Project:
Email Thread Summarization

Group Members:

1. Asma Damani
2. Saifullah Katpar

Datasets
we have reviewed a number of papers for email summarization, thread summarization, text summarization in general and other resources. So far we found following corpora mostly used by researchers:

- **Enron Corpus** (https://www.cs.cmu.edu/~enron/)
- **BC3 Dataset** (https://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/bc3html)
- CNN/Daily Mail (https://cs.nyu.edu/~kcho/DMQA/)

According to our understanding, for our task annotation is compulsory. Although we want to use email threads of our own organization, annotating those threads will take considerable time. For the proof of concept and preliminary project work, we wish to use Enron and BC3 corpora.

Queries:

1. Is annotation compulsory for our task? Is it possible to perform it with an unsupervised method?
2. What should be the minimum size of the dataset if we annotate our own dataset?
3. Is there any other annotated dataset available for email threads that we can use?
4. What should be our next step before coding if we have got the dataset in hand?