Classification Modeling of Subreddits

Post Comparison of:

A SONG OF ICE AND FIRE

GAME OF THRONES

Jeremy Opacich
Data Scientist

Words are wind.

– George RR Martin

Our Goal:

To catch wind

...and discern which subreddit post it belongs to

Project Overview

The Data:

- The posts
- Web Scraping
- Initial Observations

Initial Modeling:

- Stop_words
- Parameter Tuning
- Ensemble Models

spaCy:

- Info
- Modeling
- Next Steps

Collecting our Data

The Posts

r/gameofthrones –
Game of Thrones

r/asoiaf –
A Song of Ice and Fire

Web Scraping

- 7 days of scraping
- 2,000 posts/scrape
- Nearly 40% contained no text

Initial Concerns

Limited Postings ->

- 200/day collected
- Photos and Videos

Unbalanced Data ->

- Game of Thrones - Image heavy

Similar Content ->

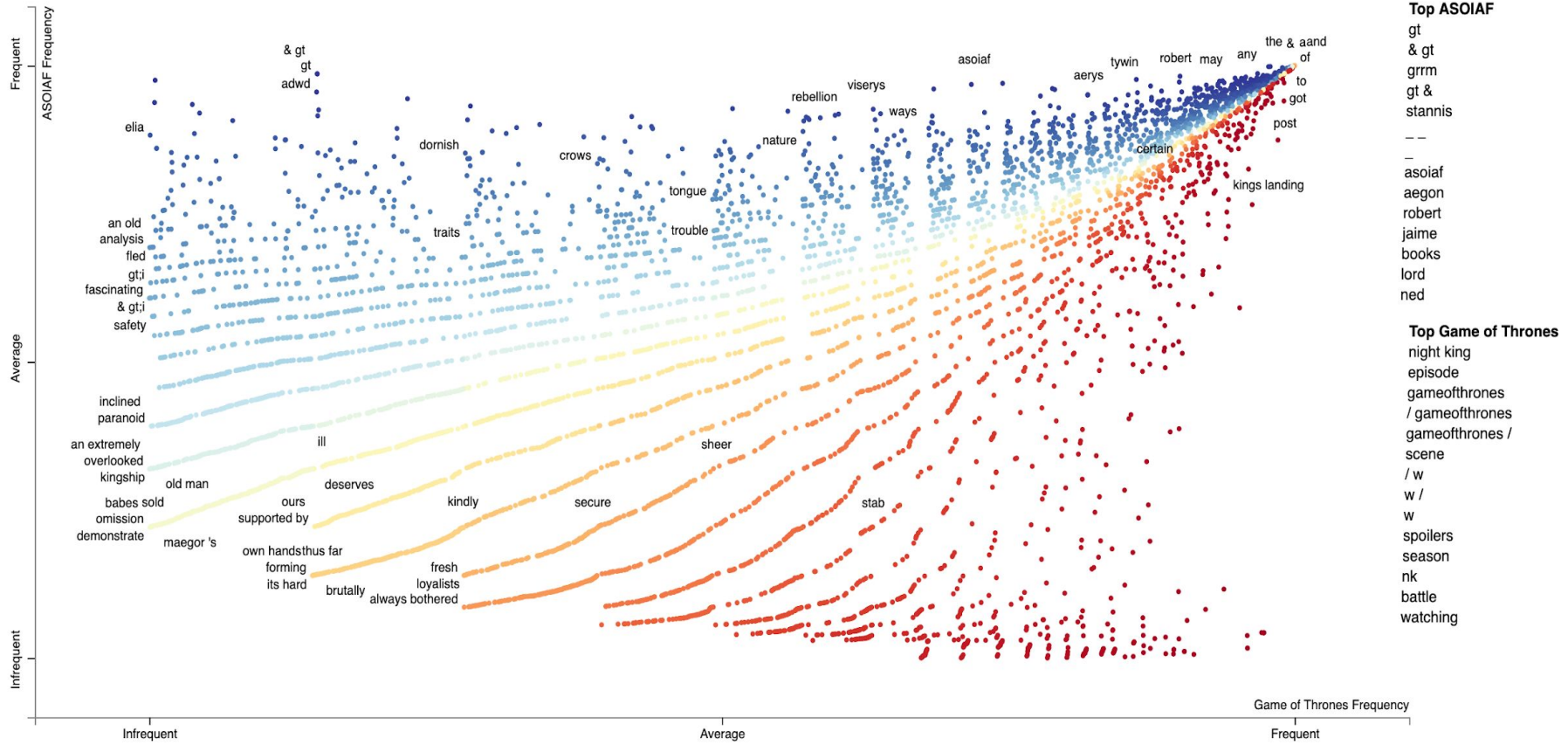
- Same characters, places
- Crossover discussion

