
Fact or Fiction?

Can predictive modeling be used
to determine if written
information is factual?

Overview

The Data

- What it is/where it originated
- Why it's been selected

Model Selection Process

- What was considered
- What was selected and why

Findings

- Confusion Matrix
- Interpreting Coefficients

Visualizations

Transferability

- How might this be used?

The Data - What subreddits were chosen

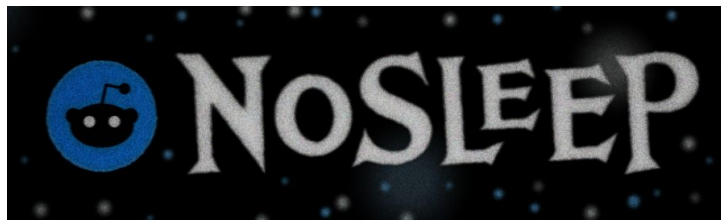
Unresolved Mysteries (Fact)



From each:

- Pulled the most recent 1250 posts
- Compiled body text of each into df

No Sleep (Fiction)



Why these were chosen:

- text heavy
- similar vein
- useful

The Model Selection Process

Classification Models Chosen:

- LogReg
- KNN
- Multinomial Naive Bayes
- Gaussian Naive Bayes

99.5%

The selected model's accuracy score

Looking at Comparisons

	LogReg	KNN	MNB	GNB
Train Score	.998	.843	.994	.998
Test Score	.995	.749	.988	.951

Logistic Regression

What it is

Most common binary classification

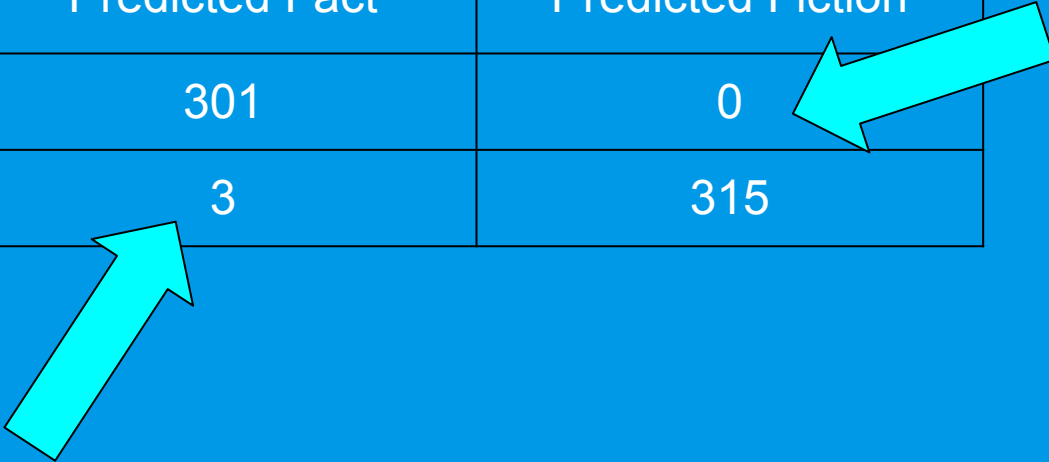
Predicts probabilities of an occurrence between 0 and 1.

Examples:

- Will someone vote?
- Is this text fact or fiction?

