

Explain:

Step 1: Prepare data:

+ Input: The data is the list of text data

+ Tokenizer\_VN: The function created to tokenize the words and create a corpus, it also includes removing punctuation function to remove unnecessary punctuation such as “#$%^”

The library used: The Underthesea

+ Stop words for Vietnamese: Use the vietnamese\_stopwords.txt to list the Vietnamese stopwords. Then the function will use this list to exclude stopwords from the corpus.

Step 2: Create TF-IDF Vectorizer:

Scikit-learn has provided the TfidfVectorizer function to transform words into TF-IDF vectors. TfidfVectorizer has 2 variables to apply to foreign languages easier: tokenizer and stop\_words

+ tokenizer: tokenizer will use the Tokenizer\_VN to tokenize the text into the corpus. However, Tokenizer\_VN has removed the punctuation function so the tokenizer can do the same

+ stop\_words: Using the list of stopwords provided which is the list of the Vietnamese stopwords in this project.

Step 3: Transform and vectorize

+TfidfVectorizer will vectorize the text using fit\_tranform

+To know how text is vectorized, the TF-IDF values need to be turned into an array list and the feature list needs to be get by using get\_feature\_names-out from vector TfidfVectorizer.

+ For evaluation, the pairwise similarity is shown by multiplying the matrix, which is vectorized from text, with the transpose of the matrix:

