Homework 5

We are turning back time and revisiting homeworks 3 and 4 this week. The goal is to create document embeddings using averaged word embeddings (word2vec, glove, and fasttext) and doc2vec. We will be loading in the word embeddings from existing resources, and creating a doc2vec model from scratch.

# Part 1 – Coding

1. Implement the following embeddings, (apply the stemming/lemmatization approach that you found to work best on each dataset to tokenize your data):
2. Average the word embeddings using the following existing vector sets:
   * glove-wiki-gigaword-100
   * word2vec-google-news-300
   * fasttext-wiki-news-subwords-300
3. Apply Doc2Vec (Distributed Memory model) to get a document vector
4. Use the new embeddings in conjunction with Naïve Bayes for both the HW3 and HW4 datasets.
5. Additionally, use the embeddings in conjunction with your two chosen models from HW4 (supervised part), which will have two parts:
   1. Use your models in the same way as before using the four different embeddings.
   2. **PERFORM HYPER PARAMETER TUNING USING A GRID SEARCH**.   
      You can tune your models using **ONE** of the embeddings and then use the same hyperparameters for the other embeddings.
6. Evaluate using F-score and accuracy

# Part 2 – Questions

1. Which combination(s) have the “best” results? Briefly explain why you think this is.
2. Compare and analyze\* the following results:
3. Tuned -- untuned models
4. GloVe – FastText – Word2Vec
5. Doc2Vec -- averaged Word2Vec embeddings
6. TFIDF/BoW – Doc2Vec/Word2Vec

\* By “analyze” I mean provide insights into the differences, you can use questions I have asked previously as guidelines on what to talk about. For example, the difficulty of the classification task (number of classes, number of instances per class, dataset size, etc.), sparse vs dense vectors, assumptions made by the models, ...

# Part 3 – Class discussion

**We will have a discussion in class on Friday about everyone’s results!**