Natural Language Programming

with Reddit

Tricia Wells • 07.12.2019

satire noun

sat·ire | \ ˈsa-ˌtī(-ə)r \

Definition of *satire*

1 : a literary work holding up human vices and follies to ridicule or scorn

2 : trenchant wit, irony, or sarcasm used to expose and discredit vice or folly

Real or Satire?

1.2k

PR Disaster: Nike Is Under Fire After It Released An Ad Featuring A Photo Of Colin Kaepernick That Was Way Too Close Up

95% Upvoted

36 Comments Share Save Hide Report

Real or Satire?

22.3k

Russian Journalist Charged for ‘Controlling Minds’ With ‘1984’ Reference

95% Upvoted

597 Comments Share Save Hide Report

Predicting Satire

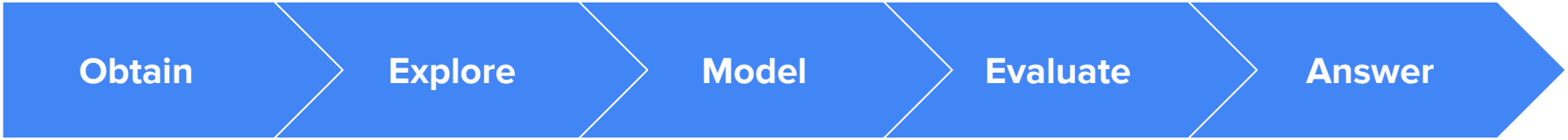
Real World Problem

Sees an absurd headline - *“This can’t possibly be real! “*

Data Science Problem

Are **words** used in the headlines of a Reddit post (r/TheOnion) predictive of ~~Satire~~?

