# NLP Project 1 Report

Dehao Wang, dw2551

Xiaoya Wang, xw1996

Jiaxin Zhang, jz3796

### Data

**10K:**

We gathered the annual 10K files of dow30 component stocks from 2010-2019.

**Noted that due to the change of component stocks in dow30, the company lists are different from year to year. We took that into account and gathered data based on each years’ components.** (Professor has asked us to use the total 41 stocks for all years, but when the notice announced, we already almost finished the project. So, we kept it this way. Hope it will not affect our grade.)

**Stock Price:**

We retrieved the ‘filed as of date’ and the following consecutive 3 days’ stock price for each company each year.

**Benchmark:**

Same as the author, we used CRSP value-weighted index as the benchmark.

### Two weighting methods:

1. Proportional Weighted Index of Negativity

This is referred as term frequency(TF) in our codes. It’s just the basic calculation of the proportion of negative words in each document.

1. tf.idf weights

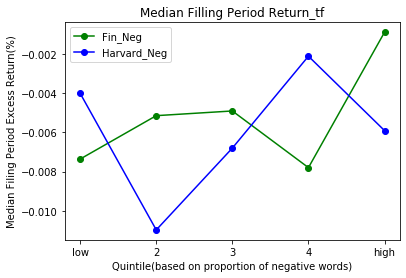
This is calculated using equation (1) in the paper.

### Two dictionaries

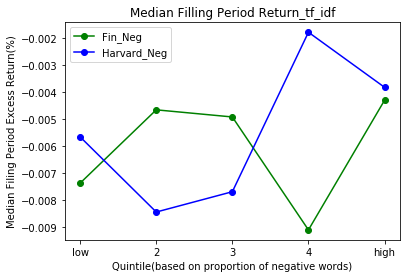
1. Harvard-IV-4-Neg
2. Fin-Neg (created by TIM LOUGHRAN and BILL MCDONALD)

### Results

Proportional Weighted Index of Negativity



tf.idf weights



### Analysis

Ideally, when proportion of negative words become higher, the median of excess return should become lower. However, we cannot discover such patterns from these two graphs. This might caused by too little sample size. In the original paper, the author used almost all 10Ks filed over 1994 to 2008. While we only used 30 companies’ 10K within 10 years, which includes around 300 files. Too small of sample size might be the main reason of failing to show a clear pattern.

There are slightly stronger downward patterns in Fin\_neg lines than H4N lines, which potentially indicates Fin\_Neg could better analysis the negative sentiment of 10K reports.

Taking a look at the difference of term frequency and tf-idf, one clear pattern is that in tf-idf graph, the difference of excess returns of each group are larger than that of term frequency graph. **Showing that tf-idf model is more sensitive to our target negative words than term frequency model.**