Sentiment Analysis

Introduction

So far, all of the analysis we've done has been pretty generic - looking at counts, creating scatter plots, etc. These techniques could be applied to numeric data as well.

When it comes to text data, there are a few popular techniques that we'll be going through in the next few notebooks, starting with sentiment analysis. A few key points to remember with sentiment analysis.

1. **TextBlob Module:** Linguistic researchers have labeled the sentiment of words based on their domain expertise. Sentiment of words can vary based on where it is in a sentence. The TextBlob module allows us to take advantage of these labels.
2. **Sentiment Labels:** Each word in a corpus is labeled in terms of polarity and subjectivity (there are more labels as well, but we're going to ignore them for now). A corpus' sentiment is the average of these.

    - **Polarity**: How positive or negative a word is. -1 is very negative. +1 is very positive.
    - **Subjectivity**: How subjective, or opinionated a word is. 0 is fact. +1 is very much an opinion.

For more info on how TextBlob coded up its [sentiment function (https://planspace.org/20150607-textblob_sentiment/)](https://planspace.org/20150607-textblob_sentiment/).

Let's take a look at the sentiment of the various transcripts, both overall and throughout the comedy routine.

Sentiment of Routine

In [1]:
```
# We'll start by reading in the corpus, which preserves word order
import pandas as pd

data = pd.read_pickle('corpus.pkl')
data
```

Out[1]:

	transcript	full_name
ali	Ladies and gentlemen, please welcome to the st...	Ali Wong
anthony	Thank you. Thank you. Thank you, San Francisco...	Anthony Jeselnik
bill	[cheers and applause] All right, thank you! Th...	Bill Burr
bo	Bo What? Old MacDonald had a farm E I E I O An...	Bo Burnham
dave	This is Dave. He tells dirty jokes for a livin...	Dave Chappelle
hasan	[theme music: orchestral hip-hop] [crowd roars...	Hasan Minhaj
jim	[Car horn honks] [Audience cheering] [Announce...	Jim Jefferies
joe	[rock music playing] [audience cheering] [anno...	Joe Rogan
john	All right, Petunia. Wish me luck out there. Yo...	John Mulaney
louis	Intro\nFade the music out. Let's roll. Hold th...	Louis C.K.
mike	Wow. Hey, thank you. Thanks. Thank you, guys. ...	Mike Birbiglia
ricky	Hello. Hello! How you doing? Great. Thank you....	Ricky Gervais

In [2]:

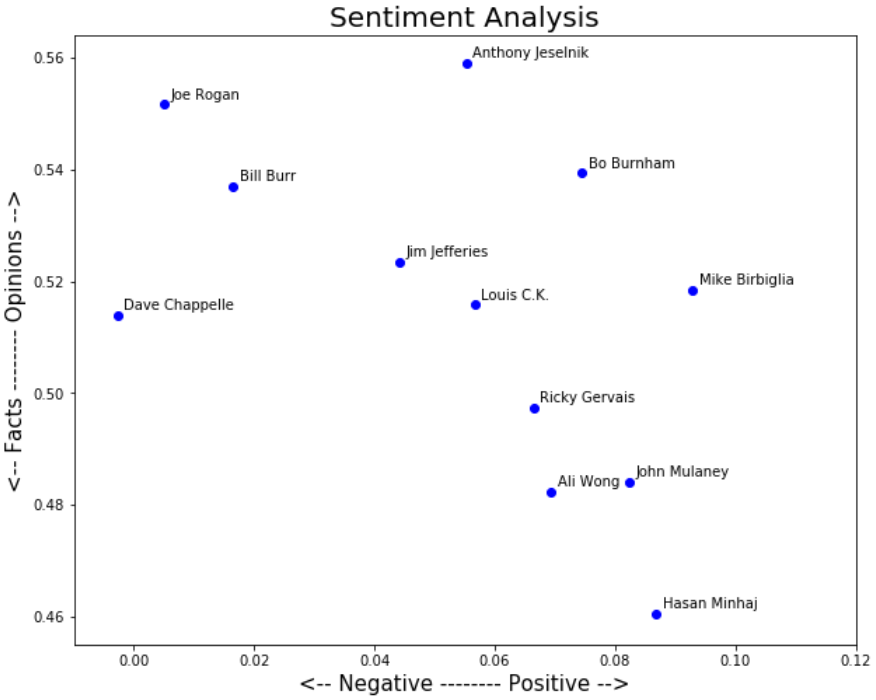
```
# Create quick lambda functions to find the polarity and subjectivity of each routine
# Terminal / Anaconda Navigator: conda install -c conda-forge textblob
from textblob import TextBlob

pol = lambda x: TextBlob(x).sentiment.polarity
sub = lambda x: TextBlob(x).sentiment.subjectivity

data['polarity'] = data['transcript'].apply(pol)
data['subjectivity'] = data['transcript'].apply(sub)
data
```