Model Outcomes

	Predicted Fact	Predicted Fiction
Actual Fact	301	0
Actual Fiction	3	315

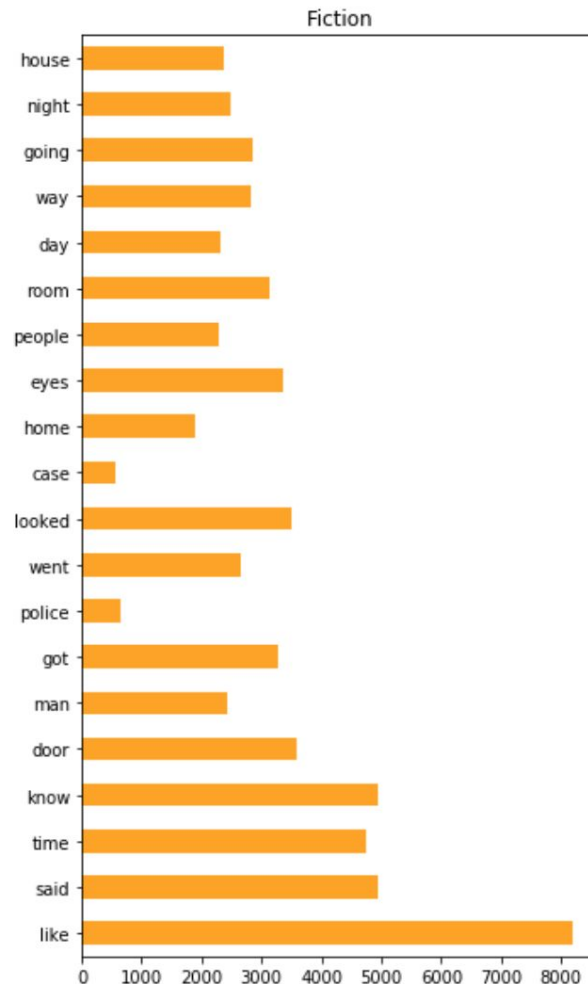
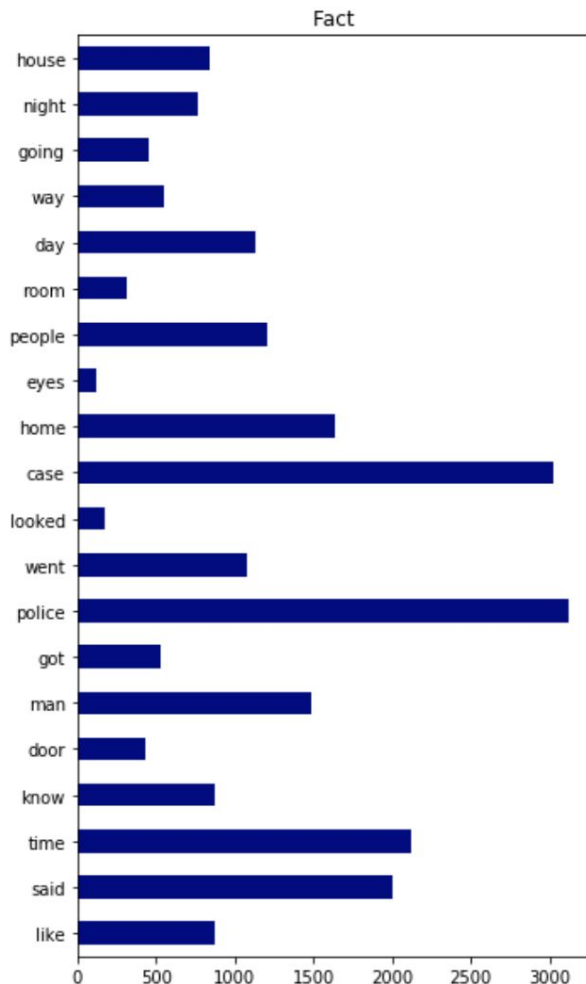
Two red arrows with black outlines are pointing to specific values in the table. One arrow points from the right towards the '0' in the 'Actual Fact' row under the 'Predicted Fiction' column. The other arrow points from the bottom-left towards the '3' in the 'Actual Fiction' row under the 'Predicted Fact' column.

In the real world, this would mean...

