 0.3434224, 'comp.os.ms-windows.misc': 0.31569117, 'rec.autos': 0.18151899, 'comp.sys.mac.hardware': 0.1832573}

Normalized score for **different groups: 0.2543929977048385**

[0.4512538, 0.37237215, 0.57102746, 0.38034502, 0.32068512, 0.5378068, 0.51333576, 0.5768151, 0.5118998, 0.34471512, 0.39120957, 0.44676933, 0.5355645, 0.48237708, 0.35537267, 0.5237335, 0.5597546, 0.41247818, 0.32921883]

Normalized score for **same groups: 0.4535123392155296**

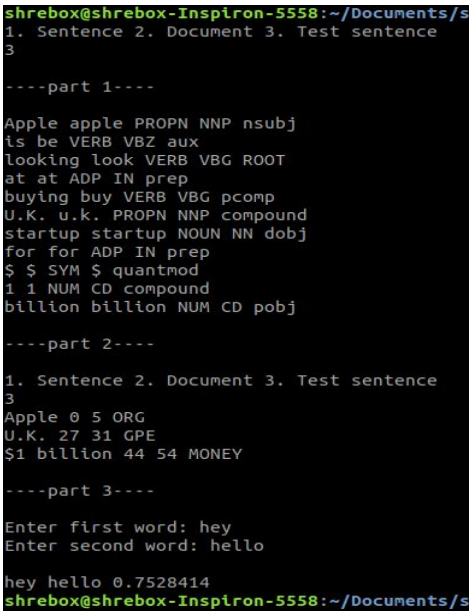
Inferences: As we can see if we compare the global file with the same group from which it is taken, it gives higher similarity (45.3%) as compared to the files from different groups (25.4%).
Link to trained model: https://goo.gl/ueECyp

3.

Spacy is a very nice tool for doing various Natural Language and Processing tasks. It has a quite intuitive documentation which is followed to perform various tasks mentioned in the questioned document.

The implementation is done for both sentence and the whole document comparisons. Also, a test sentence is given for a quick run.

Results:



References:

https://machinelearningmastery.com/develop-word-embeddings-python-gensim/
https://www.smartcat.io/blog/2017/word2vec-the-world-of-word-vectors/
https://radimrehurek.com/gensim/models/doc2vec.html
https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/doc2vec-lee.ipynb
https://radimrehurek.com/gensim/models/doc2vec.html
https://spacy.io/usage/linguistic-features