‘Speech’ Recognition:

Using NLP to Attribute Quotes to Speakers

Luke Benz, Kevin Truong, Will Langhorne

Introduction

- Speakers are remembered for idiosyncrasies
- People are able to identify speakers of famous quotes
    - “Nothing to fear but fear itself”
    - “Yes we can!”
- Can a language model do the same?
- Goal: Build a model capable of associating most likely speaker of a quote
- Train model on words and parts of speech
- Speeches and Tweets

Methodology

01 Gather and clean data

02 Construct N-gram model using the Simple Good-Turing algorithm for smoothing

03 Fit the model to text and POS tags from tweets and speeches

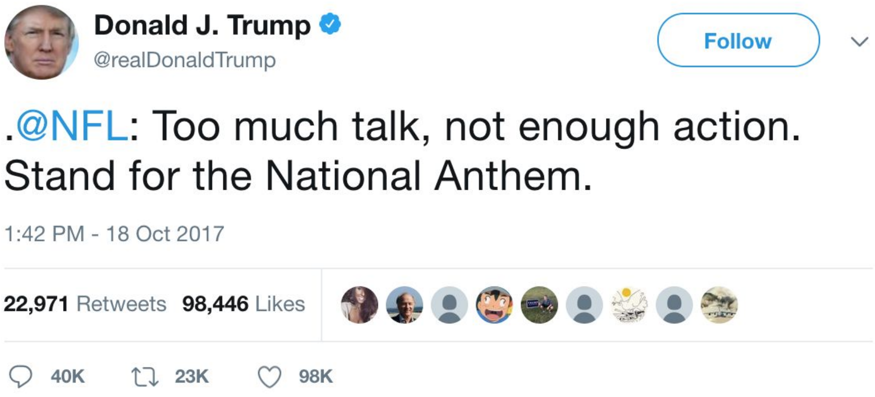
04 Test and evaluate performance on sample quotes

Data Sources and Cleaning

- Sources
  - The Grammar Lab: Speeches from 44 presidents and Hillary Clinton
  - FiveThirtyEight: Tweets from Obama, Trump and Senators
- Cleaning:
  - Lower cased
  - Punctuation removed and periods replaced with beginning and end of speech tags
  - Links and Retweets removed
  - Interview reporter questions removed
  - Information on speech date and location removed

Cleaned Tweet

[‘<s>’, u’@nfl’, u’:‘, u’too’, u’much’, u’talk’, u’not’, u’enough’, u’action’, ‘</s>’, ‘<s>’, u’stand’, u’for’, u’the’, u’national’, u’anthem’, ‘</s>’]



Donald J. Trump
@realDonaldTrump

.@NFL: Too much talk, not enough action. Stand for the National Anthem.

1:42 PM - 18 Oct 2017

22,971 Retweets 98,446 Likes

Fitting the Model

- Bigram was chosen as model
    - Limited to bigram due to large number of unseen trigrams
- Good Turing Smoothing used
- Hidden Markov Model used for POS tags
    - Trained on data from Brown Corpus

Simple Good Turing Smoothing

- Use frequency of singletons to estimate zero frequency events
- Compute N_c : number of N-grams occurring c times:
- Use N_c and N_{c+1} to estimate new counts
- N is total number of seen counts

