Sentiment Analysis of Figures in the New York Times

CHRIS LAGUNILLA

THE GOAL

- Given a person of interest, I was interested to see how the sentiment of documents written about them varied
- After processing corpus data, information relating to a specific person is aggregated into a DataFrame, which allows us to look into sentiment over time
- Given these sentiment scores, we can gather statistics about these people and visualize their data

THIS WORK SO FAR

- The people of interest in this presentation:
  - Barack Obama
  - John McCain
- The mini data set being used right now is Jan-June 2007
- I chose to look at this because there were ~ 100 articles for each person, as this was preceding the 2008 Presidential Election

THE DATA SET

- The New York Times Annotated Corpus is a licensed data set provided by the Linguistic Data Consortium
- Spans 20 years from 1987 – 2007
- It consists of 1.8 million articles
- Already annotated in an XML format

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nitf SYSTEM "http://www.nitf.org/IPTC/NITF/3.3/specification/dtd/nitf-3-3.dtd">
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
  <head>
    <title>Paid Notice: Deaths  BLUMENTHAL, MARTIN</title>
    <meta content="dn010107" name="slug"/>
    <meta content="1" name="publication_day_of_month"/>
    <meta content="1" name="publication_month"/>
    <meta content="2007" name="publication_year"/>
    <meta content="Monday" name="publication_day_of_week"/>
    <meta content="Classified" name="dsk"/>
    <meta content="7" name="print_page_number"/>
    <meta content="B" name="print_section"/>
    <meta content="3" name="print_column"/>
    <meta content="Paid Death Notices" name="online_sections"/>
    <docdata>
      <doc-id id-string="1815718"/>
      <doc.copyright holder="The New York Times" year="2007"/>
      <identified-content>
        <person class="indexing_service">BLUMENTHAL, MARTIN</person>
        <classifier class="online_producer" type="types_of_material">Paid Death Notice</classifier>
        <classifier class="online_producer" type="taxonomic_classifier">Top/Classifieds/Paid Death Notices</classifier>
      </identified-content>
    </docdata>
    <pubdata date.publication="20070101T000000" ex-ref="http://query.nytimes.com/gst/fullpage.html?res=9907E1DE1E3AF932A35752C0A9619C8B63" item-length="179" name=
  </head>
  <body>
    <body.head>
      <hedline>
        <hl1>Paid Notice: Deaths  BLUMENTHAL, MARTIN</hl1>