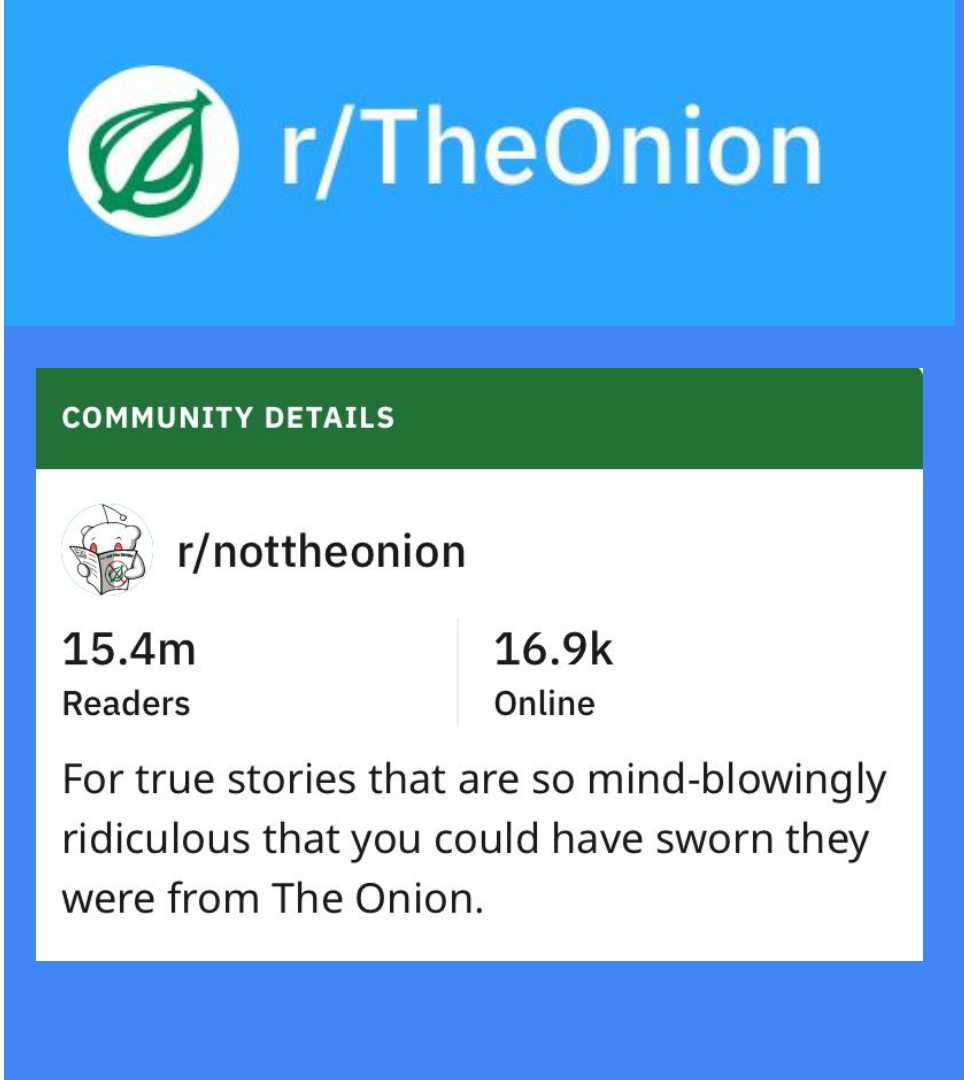
Workflow



Obtaining Data

r/TheOnion
v.
r/nottheonion

API tool: Pushshift.io



Exploratory Data Analysis

Countvectorizer

- Returns counts of words in each headline
- Classifier tuned to the top 1000 words

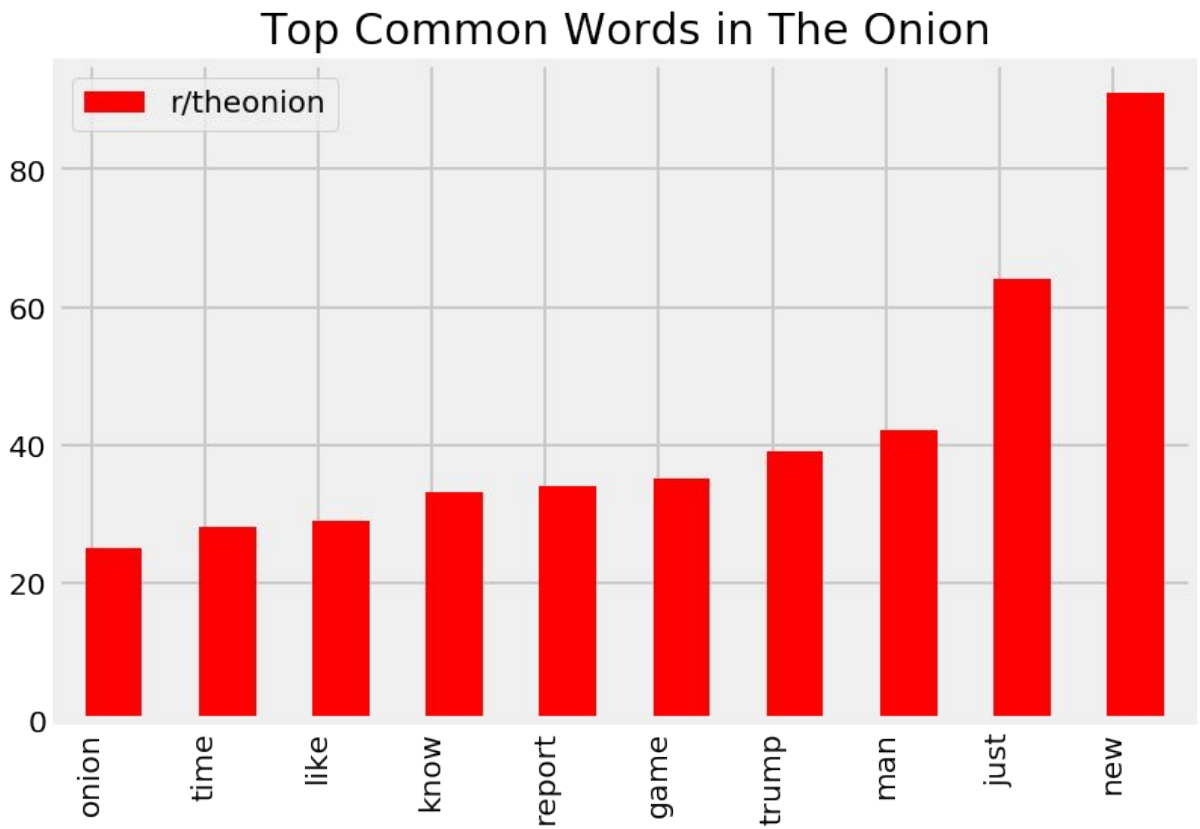
Stop Words

- English common words (sklearn) ie. "a," "and," "but," "how," "or," and "what."

	title
0	Chernobyl nuclear site to become 'official tou...
1	SEE THE NEW SITH TROOPER FROM STAR WARS: THE R...
2	Cargo ship owned by JPMorgan Chase seized by U...
3	Police: Mom drove with children in swimming po...
4	'Star Wars: The Rise of Skywalker': Meet the n...