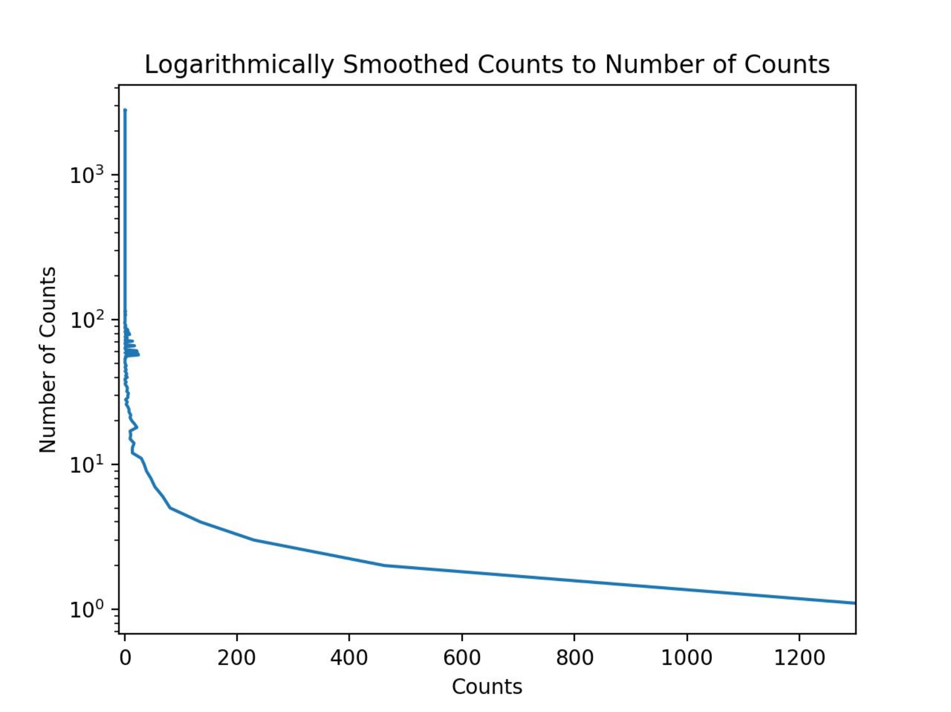
$$c* = (c+1)\frac{N_{c+1}}{N_c}$$

$$P(N - gramwithzero - count) = \frac{N_1}{N}$$

Simple Good Turing Smoothing

- Holes in raw data
- For some c, $N_c = 0$
- Filled in using linear regression function
- Map $N_c = 0$ to c in log space

$$\log(N_c) = a + b\log(c)$$



Testing the Model

- 10 quotes from each type of file (tweets/ speeches)
- Model trained on files whose speakers correspond to quotes
- Probability of each quote coming from each speaker is calculated
- Attribute quote to speaker with highest probability

$$\mathbb{P}(Q) = \prod_{i=i}^{n} \mathbb{P}(w_i|w_{i-1}) = 2^{\sum_{i=1}^{n} \log_2(\mathbb{P}(w_i|w_{i-1}))}$$

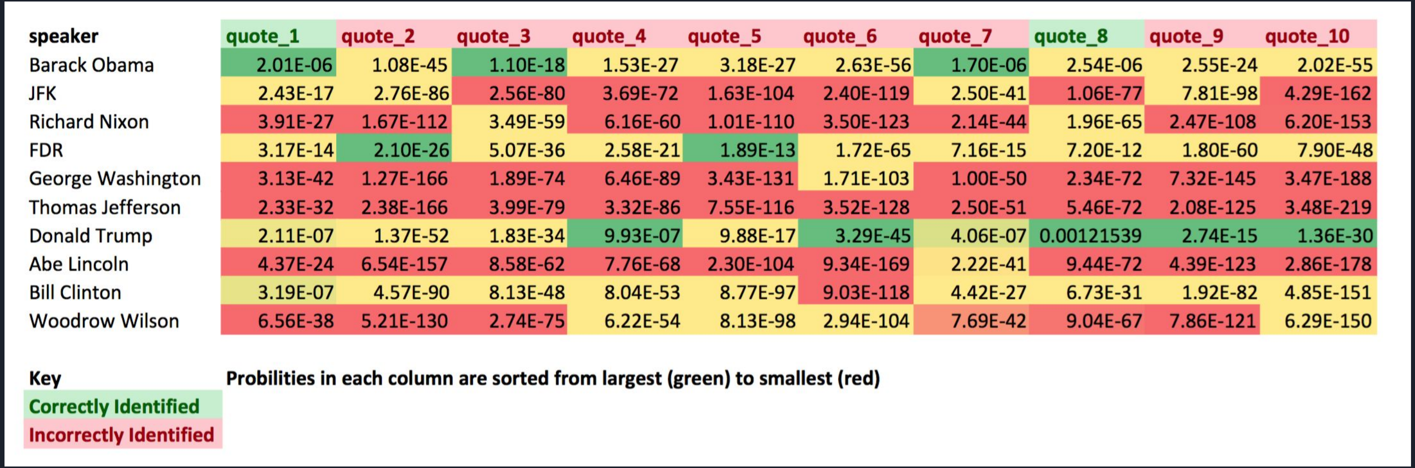
$$\hat{S} = \underset{s \in \text{ Possible Speakers}}{\arg\max} \mathbb{P}(Q|s)$$

