2.1 Theory:

1.Explain negative sampling. How do we approximate the word2vec training computation using this technique?

Training the traditional word2vec model is a huge computational task.It is considered as the classification problem with the number of classes having the size of vocabulary. Applying softmax on it would be computationally large, especially when the size of the vocabulary is large.Negative sampling is a technique used to reduce the computational complexity of word embedding algorithms.It transforms our multiclass classification of vocabulary size into binary classification task.Negative samples are random word-context pairs that are used along with actual context pairs in training.Instead of taking all possible samples,we take few negative samples to reduce the computation.In this method we compute the dot product of target between with positive and negative samples and apply log sigmoid activation function.Our objective function is to maximise the probabilities of positive samples and minimise the probabilities of positive samples.Thus,we approximate the word2vec training computation using negative sampling technique.

2.Explain the concept of semantic similarity and how it is measured using word embeddings. Describe at least two techniques for measuring semantic similarity using word embeddings

Semantic similarity is the degree to which two words are closely related with respect to their meanings.Word embeddings are representations of the words as vectors in high dimensional space,the more closer the vectors are the more similar they are.Hence ,We can measure the semantic similarity between two vectors by calculating the distance between them.

Techniques for measuring semantic similarity using word embeddings:

1.Calculating Euclidean distance: It is the distance between two vectors in high dimensional space.It ranges from zero to infinity.

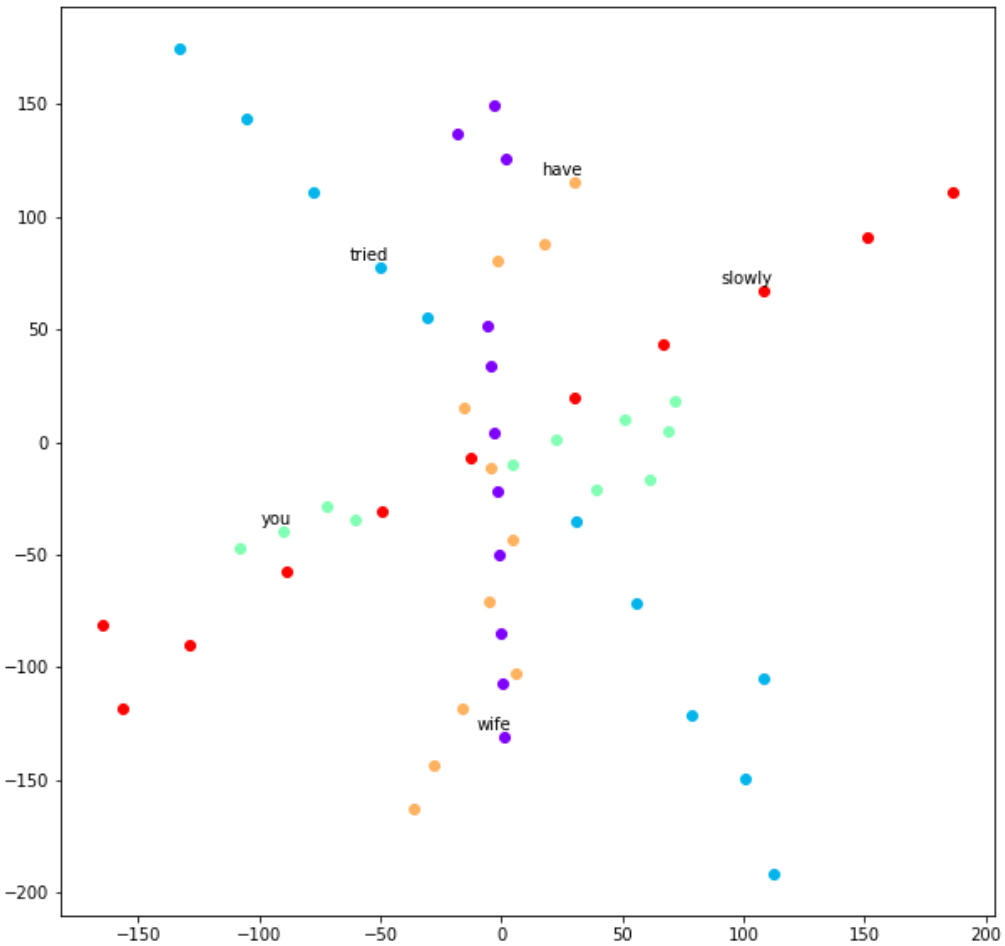
2.Calculating Cosine similarity:It measures the cosine of angle between two vectors.It ranges from -1 to 1.

