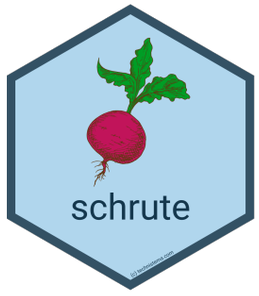
TEXT MINING AND NATURAL LANGUAGE PROCESSING
APPLICATION TO MACHINE LEARNING AND DEEP LEARNING

Leticia Han

Math 533 Fall 2020

BACKGROUND



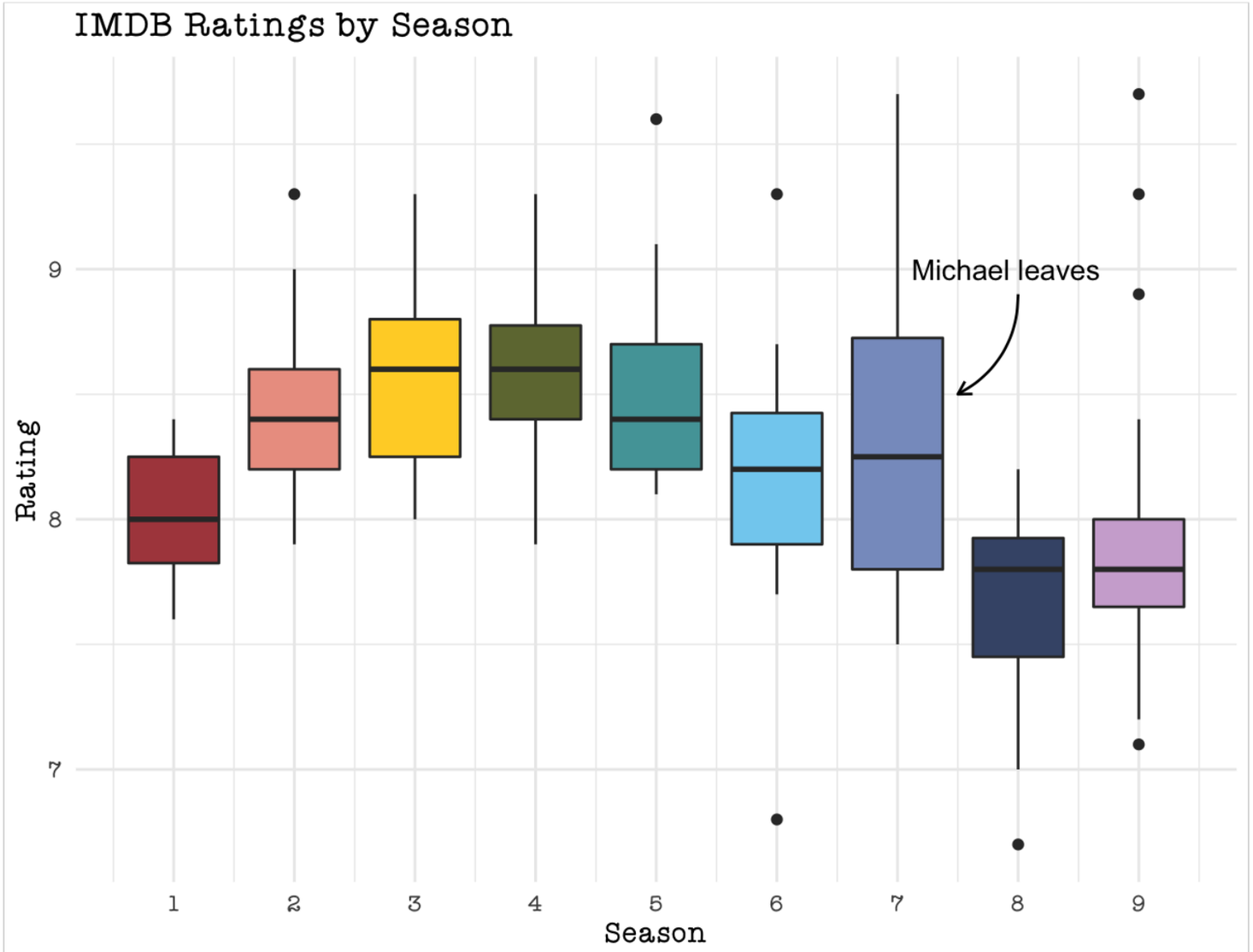
- Workplace mockumentary of employees working at a branch of the fictional Dunder Mifflin Paper Company located in Scranton, PA
- Data: The complete transcript for The Office TV show (US version), available in R, python, Julia
- "One and only one purpose": share the data to encourage exploration and learning from text data (*Brad Lindblad*)

index	season	episode	episode_name	director	writer	character	text	text_w_direction	imdb_rating	total_votes	air_date
1	1	1	Pilot	Ken Kwapis	Ricky Gervais;Stephen Merchant;Greg Daniels	Michael	All right Jim. Your quarterlies look very good. How are things at the library?	All right Jim. Your quarterlies look very good. How are things at the library?	7.6	3706	2005-03-24
2	1	1	Pilot	Ken Kwapis	Ricky Gervais;Stephen Merchant;Greg Daniels	Jim	Oh, I told you. I couldn't close it. So...	Oh, I told you. I couldn't close it. So...	7.6	3706	2005-03-24
3	1	1	Pilot	Ken Kwapis	Ricky Gervais;Stephen Merchant;Greg Daniels	Michael	So you've come to the master for guidance? Is this what you're saying, grasshopper?	So you've come to the master for guidance? Is this what you're saying, grasshopper?	7.6	3706	2005-03-24
4	1	1	Pilot	Ken Kwapis	Ricky Gervais;Stephen Merchant;Greg Daniels	Jim	Actually, you called me in here, but yeah.	Actually, you called me in here, but yeah.	7.6	3706	2005-03-24
5	1	1	Pilot	Ken Kwapis	Ricky Gervais;Stephen Merchant;Greg Daniels	Michael	All right. Well, let me show you how it's done.	All right. Well, let me show you how it's done.	7.6	3706	2005-03-24
55130	9	24	Finale	Ken Kwapis	Greg Daniels	Dwight	No, don't say it. You're fired! You're both fired!	No, don't say it. You're fired! You're both fired!	9.7	7934	2013-05-16

OUTLINE

- Exploratory Data Analysis
- Text Mining
- Sentiment Analysis
- Predict IMDB Rating
- Text Generation

