Part 7 Natural Language Processing

Table of Contents

[Steps to implement NLP on text 2](#__RefHeading___Toc26_1315846998)

# Steps to implement NLP on text

1. Import the dataset of text.
2. Clean the text using the “Bag of words algorithm”.
   * Bag of words means to create a bag of the main keywords from the text, and ignore the not relative words such as “on, the, so, ...”.
   * Remove the Punctuations such as “...”.
   * Implement Stemming on the text which taking the root of the word, such as the word “Loved, Loving” should be “Love”.
     + Do so in order to not have a big bag of words at the end.
     + And to group the same relative words together.
   * Convert all Capital letters to small letters only.
   * Tokenization process which will
     + Split all text into words.
     + Attribute one column for each word.
     + Then each column will contains the number of time that word appeared / associated with the text phrase / line.
     + We will have a lot of sparse matrices, because not all words appeared on all text phrases / utterances.
     + We will have by the end some Association Rules between each word and the associated columns.