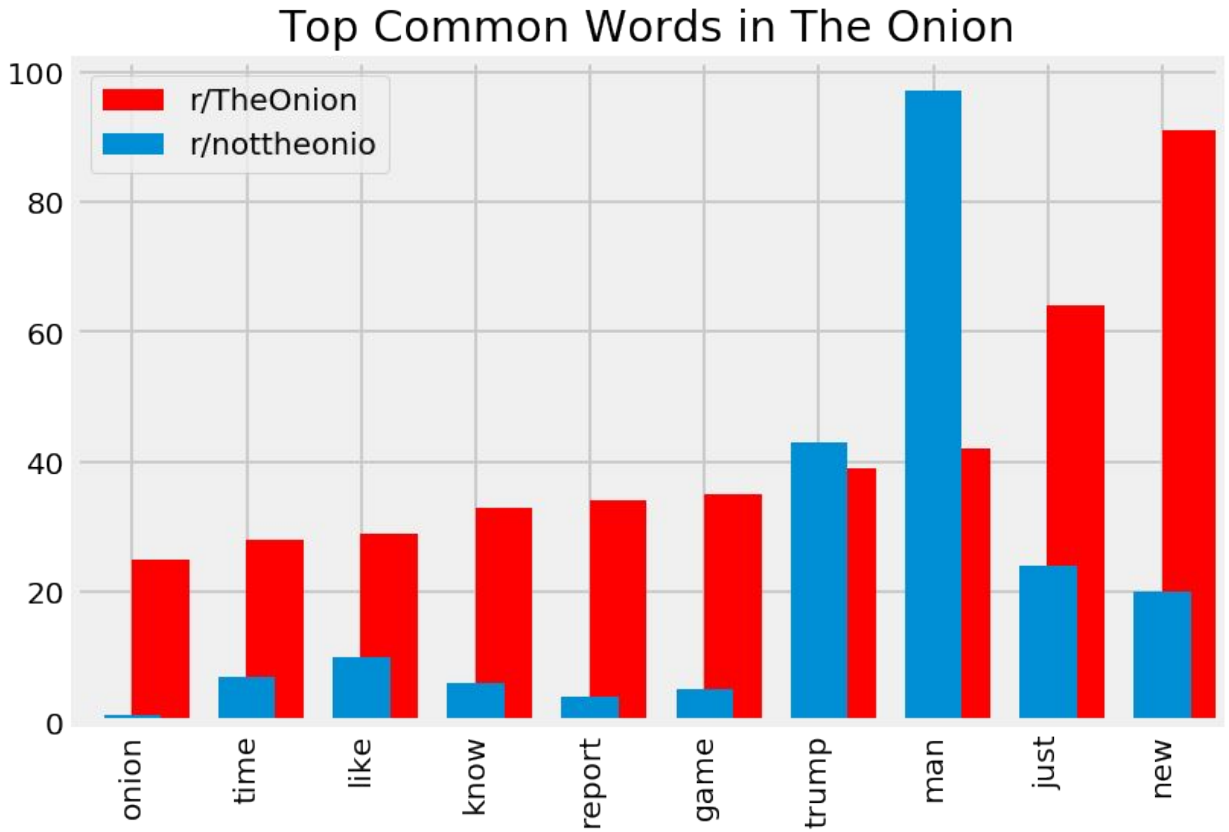
	0	1	2	3	4
sex	0	0	0	0	0
ship	0	0	1	0	0
shirt	0	0	0	0	0
shit	0	0	0	0	0

Exploratory Data Analysis



1. New
2. Just
3. Man
4. Trump
5. Game

Exploratory Data Analysis



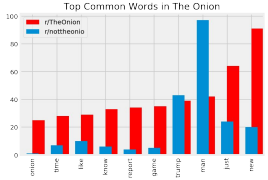
1. New
2. Just
3. Man
4. Trump
5. Game

Workflow

Vectorize

Multinomial Naive Bayes
Logistic Regression

Obtain → Explore → Model → Evaluate → Answer



Countvectorizer

Turning text into counts

	title
0	Chernobyl nuclear site to become 'official tou...
1	SEE THE NEW SITH TROOPER FROM STAR WARS: THE R...
2	Cargo ship owned by JPMorgan Chase seized by U...
3	Police: Mom drove with children in swimming po...
4	'Star Wars: The Rise of Skywalker': Meet the n...

