# Identifying Speeches of US Presidents

W266: Natural Language Processing with Deep Learning
Summer 2017

Thong Bui & Jason Vantomme

# Introduction

- **Motivation**: How does one verify authenticity of statements, policies or quotes in an era of “fake news”, click-bait headlines and articles, misquotes and plagiarism.
- **Challenge**: Identify the US President who spoke a given text.
- **Approach**: Implement a character-based model used in the PAN 2015 Author Identification task and compare its performance against other standard models.

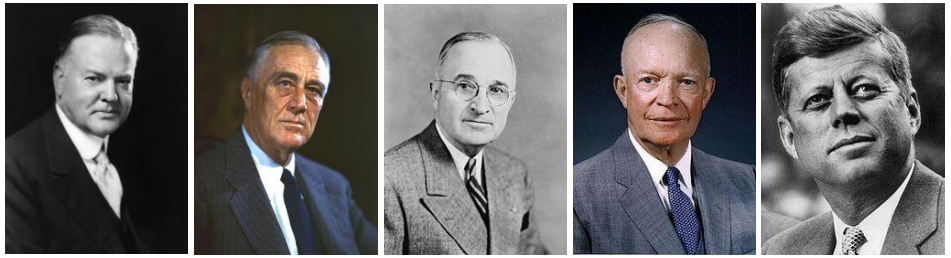
# Approach

Bagnall, Douglas. “Author Identification Using Multi-Headed Recurrent Neural Network.” In *CLEF 2015 Labs and Workshops, Notebook Papers; CEUR Workshop Proceedings*. September 8-11, 2015, 2015. http://ceur-ws.org/Vol-1391/150-CR.pdf.

- Novel approach successful in the PAN 2015 Author Identification task @ the CLEF 2015 conference
- Specific steps for text-preprocessing
- Implementation of “ReSQRT” optimizer:
- Multiple softmax output; one per author

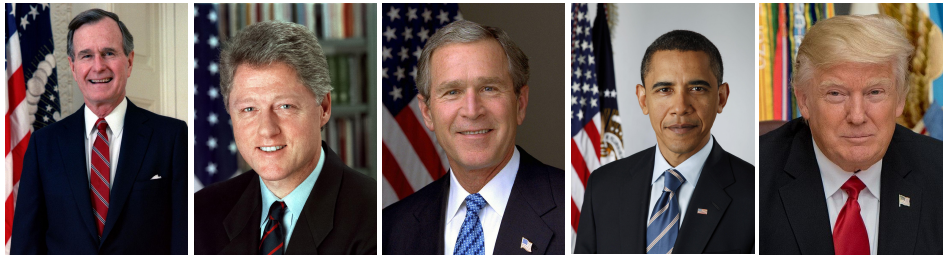
$$f(x) = \begin{cases} \sqrt{x+1} - 1 \text{ if } x \geq 0 \\ 0 \text{ otherwise.} \end{cases}$$


