Sentiment Analysis With 10-k Filings

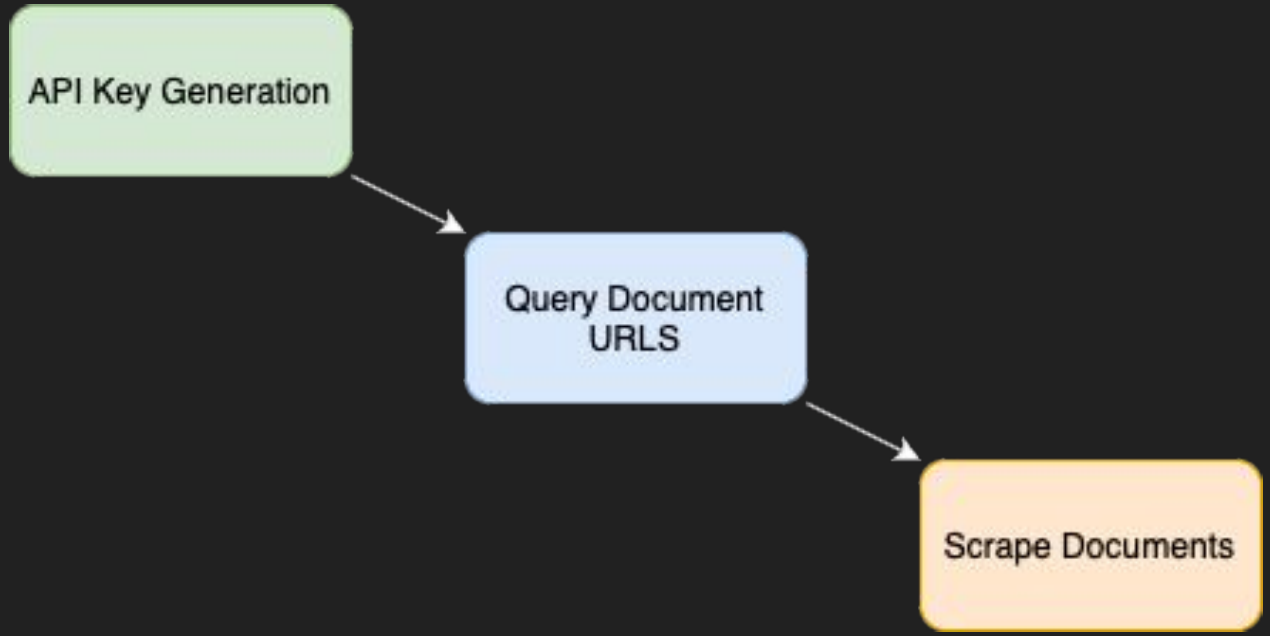
Ryan

Project Objective and Motive

- Textual data collection and document processing
- Sentiment analysis in a financial context
- Comparing document similarities
  - For a year between companies
  - For a company between years

- Learning objectives:
  - Practice and exercising data mining capabilities
    - Identifying, collecting, cleaning and formatting information
  - Explore sentiment analysis in a financial setting

Data Aggregation



Data Aggregation - Document URLS

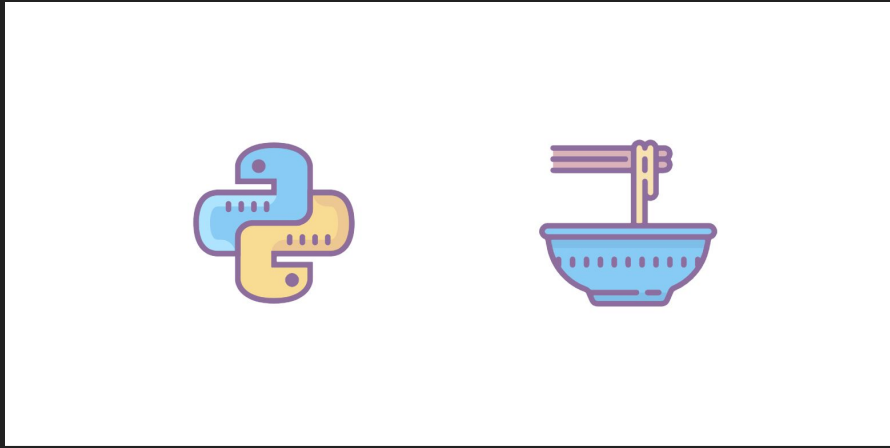
```
"query": {
    "query_string": {
        "query": "cik:1065280 AND filedAt:{1999-01-01 TO 2021-04-07} AND formType:\"10-K\""
    }
},
"from": "0",
"size": "10",
"sort": [
    {
        "filedAt": {
            "order": "desc"
```