Presidential Speech Quotes

- Quote 1: <s> yes we can </s> - Barack Obama
- Quote 2: <s> and so my fellow Americans ask not what your country can do for you ask what you can do for your country </s> - JFK
- Quote 3: <s> when the president does it that means that it is not illegal </s> - Richard Nixon
- Quote 4: <s> the only thing we have to fear is fear itself </s> - FDR
- Quote 5: <s> liberty when it begins to take root is a plant of rapid growth </s> - George Washington
- Quote 6: <s> the care of human life and happiness and not their destruction is the first and only object of good government </s> - Thomas Jefferson
- Quote 7: <s> make america great again </s> -Donald Trump
- Quote 8: <s> a house divided against itself cannot stand </s> - Abraham Lincoln
- Quote 9: <s> we must teach our children to resolve their conflicts with words not weapons </s> - Bill Clinton
- Quote 10: <s> i would rather belong to a poor nation that was free than to a rich nation that had ceased to be in love with liberty </s> -Woodrow Wilson

Speech Text Model Results

- 7 quotes attributed to either Trump or Obama
- Larger training corpuses for 2 most recent presidents
  - 50 Obama speeches to 21 Washington speeches
- Most likely quote-speaker pair: Donald Trump and “Make America Great Again”



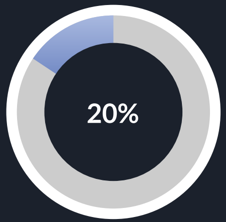
speaker	quote_1	quote_2	quote_3	quote_4	quote_5	quote_6	quote_7	quote_8	quote_9	quote_10
Barack Obama	2.01E-06	1.08E-45	1.10E-18	1.53E-27	3.18E-27	2.63E-56	1.70E-06	2.54E-06	2.55E-24	2.02E-55
JFK	2.43E-17	2.76E-86	2.56E-80	3.69E-72	1.63E-104	2.40E-119	2.50E-41	1.06E-77	7.81E-98	4.29E-162
Richard Nixon	3.91E-27	1.67E-112	3.49E-59	6.16E-60	1.01E-110	3.50E-123	2.14E-44	1.96E-65	2.47E-108	6.20E-153
FDR	3.17E-14	2.10E-26	5.07E-36	2.58E-21	1.89E-13	1.72E-65	7.16E-15	7.20E-12	1.80E-60	7.90E-48
George Washington	3.13E-42	1.27E-166	1.89E-74	6.46E-89	3.43E-131	1.71E-103	1.00E-50	2.34E-72	7.32E-145	3.47E-188
Thomas Jefferson	2.33E-32	2.38E-166	3.99E-79	3.32E-86	7.55E-116	3.52E-128	2.50E-51	5.46E-72	2.08E-125	3.48E-219
Donald Trump	2.11E-07	1.37E-52	1.83E-34	9.93E-07	9.88E-17	3.29E-45	4.06E-07	0.00121539	2.74E-15	1.36E-30
Abe Lincoln	4.37E-24	6.54E-157	8.58E-62	7.76E-68	2.30E-104	9.34E-169	2.22E-41	9.44E-72	4.39E-123	2.86E-178
Bill Clinton	3.19E-07	4.57E-90	8.13E-48	8.04E-53	8.77E-97	9.03E-118	4.42E-27	6.73E-31	1.92E-82	4.85E-151
Woodrow Wilson	6.56E-38	5.21E-130	2.74E-75	6.22E-54	8.13E-98	2.94E-104	7.69E-42	9.04E-67	7.86E-121	6.29E-150

Key

Probilities in each column are sorted from largest (green) to smallest (red)

