SI 618: Large-scale distributed computation – Yelp Sentiment Analysis

Fall 2019

Instructor: Ceren Budak

Spark Example: Yelp Sentiment Analysis

- Goal: identify words that are indicative of positive or negative reviews
- Input: yelp reviews data
- Output: A list of most positive and most negative words

How to find positive words

- Any ideas?
- We will compute the following "positivity" score for each word w.
  - Positivity(w) = log P(w in PositiveReviews) – log P(w in AllReviews)
    - where P(·) denotes probability
    - log(x) is the natural logarithm of x.
    - P(w in PositiveReviews) means the probability that a word occurs, given we are looking at the set of positive reviews.
    - P(w in AllReviews) means the probability of a word occurring, given that we are looking at all
- Same formula for negative words

Finding positive words

- Say, ‘awesome’ appears 10 times in positive reviews, and all positive reviews together contain 10000 words. Furthermore, say, ‘awesome’ appears 20 times in all reviews, and there are 100000 words in all reviews, Then, positivity score of ‘awesome’ is:

  Positivity(awesome) = log P(awesome in PositiveReviews) – log P(awesome in AllReviews) = log(0.001) – log(0.0002) = -6.9 – (-8.51) = 1.61

- Words that are more neutral, like ‘the’, that have similar probability given a positive review compared to any review, should have a positivity score close to zero.

- How to define positive reviews (and similarly negative reviews)?

Let’s code together…