TASK1- NLP- TEXT SIMILARITY:

* Import nltk
* A corpus is large set of data and texts. So importing stopwords from nltk.corpus and word\_tokenize from nltk.tokenize.
* Get two inputs in lowercase a text1 and text2.
* Tokenization : applying word\_tokenize on both the texts. Basically it splits the words of the sentence from the texts.
* Creating two lists.
* Removing stopword: remove useless words from the text that take space in our data. Appending the non stopwords in the texts sets.
* Combining both sets and creating vector: if the words are in the texts set then appending them to the lists we created earlier.
* Applying cosine formula: creating a var = 0 and adding on the product of both the lists words. Formula is product of both lists divided by the half of sum of both lists.
* This will give us the output as the percent of similarity between two texts.

SUMMARY:

1. User inputs two texts
2. Data cleaning is done by word tokenizing and removing stopwords.
3. After getting clean data, two lists are created and both the texts are compared
4. The words that are not stopwords are appended in two sets and are combined.
5. Both the texts are compared and the similar words are appended in the list.
6. Using cosine formula, we get the similarity percent of the texts.