es Headers (13) Test Results Status: 200 OK Time: 193 ms Size: 37.27 KB Save Respons

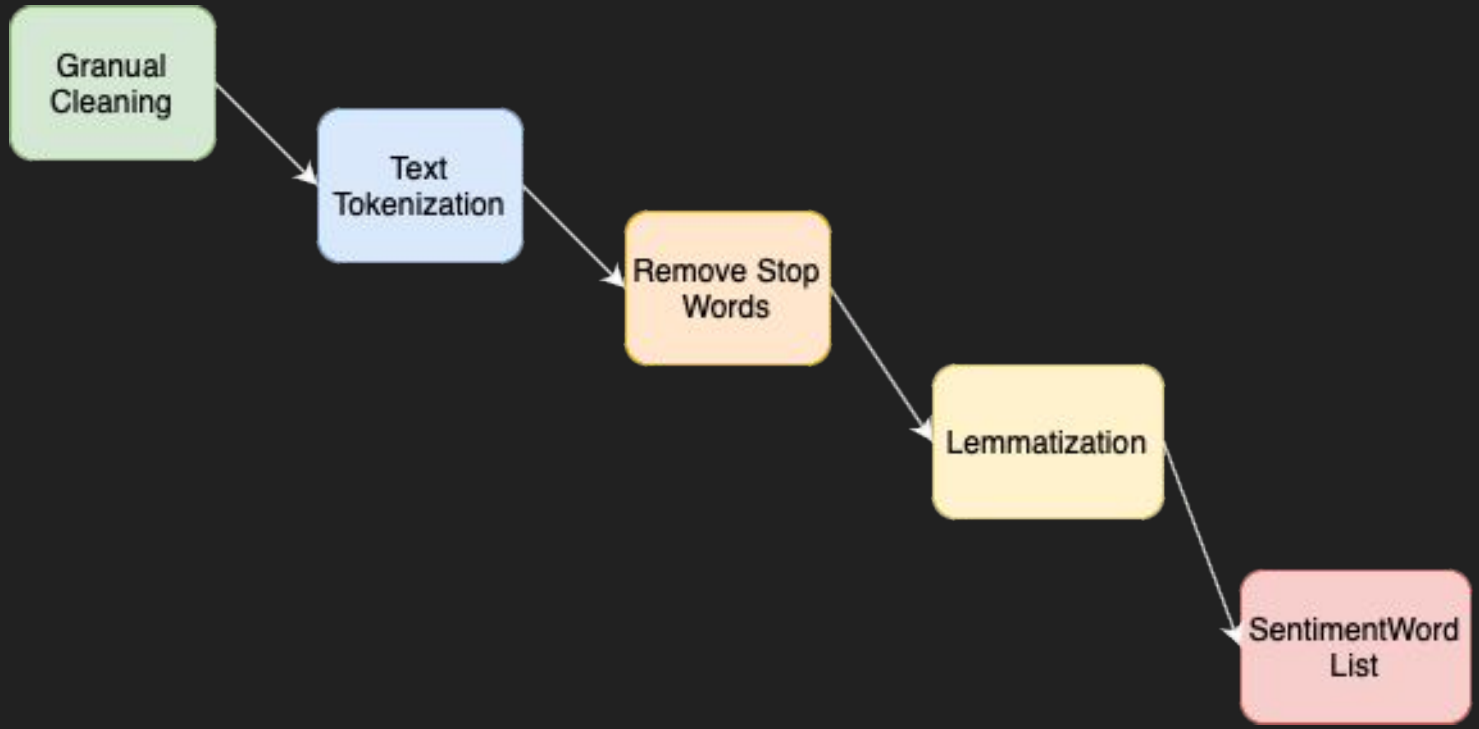
Raw Preview Visualize JSON


```
    "id": "a89c7d8e8629bf6bd1581df6dc0f1957",
    "accessionNo": "0001065280-21-000040",
    "cik": "1065280",
    "ticker": "NFLX",
    "companyName": "NETFLIX INC",
    "companyNameLong": "NETFLIX INC (Filer)",
    "formType": "10-K",
    "description": "Form 10-K - Annual report [Section 13 and 15(d), not S-K Item 405]",
    "filedAt": "2021-01-28T16:21:59-05:00",
    "linkToTxt": "https://www.sec.gov/Archives/edgar/data/1065280/000106528021000040/0001065280-21-000040.txt",
    "linkToHtml": "https://www.sec.gov/Archives/edgar/data/1065280/000106528021000040/0001065280-21-000040-index.htm",
    "linkToXbrl": "",
    "linkToFilingDetails": "https://www.sec.gov/Archives/edgar/data/1065280/000106528021000040/nflx-20201231.htm",
```