	0	1	2	3	4
sex	0	0	0	0	0
ship	0	0	1	0	0
shirt	0	0	0	0	0
shit	0	0	0	0	0

TF IDF Vectorizer

- **TF (Term Frequency)** - number of times a word appears in a document
- **IDF (Inverse Document Frequency)** - measure of how significant that word is in the corpus.
- Penalizes more common words.
- Put simply, the higher the TF IDF score (weight), the **rarer the term**.

For Example: “CAT”

TF(cat)= 12/100 = 0.12

IDF (cat) = log (10,000/300) = 1.52

∴ W(cat) = (TF*IDF) 0.12 * 1.52 = 0.182

Source :https://www.onely.com/blog/what-is-tf-idf/

Multinomial Naive Bayes

Countvectorizer

Max df = 0.3 (default = 1)
Max Features = 3000

Alpha = 2 (default = 1)

	Predicted RealNews	Predicted Satire
True News	208	49
True Satire	26	217

TF IDF Vectorizer

Max df = 0.4 (default = 1)
Max Features = 4000

Alpha = 1

	Predicted RealNews	Predicted Satire
True News	217	40
True Satire	31	212

Logit Classifier

TF IDF Vectorizer

Max df = 0.4 (default = 1)
Max Features = 4000
Alpha = 1

Logistic Regression

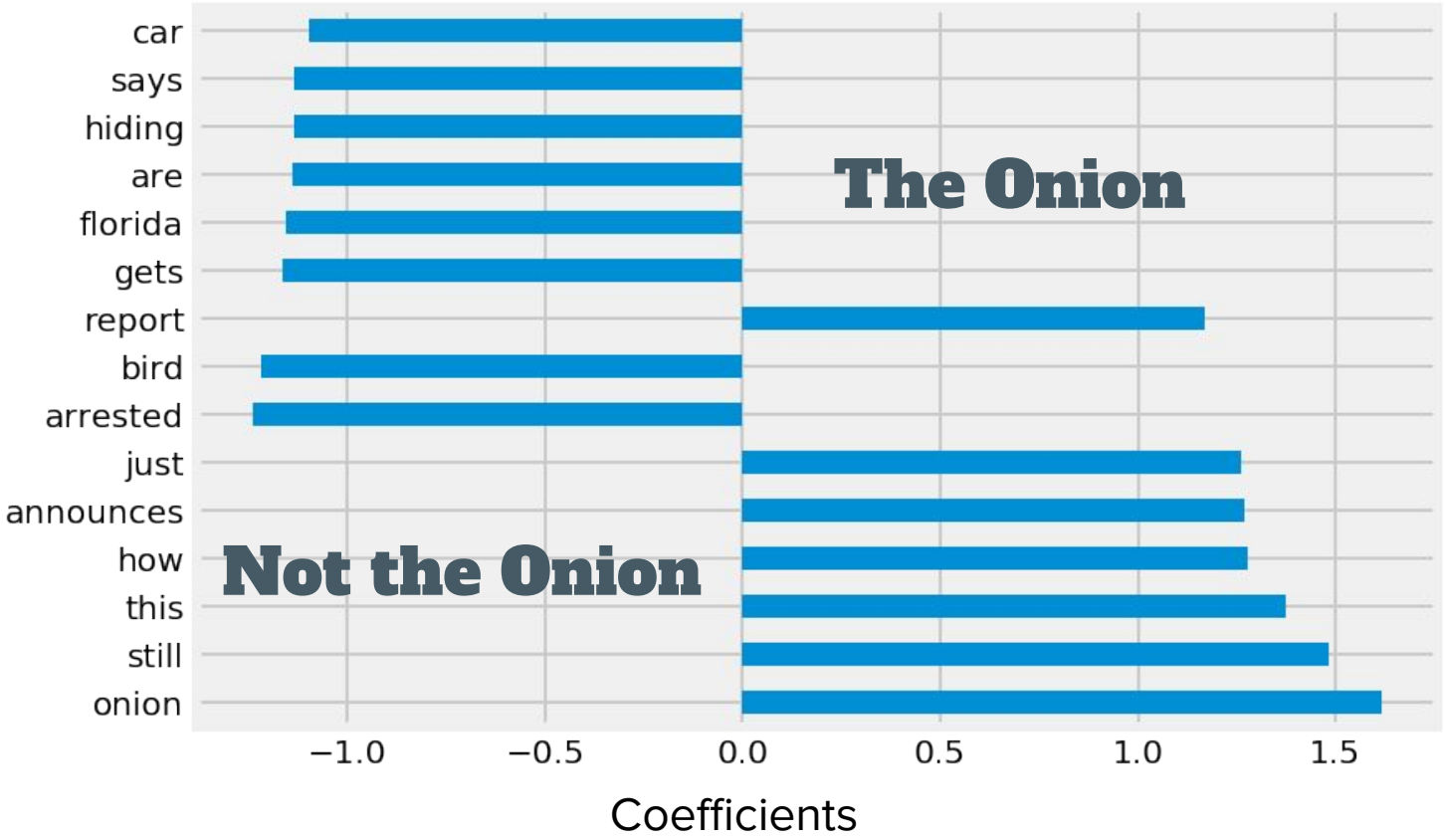
Penalty = l2
C = 1.0

Accuracy
0.82

Assumptions

1. Dependent variable (target) is Binary
2. Observations (words) are independent of each other
3. **VIOLATED** because words are collinear. They do have effect on each other.

Most Predictive Words



Naive Bayes Comparison

Sensitivity Optimization TP/(TP+FN)

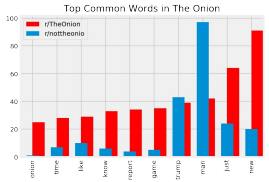
	Accuracy	Specificity	Sensitivity	Precision
Multinomial Naive Bayes with CountVectorization	.85	.809	**.893**	.816
Multinomial Naive Bayes with TF IDF	.858	.8444	.8724	.8413

Workflow

Vectorize

Multinomial Naive Bayes
Logistic Regression

Obtain → Explore → Model → Evaluate → Answer



Sensitivity
.893

CountVectorizer

Multinomial Naive Bayes

RESULTS

- Count Vectorization
- Multinomial Naive Bayes Classifier
- Sensitivity Optimization

	Results
Accuracy	.985
Specificity	.99
Sensitivity	.98
Precision	.9899

	Predicted RealNews	Predicted Satire
True News	198	2
True Satire	4	196

Predicting Satire

Data Science Problem

Are **words** used in the headlines of a Reddit post (r/TheOnion) predictive of Satire?

