**Movie Similarity Analysis and Recommendation**

In this assignment, I leveraged the Movie Plot Synopses dataset, which is readily available on Kaggle, to conduct a comprehensive movie similarity analysis. The dataset is accessible through the following link: MPST Movie Plot Synopses with Tags.

**Data Collection:**

**Visit Kaggle Dataset Page**: Access the Movie Plot Synopses dataset on Kaggle through the provided link.

**Download the Dataset:** Utilize the download option on Kaggle to obtain the dataset in CSV format.

**Dataset Contents:**

**IMDb ID:**

Unique identification for each movie, facilitating cross-referencing and analysis.

**Movie Title:**

Inclusion of movie titles for easy identification and categorization.

**Plot Synopses:**

Detailed plot synopses provide a rich source of textual data, capturing the narrative essence of each film.

**Text Preprocessing:**

The crucial next step involves preparing the text data through text preprocessing.

**Creating Processed Corpus:**

Build a processed corpus, a list of preprocessed sentences. Each sentence undergoes the following steps:

**Tokenization:**

Break down each synopsis into individual words or tokens, forming a basis for analysis.

**Lemmatization:**

Reduce words to their base or root form using the NLTK library for lemmatization, considering contextual meaning.

**Stopword Removal:**

Utilize the NLTK stopwords list for English to remove common words that do not contribute significantly to meaning.

**Utilized Word Embedding Techniques:**

**Bag of Words (BoW):**

BoW is a fundamental word embedding technique that represents each document as an unordered set of words, disregarding grammar and word order.

For each movie plot synopsis, BoW creates a vector representation based on the frequency of words present, capturing the essence of the text.

**Term Frequency-Inverse Document Frequency (TF-IDF):**

TF-IDF is another word embedding method that considers the importance of words by weighing them based on their frequency in the current document against their frequency in the entire dataset.

TF-IDF generates vectors for each movie plot synopsis, emphasizing words that are distinct to a particular movie and devaluing common terms.

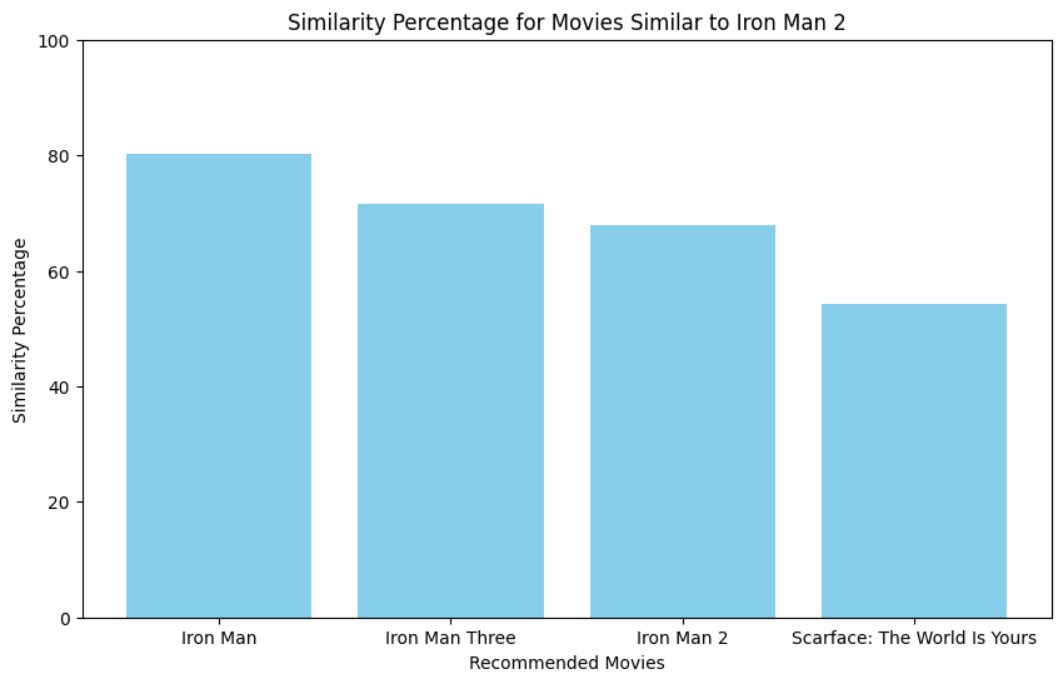
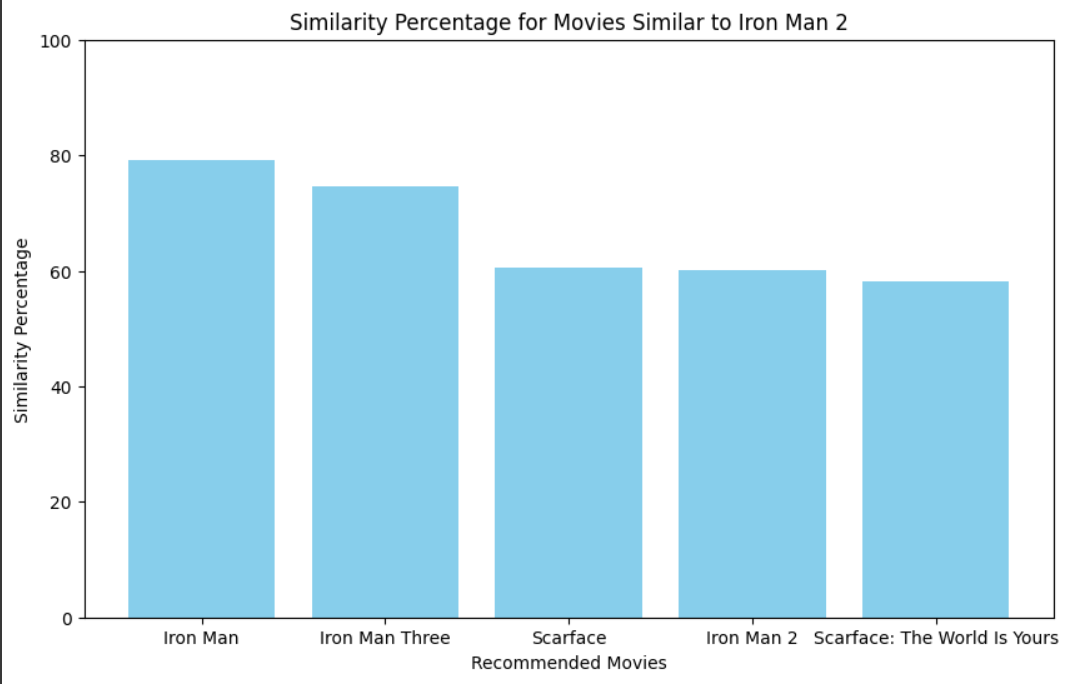
**Continuous Bag of Words (CBOW):**