Hoover
Roosevelt
Truman
Eisenhower
Kennedy

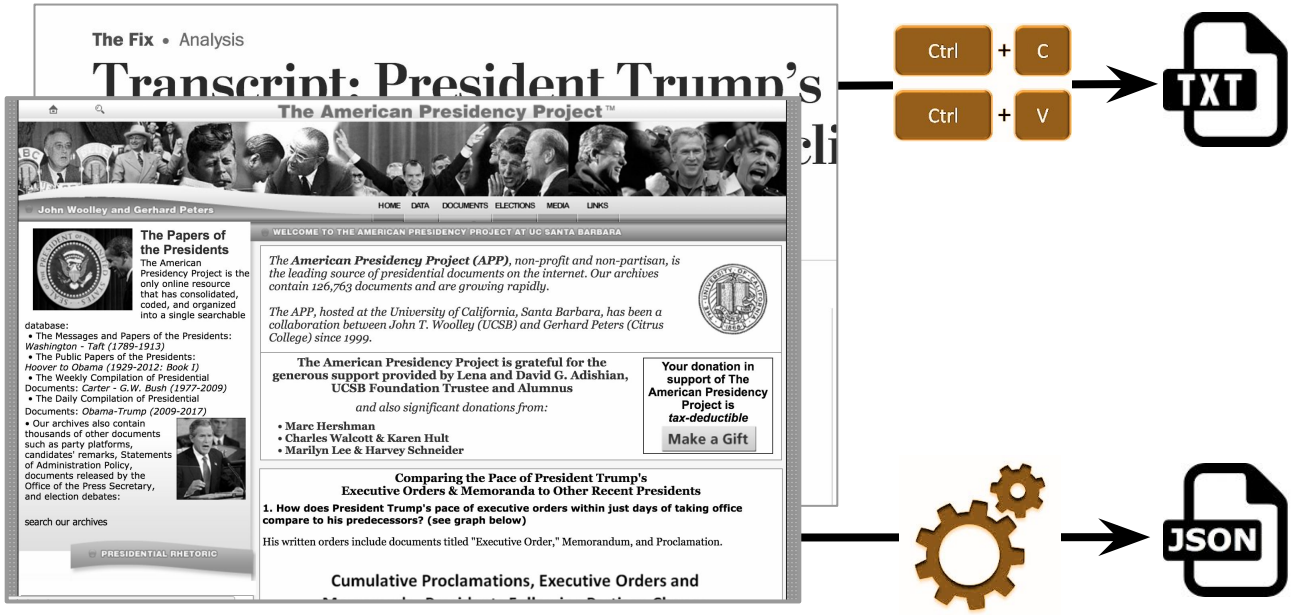


Johnson
Nixon
Ford
Carter
Reagan

G Bush
Clinton
GW Bush
Obama
Trump



# Sourcing Data



# Text Preprocessing

- Extract 15 US Presidents' speeches (primarily press conferences) since 1929
- Clean up dirty data and map inputs to Presidents
- Word-based models:
  - Split speeches into sentences
  - Tokenize the words for each sentence
- Character-based models:
  - Combine speeches into single text
  - Apply text scrubbing
  - Split text into fixed-length character sequences
  - One-hot encode character sequences

# Data Statistics

How many speeches per president?

| | | |
|---|---|---|
| 0 | : Barack Obama | 148 |
| 1 | : Donald J. Trump | 22 |
| 2 | : Dwight D. Eisenhower | 194 |
| 3 | : Franklin D. Roosevelt | 224 |
| 4 | : George Bush | 98 |
| 5 | : George W. Bush | 56 |
| 6 | : Gerald R. Ford | 40 |
| 7 | : Harry S. Truman | 302 |
| 8 | : Herbert Hoover | 268 |
| 9 | : Jimmy Carter | 60 |
| 10 | : John F. Kennedy | 64 |
| 11 | : Lyndon B. Johnson | 135 |
| 12 | : Richard Nixon | 41 |
| 13 | : Ronald Reagan | 48 |
| 14 | : William J. Clinton | 64 |

Approximately many words of text per president?

| | | |
|---|---|---|
| 0 | : Barack Obama | 860977 |
| 1 | : Donald J. Trump | 101328 |
| 2 | : Dwight D. Eisenhower | 567639 |
| 3 | : Franklin D. Roosevelt | 323789 |
| 4 | : George Bush | 346032 |
| 5 | : George W. Bush | 323772 |
| 6 | : Gerald R. Ford | 124695 |
| 7 | : Harry S. Truman | 368810 |
| 8 | : Herbert Hoover | 138249 |
| 9 | : Jimmy Carter | 224639 |
| 10 | : John F. Kennedy | 239684 |
| 11 | : Lyndon B. Johnson | 417045 |
| 12 | : Richard Nixon | 177545 |
| 13 | : Ronald Reagan | 184895 |
| 14 | : William J. Clinton | 327027 |

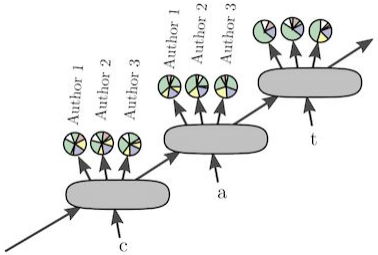
How many sentences of text per president?

