# Natural Language Processing Project - Seq NLP

## Context:

Word embedding are a type of word representation that allows words with similar meaning to have a similar representation. It is a distributed representation for the text that is perhaps one of the key breakthroughs for the impressive performance of deep learning methods on challenging natural language processing problems.

## Data:

We will use the IMDb dataset to learn word embedding as we train our dataset. This dataset contains 25,000 movie reviews from IMDB, labeled with a sentiment (positive or negative).

The Dataset of 25,000 movie reviews from IMDB, labeled by sentiment (positive/negative). Reviews have been preprocessed, and each review is encoded as a sequence of word indexes (integers). For convenience, the words are indexed by their frequency in the dataset, meaning the for that has index 1is the most frequent word. Use the first 20 words from each review to speed up training, using a max vocab size of 10,000. As a convention, "0" does not stand for a specific word, but instead is used to encode any unknown word.

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided. See the README file contained in the release for more details.

**Data reference:** https://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

**Citation**:

```
@InProceedings{maas-EtAl:2011:ACL-HLT2011,
  author    = {Maas, Andrew L.  and  Daly, Raymond E.  and  Pham, Peter T.and
Huang, Dan  and  Ng, Andrew Y.  and  Potts, Christopher},
  title     = {Learning Word Vectors for Sentiment Analysis},
  booktitle = {Proceedings of the 49th Annual Meeting of the Association for
Computational Linguistics: Human Language Technologies},
  month     = {June},
  year      = {2011},
  address   = {Portland, Oregon, USA},
  publisher = {Association for Computational Linguistics},
  pages     = {142--150},
  url       = {http://www.aclweb.org/anthology/P11-1015}
}
```

**Domain:** Customer Reviews

## Key Asks:

1. Import test and train data (5 points)
2. Import the labels (train and test) (5 points)
3. Get the word index and then Create a key-value pair for word and word_id (15 points)
4. Build a Sequential Model using Keras for the Sentiment Classification task (15 points)
5. Report the Accuracy of the model (5 points)
6. Retrieve the output of each layer in Keras for a given single test sample from the trained model you built (5 points)