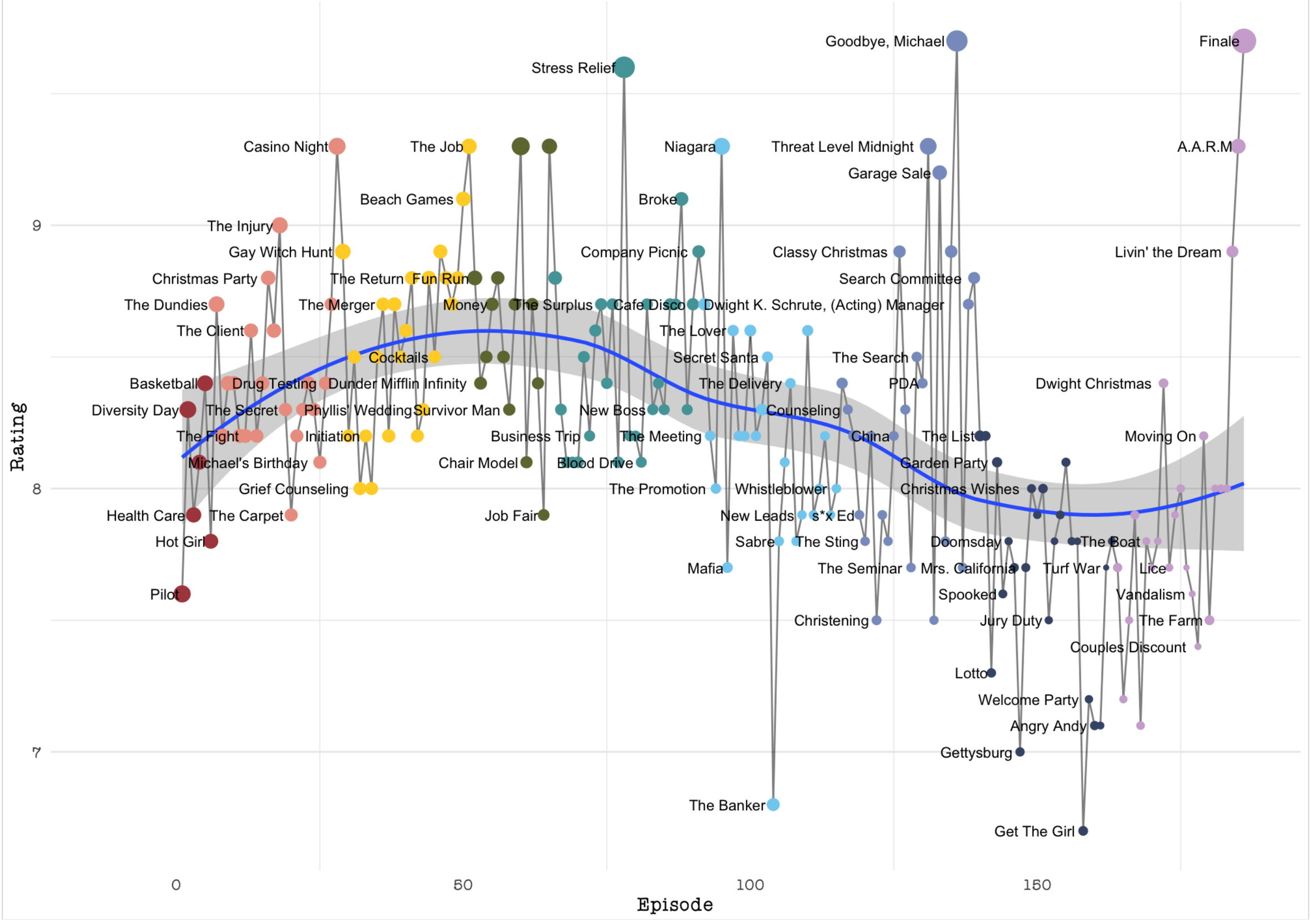
EXPLORATORY DATA ANALYSIS



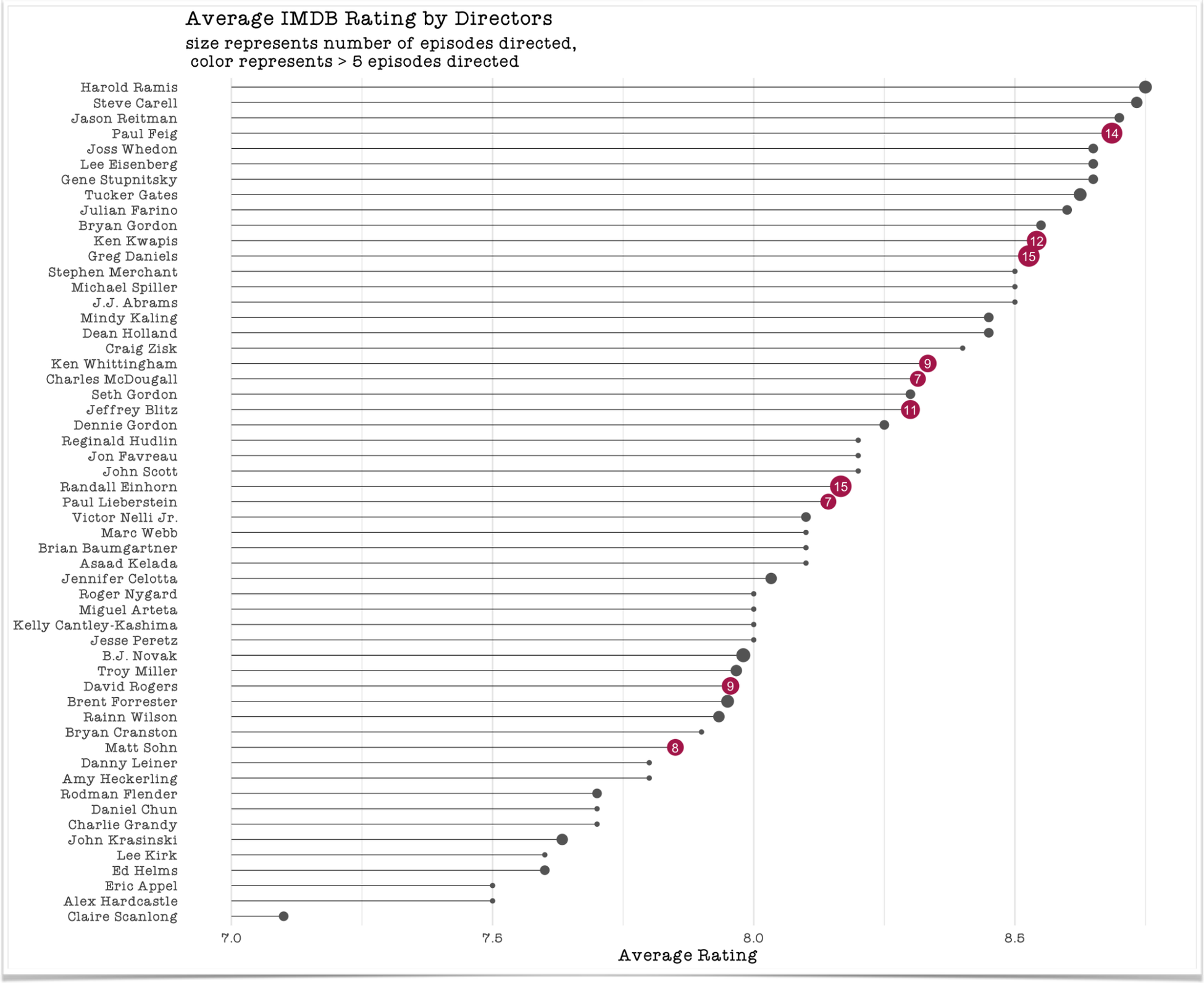
season	number of episodes	average rating
1	6	8.0
2	22	8.4
3	23	8.6
4	14	8.6
5	26	8.5
6	24	8.2
7	24	8.3
8	24	7.7
9	23	8.0

Popularity of The Office

Color represents season, size represents total votes



DIRECTORS OF THE OFFICE



Top 10 Most Involved Directors

director	n	Average Rating
Greg Daniels	15	8.53
Randall Einhorn	15	8.17
Paul Feig	14	8.69
Ken Kwapis	12	8.54
Jeffrey Blitz	11	8.30
David Rogers	9	7.96
Ken Whittingham	9	8.33
Matt Sohn	8	7.85
Charles McDougall	7	8.31
Paul Lieberstein	7	8.14

Top 10 Highest Rated Directors

director	n	Average Rating
Harold Ramis	4	8.75
Steve Carell	3	8.73
Jason Reitman	2	8.70
Paul Feig	14	8.69
Gene Stupnitsky	2	8.65
Joss Whedon	2	8.65
Lee Eisenberg	2	8.65
Tucker Gates	4	8.62
Julian Farino	2	8.60
Bryan Gordon	2	8.55