Data Aggregation - Document Scraping



Data Preprocessing



Data Preprocessing - Fine Grain Cleaning

Remove unneeded characters using regular expressions:

- whiteSpace = re.compile('\s+')
- wordCharacters = re.compile('[^A-Za-z .]+')

Data Preprocessing - Tokenization

- 1. Partition the documents into individual sentences
- 2. Further tokenize the sentences into sequences of tokens
- 3. Filter out the stop words to remove any dilution of important tokens in the text

Data Preprocessing - Lemmatization

- Using the sequences of tokens we lemmatize words into clusters of similar inflections of words to reduce them to their stems

Data Preprocessing - Loughran-McDonald Sentiment WL

86,486 Unique Words

Specially Designed for Financial Documents

Positive

Negative

Uncertainty

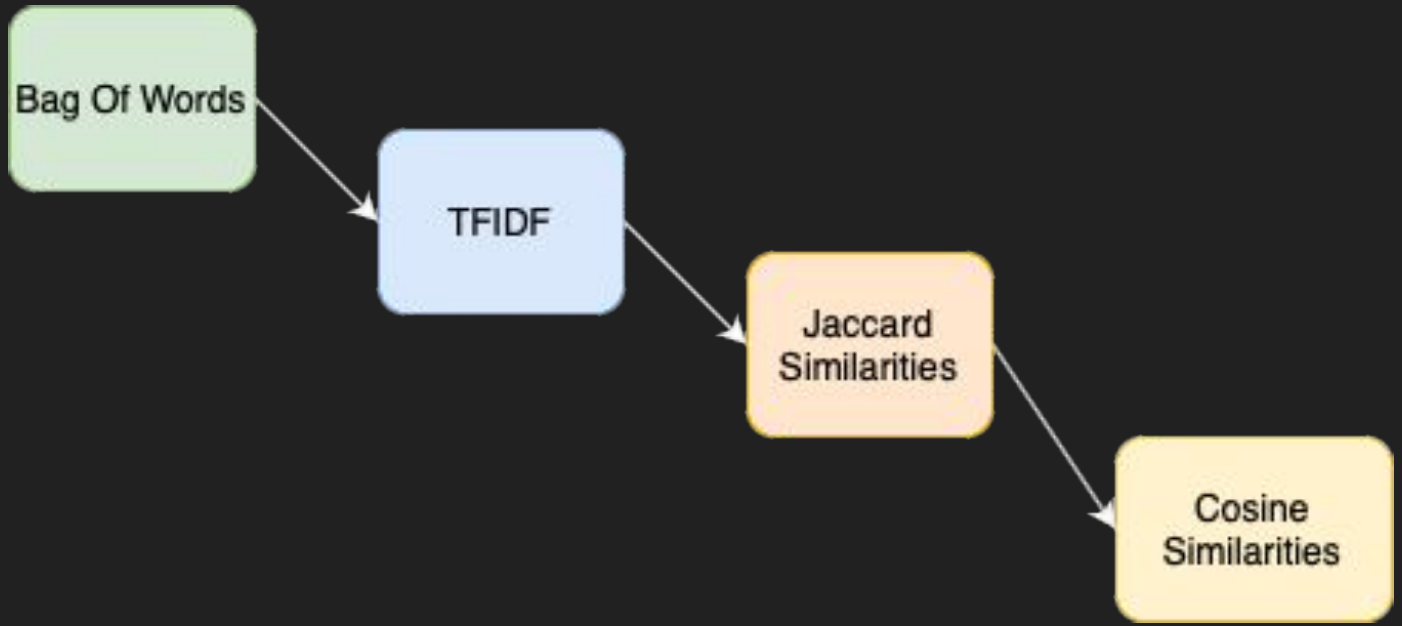
Litigious

Constraining

Superfluous

Interesting

Analysis


Bag Of Words
TFIDF
Jaccard Similarities
Cosine Similarities

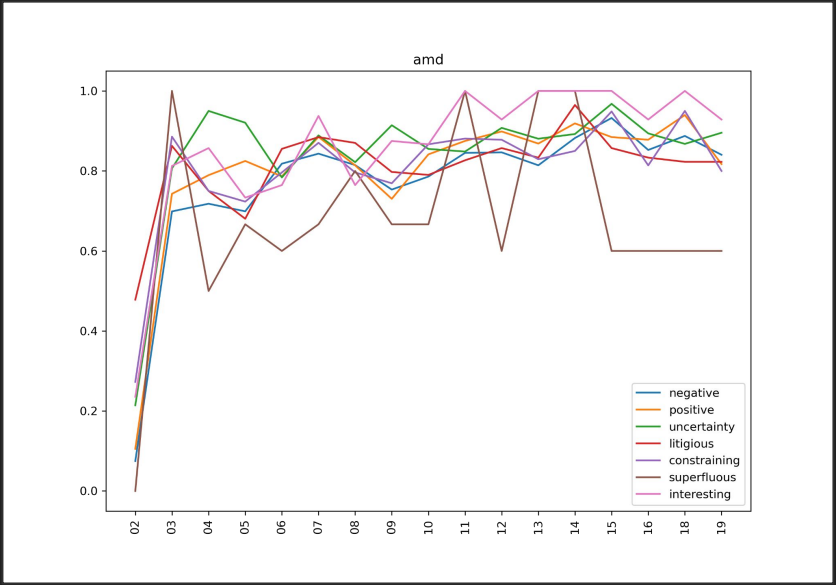
Analysis - Sentiment Bag Of Words

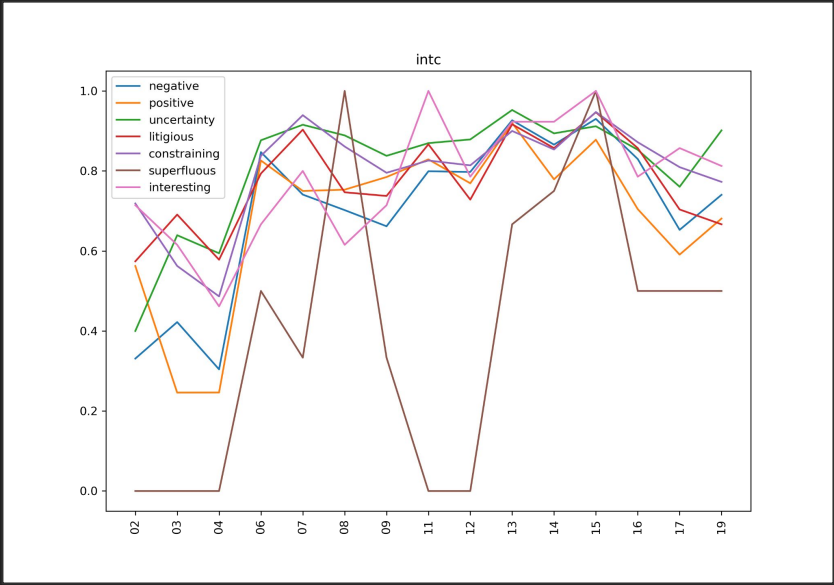
- BOW model for counting the occurrences of sentiment words in the documents
- Individual BOW for each sentiment word category for company documents
- Sklearn CountVectorizer and the Vocabulary argument