Out[2]:

	transcript	full_name	polarity	subjectivity
ali	Ladies and gentlemen, please welcome to the st...	Ali Wong	0.069359	0.482403
anthony	Thank you. Thank you. Thank you, San Francisco...	Anthony Jeselnik	0.055237	0.558976
bill	[cheers and applause] All right, thank you! Th...	Bill Burr	0.016479	0.537016
bo	Bo What? Old MacDonald had a farm E I E I O An...	Bo Burnham	0.074514	0.539368
dave	This is Dave. He tells dirty jokes for a livin...	Dave Chappelle	-0.002690	0.513958
hasan	[theme music: orchestral hip-hop] [crowd roars...	Hasan Minhaj	0.086856	0.460619
jim	[Car horn honks] [Audience cheering] [Announce...	Jim Jefferies	0.044224	0.523382
joe	[rock music playing] [audience cheering] [anno...	Joe Rogan	0.004968	0.551628
john	All right, Petunia. Wish me luck out there. Yo...	John Mulaney	0.082355	0.484137
louis	Intro\nFade the music out. Let’s roll. Hold th...	Louis C.K.	0.056665	0.515796
mike	Wow. Hey, thank you. Thanks. Thank you, guys. ...	Mike Birbiglia	0.092927	0.518476
ricky	Hello. Hello! How you doing? Great. Thank you....	Ricky Gervais	0.066489	0.497313

In [3]:

```
# Let's plot the results
import matplotlib.pyplot as plt

plt.rcParams['figure.figsize'] = [10, 8]

for index, comedian in enumerate(data.index):
    x = data.polarity.loc[comedian]
    y = data.subjectivity.loc[comedian]
    plt.scatter(x, y, color='blue')
    plt.text(x+.001, y+.001, data['full_name'][index], fontsize=10)
    plt.xlim(-.01, .12)

plt.title('Sentiment Analysis', fontsize=20)
plt.xlabel('<-- Negative -------- Positive -->', fontsize=15)
plt.ylabel('<-- Facts -------- Opinions -->', fontsize=15)

plt.show()
```



Sentiment of Routine Over Time

Instead of looking at the overall sentiment, let's see if there's anything interesting about the sentiment over time throughout each routine.

In [4]:

```python
# Split each routine into 10 parts
import numpy as np
import math

def split_text(text, n=10):
    '''Takes in a string of text and splits into n equal parts, with a default of 10 equal parts.'''

    # Calculate length of text, the size of each chunk of text and the starting points of each chunk of text
    length = len(text)
    size = math.floor(length / n)
    start = np.arange(0, length, size)

    # Pull out equally sized pieces of text and put it into a list
    split_list = []
    for piece in range(n):
        split_list.append(text[start[piece]:start[piece]+size])
    return split_list
```

