REPORT

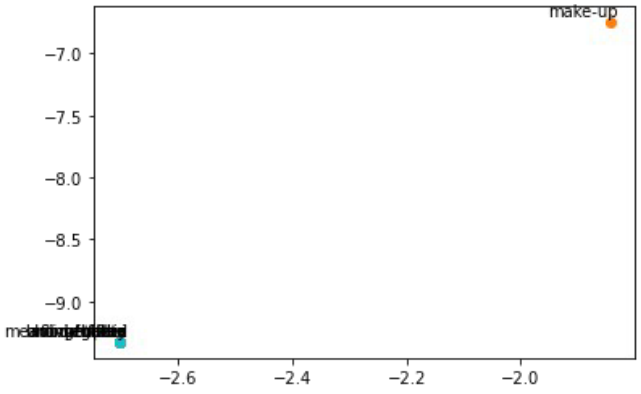
Utsav Shekhar
2021114006

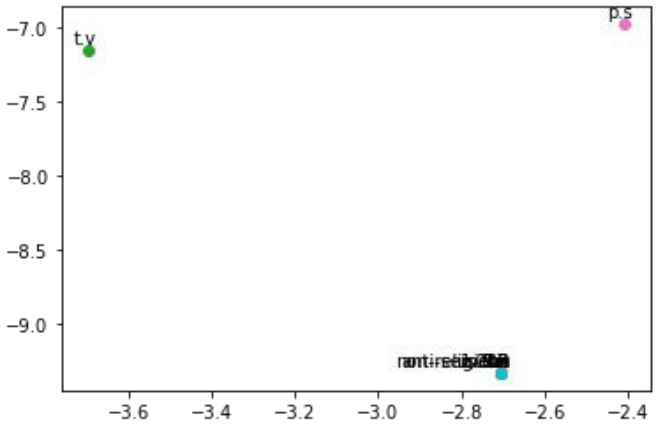
1)
Negative sampling is a method used to speed up the training process by using less computer resources while creating word embeddings, such as those created by the word2vec algorithm.
In the traditional word2vec technique, the model must determine the likelihood that every other word in the vocabulary will appear in the context of each target word in the training corpus. Due to the model's requirement to take into account every conceivable word as a negative example, this may be a highly computationally costly process, especially for big vocabularies. Contrarily, negative sampling limits the amount of computations needed by only considering a limited, randomly chosen portion of the lexicon as negative instances.
For each training example in negative sampling, the model instead of taking into account all available words, chooses a limited number of words (usually between 5 and 20) at random from the vocabulary to serve as negative examples. The probability distribution used to choose these unfavourable instances was generated using data on the frequency of each term in the training corpus. By updating the embedding vectors appropriately, the model then learns to differentiate between the positive example (the target word and its context words) and the negative instances.

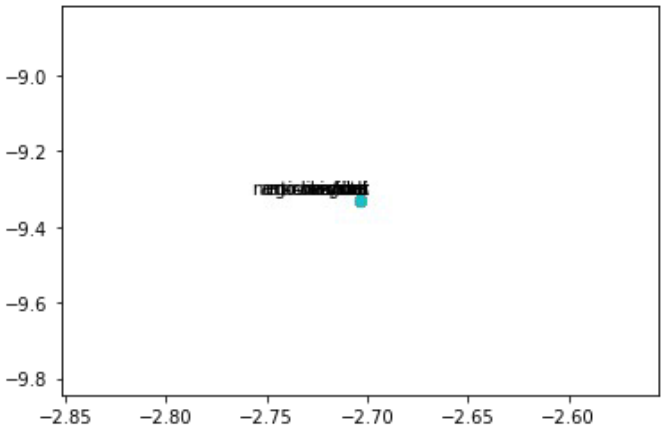
2)
Semantic similarity is a measure of how closely related or similar two words or phrases are in meaning. It is a fundamental concept in natural language processing and is important for a wide range of applications, including information retrieval, text classification, and machine translation.
Word embeddings, such as those produced by the word2vec or GloVe algorithms, provide a powerful way to measure semantic similarity between words. In these models, each word is represented as a vector in a high-dimensional space, with the distance between the vectors reflecting the similarity between the words. The cosine similarity between two vectors is commonly used as a measure of semantic similarity, with values ranging from -1 (completely dissimilar) to 1 (identical).
Two common techniques for measuring semantic similarity using word embeddings are:
Cosine Similarity:
In this technique, the cosine similarity between the vectors of two words is calculated. The cosine similarity measures the angle between the two vectors in the high-dimensional space, and is calculated as the dot product of the two vectors divided by the product of their magnitudes. Higher cosine similarity values indicate that the two words are more semantically similar.