| | | |
|---|---|---|
| 0 | : Barack Obama | 41919 |
| 1 | : Donald J. Trump | 8284 |
| 2 | : Dwight D. Eisenhower | 24607 |
| 3 | : Franklin D. Roosevelt | 19223 |
| 4 | : George Bush | 21279 |
| 5 | : George W. Bush | 20326 |
| 6 | : Gerald R. Ford | 6053 |
| 7 | : Harry S. Truman | 29432 |
| 8 | : Herbert Hoover | 5899 |
| 9 | : Jimmy Carter | 10752 |
| 10 | : John F. Kennedy | 10605 |
| 11 | : Lyndon B. Johnson | 23118 |
| 12 | : Richard Nixon | 7357 |
| 13 | : Ronald Reagan | 9120 |
| 14 | : William J. Clinton | 16222 |

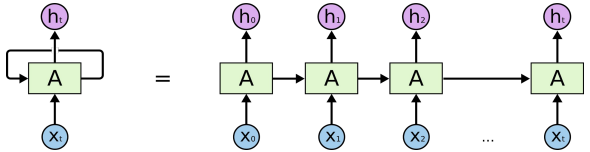
How many characters of text per president?

| | | |
|---|---|---|
| 0 | : Barack Obama | 4862045 |
| 1 | : Donald J. Trump | 554978 |
| 2 | : Dwight D. Eisenhower | 3084215 |
| 3 | : Franklin D. Roosevelt | 1743457 |
| 4 | : George Bush | 1896536 |
| 5 | : George W. Bush | 1791014 |
| 6 | : Gerald R. Ford | 687272 |
| 7 | : Harry S. Truman | 2001118 |
| 8 | : Herbert Hoover | 786184 |
| 9 | : Jimmy Carter | 1272152 |
| 10 | : John F. Kennedy | 1344012 |
| 11 | : Lyndon B. Johnson | 2288606 |
| 12 | : Richard Nixon | 972275 |
| 13 | : Ronald Reagan | 1010291 |
| 14 | : William J. Clinton | 1784166 |

