Homework 3

STAT-GB.4310: Statistics for Social Data
Instructor: Patrick O. Perry

Due February 23, 2016

Application

Create a developer account and download the Yelp Academic Dataset, then extract `yelp-nyc-business.json` and `yelp-nyc-review.json`. There are detailed instructions for how to do this at the start of the "Part-of-Speeching Tagging" lecture slides, posted on the course webpage.

For each star rating (1–5), perform the following set of actions.

1. Randomly sample 50 reviews with the given star rating.
2. Use *either* the CoreNLP or the OpenNLP tagger to tag the review texts with parts of speech.
3. Convert the parts of speech to the Universal POS tags.
4. Report the top 5 nouns, top 5 verbs, top 5 adjectives, and top 5 adverbs, according to frequency of occurrence across the sample of 50 reviews.

For each star rating, you will have 4 lists (nouns, verbs, adjectives, and adverbs).

Answer the following questions:

1. Were any of the results surprising to you? Why or why not?
2. Were any of the results interesting to you? Why or why not?
3. Do any of your results indicate potential errors in the POS tagging algorithm you used? Even if not, what are some factors that would cause the POS tagger to make an error?