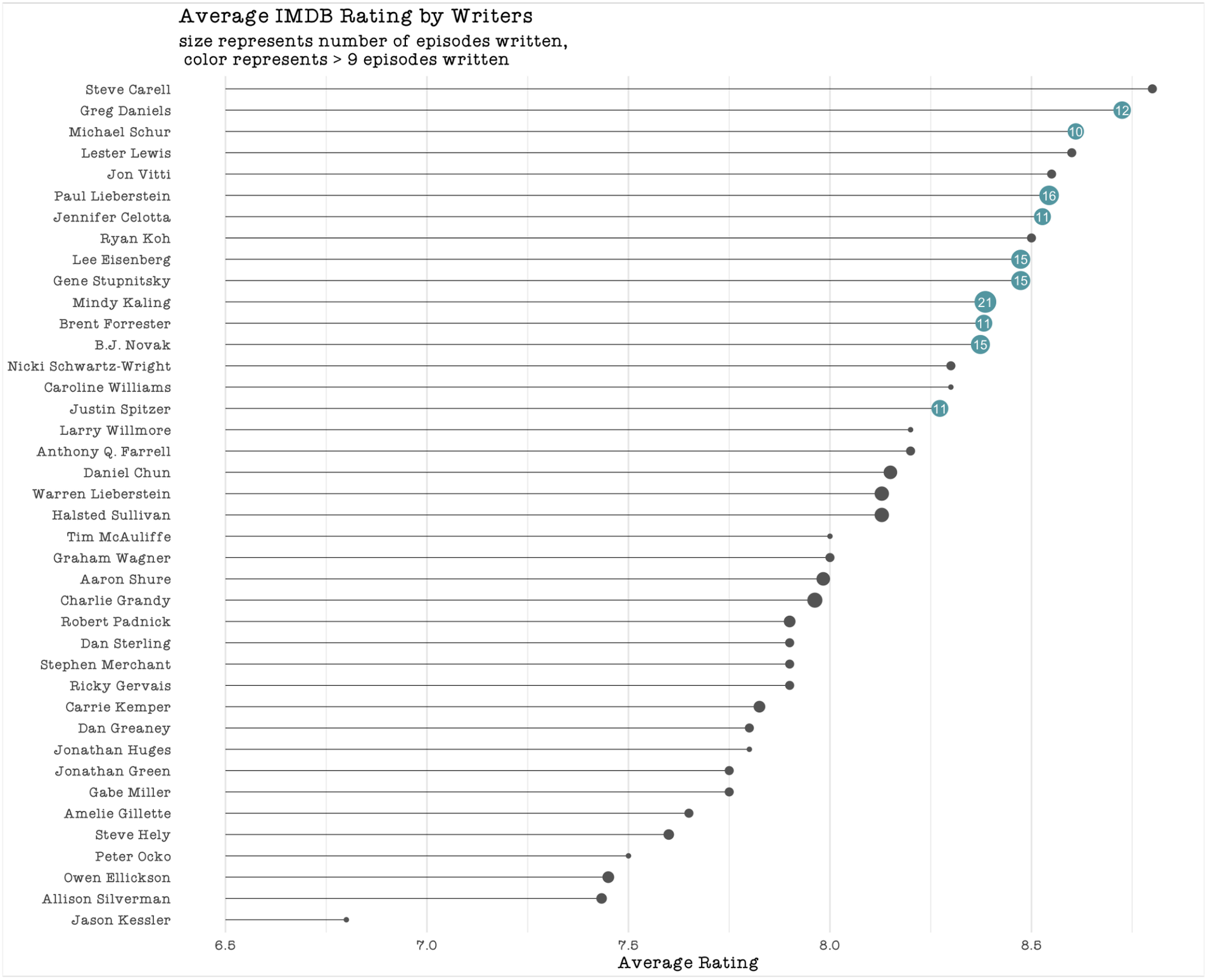
WRITERS OF THE OFFICE

Top 10 Most Involved Writers

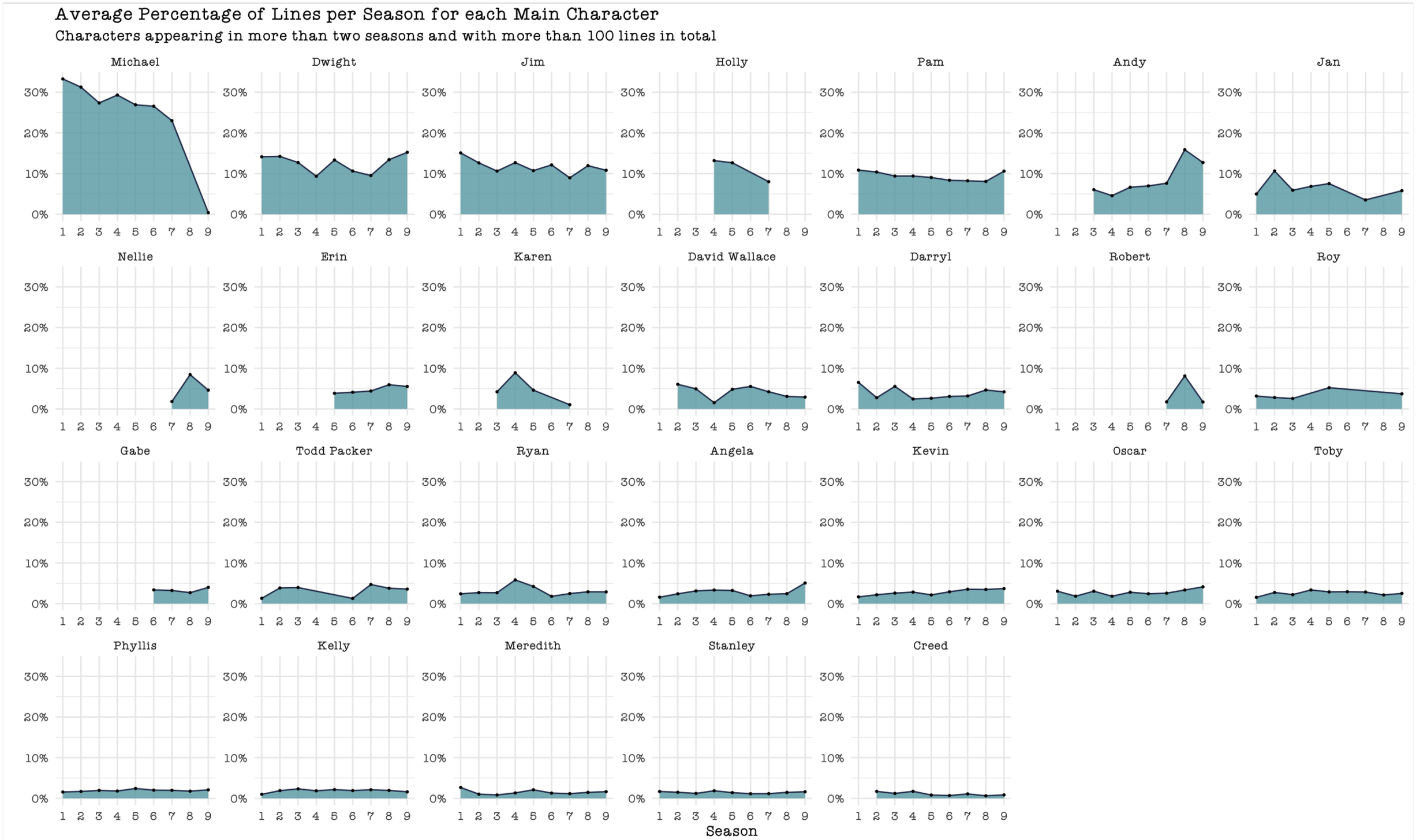
writer	n	Average Rating
Mindy Kaling	21	8.39
Paul Lieberstein	16	8.54
B.J. Novak	15	8.37
Gene Stupnitsky	15	8.47
Lee Eisenberg	15	8.47
Greg Daniels	12	8.72
Brent Forrester	11	8.38
Jennifer Celotta	11	8.53
Justin Spitzer	11	8.27
Michael Schur	10	8.61

Top 10 Highest Rated Writers

writer	n	Average Rating
Steve Carell	2	8.80
Greg Daniels	12	8.72
Michael Schur	10	8.61
Lester Lewis	2	8.60
Jon Vitti	2	8.55
Paul Lieberstein	16	8.54
Jennifer Celotta	11	8.53
Ryan Koh	2	8.50
Gene Stupnitsky	15	8.47
Lee Eisenberg	15	8.47



MAIN CHARACTERS



TEXT AS DATA

Text mining is a way to extract information from unstructured data and represent them in a vector form, which can then be used to create new feature variables for analysis.

- Preprocessing: to normalize data and reduce dimensionality. These steps also depend on the context of the data.
    - Tokenization: bag of words, n-grams
        - One of the simplest methods but at the cost of losing semantic meaning
    - Lowercase vs. case-sensitive
    - Stopword removal
        - "One man's trash is another man's treasure"
    - Punctuations & numbers
    - Stemming or lemmatization
        - Porter stemming algorithm
        - SpaCy lemmatization

- Feature Engineering: to represent text data in some meaningful way as inputs for statistical modeling

  - TF-IDF: (*term frequency-inverse document frequency*) a statistic that measures how important or relevant a word is to a document in a corpus

$$TF_{ij} = \frac{f_{ij}}{n_j} \tag{1}$$

Where f_{ij} is the frequency of term i in document j. n_j is the total number of words in document j.

$$IDF_i = 1 + log(\frac{N}{c_i}) \tag{2}$$

Where N is the total number of documents in the corpus. c_i is the number of documents that contain word i.

$$w_{ij} = TF_{ij} \times IDF_i \tag{3}$$

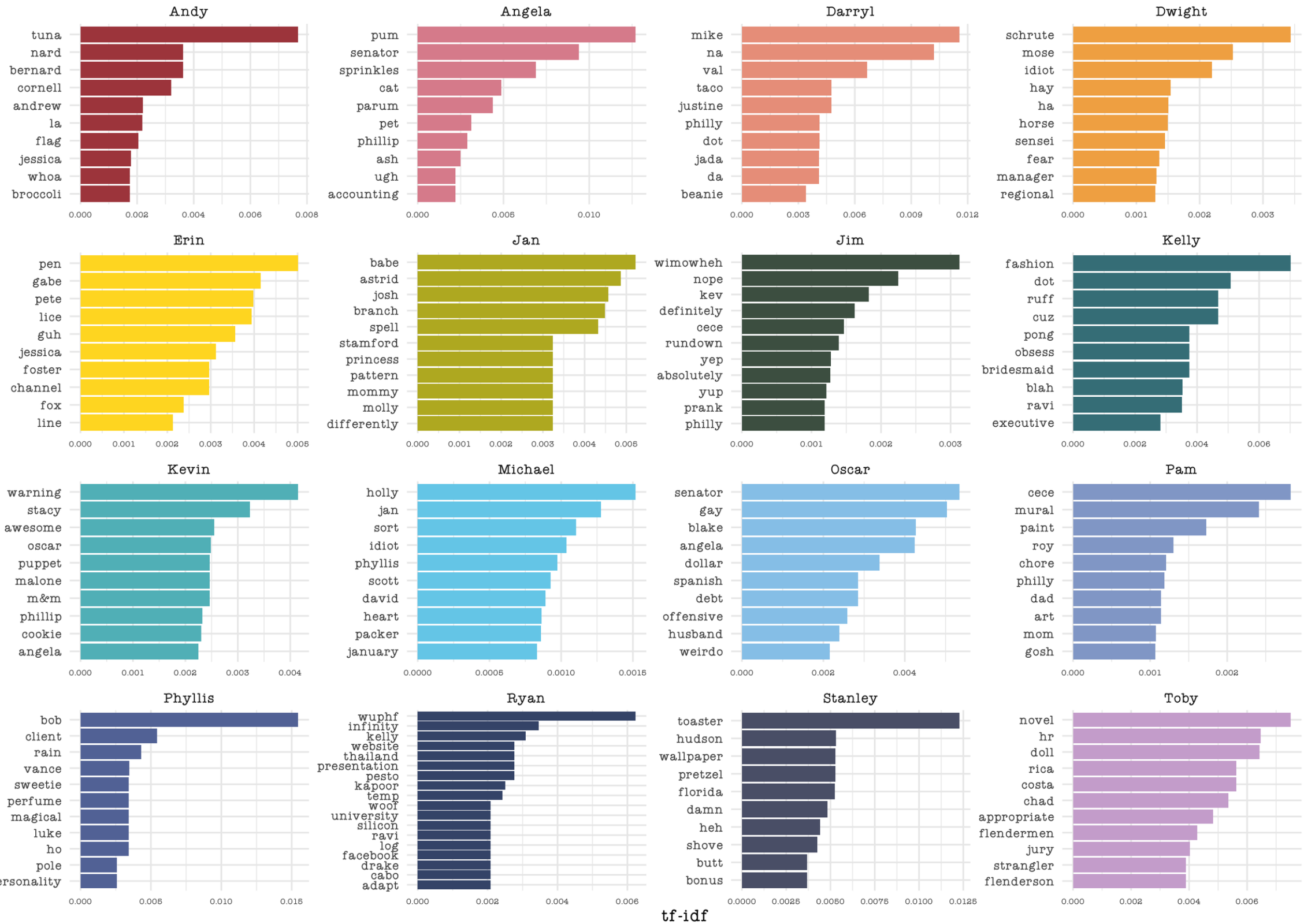
Where w_{ij} is the TF-IDF score of term i in document j.

