

CS294-1 Programming Assignment 1

Richard Hwang, David Huang

February 13, 2013

1 Our Solution

We followed advice in the assignment statement, splitting the text with the regex `"[\\s.,()!?:&\\"]+"` and indexing each word, except for stopwords such as "the", "and", "or", "a", and "of". Afterwards, we constructed a sparse matrix, with words as rows and documents as columns.

2 Smoothing/Backoff

We used Laplace Smoothing and found that an α value of TODO yielded the highest F_1 measure.

3 Performance

For some reason, indexing the words was intolerably slow. We tried using `foreach`, for loops, `String.split()` and `StringTokenizer`, and for loops and `split` turned out to be the fastest combination. It still took at least 10 minutes to index all the words, though. Constructing the word-document matrix, however, was very fast: about 7 seconds.