CE807 Lab 3
Topic Modelling and word2vec

In this lab, we will experiment with Topic Modelling and word2vec.

1. Topic Modelling using Gensim

There are several publicly available topic modeling systems, some of them available as part of complete pipelines. Gensim is one of the most widely used. Official documentation and sample code are available.
ToDo: read the description of the system on the page above
In this lab, we will see how to perform topic modeling on the 20newsgroup dataset. We aim to understand the model's working and visualize the output provided. The script "*lda_gensim.py*" provides the sample code.
After understanding the overall pipeline, try to change the different parameters and observe the difference in the topics. Some suggestions are as follows:

- Add bi-gram and tri-gram to the feature
- Use a different number of topics
- Use different parameters in the model
- Use advanced feature representations like word2vec

2. word2vec

Next, we will play with the word2vec. There are many word2vec models available; we will use Gensim one. We will have to download word2vec, which takes time (2GB download). So while it is downloading, we could play with the online demo and try to understand how word2vec works. The script "*word2vec_gensim.py*" provides the sample code.
Demos:

- http://nlp.polytechnique.fr/word2vec
- http://epsilon-it.utu.fi/wv_demo/

Good luck!