- count features: words, unique words, characters, unique characters, digits, hashtags, mentions, commas, periods, exclamations, capital letters, first person, third person, prepositions, etc.

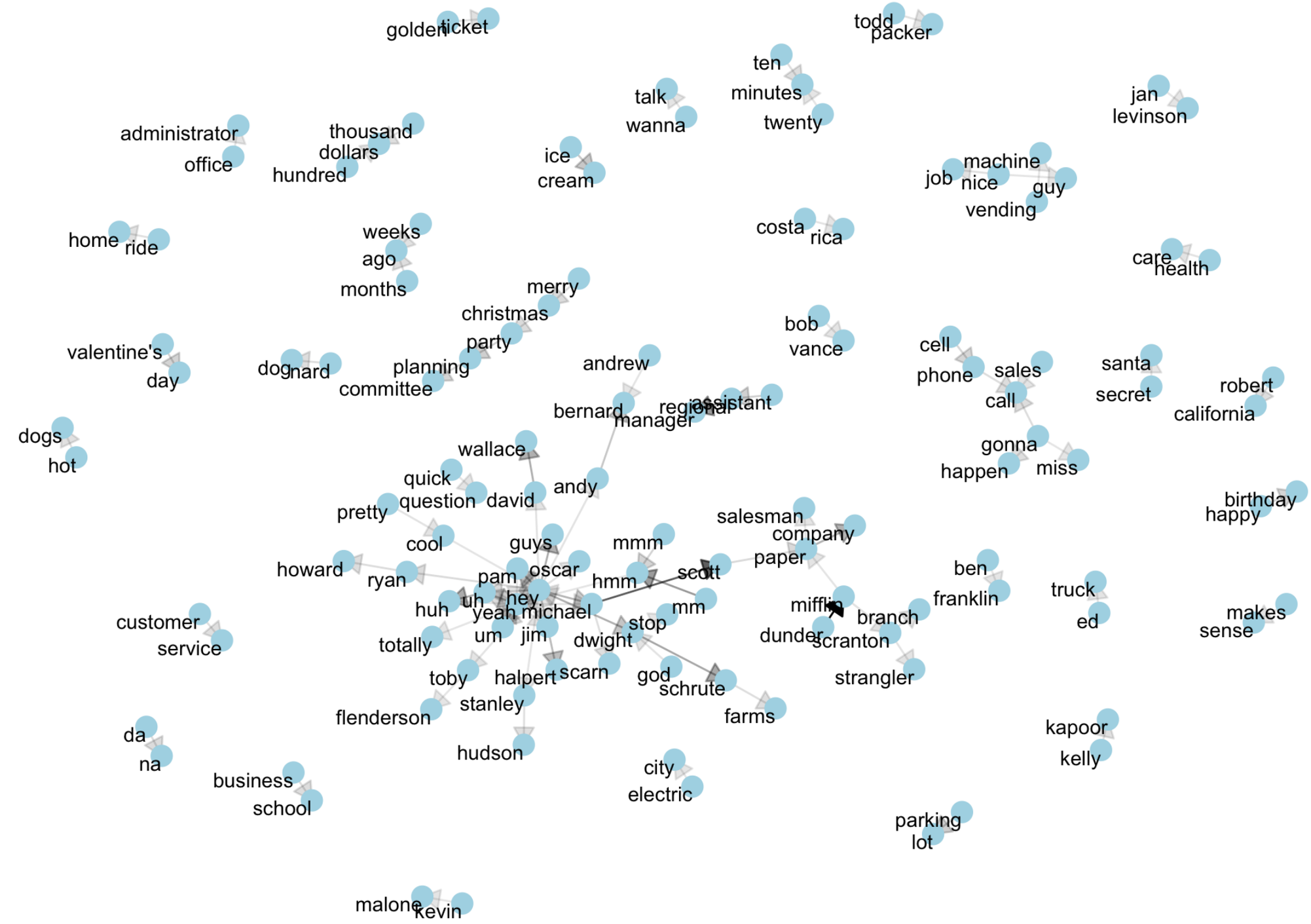
- word associations or word networks

- word embeddings: captures semantic meaning from words' context

Words specific to characters based on tf-idf

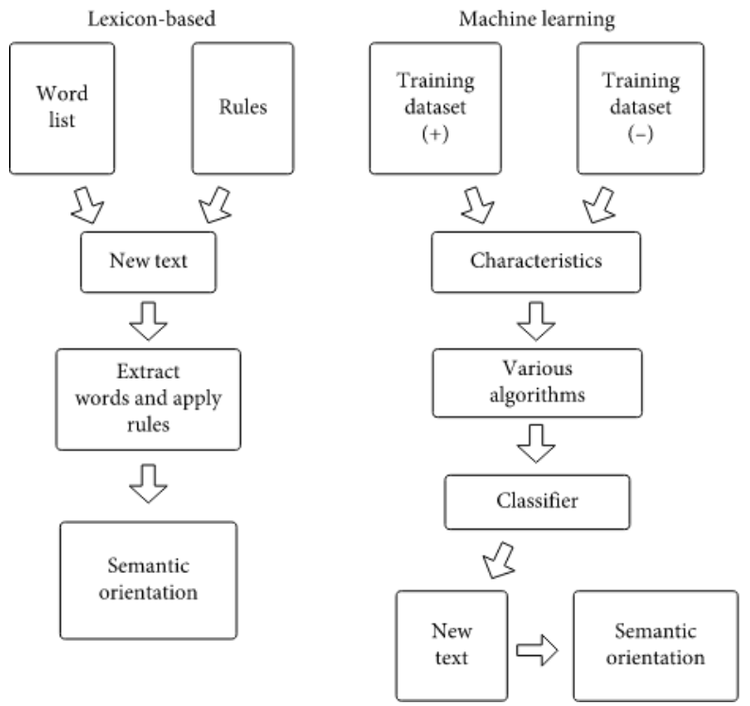


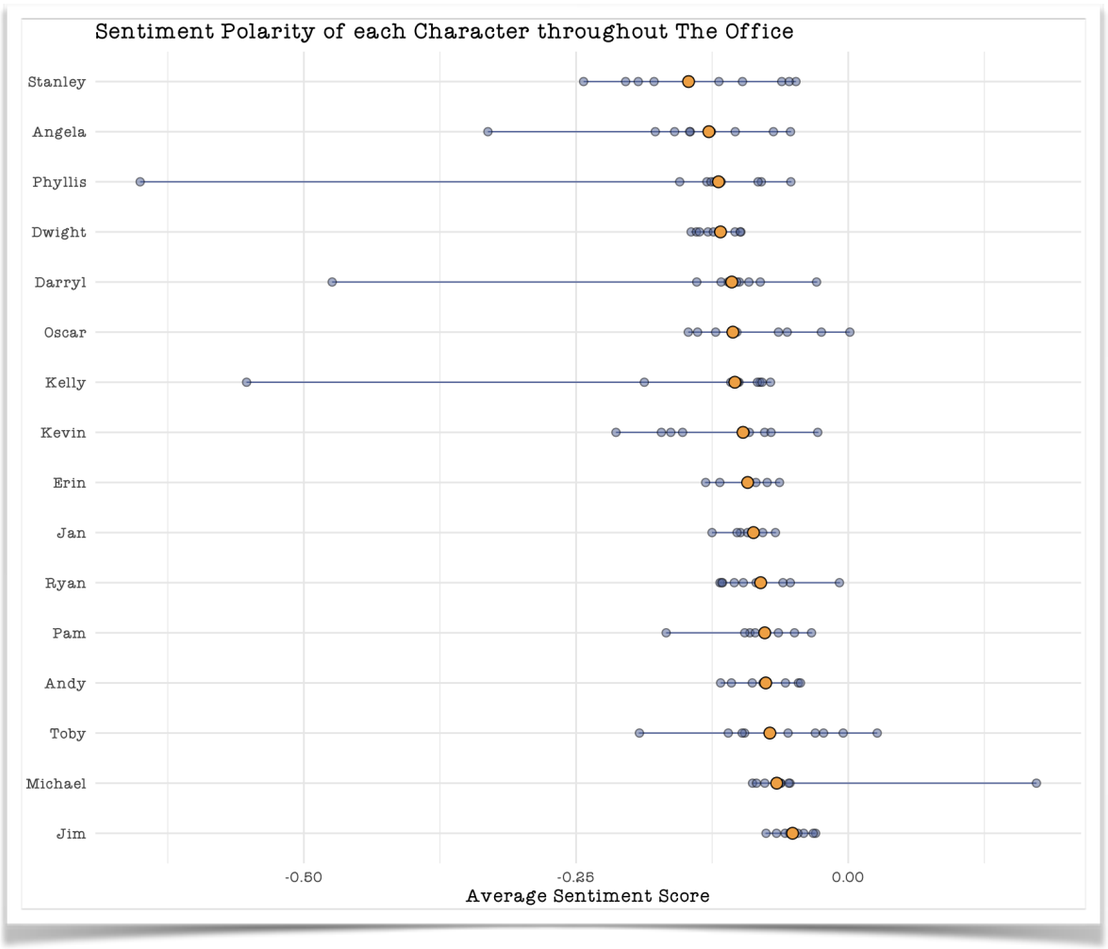
Network of Most Common Bigrams by Main Characters