      </hedline>
    </body.head>
    <body.content>
      <block class="lead_paragraph">
        <p>BLUMENTHAL--Martin. A New York business man and philanthropist, died last Saturday in Manhattan after a long illness. He was 90. Mr. Blumenthal was bor
        also active in Human Rights Watch. Mr. Blumenthal is survived by his wife, Sallie Blumenthal, his children Richard of Greenwich and David of Boston, six g
        sympathy. May his memory remain for a blessing. David H. Lincoln, Sr. Rabbi Amy AB Bressman, Chairman of the Board Menachem Z. Rosensaft, President</p>
      </block>
```

PROCESSING

PROCESSING THE DATA

- Data is already collected and annotated
- I had to collect the information I needed into an object and save that
- My tags of interest:
  - < doc-id >
  - < day >, < month >, < year >
  - < text >
  - < person >
- I adapted a jupyter notebook file into a normal python script to run over the files, generate my DataFrame, and save it as a pickled file
- “Unpacking” all of the data over 20 years takes a long time
  - (As of right now I am almost ½ done)

ANALYSIS

VADER

- Valence Aware Dictionary and sEntiment Reasoner
- Given a text, VADER assigns “polarity” scores: NEG, POS, NEU, and COMPOSITE
- The values gathered from VADER can tell if an article is skewed in a positive or negative light

ANALYSIS

- VADER did a lot of the heavy lifting
- Most work for analysis went into:
  - Aggregating the articles associated with a single person
  - Running the VADER tool over all of those articles
  - Collecting the data in order to perform operations on the resulting data

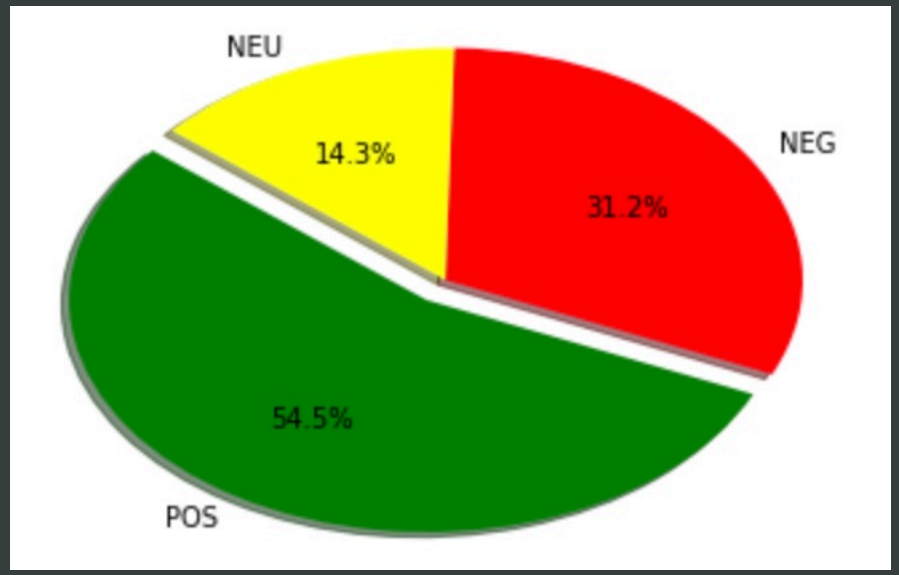
VISUALIZATION & MY RESULTS SO FAR

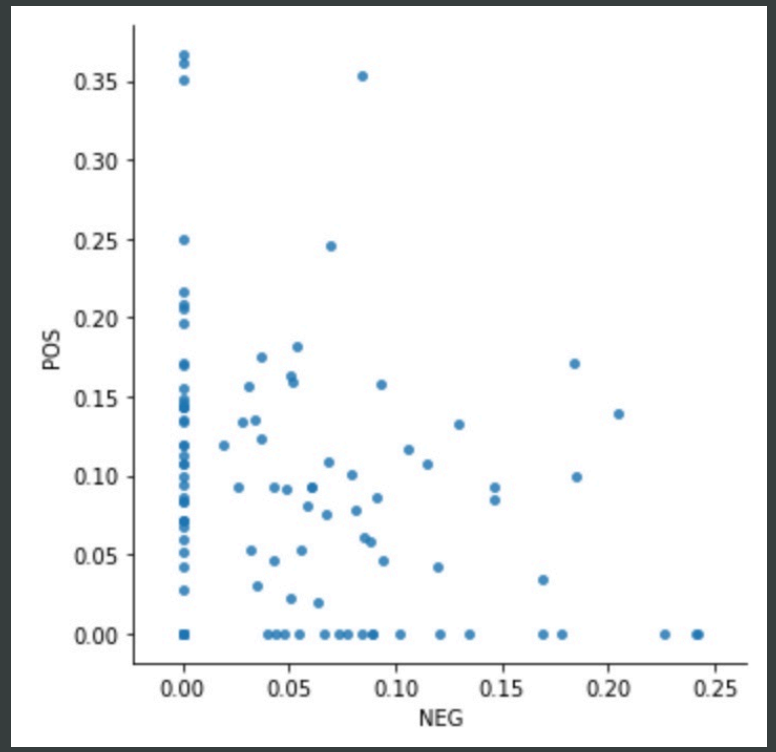
BARACK OBAMA

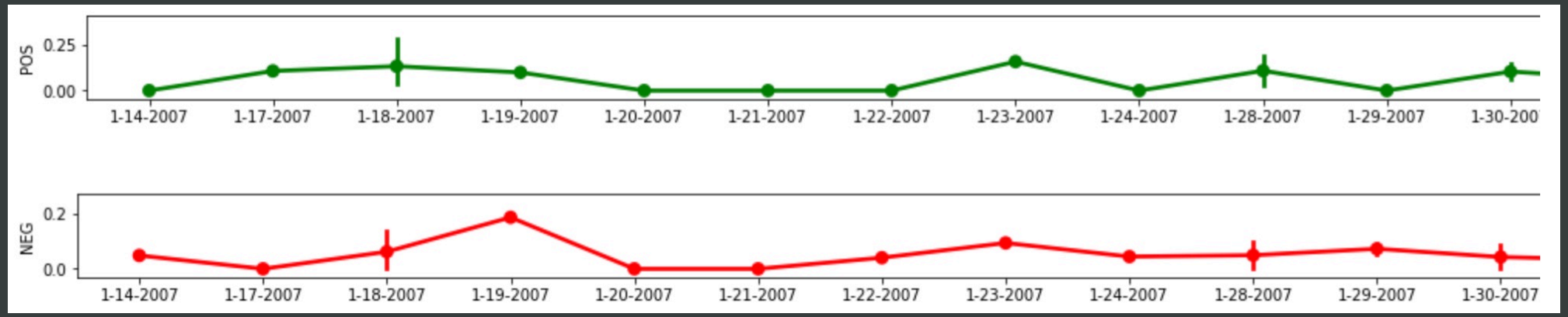
	index	DOCID	Date	Month	Year	Name	Text	Full_Date	NEG	NEU	POS
6153	0	1818955	14	1	2007	OBAMA, BARACK	on friday morning, as the capital was enmeshed...	1-14-2007	0.048	0.952	0.000
7051	0	1819443	17	1	2007	OBAMA, BARACK	two years after arriving in washington, senato...	1-17-2007	0.000	0.893	0.107
7499	0	1819626	18	1	2007	OBAMA, BARACK	there is always, it seems, a fresh new face br...	1-18-2007	0.000	0.929	0.071
7576	0	1819652	18	1	2007	OBAMA, BARACK	senator hillary rodham clinton on wednesday ca...	1-18-2007	0.178	0.822	0.000
7697	0	1819713	18	1	2007	OBAMA, BARACK	the climate here has definitely changed.	1-18-2007	0.000	0.649	0.351