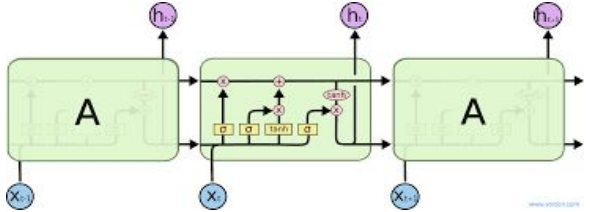
# Models




Character-level RNN
w/ multi-softmax

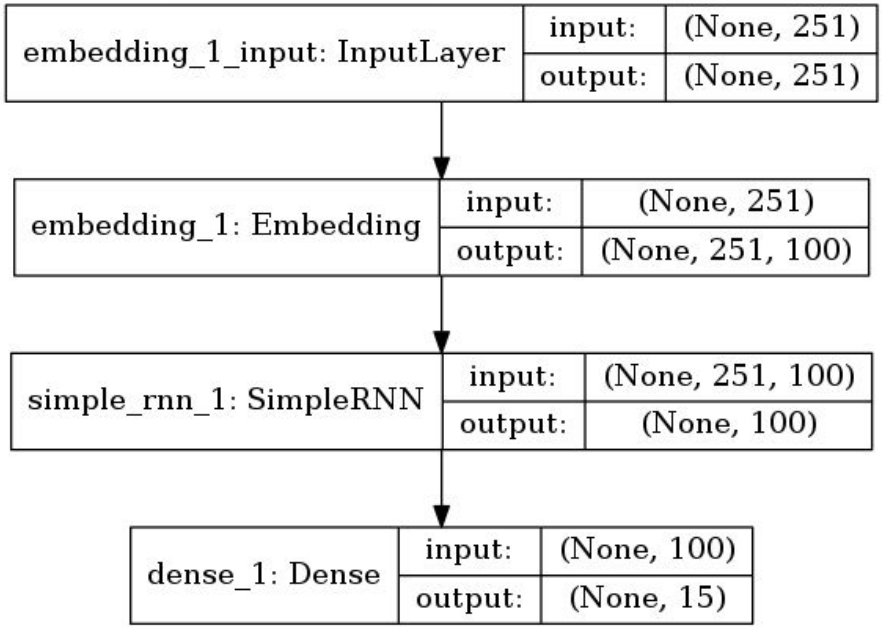



Word-level RNN
Character-level RNN

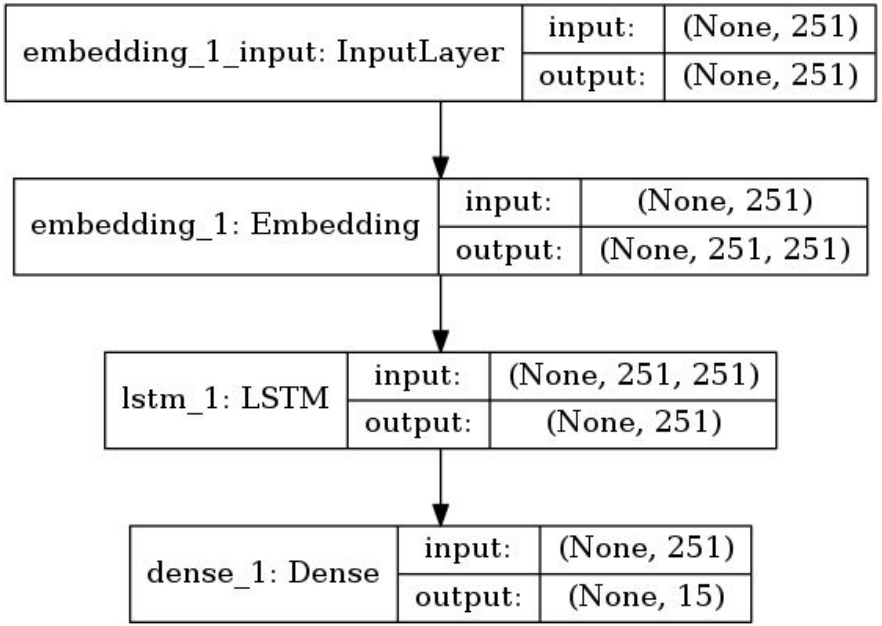



Word-level LSTM

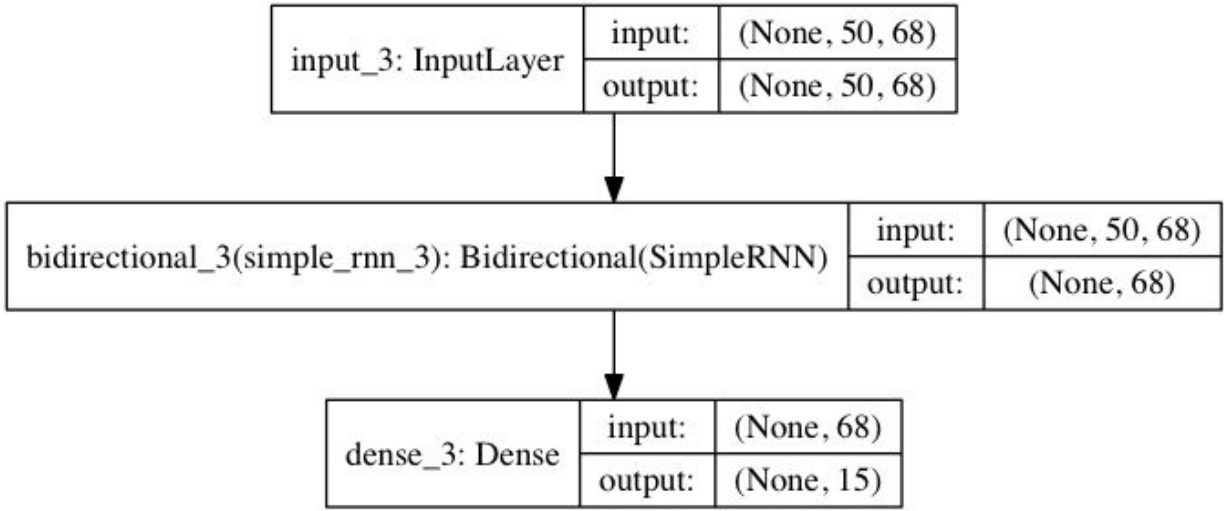
# Word-based RNN



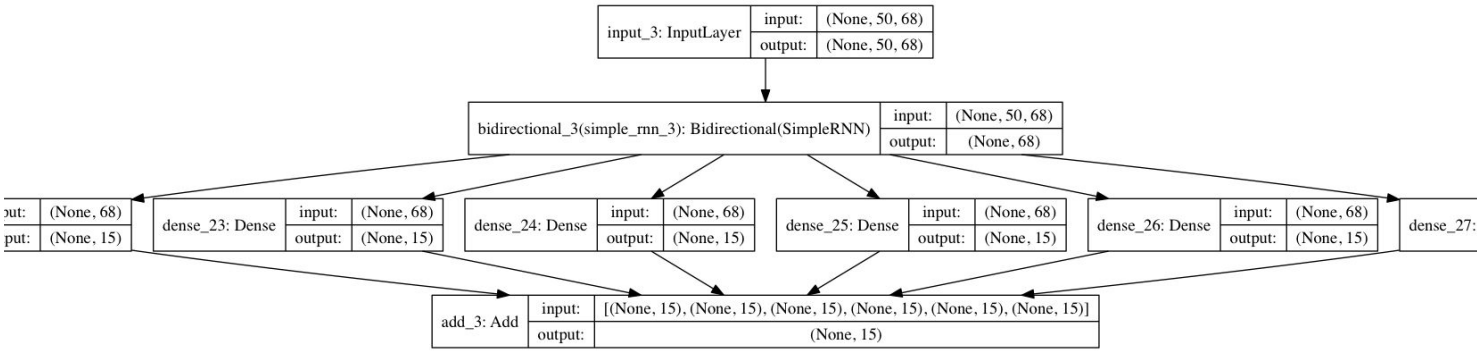