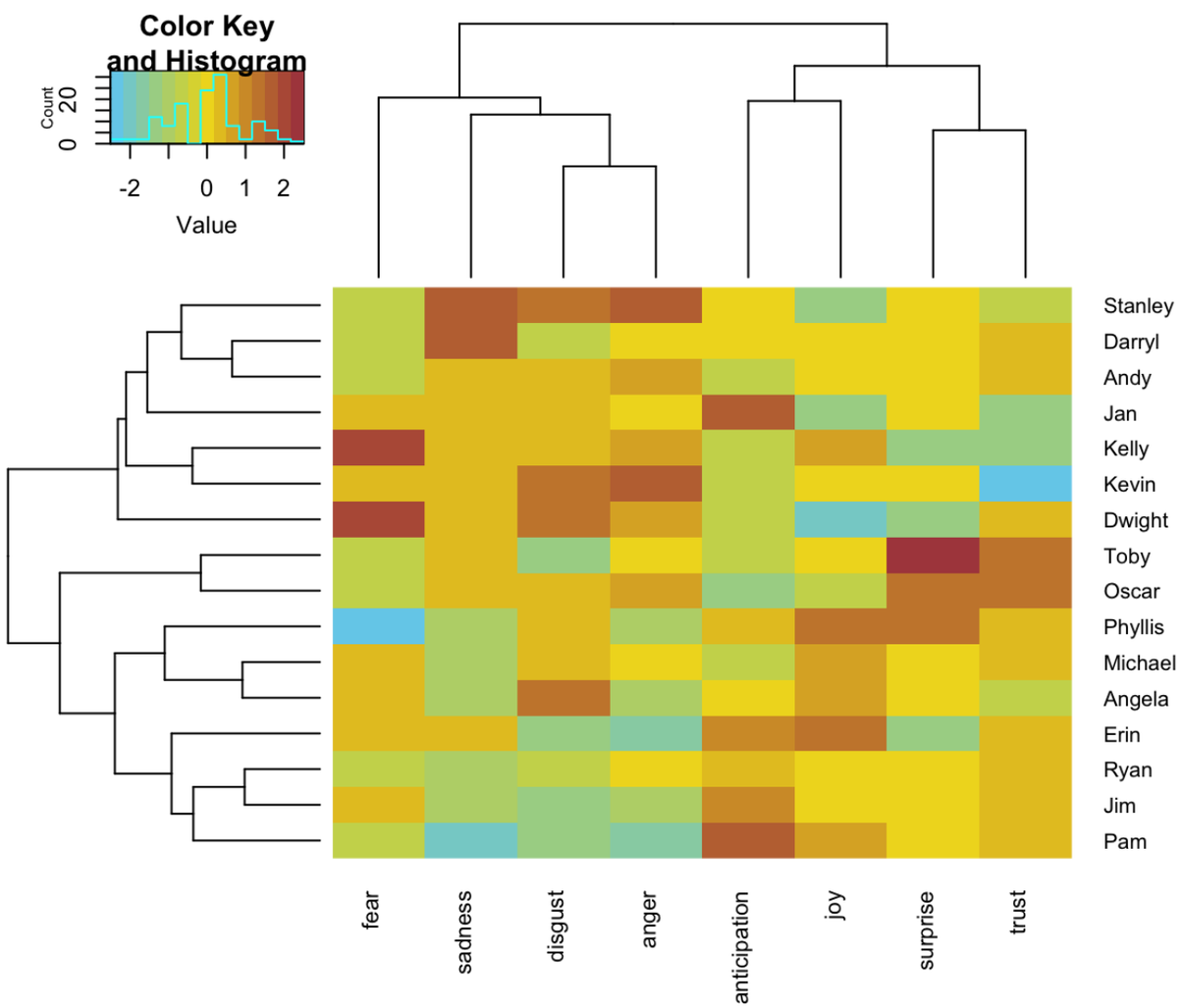


SENTIMENT ANALYSIS

- Machine Learning Approach:
  - Classify text as positive or negative using labeled data
  - Ex: IMDB movie reviews, Yelp reviews, Twitter posts, student evaluations
  - Data must be labeled
- Lexicon-based Approach:
  - Utilize a sentiment word database (lexicon) to identify any sentiment words in the text, assign a score to each of the sentiment words, aggregate the scores to determine the sentiment or polarity
  - Different lexicons available - Stanford coreNLP, AFINN, Bing, NRC, Jockers, Sentiword, emoji, internet slang
  - Different rules and scoring - binary, multi-class, numeric
  - Lexicons are context-specific
    - "I'm too *sick*" vs. "That was *so sick*"





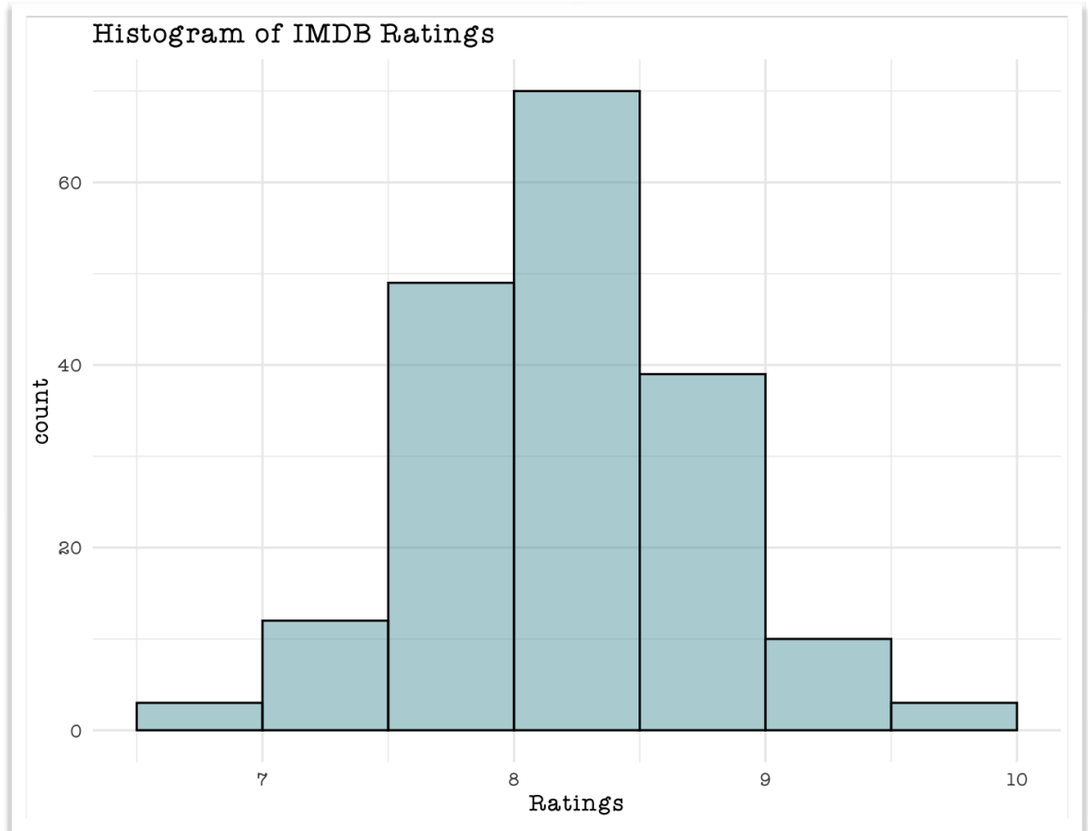


- Sentiword lexicon assigns continuous values (-1, 1)

- NRC lexicon assigns categorical values:
  - binary (positive/negative)
  - multiclass (8 emotions)

