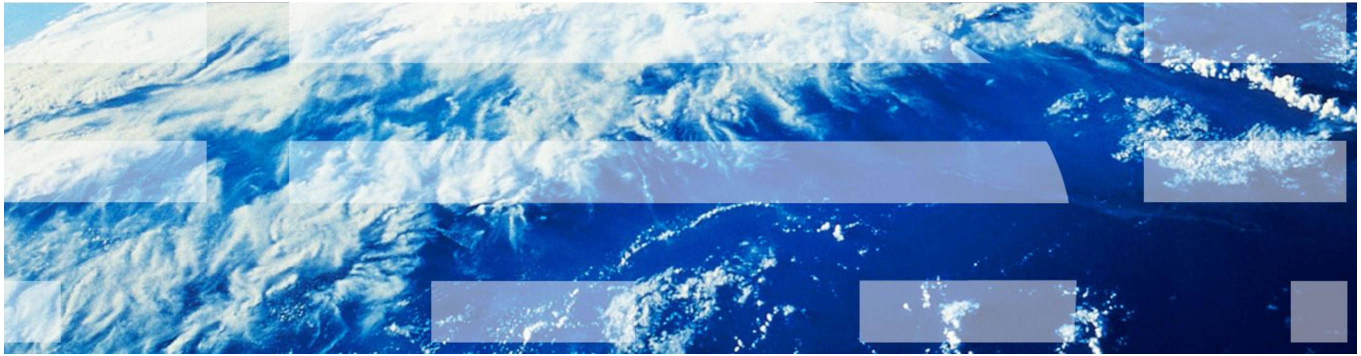
E6895 Advanced Big Data Analytics Final Project:

Recommending News Based on Twitter Profiles

Justin Tse
Cindy Xu



April 12, 2018

© CY Lin, 2018 Columbia University

Project Tasks

The overall goal of the project is to be able to recommend news articles based off of a user's tweets. This can be separated into parts.

1. Create a model that can classify the topic of a tweet.
2. Use the previous model to analyze a user's tweets and then generate relevant news topics.

© 2018 CY Lin, Columbia University

Underlying Project Structure

We will use the New York Times to generate news articles because they have a convenient API.

We will also use the Twitter API to gather data.

Data Collection

We needed to build a model that can classify the topic of a single tweet. We were unable to find a relevant labeled dataset to train our model on, so we gathered the data using the Twitter API.

For each topic, we collected roughly 40,000 tweets with the hashtag of that topic.

The topics that we focused on were:

Politics, Business, Technology, Science, Sports, Style, and Health

Before analyzing the raw tweet data, we removed the specific hashtag, all links, lower cased the text, removed mentions, removed symbols, and stemmed the words.

Bag of Words Model

Motivation: It’s extremely quick to train, and it can classify tweets quickly. It is important to be able to classify tweets quickly because we want to recommend articles in real time.

I wanted to have a model that trained quickly because Twitter data is constantly changing.

