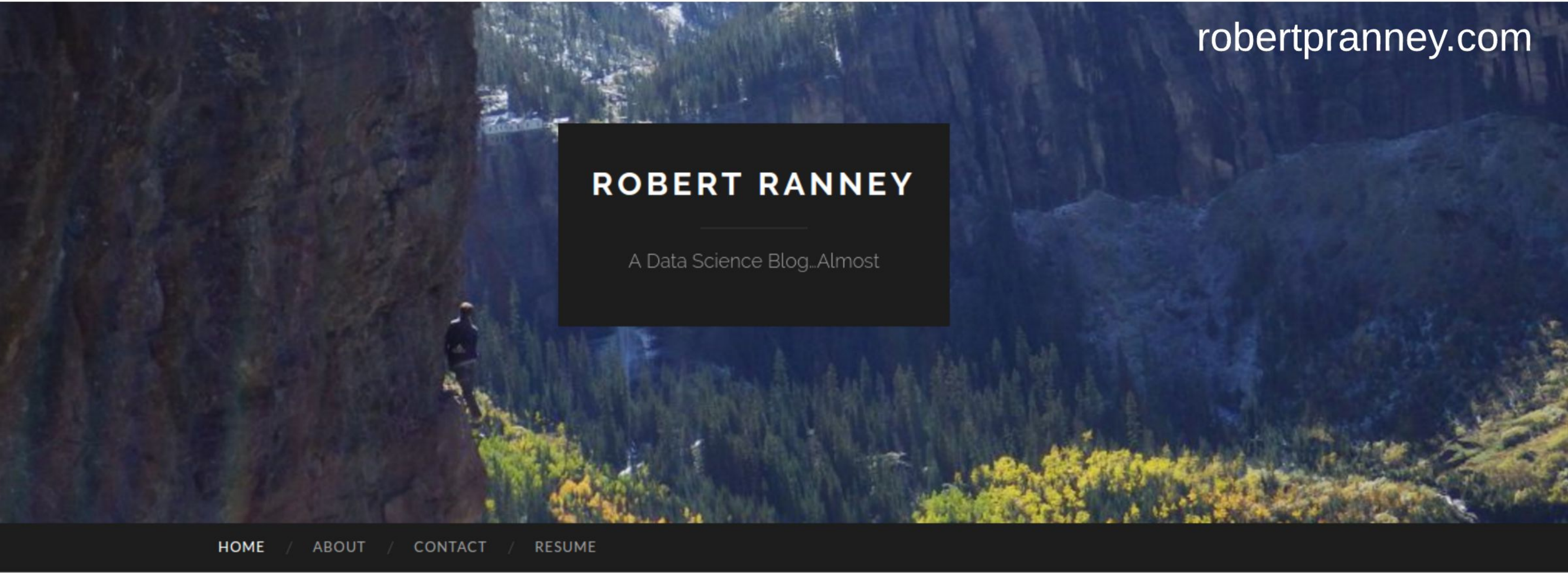
The Successful Blogger

Robert Ranney

Project Motivation

Is it possible to identifty elements that lead to both successful blogs & successful posts?

Length? Number of Images? Certain Topic? Posts on weekends?...........

