CSE 255, Winter 2015: Homework 8

Instructions

Please submit your solution **at the beginning of the next lecture (March 2)** or outside of CSE 4102 beforehand. Please complete homework **individually**.

Download the "50,000 beer reviews" data from the course webpage: `http://jmcauley.ucsd.edu/cse255/data/beer/beer_50000.json`. Code is provided on the course webpage showing how to load and perform simple processing on the data. Executing the code requires a working install of Python 2.7 with the scipy packages installed.

Tasks

Using the code provided on the webpage, read the *first 5000* reviews from the corpus, and read the reviews **without capitalization or puncuation**.

1. How many unique bigrams are there amongst all of the reviews? List the 5 most-frequently-occurring bigrams along with their number of occurrences in the corpus (1 mark).

2. The code provided performs least squares using the 1000 most common unigrams. Adapt it to use the 1000 most common *bigrams* and report the residual obtained using the new predictor (use bigrams *only*, i.e., not unigrams+bigrams) (1 mark).

3. What is the *inverse document frequency* of the words 'foam', 'smell', 'banana', 'lactic', and 'tart'? What are their *tf-idf* scores in the first review (1 mark)?

4. What is the cosine similarity between the first and the second review in terms of their tf-idf representations (considering unigrams only) (1 mark)?

5. Which other review has the highest cosine similarity compared to the first review (provide the beerId and profileName, or the text of the review) (1 mark)?