Cosine Similarity

The Cosine Similarity is a better metric than Euclidean Distance because if the two text document far apart by *Euclidean* distance, there are still chances that they are close to each other in terms of their context.

Project Plan :

- Get the news data text files
- Tokenize the data in the files to words
- Stem the words to the root word to avoid repetition of words
- Vectorize the words bag of words embedding
- Use cosine similarity and find how similar the two vectors are by finding the cosine of the angle between the two vectors.

List of improvements:

- Using TFIDF vectorizer from sklearn for better vectorization and to avoid sparse matrix .