Abstract

With the rapid growth of online social networks like tweeter, facebook users generate large amount of content of social events. Specially on Twitter, users in their microblogs generate phenomenal volume of content of real-world events on a daily basis by giving their comments on any topic. So, social sites like tweeter are useful source for information collection . Contents of any topic collected from tweeter may contain contents which may be outside of the topic or it may be unnecessarily too long. So, text summarization is needed to get a concise documentation of the real-world events.

But social context contains texts which are mostly in non-standard forms which is difficult to summarize. So, before text summarization is done, text normalization is needed to convert the non-standard words into their standard forms.

For text normalization we have collected contents of different users from tweeter and find the non-standard words in it. We categorize the non-standard words into different types and apply heurestics techniques to normalize them into standard forms.

We then did text summarization on the normalized text . In this project text summarization is performed based on senetence similarity measures. Sentence similarity measures is calculated from word form feature and word semantic similarity between the sentences.

Keywords: Text Normalization, Text Summarization, NSWs, NLP, extractive.