Final Project ChE 395

by **Spencer Hong**

"Trump speaks at level of 8-year-old, new analysis finds."
Is this true? Can we prove it?

Let's play around with the data

In [1]:
```
import pandas as pd
```

In [2]:
```
tweets = pd.read_csv('data/trumptweets-1515775693.tweets.csv', low_memory = False)
tweets.head()
```

Out[2]:

	status_id	created_at	user_id	screen_name	text	source	display_text_width	reply_to_status_id	reply_
0	x1864367186	2009-05-20 22:29:47	x25073877	realDonaldTrump	Read a great interview with Donald Trump that ...	Twitter Web Client	112	NaN	
1	x9273573134835712	2010-11-29 15:52:46	x25073877	realDonaldTrump	Congratulations to Evan Lysacek for being nomi...	Twitter Web Client	127	NaN	
2	x29014512646	2010-10-28 18:53:40	x25073877	realDonaldTrump	I was on The View this morning. We talked abou...	Twitter Web Client	139	NaN	
3	x7483813542232064	2010-11-24 17:20:54	x25073877	realDonaldTrump	Tomorrow night's episode of The Apprentice del...	Twitter Web Client	140	NaN	
4	x5775731054	2009-11-16 21:06:10	x25073877	realDonaldTrump	Donald Trump Partners with TV1 on New Reality ...	Twitter Web Client	116	NaN	

5 rows × 68 columns

What is the range of dates on this dataset?

In [3]:
```
tweets[['created_at']].sort_values('created_at').iloc[[0, -1]]
```

Out[3]:

	created_at
18	2009-05-04 18:54:25
32644	2018-01-12 13:48:49

Our analysis will pertain to Donald Trump tweets from 2009 to 2018.

How often does Donald Trump retweet vs. spit out a tweet of his own?

In [4]:
```
len(tweets[tweets.is_retweet])/len(tweets) * 100
```

Out[4]:
```
1.5627855967830377
```

He only retweets 1.6% of the time. This means we have a good enough dataset to analyze his own writing. Let's omit retweets from the dataset. We're also only worried about the text, so let's just keep the text and the date, and drop everything else.

In [5]:
```
tweets = tweets[~tweets.is_retweet][['created_at','text']]
tweets.head()
```

Out[5]:

	created_at	text
0	2009-05-20 22:29:47	Read a great interview with Donald Trump that ...
1	2010-11-29 15:52:46	Congratulations to Evan Lysacek for being nomi...
2	2010-10-28 18:53:40	I was on The View this morning. We talked abou...
3	2010-11-24 17:20:54	Tomorrow night's episode of The Apprentice del...
4	2009-11-16 21:06:10	Donald Trump Partners with TV1 on New Reality ...

If you look closely, even if we take out the retweets, there are some weird tweets (looks like not a first-person type of writing). See below for examples.

In [6]:
```
tweets.iloc[0].text
```

Out[6]:
```
'Read a great interview with Donald Trump that appeared in The New York Times Magazine: http://tinyurl.com/qsx4o6'
```

In [7]:
```
tweets.iloc[4].text
```

Out[7]:
```
"Donald Trump Partners with TV1 on New Reality Series Entitled, Omarosa's Ultimate Merger: http://tinyurl.com/yk5m3lc"
```

after some research, it is clear that even the non-retweets may not come from himself -- and to get to the heart of the question, we want to only get the tweets that represent Donald Trump as himself, no one else. Let's also avoid all the media links. Trump also sometimes retweets other accounts using "RT @account" method. Let's avoid this as well.

In [11]:
```
tweets = tweets[~(tweets.text.str.contains('http') | tweets.text.str.contains('RT @'))]
print(f'we have {tweets.shape[0]} tweets after preprocessing')
tweets.head(2)
```

```
we have 24669 tweets after preprocessing
```

Out[11]:

	created_at	text
1	2010-11-29 15:52:46	Congratulations to Evan Lysacek for being nomi...
2	2010-10-28 18:53:40	I was on The View this morning. We talked abou...

Okay, now we're ready.

Question 1

What distribution does Donald Trump's word frequency follow? For now, just plot, and find a distribution that might fit this. No parameters are needed yet, just a visual approximation.

This question is intended for you to figure out how to best 'split' tweets into data.

Question 2

Obviously, we're going to need a more rigorous way to find the distribution. First, find the maximum likelihood estimator for the parameter of the Zipf distribution. Zipf distribution only requires 1 parameter, *s*, the shape parameter.

Question 3

Using the MLE from question 2, bootstrap 10000 times and calculate this estimator 10000 times. What is the mean? the variance? Does this match with analytical solution?

Question 4

Using our hypothesis testing knowledge from class, determine if Donald Trump's tweets follow the distribution with the shape estimator from the bootstrap in question 3. You will need to set up a null/alternative hypotheses and a log-likelihood ratio statistic. Plot the p-values and the test statistic from the 10000 bootstraps.

Question 5

Now that we have a distribution that describes Donald Trump's "speaking" (writing) quality, **get the middle 30% of words in this distribution**. By doing so, we avoid very common words (called stop words) such as "the", "for", "and", but we also avoid words that are very rarely used and therefore do not represent the level of speaker well. Give us 10 words that fall in this section.

Question 6

In natural langauge processing, a method to test the reading level of a writing is called the **Flesch Kincaid Grade Level Test**. Let's apply this test to the middle 30% of words that we've identified. Give a histogram of the results.

Question 7

Calculate and present a 95% confidence interval of the grade results per tweet.

Question 8

From the probability mass distribution in Question 2, calculate random sequences of words (creating our own sentences from the bank of Donald Trump words) using a **Metropolis-Hastings** algorithm. You must recreate this from scratch (no importing custom functions!).

In [ ]:
```

```