The topic that I chose is Text Summarization.

1. For this topic, the only information that would be preferable to exist is the topic for each text/paragraph/article etc. For the 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 to the transformation process because of them being very common and having little to no meaning. Once this pre-processing step is done, the next one is 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 appear and outline the topics. Thereafter, by knowing the count for each and every word, we calculate the weight occurrence frequency of all the words.

Practically, very similar to the project, the following step is eliminating the most and least common topics using the 5-95% rule which was the result of multiple simulations. By eliminating topics that have the most amount and the least amount of appearance, not only we reduced the number of articles and topics, but also the number of words. This step is exceptionally useful because of its optimized results.

Next, we determine the range of values and where each word is located by calculating its entropy that is utilized to calculate the information gain necessary in the proceeding step. The latter is useful in depicting the connection between words and topics.

1. The 2 approaches to summarize a text are extraction-based and abstraction-based summarization. For the former, the extraction is represented by pulling only the most information out of a piece of text and combining it to create a summary. For the latter, is used to paraphrase and shorten the original text. The major difference between these is observed in the complexity and text quality making abstraction-based summarization technique more complicated and sophisticated.
2. A learning algorithm that could improve the quality of the text is inverse reinforcement learning IRL that focuses on estimating the reward function of an agent given a set of observations of that agent’s behavior. Generally, IRL provides advantages in situations where the reward function is not explicitly known or where it is difficult to define or interact with the environment directly. These situations are exactly what we observe in summarization.