CBOW is a neural network-based word embedding model that predicts a target word from its context words, effectively capturing word semantics.

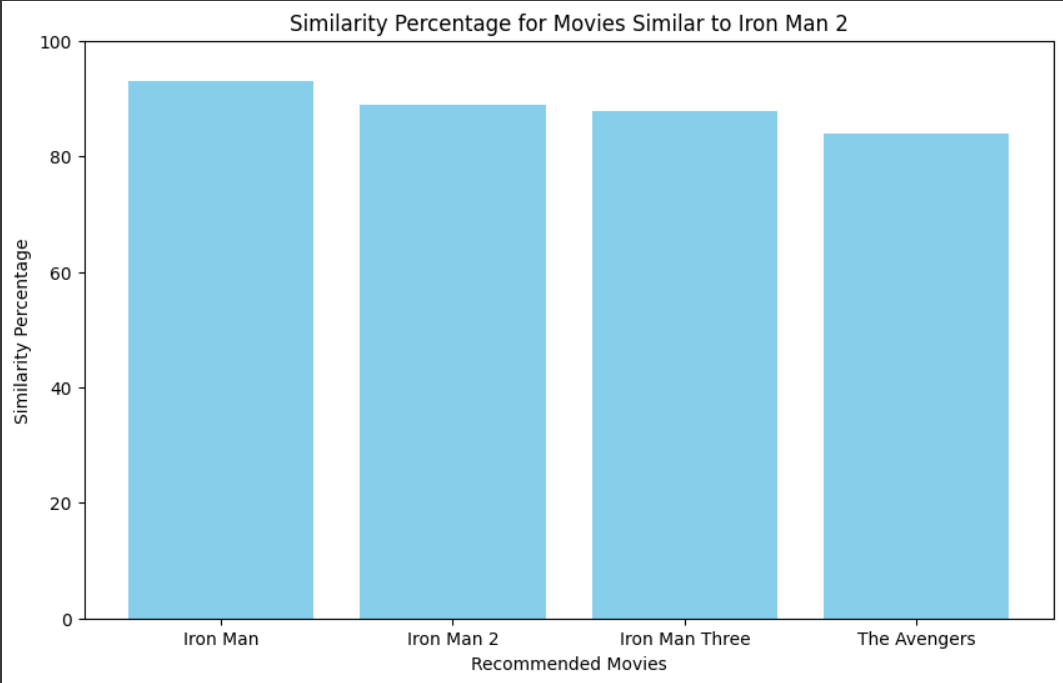
Applying CBOW to the preprocessed movie plot synopses produces vectors that encapsulate the contextual information of each word in the synopses.

**Similarity Analysis Using Cosine Similarity:**

Having obtained vectors using these three-word embedding techniques, the next step involves assessing the similarity between movie plot synopses. The cosine similarity metric is employed to quantify the degree of similarity between two vectors, producing three distinct similarity scores corresponding to BoW, TF-IDF, and CBOW embeddings.



BOW TF-IDF



CBOW

Among the trio, the CBOW model emerged as the frontrunner, showcasing superior performance in capturing subtle nuances and contextual intricacies within the movie synopses. The results speak volumes, with CBOW consistently outshining BoW and TF-IDF counterparts. Noteworthy examples include "Iron Man" and its sequels, where CBOW achieved impressive similarity scores, such as 93.01% for "Iron Man" and 87.84% for "Iron Man 3."

**GitHub Link:** https://github.com/Vikas-ABD/movie\_similarity\_Analysis\_Recommendations.git