In [5]:
```
# Let's take a look at our data again
data
```

Out[5]:

	transcript	full_name	polarity	subjectivity
ali	Ladies and gentlemen, please welcome to the st...	Ali Wong	0.069359	0.482403
anthony	Thank you. Thank you. Thank you, San Francisco...	Anthony Jeselnik	0.055237	0.558976
bill	[cheers and applause] All right, thank you! Th...	Bill Burr	0.016479	0.537016
bo	Bo What? Old MacDonald had a farm E I E I O An...	Bo Burnham	0.074514	0.539368
dave	This is Dave. He tells dirty jokes for a livin...	Dave Chappelle	-0.002690	0.513958
hasan	[theme music: orchestral hip-hop] [crowd roars...	Hasan Minhaj	0.086856	0.460619
jim	[Car horn honks] [Audience cheering] [Announce...	Jim Jefferies	0.044224	0.523382
joe	[rock music playing] [audience cheering] [anno...	Joe Rogan	0.004968	0.551628
john	All right, Petunia. Wish me luck out there. Yo...	John Mulaney	0.082355	0.484137
louis	Intro\nFade the music out. Let's roll. Hold th...	Louis C.K.	0.056665	0.515796
mike	Wow. Hey, thank you. Thanks. Thank you, guys. ...	Mike Birbiglia	0.092927	0.518476
ricky	Hello. Hello! How you doing? Great. Thank you....	Ricky Gervais	0.066489	0.497313

In [6]:
```
# Let's create a list to hold all of the pieces of text
list_pieces = []
for t in data.transcript:
    split = split_text(t)
    list_pieces.append(split)

list_pieces
```

Out[6]:
```
[['Ladies and gentlemen, please welcome to the stage: Ali Wong! Hi. Hello! Welcome! Thank you! Thank you for coming. Hello! Hello. We are gonna hav
e to get this shit over with, 'cause I have to pee in, like, ten minutes. But thank you, everybody, so much for coming. Um… It's a very exciting da
y for me. It's been a very exciting year for me. I turned 33 this year. Yes! Thank you, five people. I appreciate that. Uh, I can tell that I'm get
ting older, because, now, when I see an 18-year-old girl, my automatic thought… is "Fuck you." "Fuck you. I don't even know you, but fuck you!" 'Ca
use I'm straight up jealous. I'm jealous, first and foremost, of their metabolism. Because 18-year-old girls, they could just eat like shit, and th
en they take a shit and have a six-pack, right? They got that-that beautiful inner thigh clearance where they put their feet together and there's t
hat huge gap here with the light of potential just radiating through.\nAnd then, when they go to sleep, they just go to sleep. Right? They don't ha
ve insomnia yet. They don't know what it's like to have to take a Ambien or download a Meditation Oasis podcast to calm the chatter of regret and r
esentment towards your family just cluttering your mind. They have their whole lives ahead of them. They don't have HPV yet. They just go to sleep
in peace at night. Everybody has HPV, OK? Everybody has it. It's OK. Come out already. Everybody has it. If you don't have it yet, you go and get i
t. You go and get it. It's coming. You don't have HPV yet, you're a fucking loser, all right? That's what that says about you. A lot of men don't k
now that they have HPV, because it's undetectable in men. It's really fucked up. HPV is a ghost that lives inside men's bodies and says, "Boo!" in
women's bodies. My doctor told me that I have one of two strains of HPV. Either I have the kind that's gonna turn into cervical cancer… …or I have
the kind where my body will heal itself. Very helpful, this doctor, right? So, basically, either I'm gonna die… or you're in the presence of Wolver
ine, bitches. We'll find out. Um, I can also tell that I'm getting older, because my Kindle is turning into a self-help library. I'm not interested
in books like Fifty Shades of Grey, OK? I'm interested in The Life-Changing Magic of Tidying Up. Yes. Yes, that's right, how to declutter my home t
o achieve inner peace and my optimum level of success. That's what your 30s is all about. How can I turn this shit around? I'm a horrible person,
I'm not happy with where I am, how can I turn this shit around? Help me, Tony Robbins, help me!\nI have a hoarding problem, which I'm hoping is the
center of all of my other problems. I'm hoping that if the hoarding goes away, the HPV will also disappear. I have a hoarding problem because my mo
m is from a third world country and she taught me that you can never throw away anything, because you never know when a dictator's gonna overtake t
```