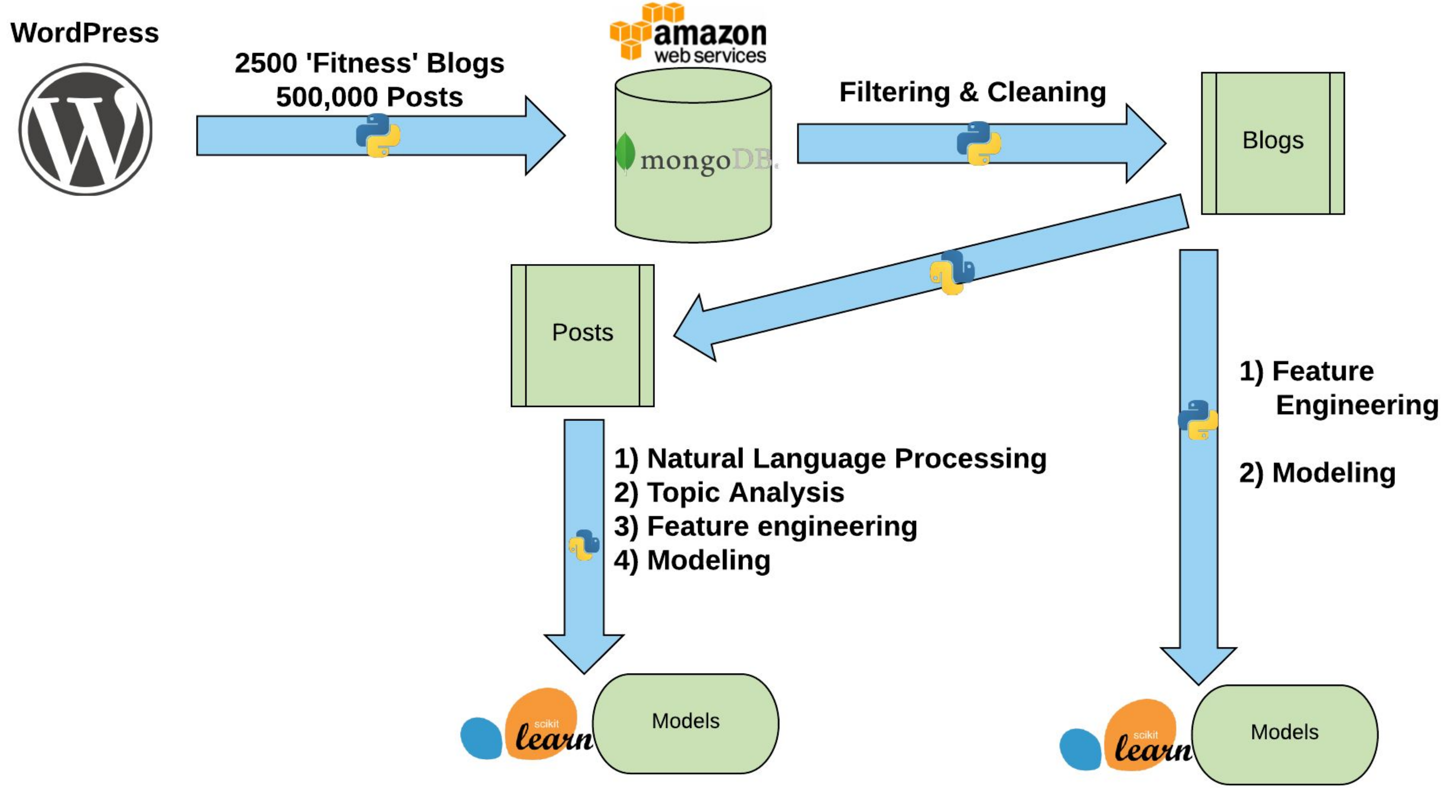


robertpranney.com

ROBERT RANNEY

A Data Science Blog...Almost

Work Flow: Executive Overview



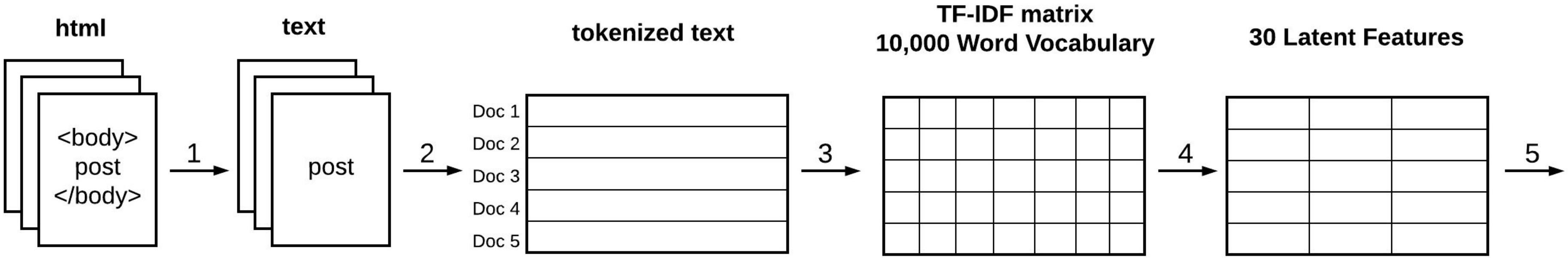
Technologies

Python
Wordpress API
AWS
MongoDB

Python Libraries

PyMongo
Requests
BeautifulSoup
Pandas
NumPy
SKLearn
cPickle
NLTK
MatPlotLib