Let's Take A Look

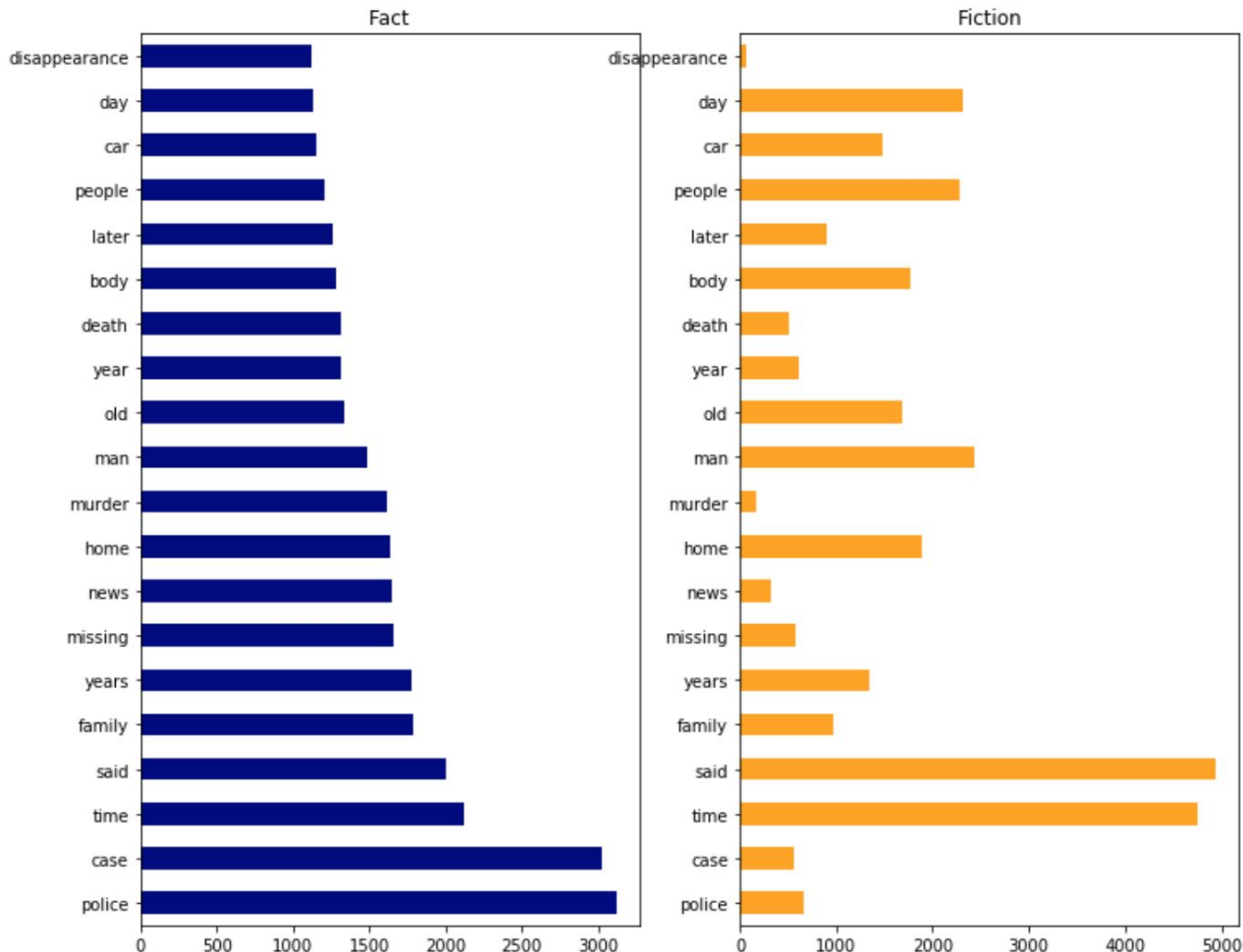
20 most common words

What was removed and why?