2.3 Analysis:

1.Co Occurrence Matrix and SVD:

Display of top 10 word vectors for five different words

```
["wife","tried","you","have","slowly"]
```



We can clearly observe that the nearest words wife(NOUN), tried(VERB), slowly(ADJ), you(PRONOUN), have(AUX) scattered in different directions as they are words that occur in different contexts.

Following are the top 10 nearest words for above five words:
word:
wife
Top-words:
brother,sister,husband,nephew,sisters,mom,dad,shook,daughter,girlfriend

word:
you
top-words:
they,we,even,anything,christians,i,your,wondering,believers,wonder,

word:
have
top-words:
had,'ve,suggest,feel,consider,never,got,think,'d,remember

word:
slowly
top-words:
grime,poignantly,globes,were,dynasties,,,corny,lovable,wastes,crew

Titanic:
Top 10 nearest words generated by the model and distance with titanic:

1. anyway 0.28207660600215045
2. damage 0.29214986685834277
3. merit 0.3275931030230981
4. glitches 0.33139029013868493
5. supermarket 0.3331766538075559

6. today 0.3380269739583446
7. stuarts 0.33927308107819676
8. nonsense 0.3430269636690162
9. whatsoever 0.34672381190342927
10. full-blown 0.35027405550331425

Top 10 nearest words generated by pre trained word2vec embeddings and their distance with titanic:

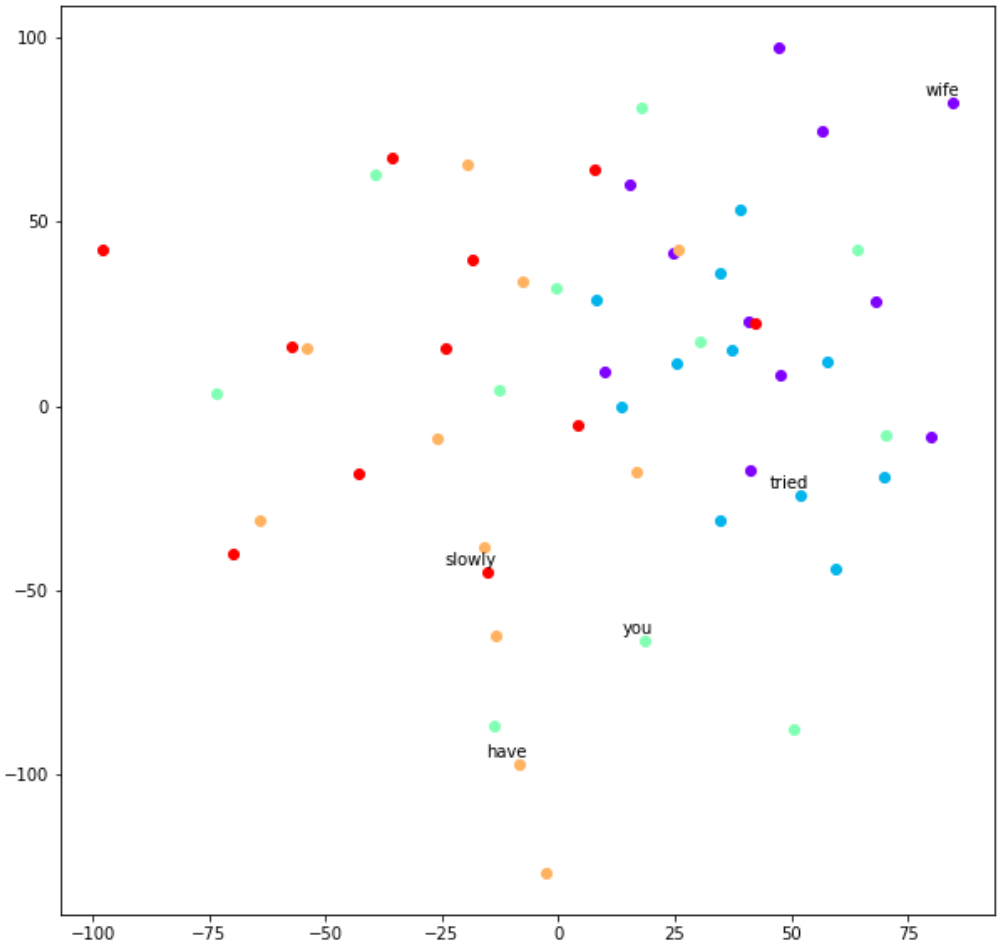
1. epic 0.600616455078125
2. colossal 0.5896502137184143
3. gargantuan 0.5718227028846741
4. titanic_proportions 0.5610266923904419
5. titantic 0.5592556595802307
6. monumental 0.5530510544776917
7. monstrous 0.5457675457000732
8. epic_proportions 0.5437003970146179
9. gigantic 0.5176911950111389
10. mighty 0.5088781118392944

