COMP7705 MSc(CompSc) Projects

NLP for annual 10-k analysis

Interim report

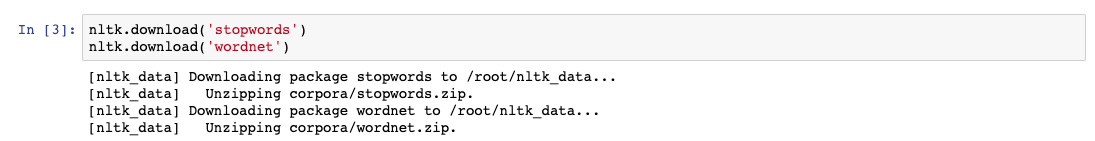
<https://wp.cs.hku.hk/2021/msp21082/>

I will be using two models in this project, BERT and FinBERT. First, the BERT model performs traditional sentiment analysis to predict negative to positive individual stock performance. I will benchmark BERT and FinBERT with typical sentiment analysis with NLTK and supervised labeled pre-trained data.

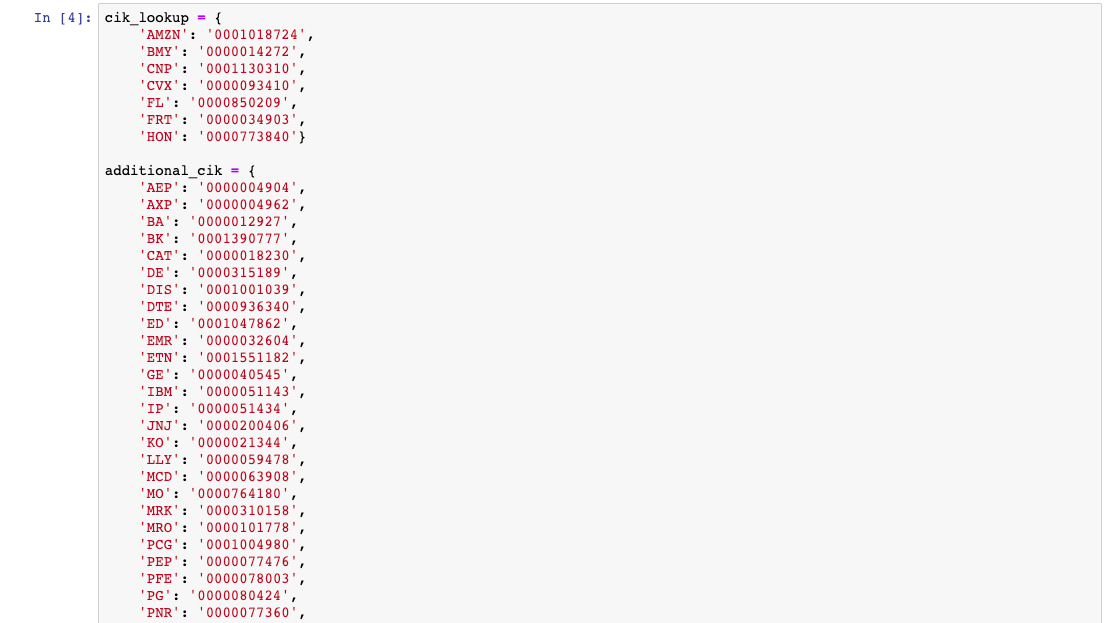
Get data from SEC

I'll be running NLP analysis on 10-k documents. To do that, we first need to download the documents. For this project, we'll download 10-ks for a few companies. To look up documents for these companies, we'll use their CIK if you would like to run this against other stocks.

preparing

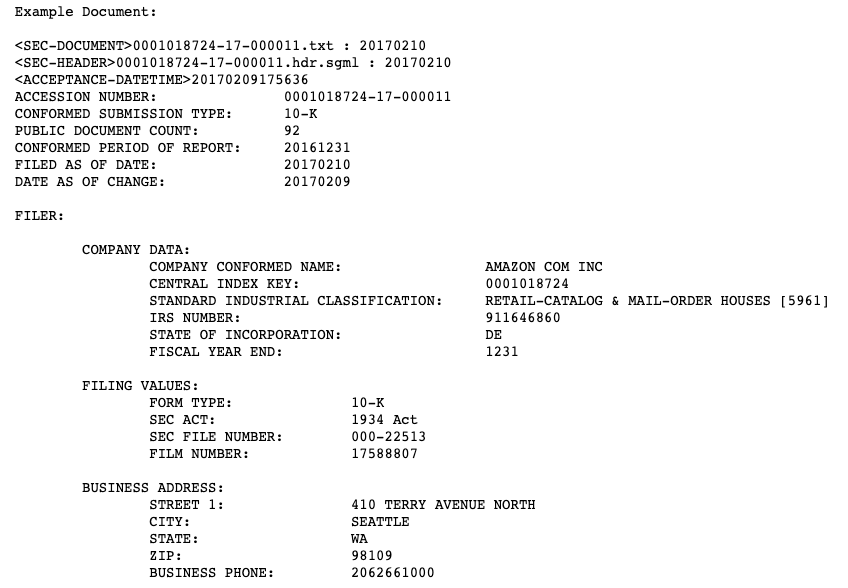


10-k stock lookup



download 10-k from SEC





i have created function to get 10-k data from SEC,

**TODO list:**

data processing, Lemmatize, remove the stop word using nltk, and i am going to do the analysis with Loughran McDonald Sentiment Word Lists for giving a result following

We'll be using the Loughran and McDonald sentiment word lists. These word lists cover the following sentiment:

* Negative
* Positive
* Uncertainty
* Litigious
* Constraining
* Superfluous
* Modal

Generate alpha factor based on sentiment analysis.

**PartII :**

build BERT and FinBERT model same dataset