Natural Language Processing



1) Get text of post from HTML

2) Tokenize text

3) Convert to TF-IDF Matrix

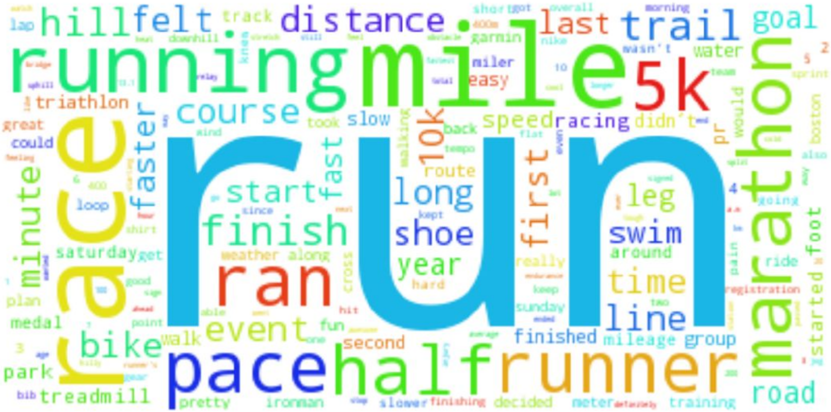
4) Reduce Dimensionality from 10,000 to 30 with NMF

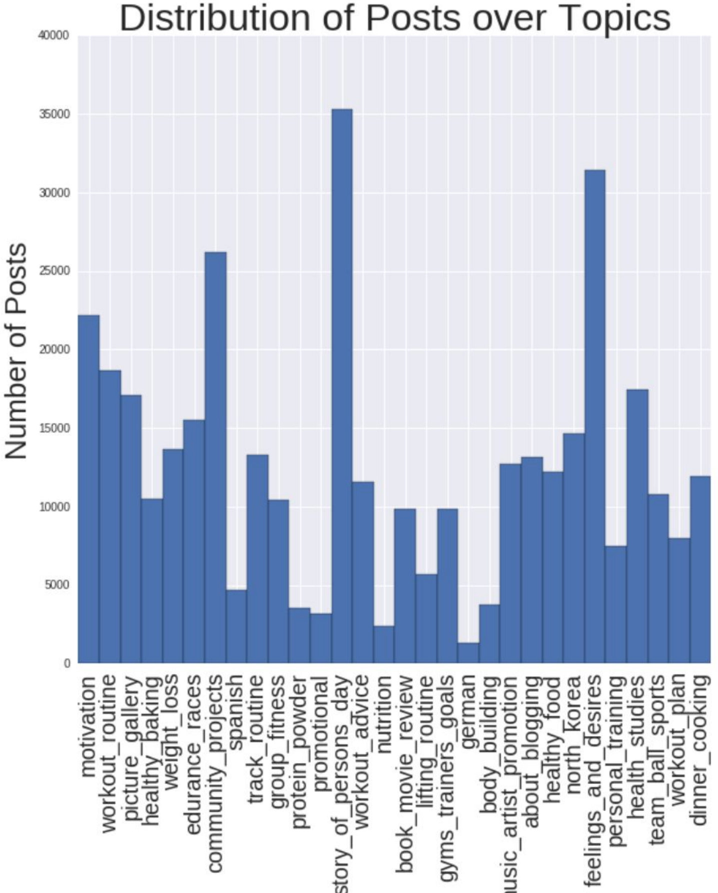
5) Use latent features to identify topics
Use latent features for modeling