In [7]:
```
# The list has 10 elements, one for each transcript
len(list_pieces)
```

Out[7]:
```
12
```

In [8]:
```
# Each transcript has been split into 10 pieces of text
len(list_pieces[0])
```

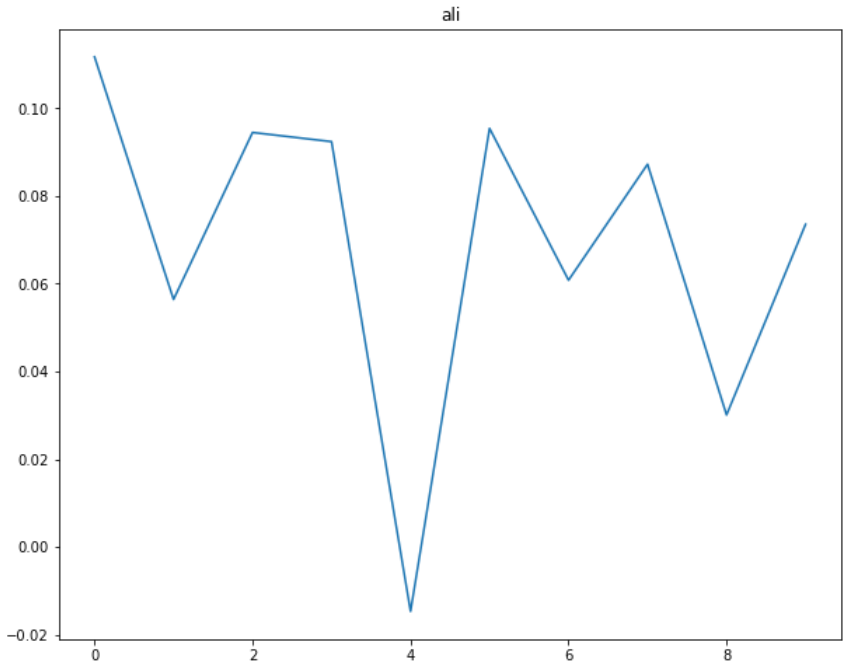
Out[8]:
```
10
```

In [9]:
```
# Calculate the polarity for each piece of text

polarity_transcript = []
for lp in list_pieces:
    polarity_piece = []
    for p in lp:
        polarity_piece.append(TextBlob(p).sentiment.polarity)
    polarity_transcript.append(polarity_piece)

polarity_transcript
```

Out[9]:
```
[[0.11168482647296207,
  0.056407029478458055,
  0.09445691155249979,
  0.09236886724386723,
  -0.014671592775041055,
  0.09538361348808912,
  0.06079713127248339,
  0.08721655328798185,
  0.030089690638160044,
  0.07351994851994852],
 [0.13933883477633482,
  -0.06333451704545455,
  -0.056153799903799935,
  0.014602659245516405,
  0.16377334420812684,
  0.09091338259441709,
  0.09420031055900621,
  0.11566683919944787,
  -0.04238582919138478,
```