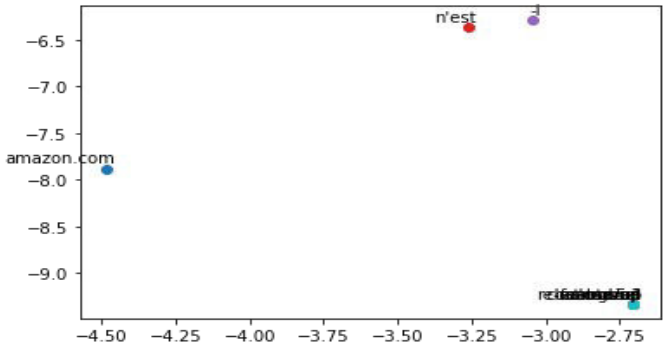
Euclidean Distance:
In this technique, the Euclidean distance between the vectors of two words is calculated. The Euclidean distance measures the straight-line distance between the two vectors in the high-dimensional space, and is calculated as the square root of the sum of the squared differences between the corresponding elements of the two vectors. Lower Euclidean distance values indicate that the two words are more semantically similar.

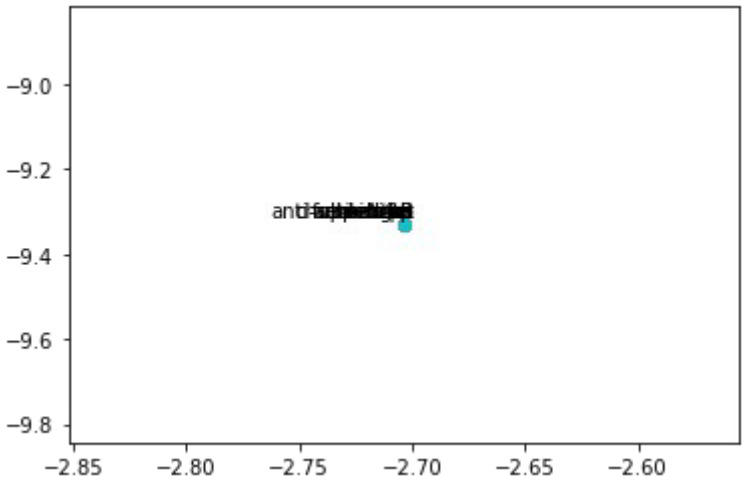
The plots of the closest words to the following words:
['helping','collection', 'the', 'class', 'great']
In sequence :
SVD model



