

CSE 158, Winter 2017: Homework 4

Instructions

Please submit your solution **by the beginning of the week 9 lecture (Nov 27)**. Submissions should be made on **gradescope**. Please complete homework **individually**.

Download the “50,000 beer reviews” data from the course webpage: http://jmcauley.ucsd.edu/cse158/data/beer/beer_50000.json. Code is provided on the course webpage (`week5.py`) 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 punctuation**.

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 MSE obtained using the new predictor (use bigrams *only*, i.e., not unigrams+bigrams) (1 mark). Note that the code performs *regularized* regression with a regularization parameter of 1.0.
3. Repeat the above experiment using unigrams *and* bigrams, still considering the 1000 most common. That is, your model will still use 1000 features (plus an offset), but those 1000 features will be some combination of unigrams and bigrams. Report the MSE obtained using the new predictor (1 mark).
4. Using the model from the previous questions which are the 5 unigrams/bigrams with the most positive associated weights, and the 5 unigrams/bigrams with the most negative associated weights (1 mark)?
5. 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 (using log base 10) (1 mark)?
6. What is the cosine similarity between the first and the second review in terms of their *tf-idf* representations (considering unigrams only) (1 mark)?
7. 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)?
8. Adapt the original model that uses the 1000 most common unigrams, but replace the features with their 1000-dimensional *tf-idf* representations, and report the MSE obtained with the new model.