Vinh Tuan Dat Nguyen (Percy)

<center>Let's Communicate 2</center>

Concept: Finding out patterns in Trump's rally speeches.

Disclaimer: I am not a supporter, but rather a critic of Trump.

<center>**Approach 1**</center>

Data:

FreelandSep10_2020.txt (**10137 words**)

Steps:

1. Load and read Freeland speech 2020.

2. Count word with 2 or more characters

3. Use POST TAG to find the types of each word

```
const language = "EN"
const defaultCategory = 'N';
const defaultCategoryCapitalized = 'NNP';

var lexicon = new natural.Lexicon(language, defaultCategory, defaultCategoryCapitalized);
var ruleSet = new natural.RuleSet('EN');
var tagger = new natural.BrillPOSTagger(lexicon, ruleSet);

console.log(tagger.tag(['the','and','we','to','it','you','that', 'of','in','they','he','re','have','was','going','said','but','is','know',
    'all', 'Jason'
    ]));
```

Legends

- ❖ CC coordinating conjunction
- ❖ DT determiner
- ❖ NN noun, singular 'desk'
- ❖ NNS noun plural 'desks'
- ❖ PRP personal pronoun I, he, she
- ❖ TO, to go 'to' the store.
- ❖ VB verb, base form
- ❖ VBD verb, past tense took
- ❖ VBG verb, gerund/present participle
- ❖ VBZ verb, 3rd person sing. presents

Words	Count #	Type
the	373	DT
and	323	CC
we	286	PRP
to	278	TO
it	237	PRP
you	227	PRP
that	197	IN
of	187	IN
in	140	IN
they	134	PRP
he	125	PRP
re	109	NN
have	90	VBD
was	81	VBD
going	78	VBG
said	77	VBD
but	68	CC
is	67	VBZ
know	64	VB
all	61	DT

Conclusion:

- ❖ The English language is very redundant. These 20 most used words do not suggest the slightest hint of a Trump's speech. Little can be learnt or hypothesized from these findings.
- ❖ Another revelation is the appearance of the illegitimate word "re". Upon later review, this made up word derives from English's contracted form of verb conjugations in present tense. **"We're"** is understood to consisted of **we** and **are**, though appears as **we** and **re** to NLP).

Approach 2

Data:

1. BemidjiSep18_2020.txt (Minnesota)
2. CharlestonFeb28_2020.txt (South Carolina)
3. CharlotteMar2_2020.txt (North Carolina)
4. ColoradorSpringsFeb20_2020.txt (Colorado)
5. DesMoinesJan30_2020.txt (Iowa)
6. FayettevilleSep19_2020.txt (North Carolina)
7. FreelandSep10_2020.txt (Michigan)
8. HendersonSep13_2020.txt (Nevada)
9. LasVegasFeb21_2020.txt (Nevada)
10. LatrobeSep3_2020.txt (Pennsylvania)

Steps:

- ❖ Load and read 10 transcripts from Trump's rallies in 2020.
- ❖ Employed TF IDF formula, then sorted the result by order from highest to lowest TF IDF (Sabine's code)

```
// Finish and calculate everything
finish(totaldocs) {

    //console.log(totaldocs)
    // calculate tf-idf score
    for (let i = 0; i < this.keys.length; i++) {
      let key = this.keys[i];
      let word = this.dict[key];
    // console.log(word);
      let tf = word.count / this.totalwords;
    //let tf = word.count;

    //console.log(totaldocs / word.docCount);
      // See:
      let idf = Math.log10(totaldocs / word.docCount);
      word.tfidf = tf * idf;
    }
  }
```

- ❖ Manually choose the top 10 words with the highest TF IDF score
- ❖ Insert these 10 words into a string, to be read individually by the sentiment analysis

```
var Analyzer = require('natural').SentimentAnalyzer;
var stemmer = require('natural').PorterStemmer;
var analyzer = new Analyzer("English", stemmer, "afinn");
// getSentiment expects an array of strings

console.log(analyzer.getSentiment(['minnesota','colorado','nevada','jason','rush','thom','pulitzer','colby','police']));
```

Result:

The highest rated word, Minnesota, has more than double the score of the 10th rated word, police. The sentiment analysis shows a score of 0 for every word from the list. This turns out to be unsurprising as 8 out of those 10 words are proper nouns, and hence remain neutral.

Word	Total Count	Doc Count	IF IDF score	Sentiment Analysis
minnesota	40	2	0.000245319342746192	0
colorado	38	2	0.0002330533756088824	0
nevada	36	3	0.00016516451693085097	0
jason	17	1	0.0001491633689863033	0
rush	15	1	0.00013161473734085586	0
lindsey	19	2	0.0001165266878044412	0
thom	19	2	0.0001165266878044412	0
pulitzer	13	1	0.00011140661056954084	0
colby	13	1	0.00011140661056954084	0
police	43	5	0.00011357728692496373	0

Conclusion:

The sentiment analysis's rigorous rules seem to calculate the average of the sentiment score among all 10 words, failing to understand the context behind them. After some tinkering and testing with proper nouns, NLP sentiment analysis appears to give all proper nouns a score of 0. This rating is dubious and debatable as the names of dictators and notorious criminals usually evoke fear and disgust, which should result in a negative score instead of a neutral score of 0.

The names of the states (Carolina, Iowa, Michigan and Pennsylvania) that were not mentioned in the top 10 were in between the top 10 to top 24 most important words. The TF DIF reveals Trump's intentional and tactical usage of the state names that he was giving the speech to. By repeating the names of the states, he appears to be addressing the targeted audience and thus appearing more caring towards the state's inhabitants. This is a psychological trick similar to how humans often call out the names of the people they love and admire. As it is a powerful method to connect to people effectively, one can argue that Trump used it to rally pride and unity.

Jason refers to Jason Miller, Trump's spokesperson.
Thom refers to Thom Tiller, a Republican Senate
Pulitzer refers to the Pulitzer Prizes, prizes for excellence in journalism.
Colby refers to Colby Collington, a martial artist and outspoken supporter of Donald Trump and the Republican party.

Data Visualization

Minnesota
Colorado
Nevada
Jason
Rush
Thom
Pulitzer
Colby
Police

TRUMP 2020

Despite disagreeing with the sentiment analysis, I chose black and white colors to represent the neutral tone of the NLP towards the top IF IDF words. Their size and placement represent the significance of each word slowly losing its importance as it goes down the list. The font chosen mimics Trump's Make America Great Again slogan.