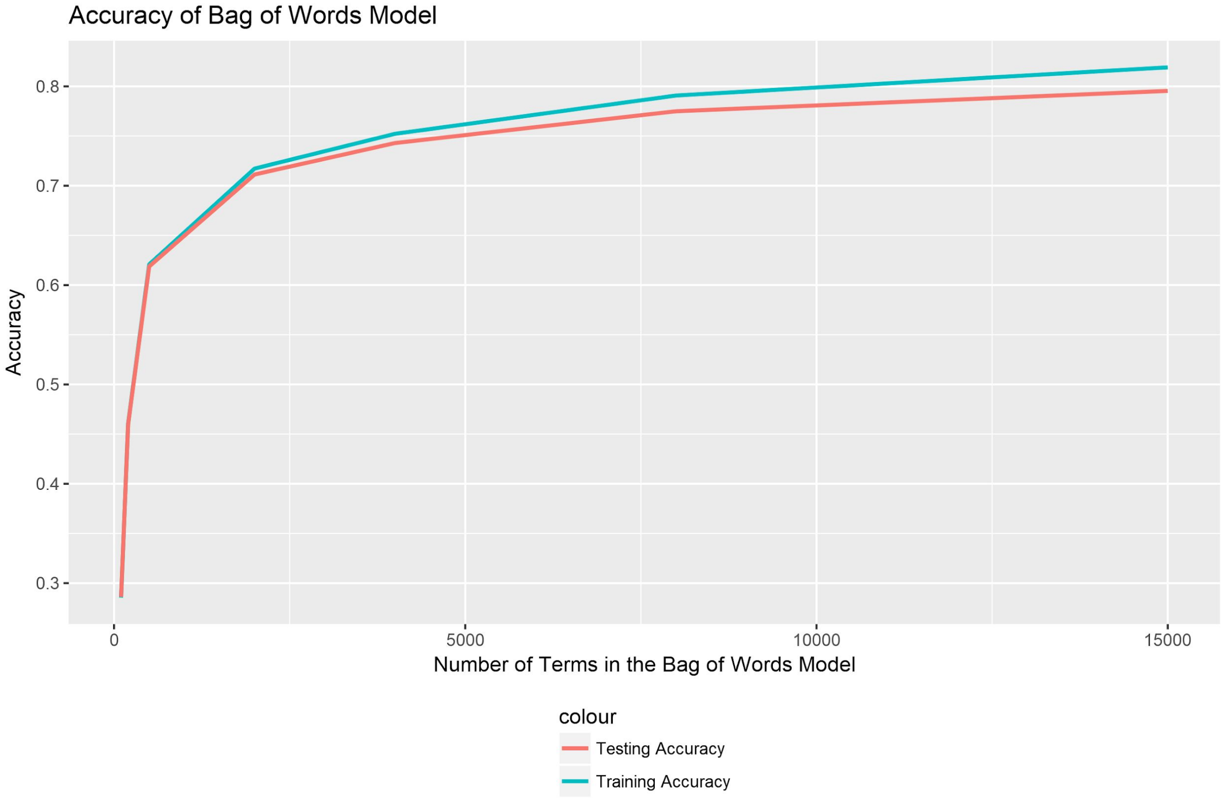
The model stores the top 15,000 terms; it takes roughly 10 seconds to analyze the past 100 tweets of a user and recommend relevant articles.

Bag of Words Model

Accuracy on the Training and Test Set:

# of Terms Saved in Model	Training Accuracy	Testing Accuracy
100	.285431	.286206
200	.460631	.459877
500	.620788	.618692
2000	.717206	.711269
8000	.790734	.774913
15000	.819089	.795407