Analysis - Sentiment TFIDF

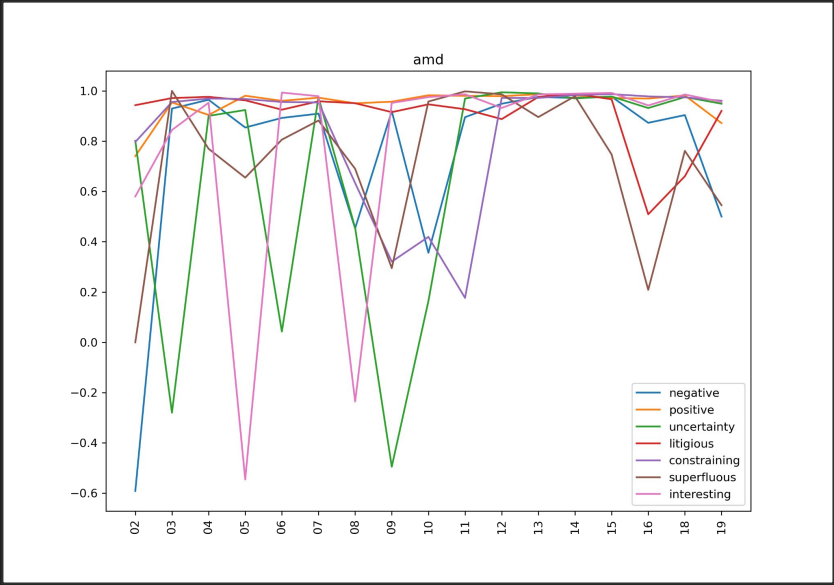
- TFIDF model for calculating the frequencies of sentiment words in the documents
- Individual TFIDF for each sentiment word category for company documents
- Sklearn TfidfVectorizer and the Vocabulary argument

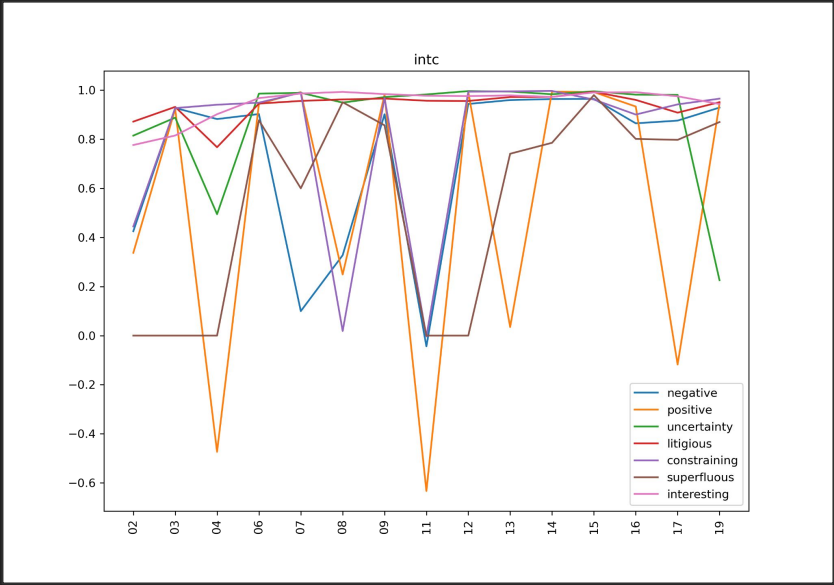
Analysis - Jaccard Similarity





Analysis - Cosine Similarity





Future Work

- Within the NLP class scope:
  - Consider other financial filings such as 10-Q, 8-K, 4A, etc …
  - Analysis of various news content and related trending stories
- Outside the NLP class scope:
  - Correlate the occurrences of sentiment words in 10-k filings with the company's stock price
  - Correlate the occurrences of sentiment words in 10-k filings with various company valuation methods
    - Factor Returns
      - Value, Size, Momentum, Quality, Volatility
    - TurnOver Analysis
      - Stability of factors over time
  - Sharpe Ratio Metric

Questions