PREDICT IMDB RATING

- Total of 186 episodes
- Model: combine text data and non-text data to predict ratings
- Variables:
    - Director
    - Writer
    - Season & episode #
    - # of lines per main character
    - TF-IDF on unigram + bigram + trigram
- LASSO, SVM (regression), XGBoost, knn, RNN, LSTM
- 70/30 split
- Cross-validation on training data to tune hyperparameters based on smallest RMSE



Recurrent Neural Network

- RNN is a variant of neural network and has been widely used in natural language processing among other applications
- It is ideal for sequential data, such as time sequences of events and language
- RNN has feedback connections (closed loops) and comes in varying architectures.
- It allows previous outputs to be used as inputs while having hidden states
- Inputs and outputs can be of any length
    - One-to-one, one-to-many, many-to-one, many-to-many
- Model size does not increase with size of input
- Model takes into account the history of its inputs
- Weights are shared across time
- Computationally slow
- Suffers from the vanishing and exploding gradient phenomena due to the backpropagation algorithm
    - Difficult to capture long term dependencies because of multiplicative gradient that can be exponentially decreasing or increasing with respect to the number of layers

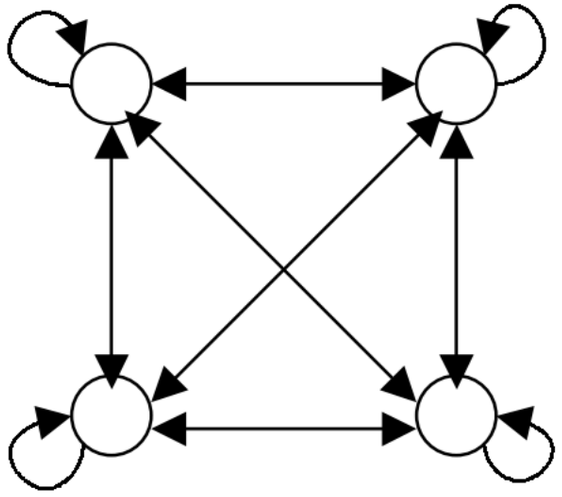


Figure 1. An example of a fully connected recurrent neural network.

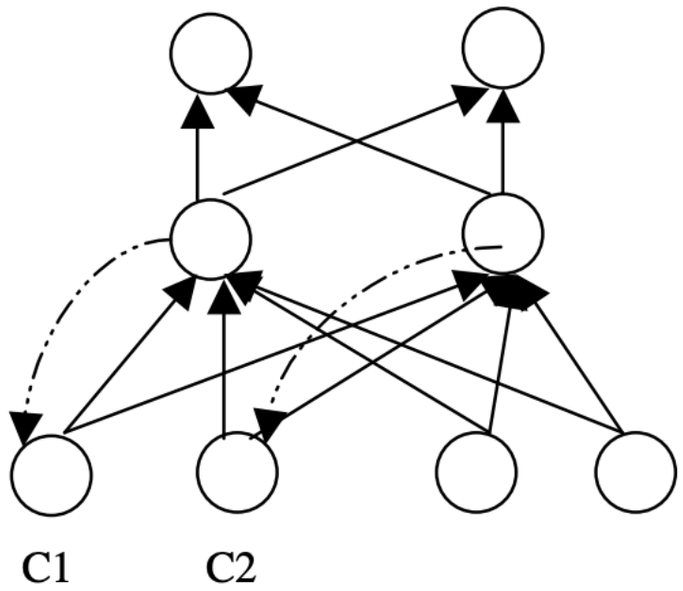
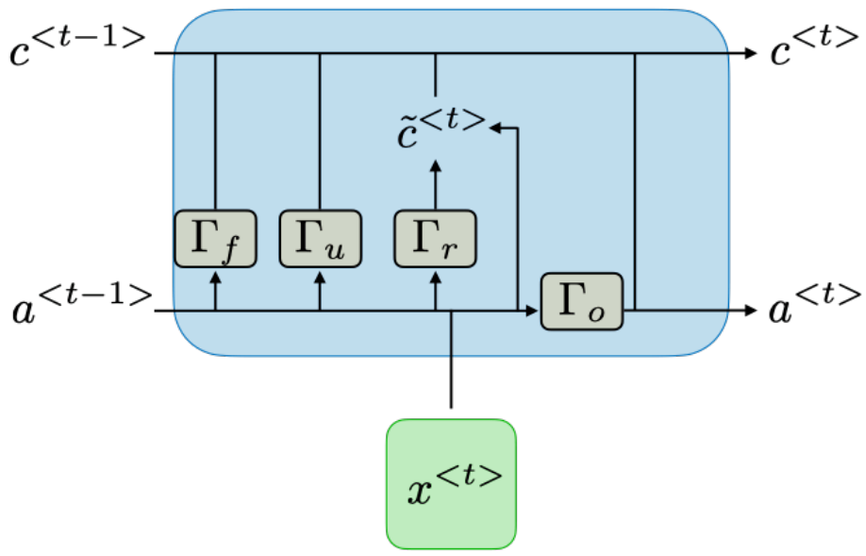


Figure 2. An example of a simple recurrent network.