Bag of Words Model

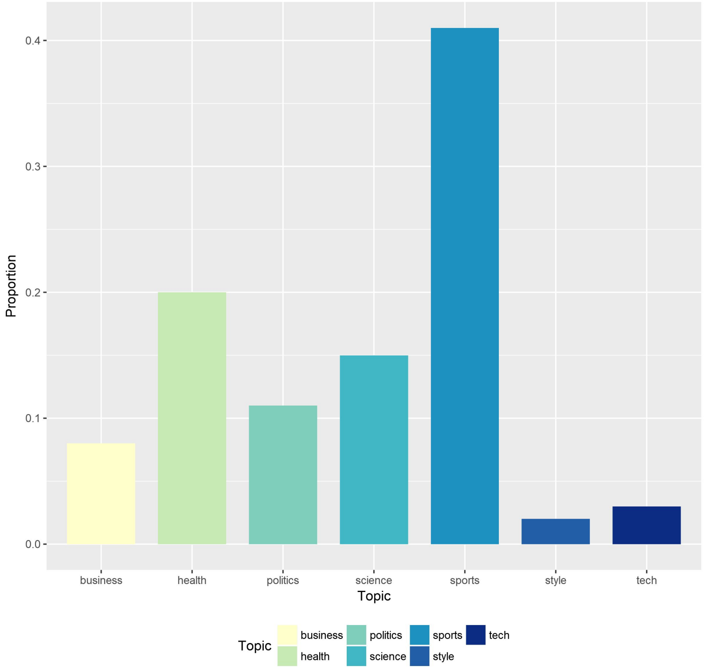


© 2018 CY Lin, Columbia University

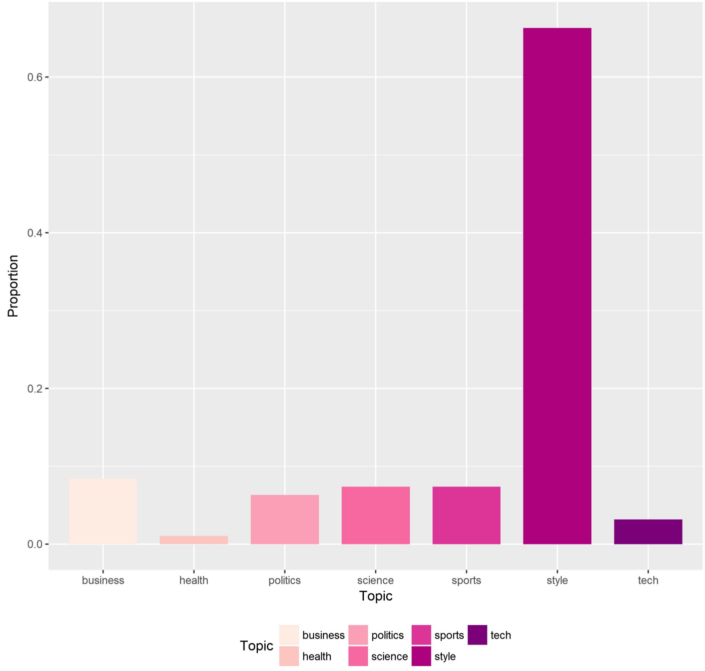
Bag of Words Model

Example of Output for Stephen Curry (@stephencurry) and Mason Margiela (@margiela)

Interests of Stephen Curry Based On His Last 100 Tweets



Interests of Martin Margiela Based On His Last 100 Tweets



Neural Network

- First randomly split data into train and test with 9:1
- Then transform text into data matrix by using tokenizer with first 3000 word. It will return a dictionary with first 3000 most frequency word in dataset and their corresponding index.
- Also transform label into categorical data. Eg: Labels='politics'   [1,0,0,0,0,0,0]
- Substitute data matrix and categorical label into neural network.
- Architecture: model=Sequential()

  model.add(Embedding(3, 10, input_length=train_x.shape[1]))

  model.add(Flatten())

  3 dropout layers

  Dense layer

© 2018 CY Lin, Columbia University

Neural Network

- *Embedding layer:*

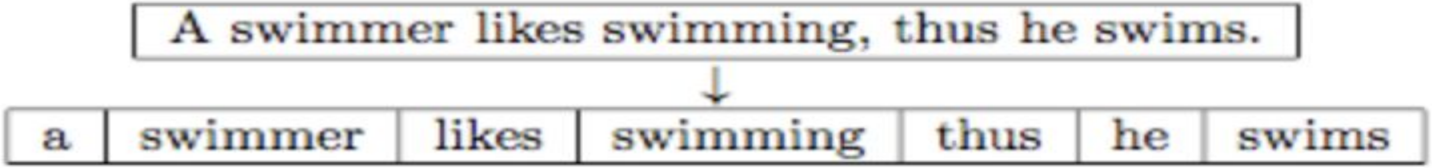
❖ Bag of Word is Sparse.

❖ Word Embedding :much lower dimensional space

- *Tokenization* :

Describes the general process of breaking down a text corpus into individual elements that serve as input for various natural language processing algorithms.

Table 2. *Example of tokenization.*



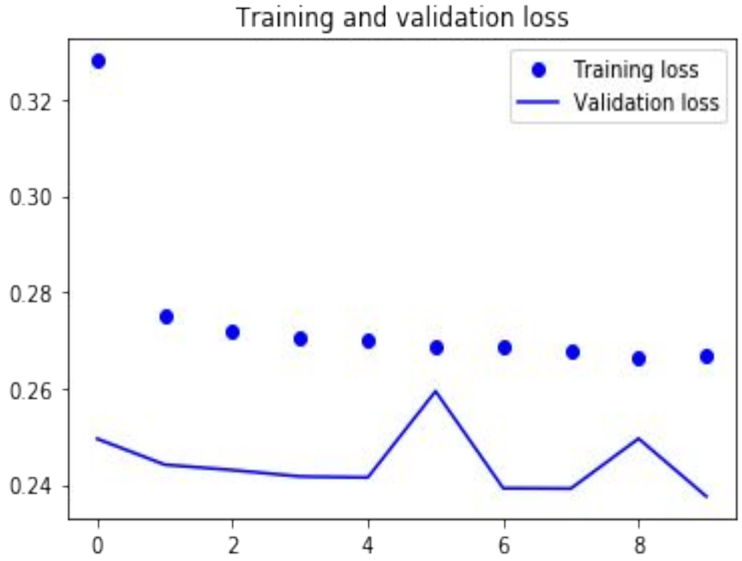
A swimmer likes swimming, thus he swims.