Topic Analysis

Topic Analysis:
30 Latent features were manually assigned names based on most important words

Example: Latent Feature 5
Endurance Races





Modeling Post Success

Defining a successful post:

Number of likes + Number of comments

Very Skewed distribution:

Binned into 4 groups

Groupings:

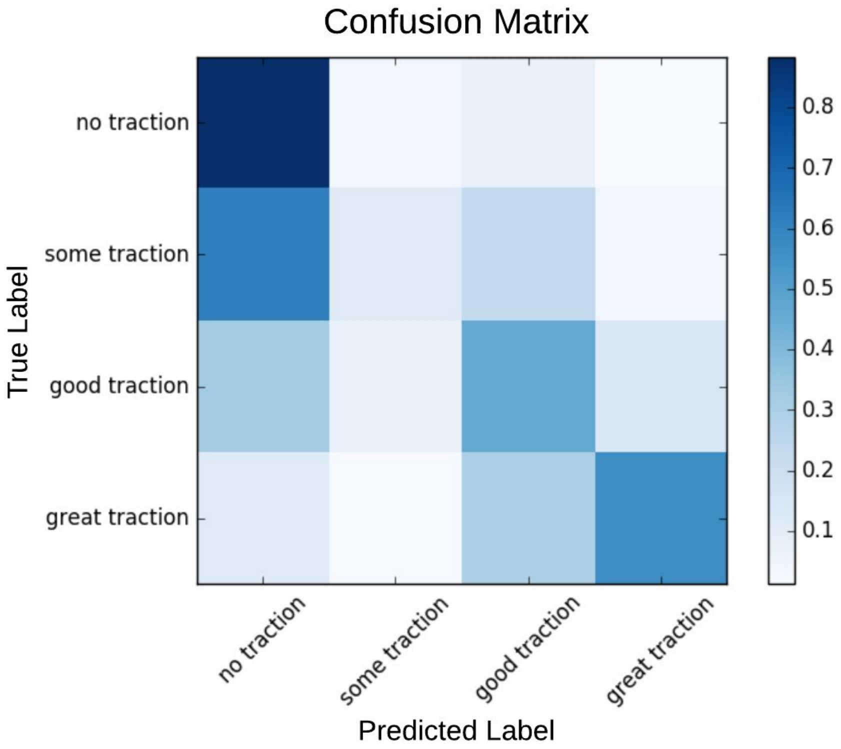
No traction (50%), Some traction(20%)
Good traction(15%), Great traction(15%)

Feature Set:

30 Engineered features &
30 latent features from text

Best Model:

Random Forest Classifier
(61% Accuracy)



Modeling Post Success

Unsurprising Results

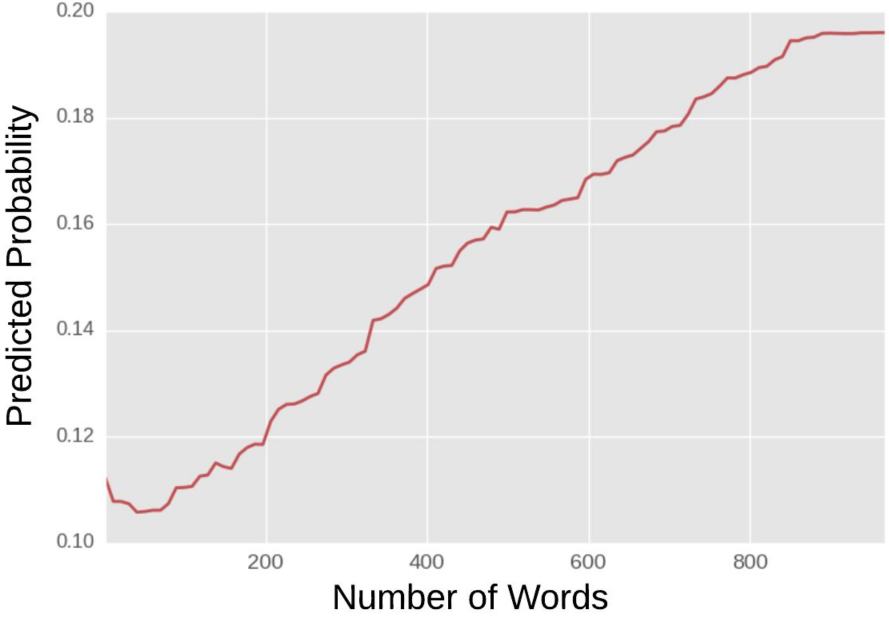
- **Longer posts do better (to the right)**
  At least up to 800 words, data doesn't say much past this point

