The topic that I am choosing is Text Summarization.

1. For this topic, the only information that it would be preferably to exist is the topic for each text/paragraph/article etc. For project, this information is already present in articles in a specific tag, but for any other article, the only information not presented is topics which make the respective article ready to be put into a category.
2. The pre-processing steps required to transform the corpus into a useful data representation are very important. The first step is converting the specific paragraph into sentences in order to be more simplistic. The next step is specifying a set of rules which all the words need to follow to have a unified and coherent set of words. These rules include eliminating punctuation, numbers, other special characters and stop words (such as “as”, “the”, “and”) that don’t contribute in the transformation process because them being very common and have little to no meaning. Once this pre-processing step is done, the next one counting all of the words in the paragraph and their appearance. This step is done because we can see the most and least words that appearance and outline the topics. Thereafter, by knowing the count for each and every word, we calculate the weight occurrence frequency of all the words. To achieve this, we calculate by the most recurrent word in the paragraph. For example, if the most appeared word is “forest” and the second one is “tree”, we have:

Word Freq Weight

Forest 3 1

Tree 2 0.67

The next step required represents substituting words found in the original sentences with their weighted frequencies and then compute their sum.