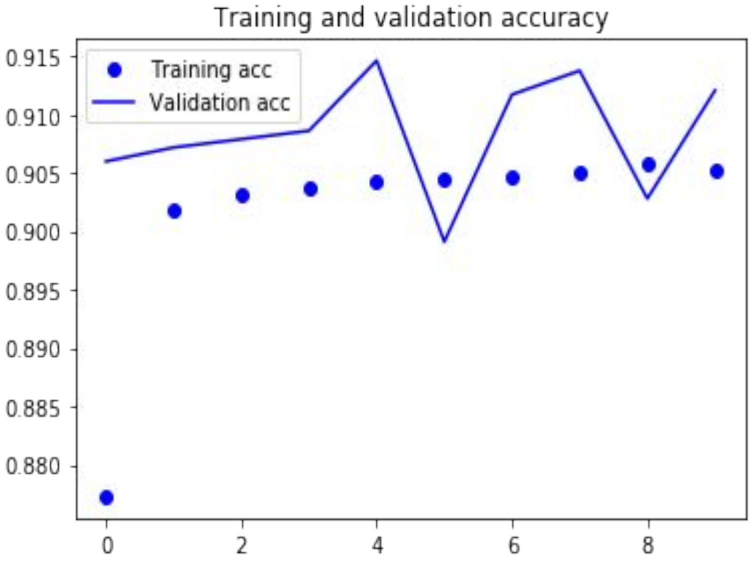
↓

a	swimmer	likes	swimming	thus	he	swims

Neural Network

test	train	validation
0.8351591446614018	0.9052	0.9121





© 2018 CY Lin, Columbia University

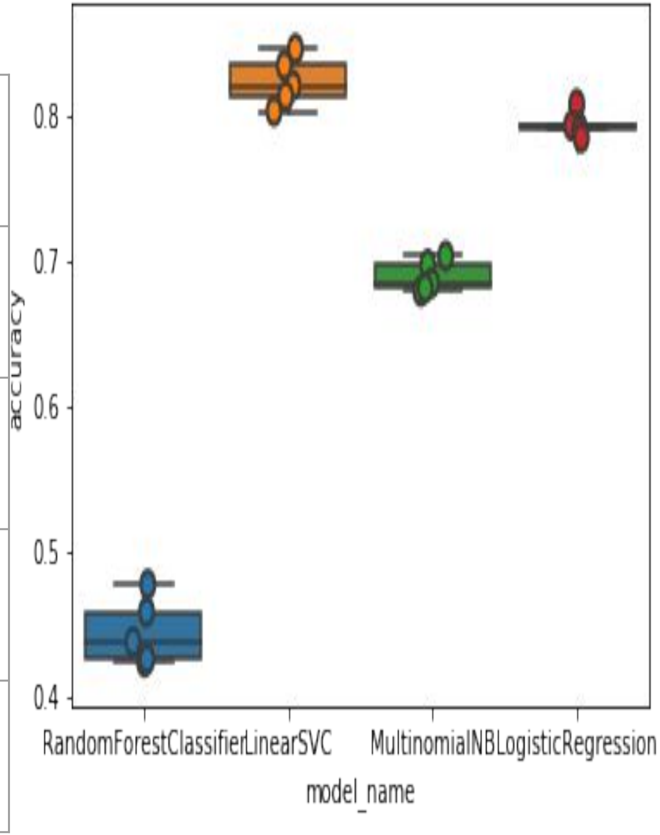
Other Traditional Machine Learning Method