Yes! We can predict r/TheOnion with 98.5% accuracy.

RESULTS

- TF IDF Vectorization
- Multinomial Naive Bayes Classifier
- Accuracy Optimization

	Results
Accuracy	.995
Specificity	1.0
Sensitivity	.99
Precision	1.0

	Predicted RealNews	Predicted Satire
True News	200	0
True Satire	2	198

Real or Satire

Real or Satire?

1.2k

PR Disaster: Nike Is Under Fire After It Released An Ad Featuring A Photo Of Colin Kaepernick That Was Way Too Close Up

95% Upvoted

36 Comments Share Save Hide Report

Real or Satire?

the ONION®

1.2k

r/TheOnion · Posted by u/GriffonsChainsaw 10 months ago

PR Disaster: Nike Is Under Fire After It Released An Ad Featuring A Photo Of Colin Kaepernick That Was Way Too Close Up

clickhole.com/pr-dis...

95% Upvoted

36 Comments Share Save Hide Report

Satire

Real or Satire?

22.3k

Russian Journalist Charged for ‘Controlling Minds’ With ‘1984’ Reference

95% Upvoted

597 Comments Share Save Hide Report

Real or Satire?

22.3k

r/nottheonion · Posted by u/peter_bolton 1 day ago

Russian Journalist Charged for ‘Controlling Minds’ With ‘1984’ Reference

themoscowtimes.com/2019/0...

95% Upvoted

597 Comments Share Save Hide Report

The Moscow Times

REAL

Real or Satire?

60

Female journalist told she needs male chaperone to cover politician's campaign

87% Upvoted

30 Comments Share Save Hide Report

Real or Satire?

The Guardian

60

r/nottheonion · Posted by u/sslloooww 3 hours ago

Female journalist told she needs male chaperone to cover politician's campaign

theguardian.com/us-new...

87% Upvoted

30 Comments Share Save Hide Report

REAL

Real or Satire?

1

Outraged Trump Declares He Would've Gotten Jeffrey Epstein Way More Lenient Plea Deal

100% Upvoted

1 Comment Share Save Hide Report

Real or Satire?

the ONION®

1

r/TheOnion · Posted by u/emotionalraven 5 hours ago

Outraged Trump Declares He Would've Gotten Jeffrey Epstein Way More Lenient Plea Deal

politics.theonion.com/outrag...

100% Upvoted

1 Comment Share Save Hide Report

Satire

Natural Language Programming

with Reddit

Tricia Wells • 07.12.2019