🟩 **Correctly Identified**

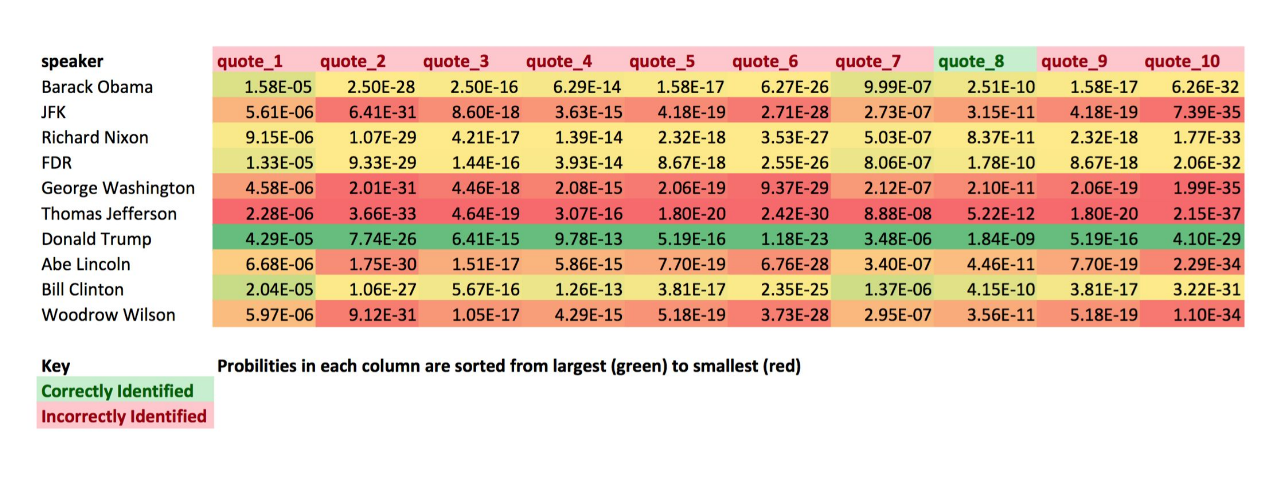
🟥 **Incorrectly Identified**



Text Model Accuracy

Speech POS Model Results

- Probabilities of each speaker-tag sequence pair are much larger than each speaker-quote pair
- Due to fewer possible tags - few possible bigrams
- Model attributes each POS tag to Donald Trump
- Little variation in POS tags between quotes
- Trump has most speeches
- Little discernibility between speakers and POS tags



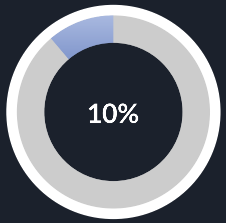
speaker	quote_1	quote_2	quote_3	quote_4	quote_5	quote_6	quote_7	quote_8	quote_9	quote_10
Barack Obama	1.58E-05	2.50E-28	2.50E-16	6.29E-14	1.58E-17	6.27E-26	9.99E-07	2.51E-10	1.58E-17	6.26E-32
JFK	5.61E-06	6.41E-31	8.60E-18	3.63E-15	4.18E-19	2.71E-28	2.73E-07	3.15E-11	4.18E-19	7.39E-35
Richard Nixon	9.15E-06	1.07E-29	4.21E-17	1.39E-14	2.32E-18	3.53E-27	5.03E-07	8.37E-11	2.32E-18	1.77E-33
FDR	1.33E-05	9.33E-29	1.44E-16	3.93E-14	8.67E-18	2.55E-26	8.06E-07	1.78E-10	8.67E-18	2.06E-32
George Washington	4.58E-06	2.01E-31	4.46E-18	2.08E-15	2.06E-19	9.37E-29	2.12E-07	2.10E-11	2.06E-19	1.99E-35
Thomas Jefferson	2.28E-06	3.66E-33	4.64E-19	3.07E-16	1.80E-20	2.42E-30	8.88E-08	5.22E-12	1.80E-20	2.15E-37
Donald Trump	4.29E-05	7.74E-26	6.41E-15	9.78E-13	5.19E-16	1.18E-23	3.48E-06	1.84E-09	5.19E-16	4.10E-29
Abe Lincoln	6.68E-06	1.75E-30	1.51E-17	5.86E-15	7.70E-19	6.76E-28	3.40E-07	4.46E-11	7.70E-19	2.29E-34
Bill Clinton	2.04E-05	1.06E-27	5.67E-16	1.26E-13	3.81E-17	2.35E-25	1.37E-06	4.15E-10	3.81E-17	3.22E-31
Woodrow Wilson	5.97E-06	9.12E-31	1.05E-17	4.29E-15	5.18E-19	3.73E-28	2.95E-07	3.56E-11	5.18E-19	1.10E-34

