# Bag-of-Concepts:

Ever wonder bag full of concepts? Well, it is one of the research based approach where it has no “best solution” yet. Often we come across information from various categories but, categorizing them into concepts have always been challenging. There are many classification algorithms, machine learning techniques to start with. In a similar way here we can see Natural language processing technique and Naïve Bayesian Classification used together as one of the solution.

A set of common words that falls into various concepts is used to train the algorithm. The keywords to search for the information are fetched from user. The program also includes a spell checker, where invalid or misspelled words are targeted and further corrected. This is necessary as the wrong spelled words may fetch irrelevant urls in the process.

With the help of Beautiful Soup library in python, the urls of the keywords are fetched. Formatting of the text is removed and the text is cleaned.

Tokenization is task of breaking information into pieces which is very simple technique provided by natural language processing tool kit where huge data can be break into pieces and later process individually. Here the text is tokenized into words appended in a list. To make it easy, most frequently occurring words in the text is collected and allowed to classify.

With the help of classification technique and the training data set, words or sentences that occur in the text has a match in the data set are categorized into respective concepts.

The approach used here can still be better, as the field is vast and the combination of various algorithms which works precisely could possibly result into a best way.