```
========== SENTIMENT RESULTS ==========
POI: OBAMA, BARACK
POS SKEWED ARTICLES: 61
NEG SKEWED ARTICLES: 35
ENTIRELY NEUTRAL: 16
```

BARACK OBAMA





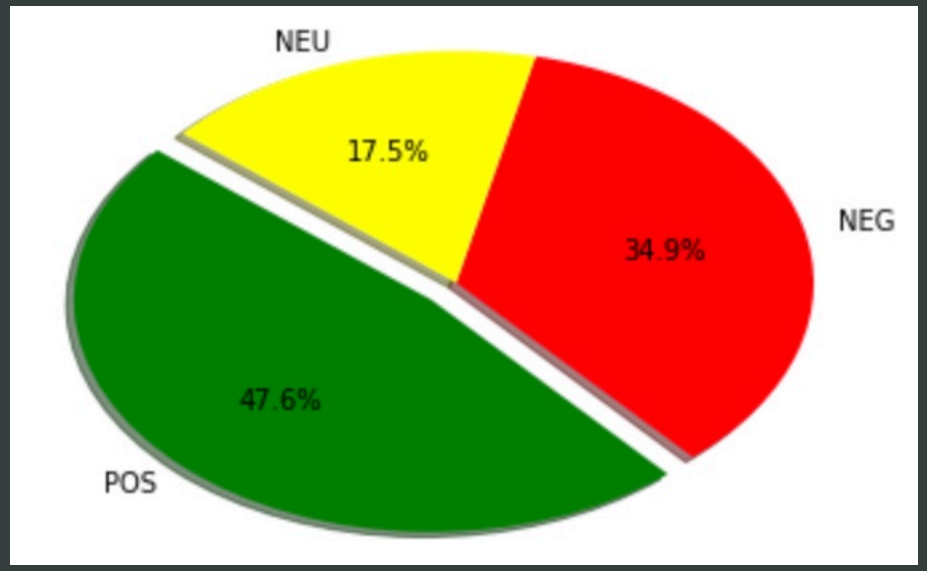


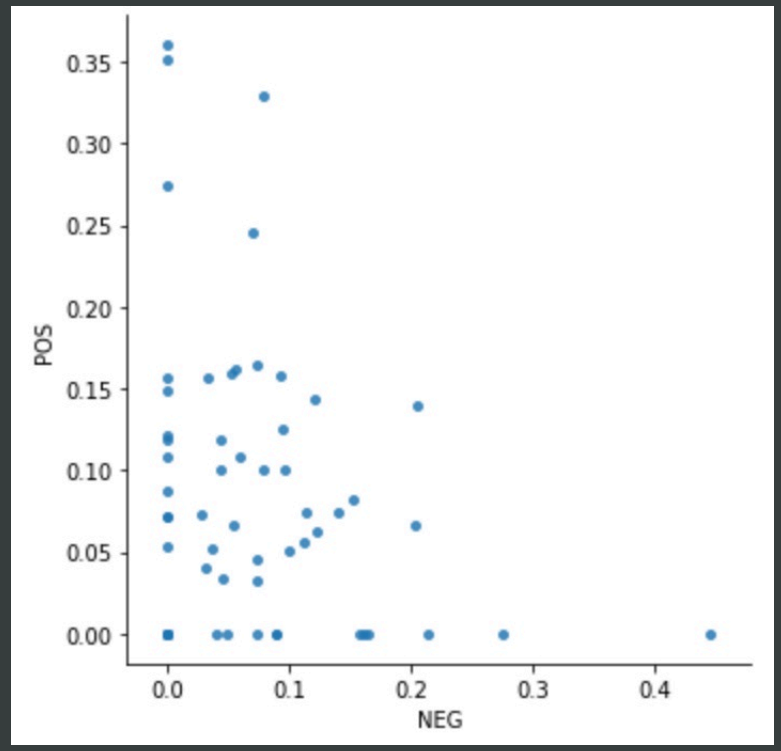
JOHN MCCAIN

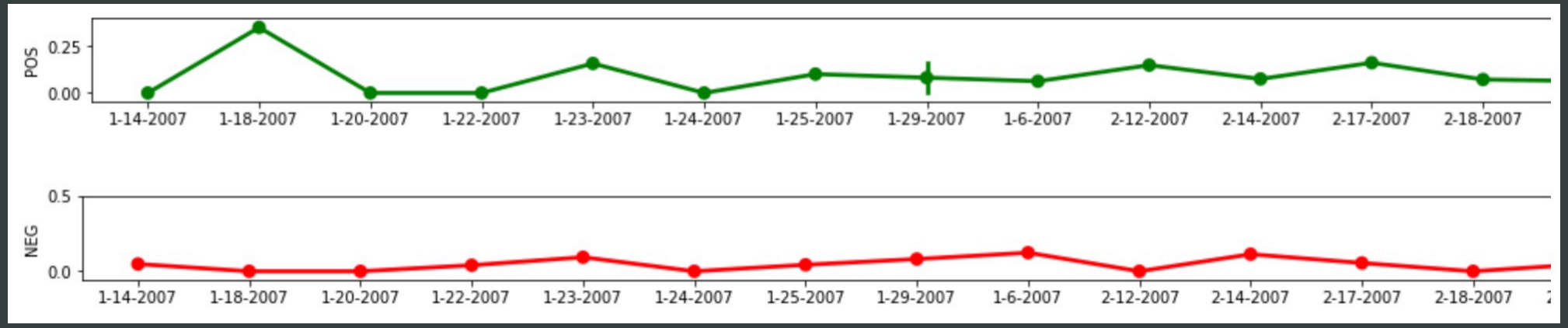
	index	DOCID	Date	Month	Year	Name	Text	Full_Date	NEG	NEU	POS
6152	0	1818955	14	1	2007	MCCAIN, JOHN	on friday morning, as the capital was enmeshed...	1-14-2007	0.048	0.952	0.000
7698	0	1819713	18	1	2007	MCCAIN, JOHN	the climate here has definitely changed.	1-18-2007	0.000	0.649	0.351
8467	0	1820058	20	1	2007	MCCAIN, JOHN	senator john mccain of arizona has enlisted st...	1-20-2007	0.000	1.000	0.000
9693	0	1820735	22	1	2007	MCCAIN, JOHN	two years before the next president is inaugur...	1-22-2007	0.040	0.960	0.000
10124	0	1820948	23	1	2007	MCCAIN, JOHN	the public financing system for presidential c...	1-23-2007	0.093	0.749	0.158

