This project is about identifying the answer for a multiple-choice question using NLP methods. There are different packages must be installed in this problem like Spacy, Scikit-learn package. I have tried using different methodologies which help us understand better and find out correct answer.

Data Collection:

Data has been collected from a website with 31 multiple choice questions

Data Preparation:

Data must be prepared for training and testing purposes

Data Preprocessing:

Convert data from list to a pandas data frame and split with special characters and split paragraph, questions in each column. Dropping null values from columns Tokenized the data and remove punctuations and stop words. While removing stop words some options will be empty so need to remove those specific words from the stop words list. Used Lemmatization to map with the root words.

Using Scikit-Learn package found TFIDF vectorization and calculated the vector values for each statement. TFIDF is the product of two values TF and IDF which gives us information about the weightage of the documents or words. Using TFIDF calculated each statement vector and stored in a Data Frame. In the Same way calculated the TDIDF vector for question, after created a function which calculates the Cosine similarity of two TDIDFs.

Named Entity Recognition:

NER is nothing but recognizing entity of each word in the text. Using NER identify the entities of Question and final answers and options.