Researching progress

At first, I created a class TextMachine with cosine similar function to build a text classification for comment moderator. With the use of StratifiedKFold, I received an accuracy as 80%.

Following the development of TextMachine, I created a class TCWBuilder based on it for making a TCW matrix for features extracting. The TCWBuilder do the same jobs as TextMachine, without using a cosine similar function to decide the classes. But there are some problem. The dataset loaded (SOCC) has 1209 rows of data, through the TCWBuilder, the tcw matrix was initialized with size 6153 x 1209. That means we have 6153 tokens (words) and 1209 documents (sentences).

However, the problem is that the SVDExtracter class which is using SVD method to extract it into three singular matrices, follow the guide, I use the first matrix (U, constructed by left singular vectors of the tcw matrix) as the features matrix. And I got the final matrix with the size 6153 x 6153, but my expected is 1209 x a number of features.

The first try, I will try to fix the problem and remove the dependencies of TextMachine in TCWBuilder, what is not really necessary, that TrainingSet class.

Next, note that SVDExtractor generates a features matrix with the values inside [-1, 1] following the operation of SVD. But currently, to pass these numerical datas into the GFMM algorithms, necessary to normalize it to the range [0, 1] by using MinMaxScaler.