# Word-based LSTM



# Character Model: Simple RNN

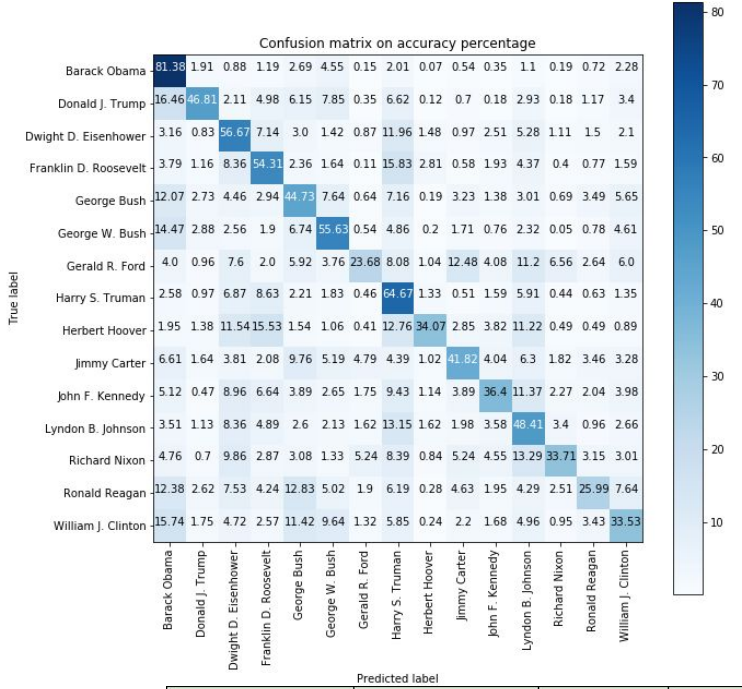


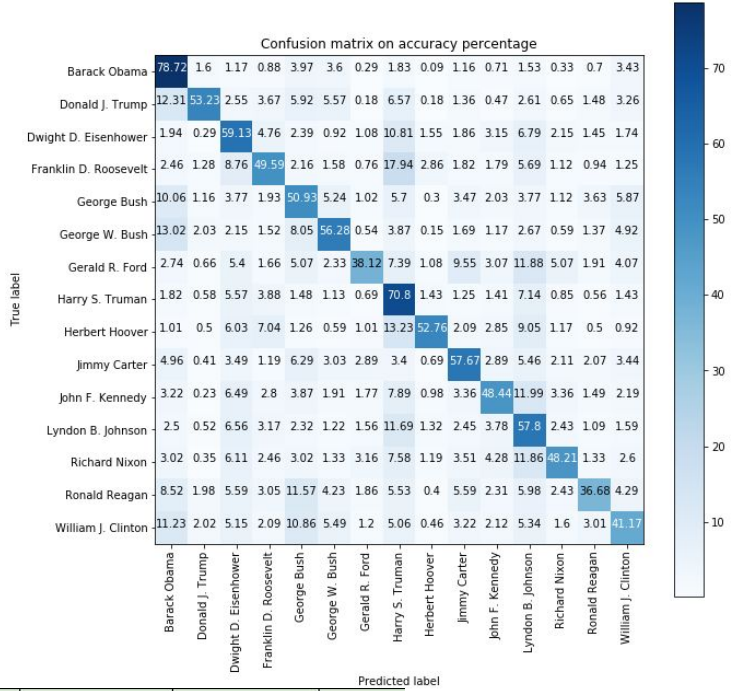
# Character Model: Multi-Softmax




*(diagram only shows 6 softmax layers to fit on screen; actual is 1 per president)*

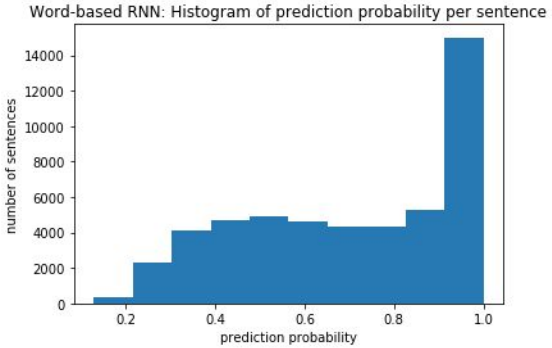
# Results: Word-based RNN vs LSTM

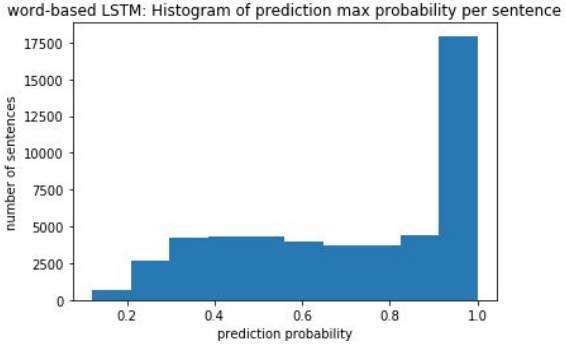


