**KEYWORD EXTRACTION USING SVM:**

**A. Creation of Vector Space Model:**

Vector space models (VSMs) of word semantics use large collections of text to represent word meanings. Each word vector is composed of features, where features can be derived from: global corpus co-occurrence patterns - how often a word appears in each document .

The vector space model used in SVM is created using 80 documents present in the multi labeled wiki dataset representing the rows in the matrix.The total number of columns in this matrix is the total unique words that appear in the document.

**B. Term Frequency Document:**

TFIDF is a commonly used feature in information retrieval. TF-IDF is used to figure out if the word that appears more often in a document but not very often in the corpus is more likely to be a keyword. However, in our problem for snippets,TF for most keyword candidates is close to 1, this feature did not help in the purpose of snippet extraction.

**C. One Vs. Rest Classifier:**

Given that the vector space model creates an equal feature set for each document, we can use it to represent the training set while the tags associated to it are represented as the predicted outcomes of the given document.One-vs-all consists of fitting one classifier per class. For each classifier, the class is fitted against all the other tags present in the document. Since each class is represented by one and one classifier only, it is possible to gain knowledge about the document by inspecting the tag associated to it.

**D. Creation of document window:**

For the purpose of snippet extraction, a window of 15 words is extracted to create a test set.This test set is again converted into a vector space model to identify the relevant tag associated with it with a test error estimate.This window is slided through the document to get the snippet with the least test error associated to the query tag.