1. Tokenizing words with Separators : \n,\b,\t, .:

Exceptions:

* Adjusting Words that start with capital letters will be contacted.
* Words inside “” marks will be contacted.
* Words with – between them will be contacted.
* If there is a \n but the 2 words between the \n both start with capital letters, or

There is a – between them then contact the words.

1. Case Folding(by using the nlp library)
2. Remove Punctuation(check for list)
3. Remove stop words(by using nlp library)
4. Perform Stemming or Lemmatization(by using nlp library)