Word Cloud

<https://www.wordclouds.com/>

# **Welcome to Project Gutenberg**

<https://monkeylearn.com/text-analysis/>

<https://www.gutenberg.org/>

Dimension will not change as we have pre-decoded on aster list of words

Is bag a data structure like arrays, sets, tuple, dict, list etc.

dimensions, a predefined list of master list of words, will be captured from the training set?

Millions of dim (dict size)=Millions of dim (dict size

If each word in bag is considered a dimension, imagine trying to classify books into their geners we would end up with thousands of dimensions .

we would need to be careful in choosing stop words, some words which are stopwords might actually be a key word.

This will be a vector with all 0’s and a single 1 for that word in frequency distribution

[1 0 1 1 0 1 1 0 0] -> Andrew is a tall boy

for each word we are assuming a document with only that word and trying to represent it in the vector form

if the word Andrew comes twice what happens to the representation? Does the histogram change? TF is used for normalization of frequency of words in a document before comparison of documents?

master list has words from N number of documents. Here we are calculating TF from each document (document instance i). My understanding is correct sir?

For two docs, the bag of words may be very large. To avoid excessive dimensions , we will use only few words from document which has higher frequency.(because we presume that higher frequency words are important).