- **Post from more recent years do better**
  i.e. 2016 blogs posts vs 2012

- **More images is better**

- **More links is better**

Probability for being in most successful group



Modeling Post Success

More Surprising Results

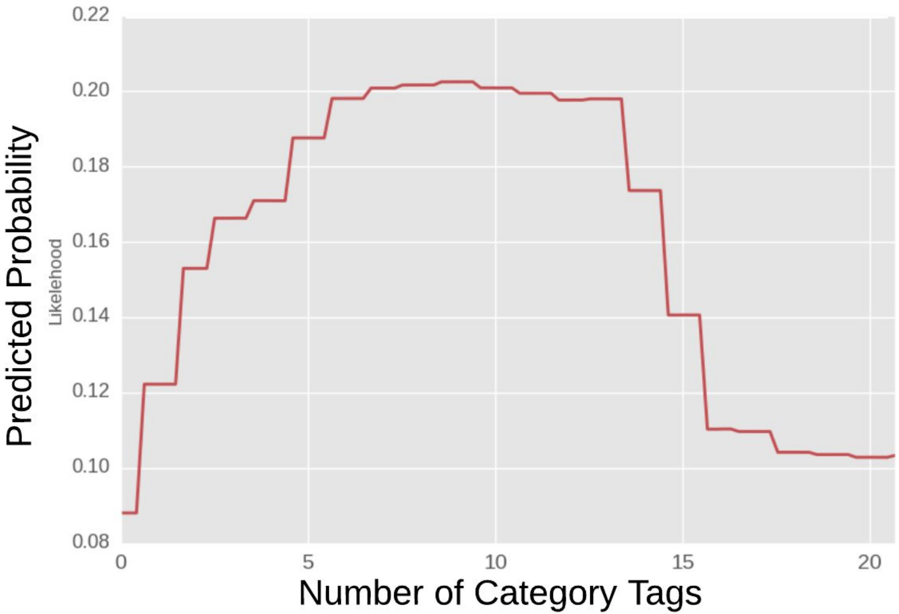
- **More tags only help to a point (to the right)**

  Used to list posts in relevant topics

- **30 latent features contribute very little**

- **Weekend vs. Weekday = no effect**

Probability for being in most successful group



Modeling Blog Success

Defining a successful blog:

Number of subscribers

Feature Set:

~30 Features

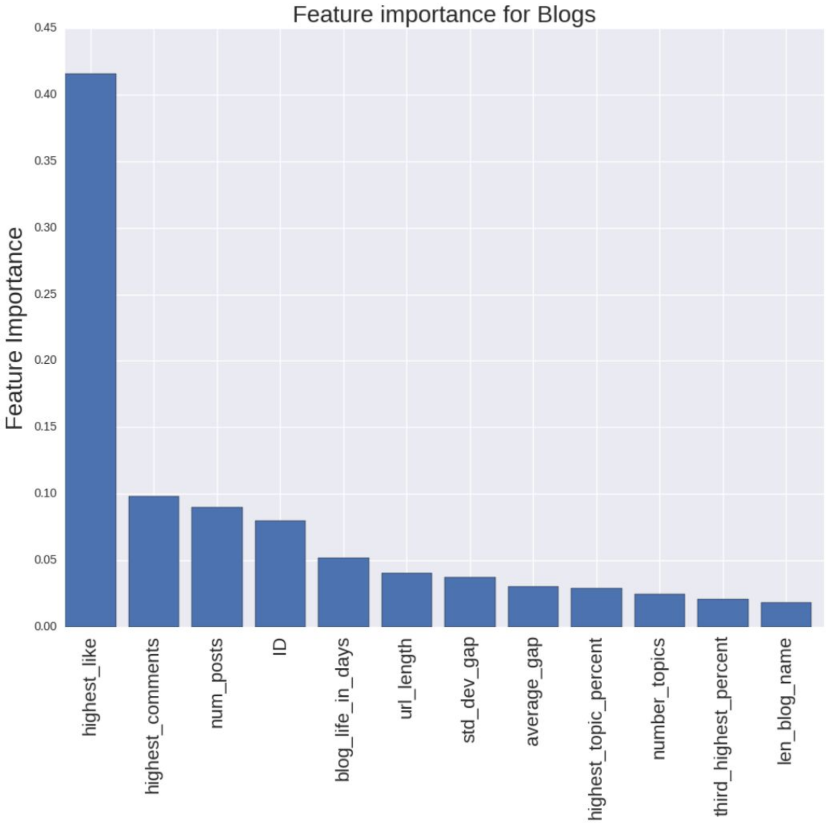
Best Model:

Gradient Boosting Regressor

Modeling Blog Success

Unsurprising Results

- **A highly liked post = More subscribers**
- **A highly commented post = More subscribers**
- **More posts = More subscribers**
- **Longer blog existence = More subscribers**

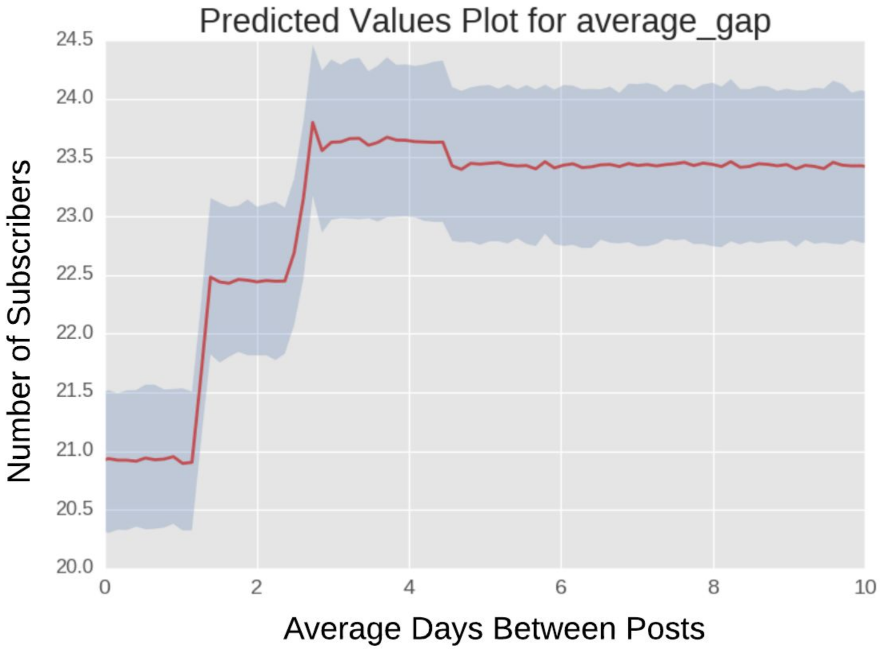


Modeling Blog Success

More Surprising Results

- **Days between posts > 3:**
  More subscribers (shown at right)

- **Paying to remove wordpress.com from url:**
  Very little affect



Future Things to Work On

- Find a less naive way to filter out irrelevant blogs

- Incorporate sentiment analysis on comments

Latent Feature 18: German

Contact Info

Email: robertpranney@gmail.com

Github: www.github.com/robertpranney

Linkedin: https://www.linkedin.com/in/robertpranney

Blog: https://robertpranney.com/

Thanks for your time