```
========== SENTIMENT RESULTS ==========
POI: MCCAIN, JOHN
POS SKEWED ARTICLES: 30
NEG SKEWED ARTICLES: 22
ENTIRELY NEUTRAL: 11
```

JOHN MCCAIN


NEU
17.5%
NEG
34.9%
47.6%
POS


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
POS
0.0
0.1
0.2
0.3
0.4
NEG


POS
0.25
0.00
1-14-2007
1-18-2007
1-20-2007
1-22-2007
1-23-2007
1-24-2007
1-25-2007
1-29-2007
1-6-2007
2-12-2007
2-14-2007
2-17-2007
2-18-2007
NEG
0.5
0.0
1-14-2007
1-18-2007
1-20-2007
1-22-2007
1-23-2007
1-24-2007
1-25-2007
1-29-2007
1-6-2007
2-12-2007
2-14-2007
2-17-2007
2-18-2007

FURTHER WORK

FURTHER WORK

- Now that I have visualizations, it gives me a point of reference to start looking into the articles themselves for further analysis
- I also want to look at various n-grams and their sentiments to see if the presence of particular phrases trigger a positive or negative sentiment
- Additionally, I want to look at what kinds of articles align with which sentiment
- One thing I would like to explore would be the same procedure, but scraping from social media websites in real time

QUESTIONS?