Frequency of titanic in the training data used is 7 which is very less because of which the no of contexts it occurred would be very less whereas pretrained embeddings are trained on huge data.Hence words given by svd model does not seem any similar context with 'titanic' as compared to pre trained word2vec embeddings.

2.CBOW with negative sampling

Display of top 10 word vectors for five different words

```
["wife","tried","you","have","slowly"]
```



word:
wife
top-words:

1. see
2. just
3. life
4. lost
5. anything
6. about
7. not
8. ;
9. he
10. a

word:
tried

top-words:

1. chance
2. understanding
3. its
4. beat
5. biblical
6. option
7. final
8. testament
9. joke
10. wonder

word:
you
top-words:

1. that
2. the
3. film
4. if
5. great
6. for
7. i
8. watch
9. ,
10. n't

word:
have
top-words:

1. ,
2. and
3. from
4. a
5. we
6. this
7. in
8. the
9. that
10. for

word:
slowly
top-words:

1. edition
2. garden
3. suitable
4. union
5. mandarin

6. mention
7. issue
8. devout
9. accept
10. shape

We can observe that results given by cbow with negative sampling is not as good as compared to the results of svd model.

Titanic:

Top 10 nearest words generated by the model and distance with titanic:

1. ambassadors 0.4515478014945984
2. devotees 0.46002763509750366
3. untruth 0.4708781838417053
4. nikos 0.48473966121673584
5. sterility 0.4886327385902405
6. overplay 0.49272429943084717
7. pursuade 0.4943332076072693
8. ansioso 0.5132968425750732
9. slant 0.514819473028183
10. sizeable 0.5294725000858307

Top 10 nearest words generated by pre trained word2vec embeddings and their distance with titanic:

11. epic 0.600616455078125
12. colossal 0.5896502137184143
13. gargantuan 0.5718227028846741
14. titanic_proportions 0.5610266923904419
15. titantic 0.5592556595802307
16. monumental 0.5530510544776917
17. monstrous 0.5457675457000732

18. epic_proportions 0.5437003970146179
19. gigantic 0.5176911950111389
20. mighty 0.5088781118392944

Similar to svd model,embeddings generated by cbow model are not as good as pre trained word2 vec embeddings.This is again due to low occurrence of the word titanic in the given data.