Assignment-3

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

Negative Sampling is a technique in which we update fewer weights rather than updating all the weights and it helps us in making our training process faster. In case of Word2Vec, we've a context window (list of words), target word. Here, we find the mean of context window embeddings and also the target word embeddings. Our aim is to minimise the separation between these 2 vectors. When we chose to pass the vector (mean of context window word embeddings) to fully connected layer, we get probability distribution over the whole set of vocabulary and we find loss using cross entropy function. Whereas in case of negative sampling, we would sample the words that are not neighbors of target words (generally 5-10 negative samples) and try to maximise the separation between target word and negative samples.

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 based on the similarity of meanings i.e, it is a metric in which distance between words is dependent on similarity of their meaning. Embedding layer projects words on a lower dimensional space based on the semantic similarities between the words and this can be measured by considering dot product of 2 vectors.

→Eucledian metric: It is computed in the same way as we compute distance between vectors. If the distance between them is less, that means they're semantically similar.

→cosine similarity: It is computed as follows:

$d = \frac{u.v}{||u||\,||v||}$ie., the dot product divided by the product of norms of individual vectors.

As we know that dot product signifies the orientation between vectors i.e, lower angle between two vectors implies higher dot product and lower dot product implies more separation between them. So, this is also one of the metrics to measure semantic similarity.

Nearest words for 'Titanic' (using SVD)

```
[(0.058836394278626525, 'planets'),
(0.06864596007943524, 'trilogy'),
(0.07616190386964006, 'inner'),
(0.07812271628225043, 'monsters'),
(0.0782223646090615, 'moon'),
(0.07833297693176866, 'war'),
(0.0816677089633483, 'park'),
(0.0828145295604249, 'opening'),
(0.08324558875030852, 'valley'),
(0.08401008105602892, 'holy')]
```

```
Nearest words for ‘wife’
[(0.04386888012760859, 'sister'),
(0.06408785007467144, 'brother'),
(0.06767215152681572, 'girlfriend'),
(0.08374914859771099, 'husband'),
(0.09312921562253007, 'mom'),
(0.10169765148815357, 'fiance'),
(0.11292399821520749, 'eldest'),
(0.12290456441823638, 'aunt'),
(0.13137546742490658, 'dad'),
(0.1318007873149737, 'daughters')]

Nearest words for ‘tried’:
[(0.04061987981106874, 'wanted'),
(0.05181968535134274, 'decided'),
(0.05332681522524507, 'listened'),
(0.08980310376226752, 'plan'),
(0.09426337295685006, 'prefer'),
(0.1053193971180807, 'failed'),
(0.10978804352204319, 'admire'),
(0.10991228114755558, 'proceed'),
(0.11521986152595931, 'chose'),
(0.11718087696413593, 'compelled')]

Nearest words for ‘forces’:
[(0.04745303124515765, 'power'),
(0.051116092353151066, 'authority'),
(0.05242331830753155, 'masses'),
(0.05293247712033322, 'realism'),
(0.05623053994022176, 'core'),
(0.05784044631774099, 'tone'),
(0.05905654312225639, 'details'),
(0.060264070577890405, 'atrocities'),
(0.06036795721122945, 'land'),
(0.06086950885858944, 'roman')]

Nearest words for ‘they’:
[(0.03416143529062576, 'we'), (0.13577442326465927, 'christians'),
(0.1363850433436724, 'others'), (0.1394755948566706, 'explained'),
(0.1473303527654597, 'you'), (0.14733077946876116, 'either'),
(0.14955953231004548, 'i'), (0.1542803098806711, 'he'),
(0.15701055333955405, 'sometimes'), (0.15804357019021287, 'humans')]
```

Nearest words for ‘titanic’ using CBOW:

```
(0.08306554800679589, 'villains'),
(0.08737636288179496, 'thieves'),
(0.09210101068327392, 'newspapers'),
(0.09777262097230477, 'casino'),
(0.09901831048816756, 'fog'),
(0.10265381204982826, 'max'),
(0.10507525435574294, 'rebels'),
(0.10539126245940467, 'spider-man'),
(0.11318473872366319, 'water'),
(0.11409108123845779, 'bikes')
```

Top 10 closest words to "titanic" in the pre-trained gensim model

```
titanic: 1.0000
epic: 0.6006
colossal: 0.5897
gargantuan: 0.5718
titanic_proportions: 0.5610
titantic: 0.5593
monumental: 0.5531
monstrous: 0.5458
epic_proportions: 0.5437
gigantic: 0.5177
mighty: 0.5088781118392944
```