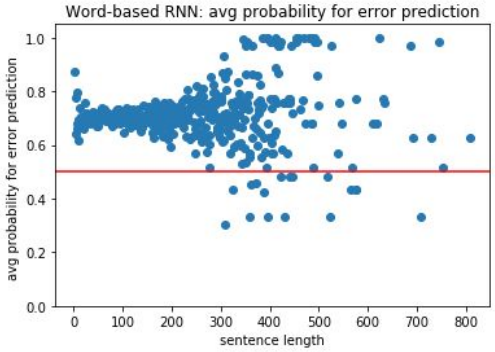



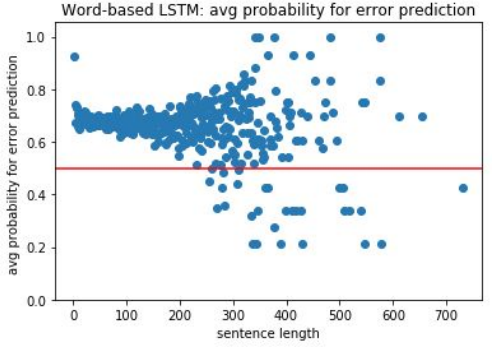

| | Training hours | Train loss | Test loss | Train accuracy | Test accuracy | AUC |
|---|---|---|---|---|---|---|
| Word-based RNN | 40 epochs, 3hrs | 0.6221 | 1.876401 | 0.8165 | 0.52046 | 0.8754 |
| Word-based LTSM | 30 epochs, 20hrs | 0.7485 | 1.505552 | 0.7591 | 0.58316 | 0.9116 |

