**Text Analytics**

**Assignment 4 Topic Models**

This assignment will give you hands-on experience in building topic models and clustering in text mining. Your input file are fashion reviews from SS2016 runway: ***fashion.csv***. This is the same dataset you used in Lab 4.

**Question 1: Topic Inference**

1. You will infer topics from word documents using two approaches:
2. LDA and 2) LSI

Generate 10 topics from fashion reviews. Select the approach that gives you the better results from LDA or LSI. Label the topics if you find semantically meaningful concepts associated with them (Note, you may not find all topics to be meaningful).

1. Result improvement:

Perform additional steps, such as stop word removal, bigram representation, etc. Do these steps improve the quality of topics? If so, update the 10 topics with new labels.

**Question 2: Compare Clusters**

1. Pick the best Topic Model result from Question 1, use it as the input and perform KNN clustering on all the review documents (your input is the U-matrix).
2. In comparison, use the original term-document matrix as the input for KNN clustering. Use the same number of K.
3. Compare the clustering results. Based on your observation, which one gives you better result? (Note, TM doesn’t guarantee to give you better clustering result. There is no need to calculate measures, just observe the clusters)

Submission:

1. Word Report
2. Python program. Please make sure your python program can run successfully.

Other instructions:

1. DO NOT submit your dataset. Only submit Word and python program.
2. Do not use absolute path to read your input data (it won’t run on your TA’s computer)
3. Name all your files FirstName\_LastName.xxx. This will make our grading easier.
4. Do not zip your file. Submit two files directly.

Thank you!