Key

Probilities in each column are sorted from largest (green) to smallest (red)

Correctly Identified

Incorrectly Identified



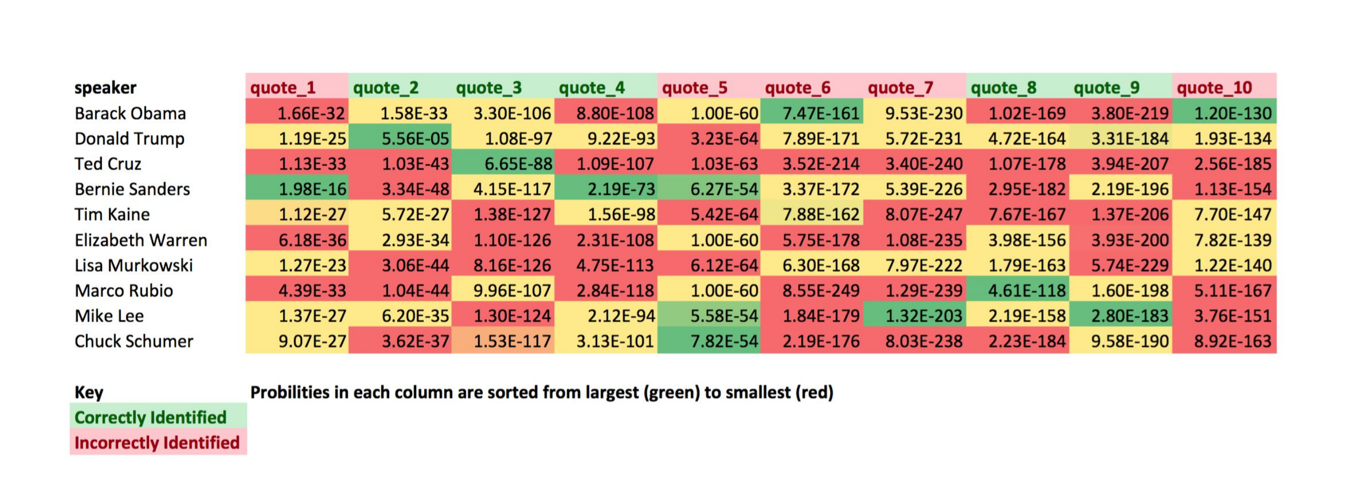
POS Model
Accuracy

Tweet Quotes

- Quote 1: <s> yes we can </s> - Barack Obama
- Quote 2: <s> make america great again </s> - Donald Trump
- Quote 3 <s> stark reminder of the threat radical islamic terror still poses to our homeland </s> -Ted Cruz
- Quote 4: <s> it’s not by giving massive tax breaks to your billionaire friends </s> - Bernie Sanders
- Quote 5: <s> tax bill equals swamp creature </s> - Tim Kaine
- Quote 6: <s> @SenBernieSanders & I are talking about things democrats are fighting for in this spending bill that would make a big difference to working families </s> - Elizabeth Warren
- Quote 7: <s> whether you are in the media politics or anywhere else abuse of power is unacceptable & shouldn’t be tolerated at any place at any level </s> - Lisa Murkowski
- Quote 8: <s> in a foreign relations committee hearing today i discussed how u.S. nuclear forces protect our nation and allies in the 21st century </s> - Marco Rubio
- Quote 9: <s> president trump did the people of utah a great favor today by rolling back harmful land use restrictions in southern utah #utpol </s> - Mike Lee
- Quote 10: <s> better #GetCovered it’s the last weekend to sign up for 2018 health coverage before the 12/15 deadline </s> - Chuck Schumer