# Error Analysis: Word-based RNN vs LSTM



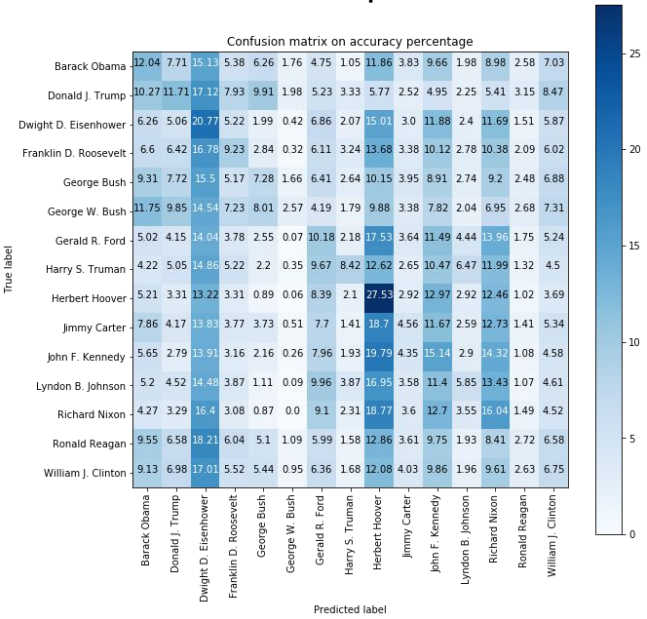



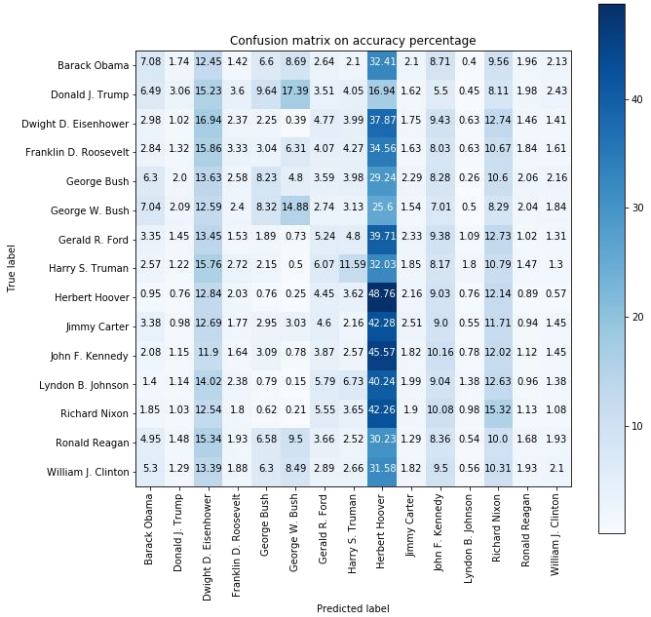



# Results: Character-based Simple & Multi

## Simple

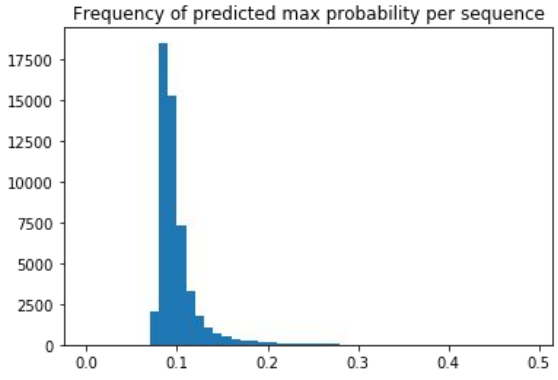



## Multi

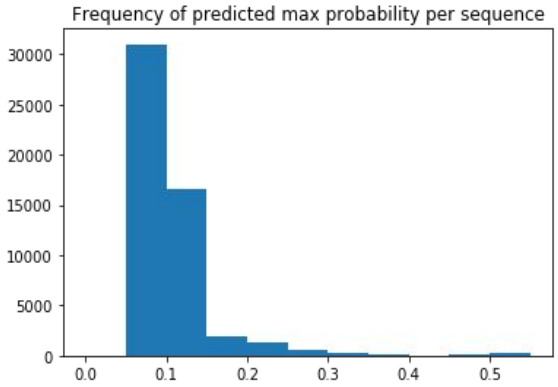


# Error Analysis: Character-based

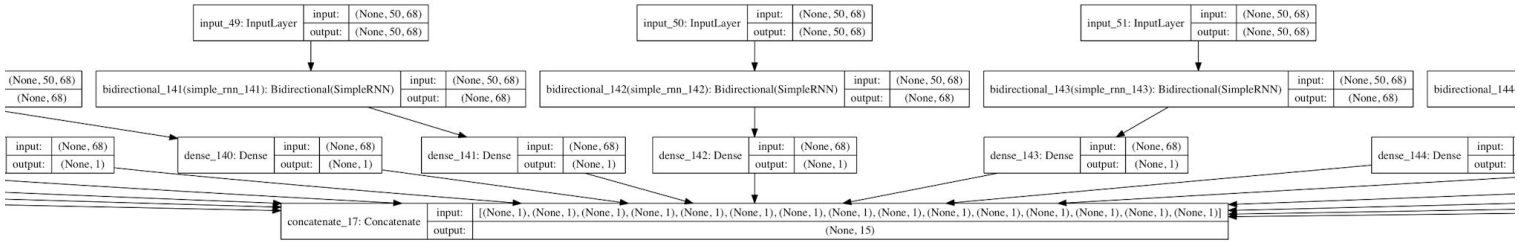
Simple



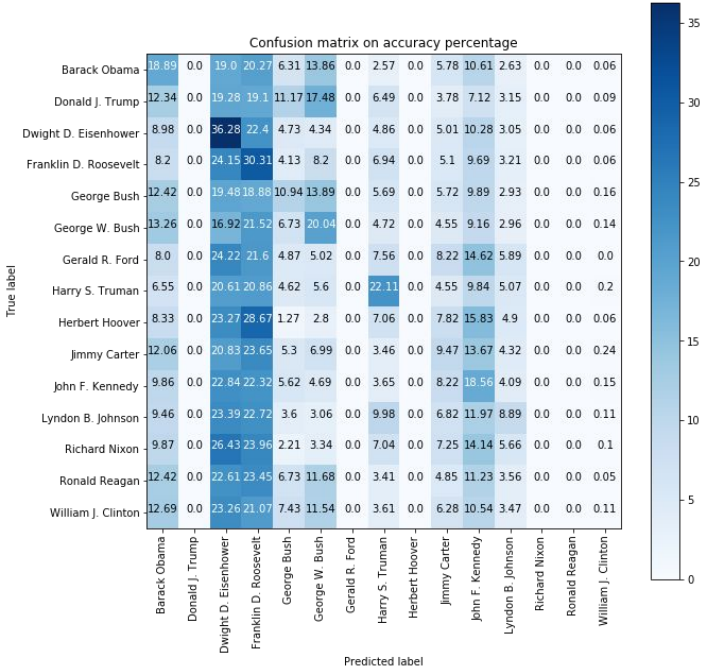

Multi



# Results: Char-based Multi-RNN (take 2)



# Results: Char-based Multi-RNN (take 2)



# Challenges

- Source text size differences presented tuning challenges.
- Target character-based approach was light on details in paper and  available code was written from scratch in C. Most implementation examples found for char-based models are for generative purposes.
- Activation function not stable (loss -> NaN).
- Implementation and tuning for character-model more complex than word-based model.  Non-zero chance our implementation is flawed.

# Conclusions

- Both word-based models significantly outperformed character-based models on current data set & context.
- Simple character-based architectures may not be effective; high-performing models are vastly more complicated (e.g., including CNNs, highway networks, RNN+LSTM layers, etc.).
- In this context, fifteen classes may be too many classes for a simple character-model to effectively differentiate.

# QUESTIONS