Text Analytics

Assignment 1

This is an exercise to generate insights from input text dataset. Please use your Lab 1 and Lab 2 materials as reference.

Input data - The input dataset is a sample of fashion reviews crawled from Vogue during Fashion Week. The main content is the review text. The file also contains meta data such as "year", "season", "brand", "author of review."

Task – Your task is to analyze the dataset based on *keyword frequency* to understand fashion trend. Present top 30 keywords (concepts) you extracted and plot them in a distribution chart. You will be using the NLTK tool. You are required to analyze the data with 4 approaches:

1) Use a simple bag-of-words approach
2) Use a bag-of-words approach with stemming and stop words removal
3) Use POS approach and focus on all the noun forms (NN, NNP, NNS, NNPS)
4) Use POS approach and only focus on NNP

(You may perform other analysis as bonus point. You can also make use of the meta data)

Submission – You need to submit the following document on the Blackboard:

1. A report in word document.
   In the report, you need to present the results of 4 approaches (in terms of frequency and plot). Compare the results, and provide your insights about
   1) Performance of each approach (which one give you best result?)
   2) What do you see is a fashion trend in 2016. Figures and tables are encouraged in the report.
   3) Brief description of other improvements you can make.
2. Your Python code (Both .ipynb and .py files are accepted).