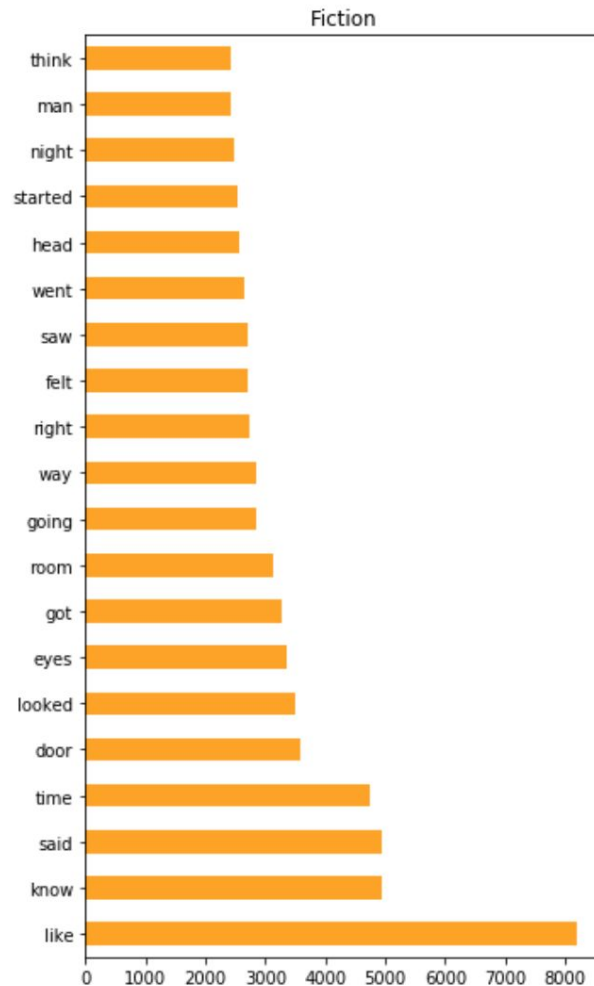
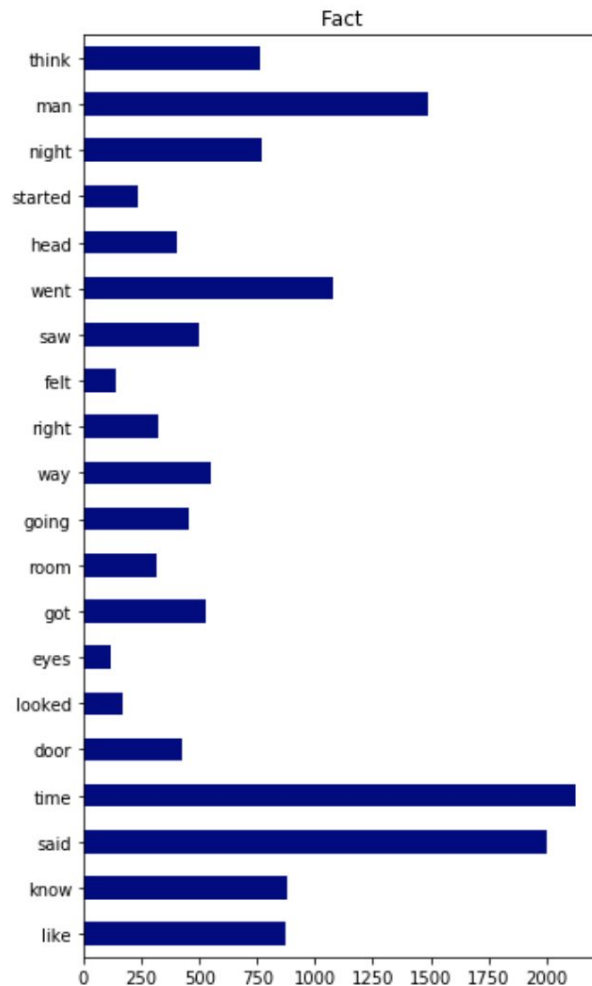
Let's Take A Look