- MultinomialNB
- LogisticRegression
- LinearSVM
- RandomForest

© 2018 CY Lin, Columbia University

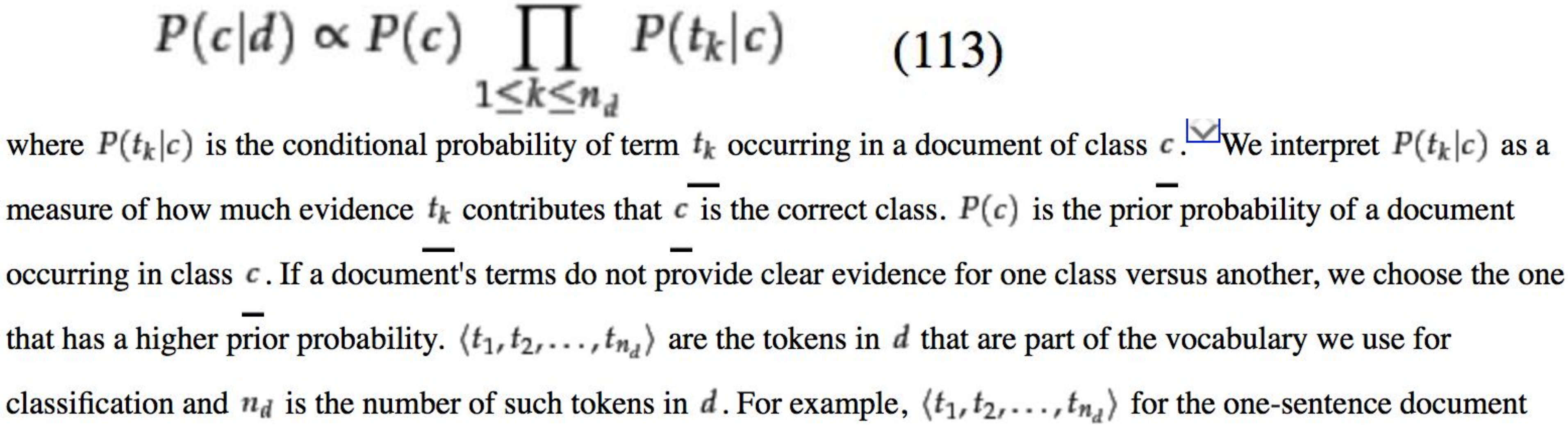
Other Traditional Machine Learning Method

Model	train	test
MultinomialNB	0.8486141216621316	0.8040192314344021
LogisticRegression	0.8768197820104844	0.8161253502127218
LinearSVM	0.9383349475011146	0.8602262114766006
RandomForest	0.2685975187935249	0.2721801390474214



© 2018 CY Lin, Columbia University

MultinomialNB



$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \qquad (113)$$

where $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c. We interpret $P(t_k|c)$ as a measure of how much evidence t_k contributes that c is the correct class. $P(c)$ is the prior probability of a document occurring in class c. If a document's terms do not provide clear evidence for one class versus another, we choose the one that has a higher prior probability. $\langle t_1, t_2, \ldots, t_{n_d} \rangle$ are the tokens in d that are part of the vocabulary we use for classification and n_d is the number of such tokens in d. For example, $\langle t_1, t_2, \ldots, t_{n_d} \rangle$ for the one-sentence document

LogisticRegression

- Compare to project: Recommendation Product for Customers in Santander Bank.

- Logistic Regression works well for both financial dataset and text dataset.

- But Compare to Randomforest, it gives a really poor accuracy in text data.

© 2018 CY Lin, Columbia University

RandomForest

- Reason why it performs bad: Text data is special

❖ Shortness:

there is not enough information to measure relevant similarities between several texts.

❖ Sparkness

text data also lacks of contextual information, especially when using vector presentation and combined with RandomForest

- Improve: use topic model technique LDA to combine features, reduce feature space size significantly

© 2018 CY Lin, Columbia University

LinearSVM

- High dimensional Input Space:

SVMs use overfitting protection which does not necessarily depend on the number of features.

- Few irrelevant features:

It is not easy to do feature selection.

- Document vectors are sparse:

SVMs are well suited for problems with dense concepts and sparse instances.