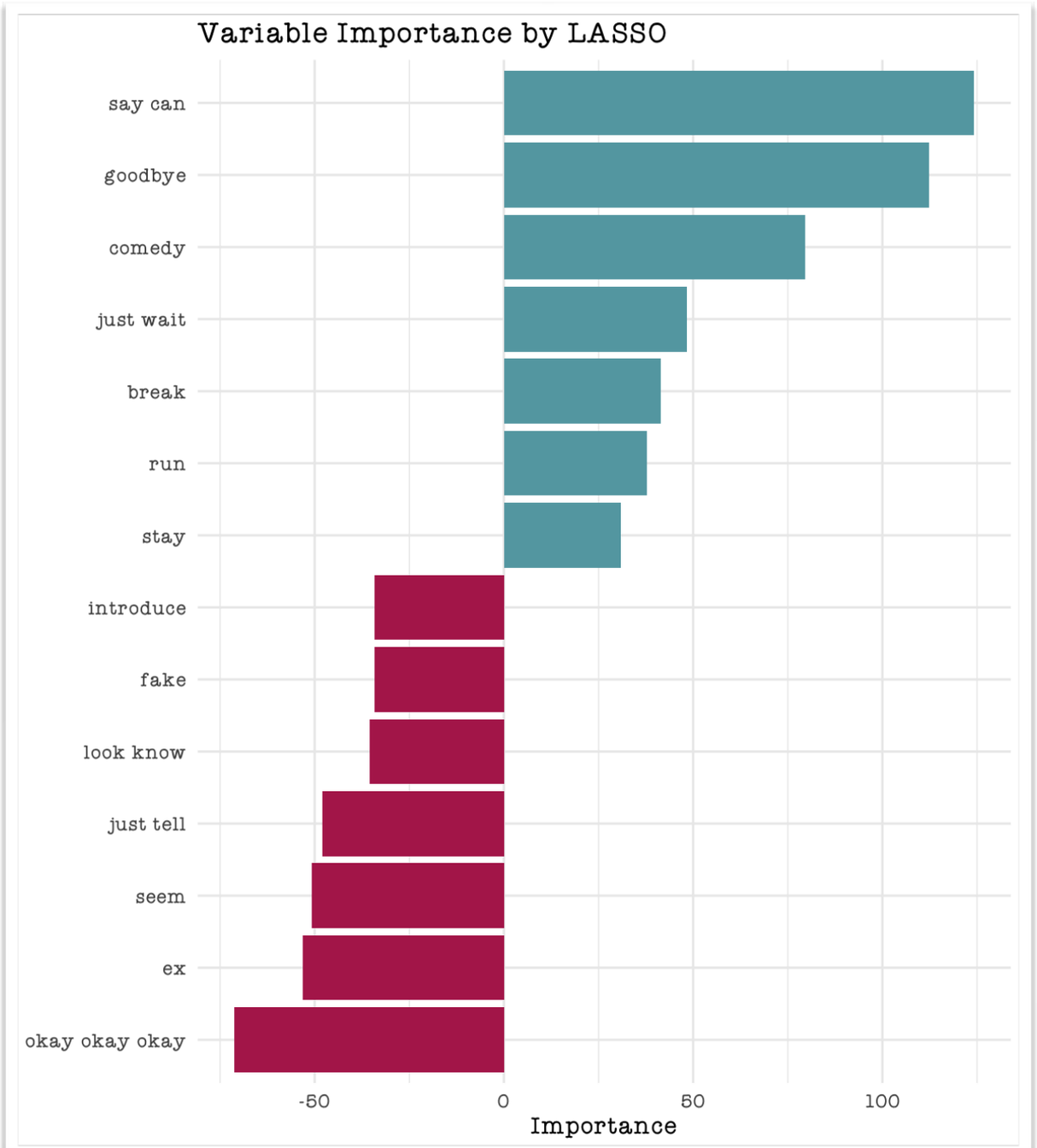
Long Short-Term Memory

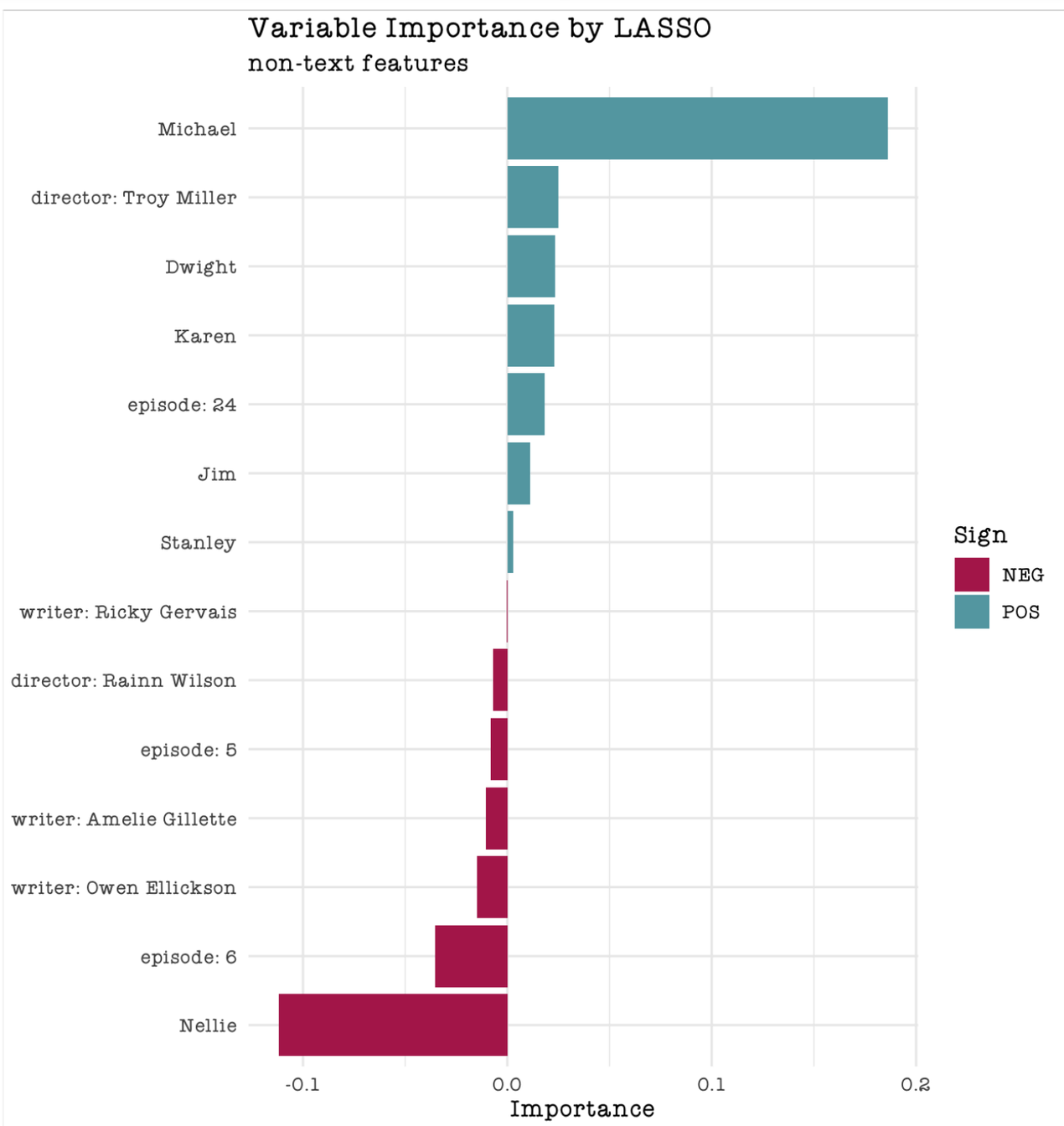
- LSTM addresses the vanishing gradient problem by incorporating specific gates
    - Idea is that it regulates the gates by allowing some new information to flow in and some to flow out which allows it to be capable of handling long term dependencies
- Google uses LSTM for Google Voice and Google Translate (2015; 2016)
- Apple uses LSTM for Siri and quicktype (2016)
- Amazon also used it for Alexa (2016)
- Facebook used it for translations (2017)

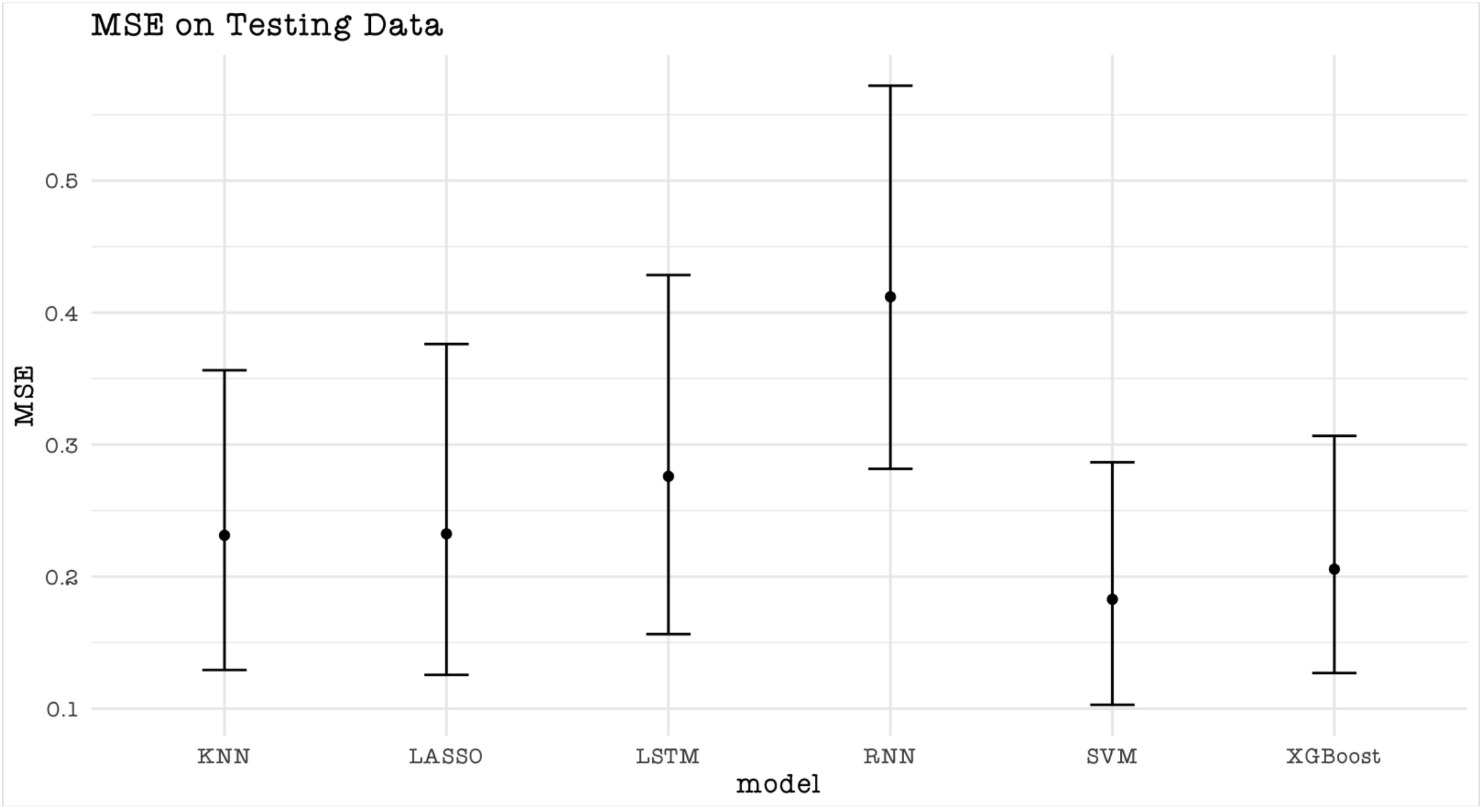
Type of gate	Role	Used in
Update gate Γ_u	How much past should matter now?	GRU, LSTM
Relevance gate Γ_r	Drop previous information?	GRU, LSTM
Forget gate Γ_f	Erase a cell or not?	LSTM
Output gate Γ_o	How much to reveal of a cell?	LSTM



