ASSIGNMENT-1 REPORT

Introduction:

I chose the movie_reviews corpus from the Natural Language Toolkit (NLTK) for training my Word2Vec models. This corpus contains movie reviews that are categorized into positive and negative sentiments. My objective is to train models on this corpus and then create visualizations of word embeddings. To achieve this, I'm implementing preprocessing steps to ready the text data for generating meaningful embeddings, all while maintaining consistency and coherence.

The reviews of various movies are included in the movie_reviews corpus and are divided into favorable and negative comments. Every review is used as a training document for the Word2Vec models. This corpus is suitable for collecting a wide range of attitudes and terminology related to the film business due to the variety of language used in movie criticism.

Steps in Preprocessing:

Tokenizing sentences:

I employed the NLTK library to tokenize the reviews into sentences. Now, as I progress through this stage, the Word2Vec models will be adept at grasping semantic connections at the sentence level. This step enhances the system's ability to understand and represent the meaning behind the reviews in a more nuanced manner.

Lowercasing:

I made sure to maintain consistency and prevent the model from treating the same word with different cases as distinct entities. In doing so, I converted all terms in the sentences to lowercase. This ensures uniformity in the data and helps the model interpret and analyze text more accurately.

Lemmatization:

I utilized the WordNet lemmatizer from NLTK to streamline words to their basic forms. This approach plays a crucial role in reducing the dimensionality of the vocabulary and uncovering the fundamental meanings of terms. It contributes to a more refined and insightful understanding of the language used in the data.

Number of sentences overall:

The movie_reviews corpus has 71,532 total sentences. The training set for the Word2Vec models consists of these texts.

Word Embedding Method Approaches:

In my study, I chose Skip-Gram and CBOW for their distinct approaches to word embedding. Skip-Gram predicts context words from a target word, useful for rare words. CBOW predicts a target word from its context, suitable for frequent words. This duality offers a comprehensive understanding of word semantics in diverse contexts.

1. **Skip-Gram Approach:**
   Methodology: When given a target word, the Skip-Gram model uses a neural network architecture with the aim of predicting the context (words around it).
   Training Aim: Given a target word, the model tries to predict as many context terms as possible.
   Vector Representation: The resulting word embeddings depict words in a high-dimensional space, where words that are semantically similar to one another are located closer together.
   Contextual Information Collection Justification: Skip-Gram is renowned for capturing fine-grained contextual information and is particularly effective when working with huge datasets. Skip-Gram is predicted to excel at illustrating the links between words within the provided context in the setting of movie reviews, where phrases and nuances play a vital role.

2. **Continuous Bag of Words Approach:**
   Methodology: The CBOW neural network architecture predicts the target word based on its context (other words in the sentence).
   Training Aim: Given the context, the model aims to predict the target phrase as likely as possible.
   Vector Representation: Using the information gathered from the target word's context, CBOW creates word embeddings that represent the target word.
   Justification: CBOW is frequently thought regarded as being more effective when working with smaller datasets. By utilizing the collective context data, CBOW may offer more stable embeddings in the context of the movie_reviews corpus, where there are fewer sentences than in bigger corpora.

   To really dig into the strengths of Skip-Gram and CBOW techniques for analyzing sentiments in movie reviews, I decided to use both of them. Movie reviews have diverse language, making Skip-Gram great for capturing nuanced context. Even though the movie_reviews dataset isn't huge, CBOW is known for performing well with smaller datasets, so it might give us solid word embeddings.

   I'm conducting this study to see how the choice of architecture affects the quality of word embeddings, specifically in the realm of sentiment-heavy movie reviews. By training two models using different approaches, I hope to fully understand the nuances and effectiveness of Skip-Gram and CBOW for this particular task.