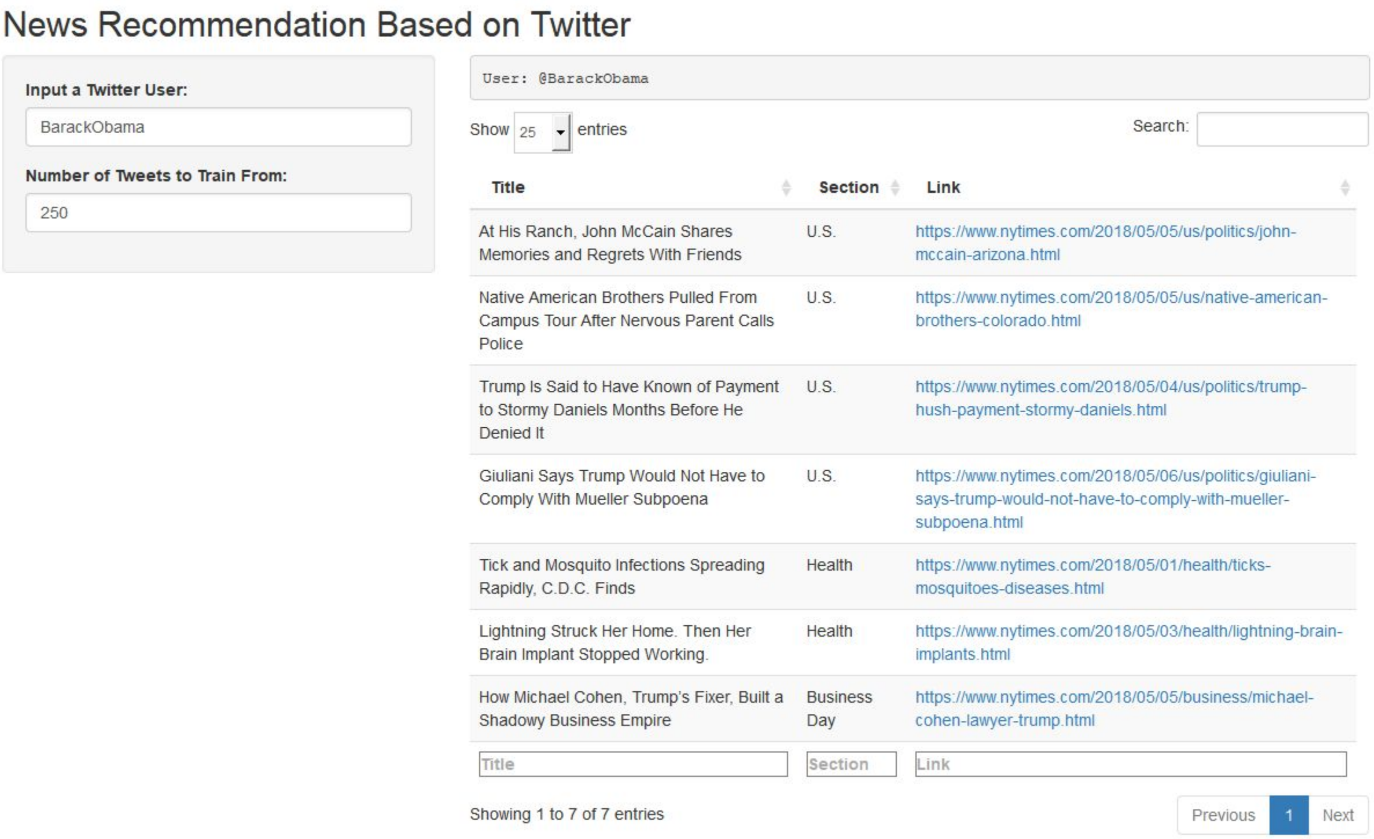
© 2018 CY Lin, Columbia University

Recommending News

We use the output of the distribution given by the Bag of Words model to randomly choose topics. We then use the New York Times API to choose popular articles in those topics.

Demo: https://justintse.shinyapps.io/NewsRecTwitter/

© 2018 CY Lin, Columbia University



News Recommendation Based on Twitter

Input a Twitter User: BarackObama

Number of Tweets to Train From: 250

```
User: @BarackObama
```

Show 25 entries

Search:

Title	Section	Link
At His Ranch, John McCain Shares Memories and Regrets With Friends	U.S.	https://www.nytimes.com/2018/05/05/us/politics/john-mccain-arizona.html
Native American Brothers Pulled From Campus Tour After Nervous Parent Calls Police	U.S.	https://www.nytimes.com/2018/05/05/us/native-american-brothers-colorado.html
Trump Is Said to Have Known of Payment to Stormy Daniels Months Before He Denied It	U.S.	https://www.nytimes.com/2018/05/04/us/politics/trump-hush-payment-stormy-daniels.html
Giuliani Says Trump Would Not Have to Comply With Mueller Subpoena	U.S.	https://www.nytimes.com/2018/05/06/us/politics/giuliani-says-trump-would-not-have-to-comply-with-mueller-subpoena.html
Tick and Mosquito Infections Spreading Rapidly, C.D.C. Finds	Health	https://www.nytimes.com/2018/05/01/health/ticks-mosquitoes-diseases.html
Lightning Struck Her Home. Then Her Brain Implant Stopped Working.	Health	https://www.nytimes.com/2018/05/03/health/lightning-brain-implants.html
How Michael Cohen, Trump's Fixer, Built a Shadowy Business Empire	Business Day	https://www.nytimes.com/2018/05/05/business/michael-cohen-lawyer-trump.html
Title	Section	Link

Showing 1 to 7 of 7 entries