Assignment 7: Text Clustering and Topic Modeling

This assignment needs dataset "ydata_3group.json". This dataset contains samples in the format of (**text**, **first_label**, **all_labels**). For each sample, "text" element contains the text of the sample, "first_label" is the top class assigned to the sample, and "all_labels" is a list of classes assigned to the sample (including "first_label"). In total, there are three classes (labels): T1, T2, T3.

Task 1: K-Mean clustering

- Use KMeans to cluster the text into 3 clusters by cosine similarity
- Apply majority vote rule to map clusters to ground-truth values in "first_label"
- Calculate precision/recall/f-score
- Based on cluster centroids/samples, give a **meaningful name** (instead of T1, T2, T3) to each cluster.

Task 2: LDA (single-label)

- Use LDA to cluster the text into 3 topics
- For each sample, select only the top one topic (i.e. the topic with highest probability)
- Apply majority vote rule to map topics to ground-truth values in "first_label" values
- Calculate precision/recall/f-score and compare them with the results in Task 1.
- Based on word probabilities in each topic, give the topic a **meaningful name**.

Task 3 ***(Bonus):*** *LDA (multiple-label)*

1) For LDA model in Task 2, use a threshold to assign topics to samples. Assign a topic to a sample *only if the probability of the topic given the sample is greater than the threshold*.
2) After topics for each sample are determined, calculate precision/recall/f-score macro against "all_labels" values in the dataset.
3) To determine the best threshold, vary the threshold from 0 to 1 with a step of 0.05, (i.e. 0, 0.05, 0.1, 0.15, …, 1), assign topics for each sample, and then calculate precision/recall/f-score macro (i.e. apply steps (1)-(2)).
4) Plot the results of precision/recall/f-score macro vs. different threshold values. Write your analysis about the results, and conclude what can be the best threshold value.

Task 4: Reading

We're going to learn neural network and deep learning in our next lecture. Neural networks are very powerful, but also quite complicated with lots of mathematical notations! To prepare for this lecture, please read the following references:

- http://ufldl.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/
- Le Q. and Mikolov, T. Distributed Representations of Sentences and Documents https://arxiv.org/pdf/1405.4053v2.pdf

Please get yourself familiar with the mathematical notations used in these materials. You may find that it is difficult to understand some part of the materials. Don't worry about it. I'll explain it in class and feel free to ask.

Submission Guideline:

Write a block of code for Tasks (1-3) in **a python file (.py)**. Meanwhile, submit **a pdf document** (.pdf) to show your model training results, necessary print-out, plot, and your analysis.