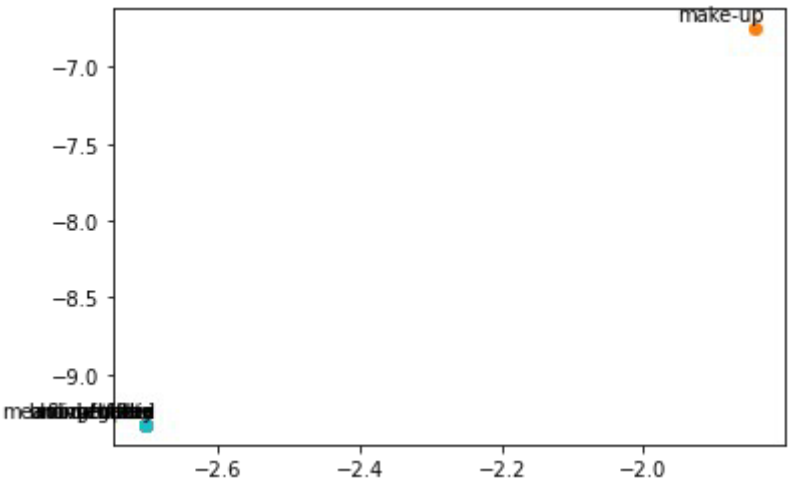


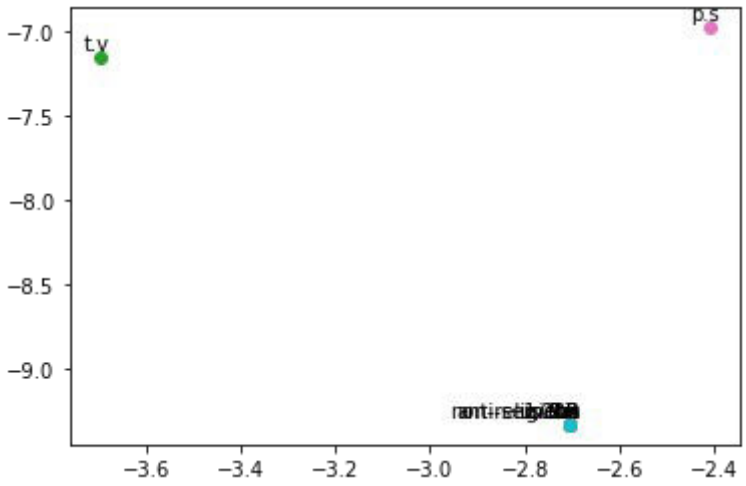


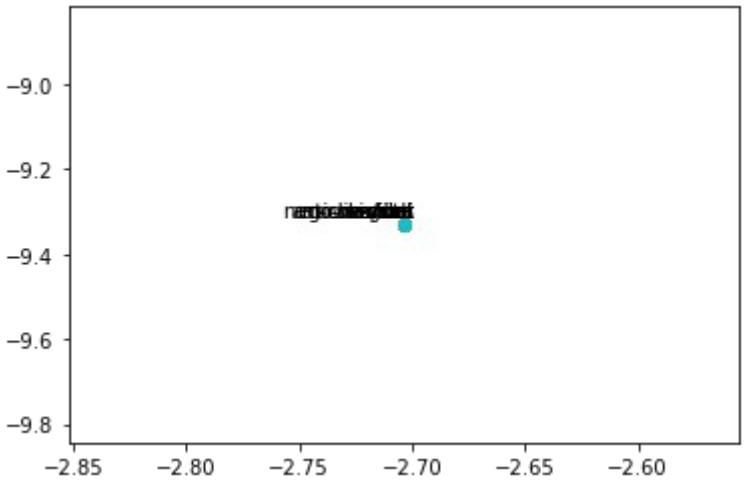


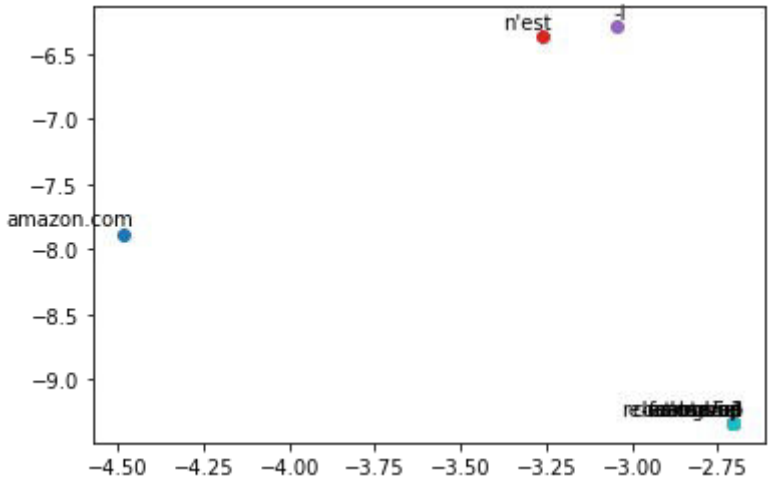


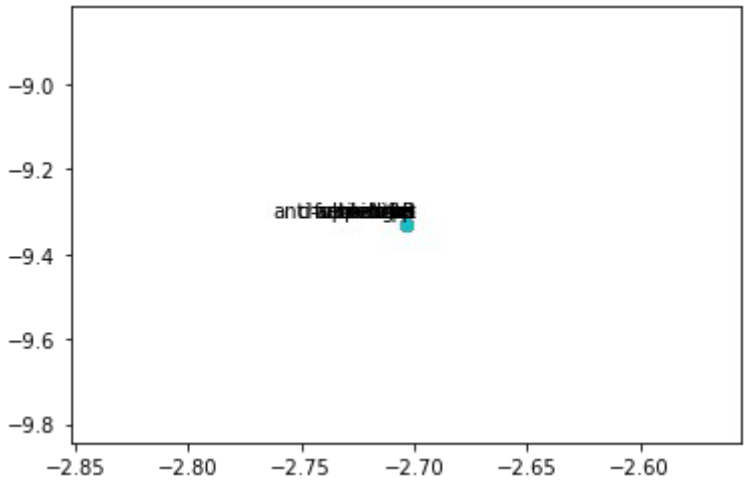
Through the CBOW model:











-> The words obtained through pretrained model are:
'epic','colossal','gargantuan','titanic_proportions','titantic','monumental',
'monstrous','epic_proportions','gigantic','mighty',

through the constructed cbow model :
'life-long', 'jesus.this', 'well-rounded', 'multi-colored', 'god-given', 'movie.my', 'film-maker', 'pro-jewish', 'sub-titles', 'pro-christian'
through initial SVD model
['first-century', 'movie.they', 'st.', 'season.i', 'modern-day', 'people.i', 'anti-semites', 'in-your-face', 'wake-up', 'on-line']

The differences can be seen clearly that, the pretrained model relates it to the physical properties of a ship whereas the trained model relates it to some what godly or movie like, but both the models appreaciate the physical features of titanic.