Tweet Results

- Most likely tweet-speaker pair: Donald Trump: “Make America Great Again”
- Over 10,000,000,000 times more likely than any other tested combination
- Partially due to shortness of this tweet
- Demonstrates unprecedented use of Twitter by Trump



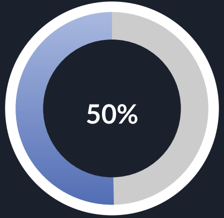
speaker	quote_1	quote_2	quote_3	quote_4	quote_5	quote_6	quote_7	quote_8	quote_9	quote_10
Barack Obama	1.66E-32	1.58E-33	3.30E-106	8.80E-108	1.00E-60	7.47E-161	9.53E-230	1.02E-169	3.80E-219	1.20E-130
Donald Trump	1.19E-25	5.56E-05	1.08E-97	9.22E-93	3.23E-64	7.89E-171	5.72E-231	4.72E-164	3.31E-184	1.93E-134
Ted Cruz	1.13E-33	1.03E-43	6.65E-88	1.09E-107	1.03E-63	3.52E-214	3.40E-240	1.07E-178	3.94E-207	2.56E-185
Bernie Sanders	1.98E-16	3.34E-48	4.15E-117	2.19E-73	6.27E-54	3.37E-172	5.39E-226	2.95E-182	2.19E-196	1.13E-154
Tim Kaine	1.12E-27	5.72E-27	1.38E-127	1.56E-98	5.42E-64	7.88E-162	8.07E-247	7.67E-167	1.37E-206	7.70E-147
Elizabeth Warren	6.18E-36	2.93E-34	1.10E-126	2.31E-108	1.00E-60	5.75E-178	1.08E-235	3.98E-156	3.93E-200	7.82E-139
Lisa Murkowski	1.27E-23	3.06E-44	8.16E-126	4.75E-113	6.12E-64	6.30E-168	7.97E-222	1.79E-163	5.74E-229	1.22E-140
Marco Rubio	4.39E-33	1.04E-44	9.96E-107	2.84E-118	1.00E-60	8.55E-249	1.29E-239	4.61E-118	1.60E-198	5.11E-167
Mike Lee	1.37E-27	6.20E-35	1.30E-124	2.12E-94	5.58E-54	1.84E-179	1.32E-203	2.19E-158	2.80E-183	3.76E-151
Chuck Schumer	9.07E-27	3.62E-37	1.53E-117	3.13E-101	7.82E-54	2.19E-176	8.03E-238	2.23E-184	9.58E-190	8.92E-163

Key

Correctly Identified

Incorrectly Identified

Probilities in each column are sorted from largest (green) to smallest (red)



Tweet Model Accuracy

Discussion

- Model performs best on tweets
- Suggest more formality in speeches - generality in language used
- While words might change over time POS tags remain relatively constant
- Compared to other (published) attempts
    - Some have produced 70 -95% accuracy rate on speaker identification
        - Neural Networks
        - Support Vector Machines
    - Basyesian Inference used in 1964 to attribute 12 of Federalist Papers to John Madison
- Low accuracy due to simplicity of bigram model
- Improvements
    - Use more data- Fewer unseen bigrams
    - Expand to trigram model
    - Explore other types of smoothing: Katz's-Backoff or add-lambda

Works Cited

- Gale, W. and G. Sampson. Good Turing Estimation Without Tears. Journal of Quantitative Linguistics, vol. 2, 217-237, 1995.
- Japi, A. Mimicking Writing Style With Markov Chains. The Sopranos, Silicon Valley, and Summer Afternoons. http://aakashjapi.com/mimicking-writing-style-with-markov-chains/
- Mosteller, F. and D. L. Wallace. Inference and Disputed Authorship: The Federalist. Reading, MA., 1964.
- Zheng, R., Li, J., Chen, H. and Huang, Z. A framework for authorship identification of online messages:
- Writing-style features and classification techniques. J. Am. Soc. Inf. Sci., 57: 378–393, 2006.