Most common words in *fact*
Very predictive v. not predictive



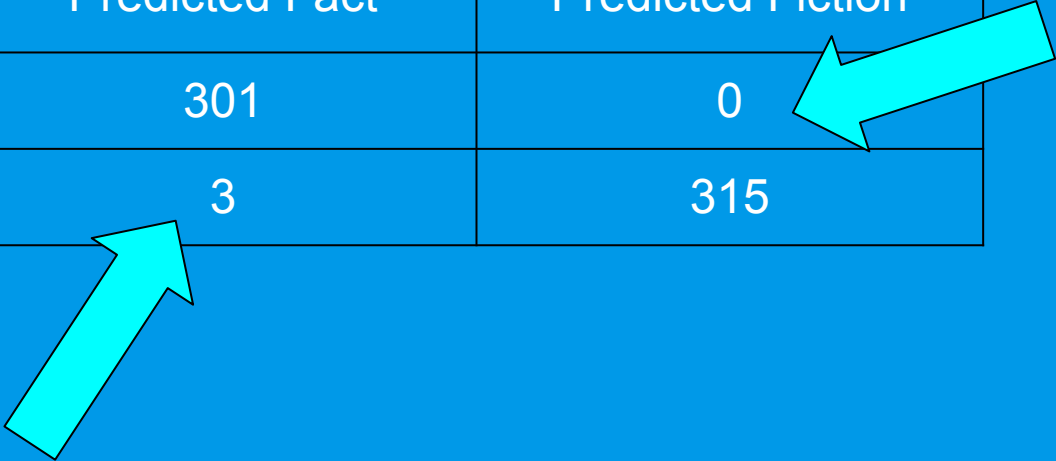
Let's Take A Look

Most common words in *fiction*
Implication of word count



Model Outcomes

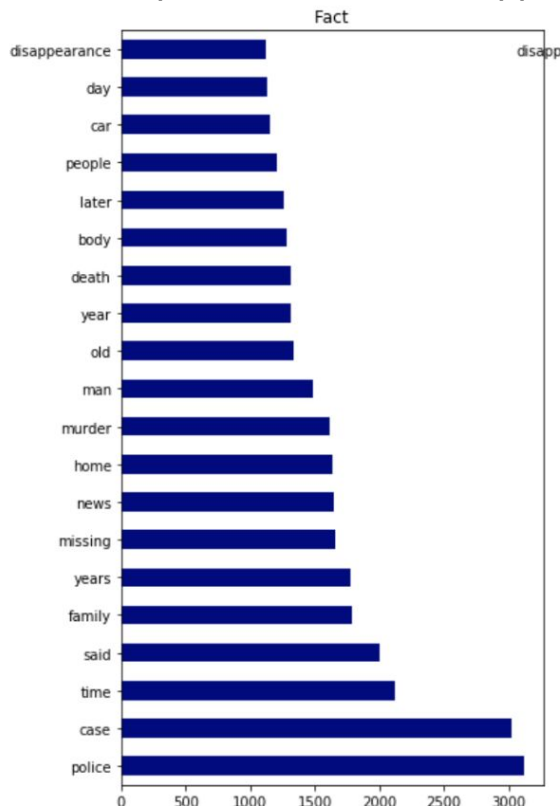
	Predicted Fact	Predicted Fiction
Actual Fact	301	0
Actual Fiction	3	315



In the real world, this would mean...

Digging In

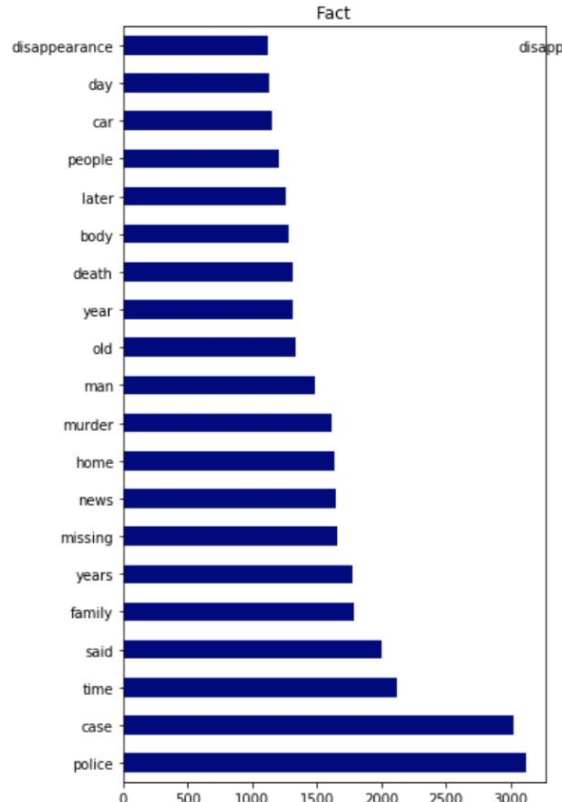
Incorrect predictions - what happened?



"I was hoping someone will remember the case i am trying to. It's a cold case, i am unsure if it was some thing that happened to more than one kid, most of the information i remember isn't much, i did read about it on unresolved mysteries so hopefully others who read it know much more than i do.\n\nThe case was about 60/70s, a teenage girl was contacted to babysit and it was a man who contacted her and picked her up and was never seen again. I think there were other similar cases but I'm not sure if they were connected. It's not the Amy Mihaljevic case. The information i remember the most is that she was asked to babysit and the man who contacted her picked her up and i think he may of picked her up outside her own home. \n\nHopefully others remember the post i also read. I don't remember enough to actually search for it on Google or on unresolved mysteries, i just keep getting cases of babysitter who kill. \n\nhttps://en.m.wikipedia.org/wiki/Murder_of_Amy_Mihaljevic"

Digging In

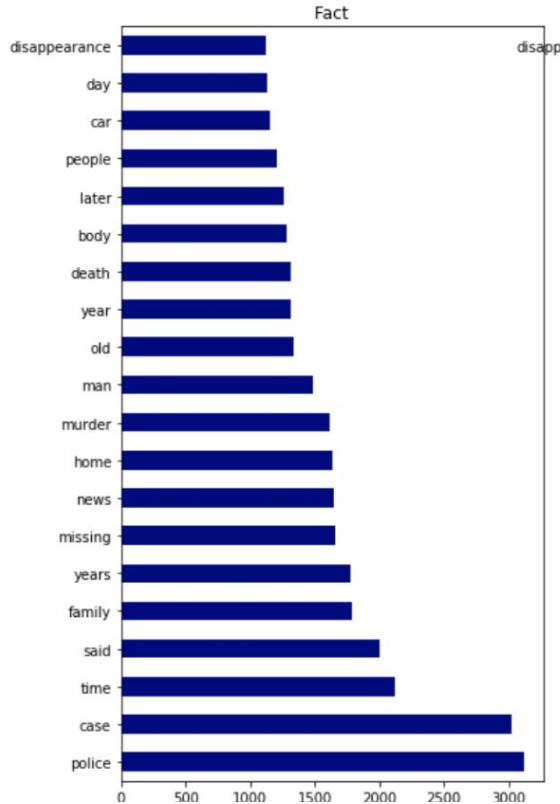
Incorrect predictions - what happened?