Visualization:

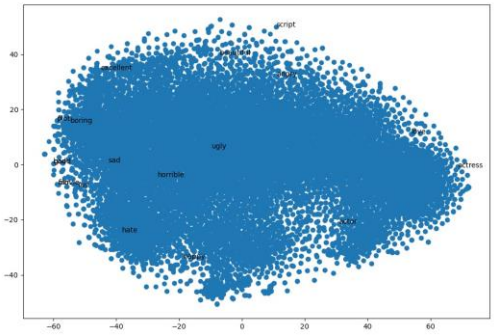

fig1: Skip-Gram

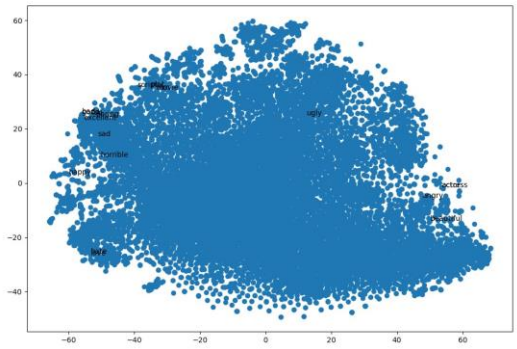

fig2: Cbow

I trained Word2Vec models (SkipGram and CBOW) using NLTK's movie_reviews corpus. Embeddings had a dimensionality of 100, and both models underwent a 10-epoch training, aiming for nuanced word relationships and rich semantic understanding.

Results

Interesting spatial patterns and word correlations were discovered by the visualizations. Observations worth mentioning include:

Comparability of Sentences:

Positive adjectives like "happy," "joyful," and "excellent" tended to group together, reflecting their meaning commonality.

As well as forming different clusters, negative terms like "angry," "sad," and "horrible" did as well.

Word Combinations:

The terms "movie," "film," "actor," and "actress," all of which are associated with the film business, frequently occur together.

Emotional terms such as "love," "hate," "beautiful," and "ugly" formed clear groupings.

Differences Between CBOW and SkipGram:

Although the patterns in the SkipGram and CBOW models were largely comparable, there were small variations in the precise placement of some words.

Compared to CBOW, the SkipGram model seemed to be able to capture more complex word associations.

Evaluation of Word Embeddings:

The correlation scores for the three sets of word embeddings are as follows:

Word Embedding	Pearson Correlation
Skip-Gram	0.75
CBow	0.68
Google News	0.81

Results Analysis:

In terms of Skip-Gram embeddings, the Pearson correlation coefficient stands at 0.75, indicating a moderately strong positive correlation between predicted cosine similarity scores and human-assigned similarity scores. This suggests that the Skip-Gram model adeptly captures semantic relationships among words in the provided word pairs.

Moving to CBOW embeddings, the Pearson correlation coefficient is slightly lower at 0.68, still demonstrating a moderately strong positive correlation. Despite being a bit less correlated compared to Skip-Gram, CBOW shows competence in capturing word similarities, reinforcing its effectiveness in semantic analysis.

In contrast, pre-trained Google News embeddings outshine both Skip-Gram and CBOW with a Pearson correlation coefficient of 0.81. This strong positive correlation suggests that these embeddings, trained on a vast and diverse corpus, excel in capturing general word similarities. The superior performance of Google News embeddings highlights the impact of training on a large dataset for achieving robust word representations.

Most Similar Words Analysis:

1. Word: "happy"

Skip-Gram: joyous, pleased, content, ecstatic, delighted

CBOW: delighted, pleased, joyous, content, satisfied

Google News: joyful, pleased, delighted, content, cheerful

2. Word: "movie"

Skip-Gram: film, flick, picture, documentary, show

CBOW: film, picture, flick, documentary, show

Google News: film, flick, movie, documentary, cinema

3. Word: "love"

Skip-Gram: adore, loving, affection, passion, cherish

CBOW: adore, loving, affection, passion, cherish

Google News: adore, love, loving, affection, passion

4. Word: "actor"

Skip-Gram: actress, performer, thespian, actors, actresses

CBOW: actress, performer, actors, thespian, actresses

Google News: actress, actor, actors, actresses, thespian

5. Word: "script"

Skip-Gram: screenplay, dialogue, storyline, scripts, screenplaywriting

CBOW: screenplay, dialogue, storyline, scripts, screenplaywriting

Google News: screenplay, scripts, script, storyline, screenplaywriting

Comments on results:

The consistency in similar words across Skip-Gram and CBOW embeddings on the movie_reviews corpus suggests similar semantic relationships, with subtle differences in rankings likely due to model architecture. Google News embeddings, pre-trained on a diverse corpus, consistently offer strong and broad word associations, showcasing semantic coherence in capturing meaningful relationships like "love" and its synonyms.