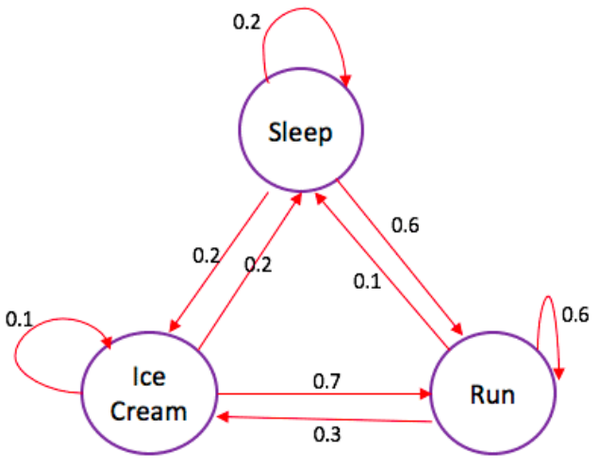
VARIABLE IMPORTANCE





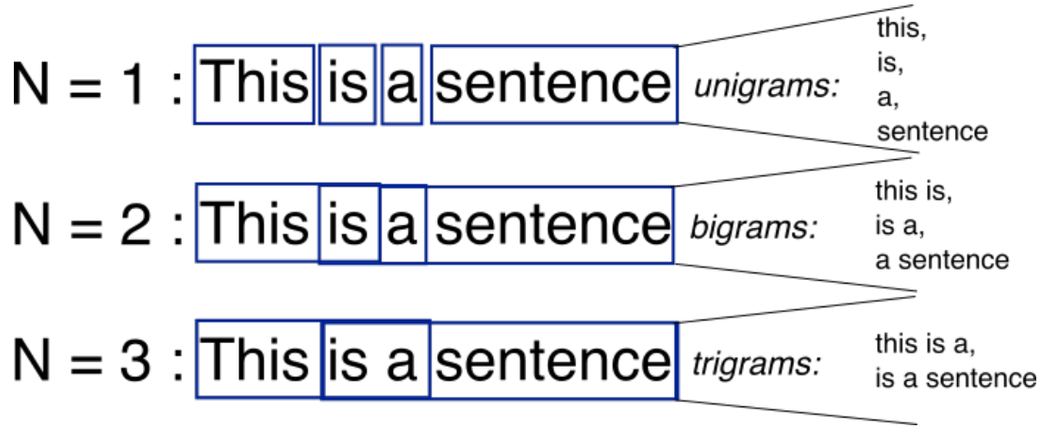

MSE on Testing Data
MSE
0.5
0.4
0.3
0.2
0.1
KNN
LASSO
LSTM
RNN
SVM
XGBoost
model

THAT'S WHAT HE SAID



- Goal: To generate text based on Michael Scott's dialogues
- Compare two language models: Markov Model & LSTM language model
- Markov Language Model
    - "A Mathematical Theory of Communication" Claude Shannon (1948)
        - proposed using Markov Chain to produce text consisting of a sequence of words
- Given fixed k previous word(s), "generate" the next word by calculating the word transition probabilities from the corpus
    - Current word: "I"
    - Find all the words that come right after "I" and create a dictionary with all the words associated it
    - Add a count to the words in the dictionary every time any of those words appears next to "I"

- Markov Bigram Language Model:
  - Previous 2 words to generate the next word



Result:

Move in here. Jim and David Wallace would know what?
I've fallen and I don't know.
"Those of you to find out if there's downsizing, okay?"
"I score. Well, that's like six years ago."
"Yes. Karen, do I do."
"Goooood morning, baby."
"Names, numbers. Okay, well, I would like to have a word with you. The Gould has been a big Fear Factor fan."
Jim becomes a manager.
"I loved it, right guys? See? I don't know about that."
And those are the number one cause of death. It's just potatoes and mayonnaise. Just gimmee -
Lonny. And now you can't. You're too young. And I'm very sorry for me to be a baby in this office. That's terrible... terrible news for both of us.

- Word-level LSTM Language Model:
  - 2 hidden layers with 128 units each
  - Computationally slow

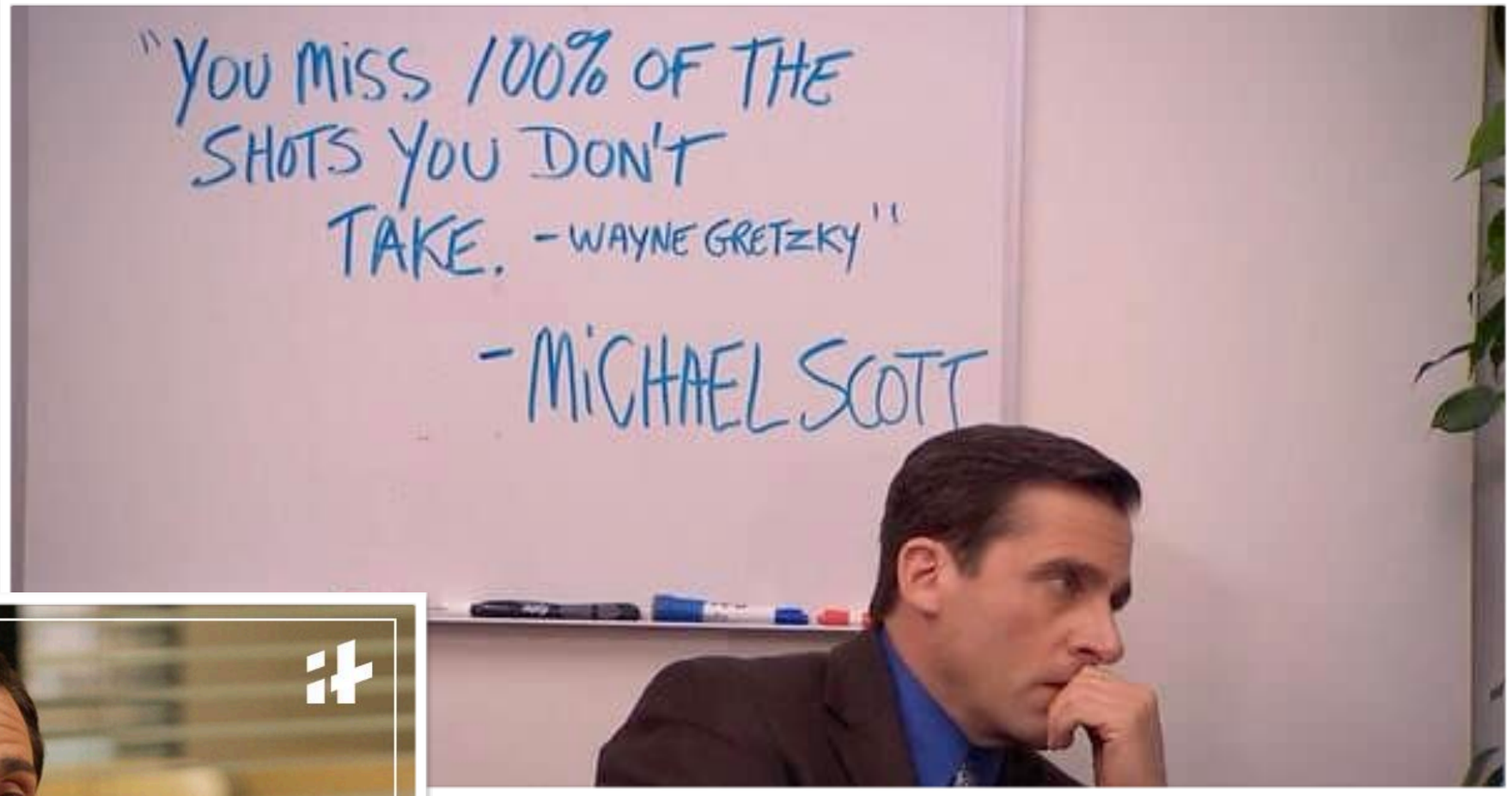
Result:

gonna come back with this. i live around a fantasy world. i need to do something about your coffee breath " " - ok. shut it up, change my own friends. connect! here we go. ugh... i'm in. oh yeah! yes he is. just a
is married, i will be in my office making toys two feet. okay. i would like to see those please. here we go! who's on him? somebody get him! yeah? that's good and i will kill you in any man, and i am the boss and apparently
going to tell her. number one what is up to these people like this company, and i heard that, "" i pledge to always keep an open mind and an open heart." " i do believe... the honors. i don't know. i don't know.
month. and i think she wants to talk why you're going to do. dwight, do you have the box? you made it, i'm not gonna be in town. i'm going to be thinking about it. i've already got my name picked out, lord of
once again, i am just thankful that i am... hey, hey, you. thanks for him and i'm keeping him in. take what? it's a joke. when you do this we and? whoa, what's up? hey hey. josh porter,
part of your life today, you know what, it's kind of our coming out party. really. and that's just... it's a big red trash compactor! okay, okay! i got it. dwight, you saw darryl move it, right? if you had
. no, nothing like that at all. hey, what's going on? what is it? no, just tell ya what it means, you know the dundies, the real thing, i will wait. no, it's nobody. hmm. he sounds on the booze. no,

want you to go find firecrackers. and a chihuahua. pam, in the frozen food section - o gonna see the sales department. oh, wow. whew, oh god. laying a base. laying. yes, and you know who i end up with her. and it's just men'
posse guys are gonna by pam and phyllis is very smart. and she was my girl now. she feels like you think you can't catch anything. here we go. it is time, thank you. ok, come on. let's go. let's do that. that'

CONCLUSION

- Markov model seem to be slightly more coherent
- Data had more features than observations (186 episodes)
    - Simpler machine learning tools can be more efficient like LASSO than neural networks
- Attempted text classification based on characters with poor results (~~predict who said it~~)
- Clustering and topic modeling (*work in progress*)
- Future work:
    - Consider other methods of representing text data (word embedding)
    - Consider text with direction
    - More cleaning and preprocessing


"YOU MISS 100% OF THE SHOTS YOU DON'T TAKE. -WAYNE GRETZKY"
-MICHAEL SCOTT


"And I knew exactly what to do.
But in a much more real sense,
I had no idea what to do."
- Michael Scott

THANK YOU