Project Information

Team Member 1: Paarth Neekhara (pneekhar@eng.ucsd.edu)
Team Member 2: Prakhar Pandey (p4pandey@eng.ucsd.edu)
Team Member 3: Shehzin Hossain (ssh028@eng.ucsd.edu)

Goal:
Summarization of Document: We want to summarize a given document such that the summary contains the most important or informative sentences. We focus on extractive summary instead of abstractive summary meaning we summarise the document by extracting sentences from the original document which act as representative for the entire document.

Initial Solution and Thoughts:
We want to employ a RNN based model which can selectively make decision about which sentences to include in a summary or not. So we would like to make a binary decision for each sentence as to whether or not we want to include it in our summary. This is a problem related to sequences because at each point we make decision based on the sentences that we have seen previously and which appear after the current sentence depending on the content they have already presented and the content that is going to be presented to make a decision on the current sentence.

We would probably include a bi-directional RNN to capture the entire content of a document and summarize it in its hidden state. We would use a two layer bidirectional RNNs first at word level to encode the words in individual sentences and then at sentence level to encode the summary. Something very similar to as shown in the paper: "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents" : By Nallapati, Zhai and Zhou.

[Ref: Nallapati, Zhai and Zhou. as mentioned above]

We would use the dataset described in the paper i.e. CNN/DailyMail corpus originally constructed by (Hermann et al. 2015). We would use an unsupervised approach to calculate a ROGUE score of each sentence with the gold summary to create a supervised dataset with binary labels for each sentence.