*"**I would like to put focus on one of the most famous death through out history and has been the season finale of** [**buzzfeed unsolved.**](<https://www.youtube.com/watch?v=Hrsdntds9kM>)\n\n& mp;\#x200B;\n\n**The HIStory:** Vincent van gogh left the inn one hot summer to paint on the wheatfield...There he shot himselfwalk back to the inn he was staying and live for 30 more hours and die on his loving\ 's brother\ 's arm. \n\n& mp;\#x200B;\n\n(This is the most popular version of the story that even has been adapted to numerous books and even has it\ 's own movie)*\n\n& mp;\#x200B;\n\n**The reason why we think it\ 's not possible.**\n\n& mp;\#x200B;\n\nThe gun used and close inspection of shooting suggest it was by someone else and Vincent left the inn with a lot of materials yet there was not one single evidence that was found in the wheatfield.\n\n& mp;\#x200B;\n\n**The kids who might have killed the legend.**\n\nIn 2011, authors Steven Naifeh and Gregory White Smith published a biography, Van Gogh: The Life, in which they challenged the conventional account of the artist\ 's death. In the book, Naifeh and Smith argue that it was unlikely for van Gogh to have killed himself, noting the upbeat disposition of the paintings he created immediately preceding his death; furthermore, in private correspondence...*

Digging In

Incorrect predictions - what happened?



'Vatican officials opened two tombs in an attempt to find the remains of 15 year old missing girl Emanuela Orlandi. [Wikipedia article here](https://en.m.wikipedia.org/wiki/Disappearance_of_Emanuela_Orlandi)\n\nToday the AP reports:\n\n>Experts were looking for the remains of Emanuela Orlandi, the daughter of a Vatican clerk who failed to return home following a music lesson in Rome. Her disappearance has been the subject of wild speculation in the Italian media for years.\n\n>Exhumation work began after a morning prayer in the Teutonic Cemetery, a burial ground just inside the Vatican walls used over the centuries mainly for Church figures or members of noble families of German or Austrian origin.\n\n>Officials were expecting to find at least the bones of Princess Sophie von Hohenlohe, who died in 1836, and [Princess Carlotta Federica of Mecklenburg](https://en.m.wikipedia.org/wiki/Duchess_Charlotte_Federica_of_Mecklenburg-Schwerin), who died in 1840, but there was no trace of either.\n\n>“The result of the search was negative. No human remains or funeral urns were found,” Vatican spokesman Alessandro Gisotti said.\n\n>Gisotti said the Vatican would now examine records structural work done in the cemetery at the end of the 19th century...

Coefficients



	mnb coefs	abs
supermercado_1169293422_72276007_667x375	-13.416252	13.416252
shtml	-13.416252	13.416252
admits	-13.416252	13.416252
zealander	-13.416252	13.416252
zealand	-13.416252	13.416252
admission	-13.416252	13.416252
shultz	-13.416252	13.416252
adolescence	-13.416252	13.416252
shuglie	-13.416252	13.416252
shud	-13.416252	13.416252
adolf	-13.416252	13.416252
shsc	-13.416252	13.416252
shunting	-13.416252	13.416252
shrubs	-13.416252	13.416252
adoptive	-13.416252	13.416252



	lr coefs	abs
case	-0.678954	0.678954
mystery	-0.431941	0.431941
police	-0.394047	0.394047
article	-0.386907	0.386907
missing	-0.368426	0.368426
weekly	-0.348838	0.348838
thread	-0.346183	0.346183
murder	-0.343810	0.343810
http	-0.343355	0.343355
recently	-0.339139	0.339139
discussion	-0.338748	0.338748
offtopic	-0.337889	0.337889
listened	-0.336195	0.336195
youtube	-0.320218	0.320218
news	-0.302773	0.302773
like	0.294441	0.294441
read	-0.292156	0.292156
watched	-0.292146	0.292146
html	-0.288456	0.288456
door	0.277459	0.277459

—

Why is this useful?

Testimony

Confessions

Statements

Moving Forward

Using a pipeline and gridsearch

Collecting 'better' data
