CSE556 - Natural Language Processing
Assignment 3 Report

Aditya Chetan (2016217)

November 17, 2018

1 Assumptions

- I has assumed that in Question 2, the model had to be trained on the 20 Newsgroups first and then used for measuring similarity
- I have compared document similarities on the basis of mean

2 Question 1

2.1 Part I

For this question I simply loaded the model using `gensim`. After that, using the `most_similar()` function I extracted the top 10 words similar to the given word.
Since I had to find analogoue of China as Delhi is to India, I simply extracted the words with vectors most similar to the vector:

$$vector(Delhi) - vector(India) + vector(China)$$

Similarly, for the second part,

$$vector(ISRO) - vector(India) + vector(USA)$$

The output of the word similarity experiment is given below:

```
1.  [('Beijing', 0.7975110411643982), ('Shanghai', 0.6384025812149048),
('Beijng', 0.6233851909637451), ('Guangzhou', 0.6154200434684753),
('Shenyang', 0.6146994233131409), ('Chinese', 0.6092808246612549),
('Guangdong', 0.6081507205963135), ('Tongzhou', 0.6061089038848877),
('Bejing', 0.6039618253707886), ('Nanjing', 0.5980162620544434)]

2.  [('STScI', 0.42706042528152466), ('Orbital_Sciences_Corporation',
0.4269925355911255), ('GPS_IIR', 0.4206993579864502), ('Thales_Alenia',
0.41337308287620544), ('AMSAT', 0.411637544631958), ('RSC_Energia',
0.41066551208496094), ('Agency_JAXA', 0.410483181476593), ('NOAA_GOES',
0.41036567091941833), ('NROL', 0.4101548194885254), ('NASA', 0.4088003933429718)]
```

2.1.1 Part II

From my observations, I could figure out the following:

- First I visualized **question** with **questions**. I took a screenshot and drew a vector **question** $\longrightarrow$ **questions**
- Then I visualized **answer** and **answers**. I did the same thing to this screenshot.
- Seeing these 2 screenshots side-by-side I realised:
  - The relative positions of the word and their purals was almost the same. Even though the singular words were different, both of them were positioned "below" their plurals in a sense.
  - The vector joining the words to their plurals are also almost identical-looking.
  - This seems to imply that the notion of **plurality** is preserved in the vector space. The vectors drawn the in figure must be the ones that contain the sense of plurality and they look the same irrespective of the singular word.

The screenshots I took are shown in Figure 1 and Figure 2.

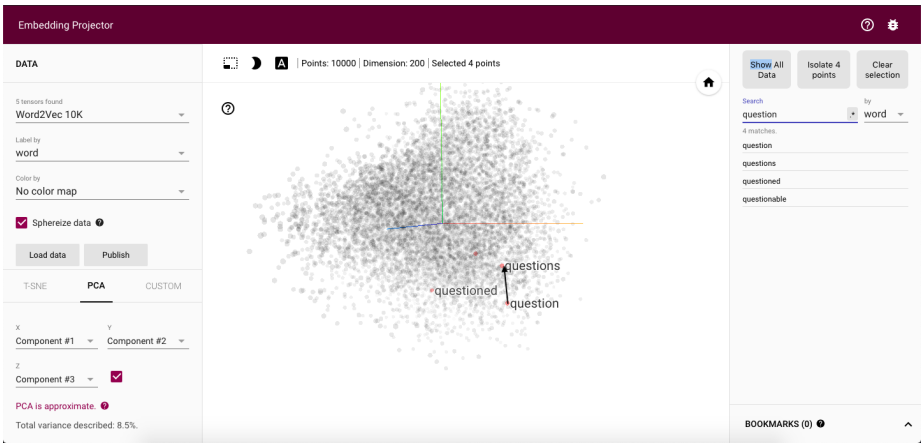

Figure 1: Visualization for **question** and **questions**

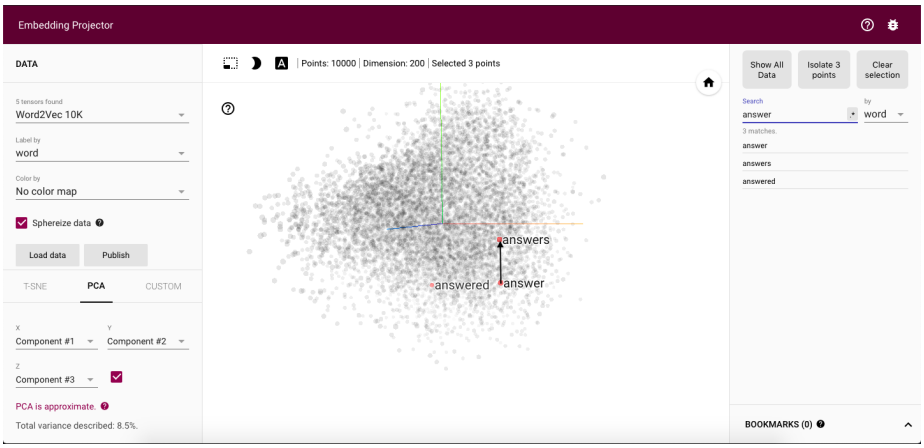

Figure 2: Visualization for **answer** and **answers**

3 Question II

For this question, I first trained my doc2vec model on the 20 Newsgroups dataset. I kept the vector dimensionality as 100 and the number of epochs as 100. Then, I extracted 19 random `comp.graphics` documents, inferred their vectors from my trained model and calculated their cosine similarities with a fixed `comp.graphics` document. I did the same for 19 documents all drawn from different folders. I then calculated the mean of both these sets of cosine similarities and compared them.

The output of the experiment is given below:

```
Mean cosine similarity of graphics docs:  0.011767632
Mean cosine similarity of diverse docs:  -0.03629752
Are graphics documents more similar to a graphics doc on an avg.?:
True
```

4 Question III

In this part, I first tokenised the given data using `SpaCy`. Then in those toks, I accessed their POS tags using `.pos_` and their lemmas using `.lemma_`.

For NER, I accessed the names entities in the tokens using the `.ents` attribute. Then I accessed each of the entities text and label using `.text` and `.label_` respectively.

Lastly, for word similarity, I used the `similarity()` function of the token object to calculate the similarities between the tokens.

The results of Question III are given below:

```
----------------------------------------------------------------


DATA GIVEN
----------------------------------------------------------------

John has a nice house in India
----------------------------------------------------------------


POS TAGGING

John   PROPN
has   VERB
a   DET
nice   ADJ
house   NOUN
in   ADP
India   PROPN
----------------------------------------------------------------


LEMMATIZATION

John   john
has   have
a   a
nice   nice
house   house
in   in
India   india
----------------------------------------------------------------


NER
```

```
John   PERSON
India   GPE
-----------------------------------------------------------------


WORD SIMILARITY EXPERIMENT

Word 1: cat   Word 2: dog  Similarity:  0.010874684
Word 1: apple   Word 2: dog  Similarity:  -0.054777365
-----------------------------------------------------------------
```

5 References

- https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/doc2vec-lee.ipynb
- https://spacy.io/usage