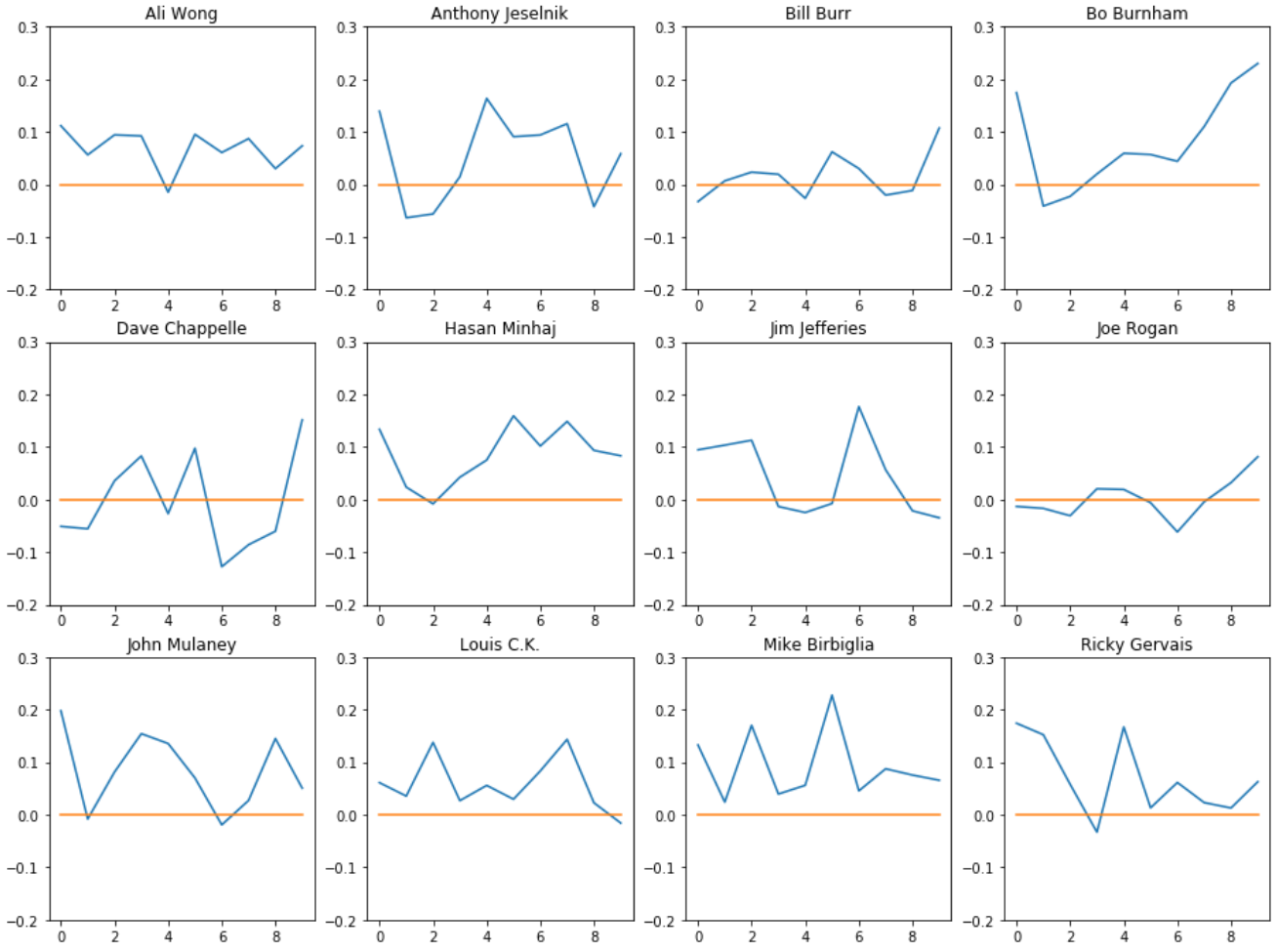
In [10]:

```
# Show the plot for one comedian
plt.plot(polarity_transcript[0])
plt.title(data['full_name'].index[0])
plt.show()
```



In [11]:

```
# Show the plot for all comedians
plt.rcParams['figure.figsize'] = [16, 12]

for index, comedian in enumerate(data.index):
    plt.subplot(3, 4, index+1)
    plt.plot(polarity_transcript[index])
    plt.plot(np.arange(0,10), np.zeros(10))
    plt.title(data['full_name'][index])
    plt.ylim(ymin=-.2, ymax=.3)

plt.show()
```



Ali Wong stays generally positive throughout her routine. Similar comedians are Louis C.K. and Mike Birbiglia.

On the other hand, you have some pretty different patterns here like Bo Burnham who gets happier as time passes and Dave Chappelle who has some pretty down moments in his routine.

Additional Exercises

1. Modify the number of sections the comedy routine is split into and see how the charts over time change.

In [ ]:
```
1
```