NLP

Initial Modeling

And Visualization

Word Frequency per Subreddit

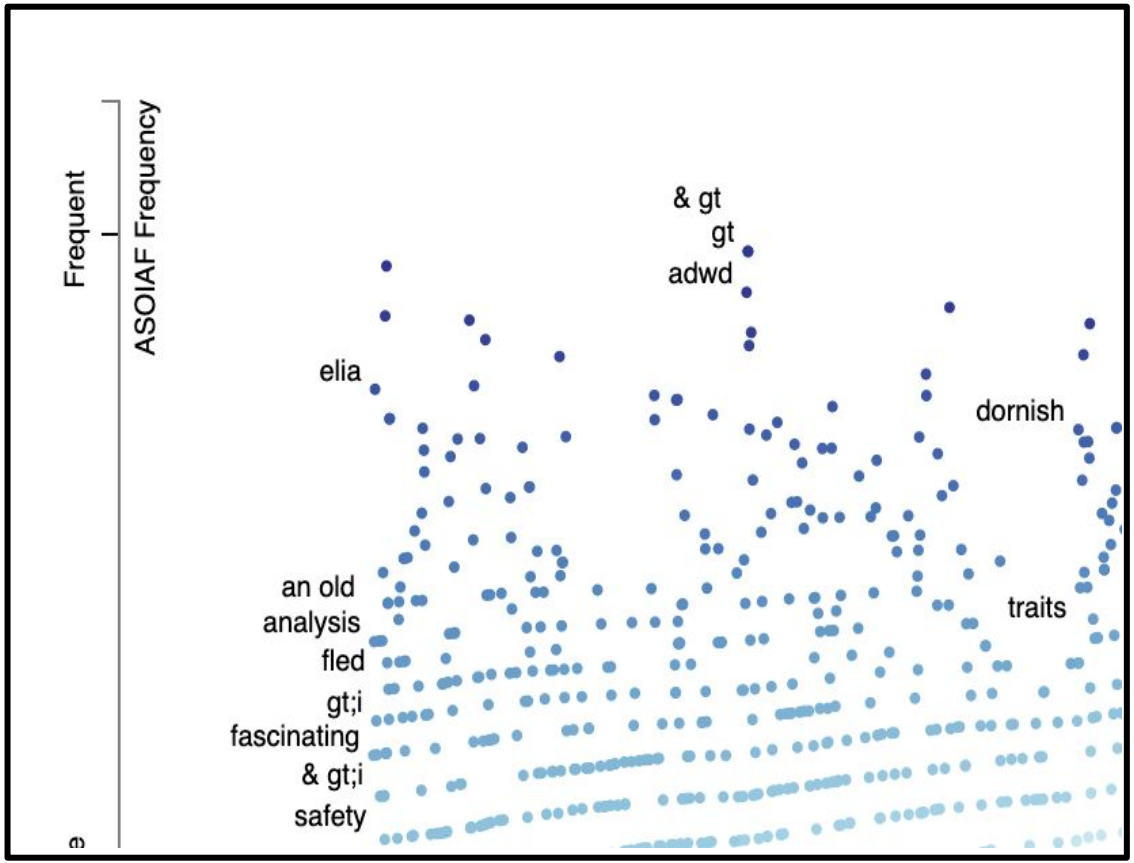


Top ASOIAF

gt
& gt
grrm
gt &
stannis
_ _
_
asoiaf
aegon
robert
jaime
books
lord
ned

Top Game of Thrones

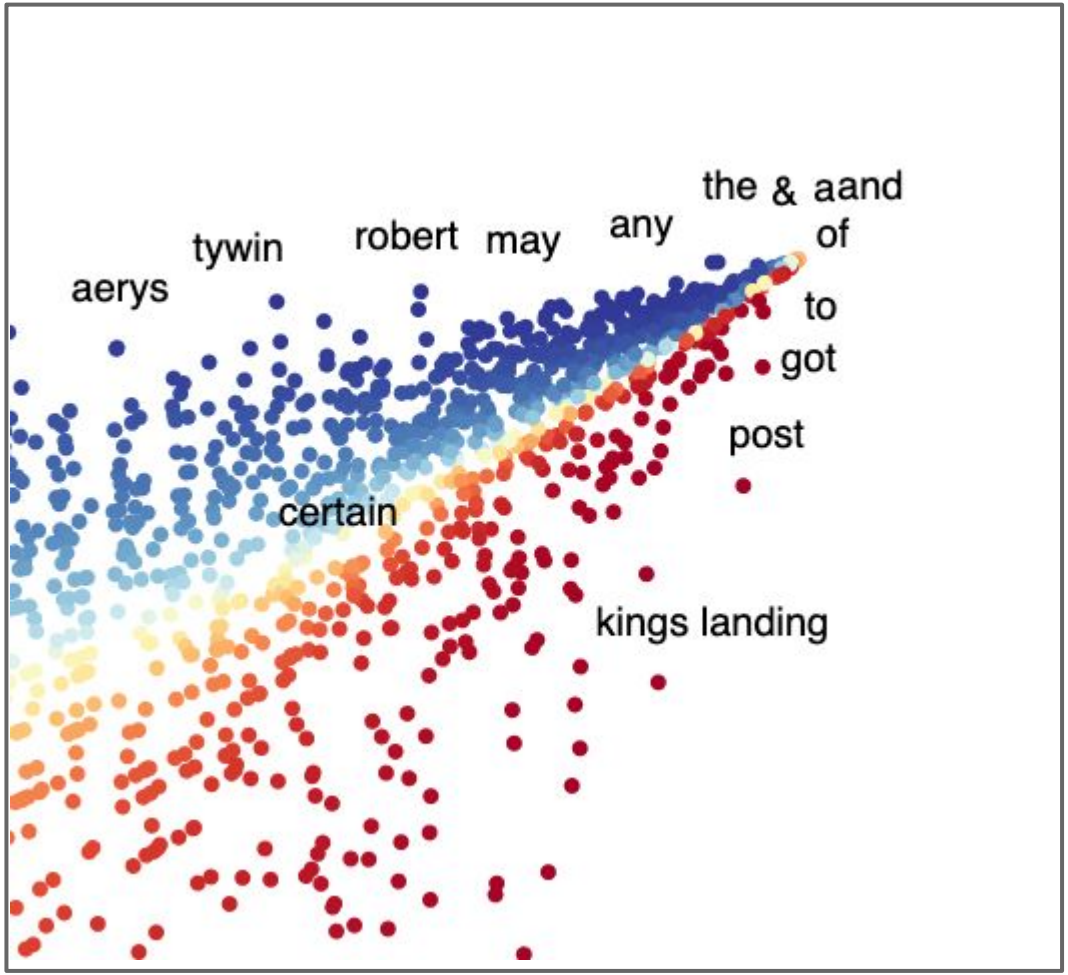
night king
episode
gameofthrones
/ gameofthrones
gameofthrones /
scene
/ w
w /
w
spoilers
season
nk
battle
watching



Most Frequent ASOIAF

Key Words

- Elia
- analysis
- adwd
- fled
- fascinating



Most Frequent Shared Words

- GoT
- Robert
- Kings Landing
- Tywin
- the, to, post

Building out our Model

- Baseline Model
  - **60.9%** – ASOIAF
- Scattertext / Stop Words
  - Frequency scaled words
  - 14,000 parses
- Gridsearch
  - Multinomial Bayes
  - Logistic Regression
  - SVC
  - RandomForest
- Ensemble
  - VotingClassifier
- TF-IDF & CountVectorizers

Over 3600 fits

Best Model: 78.2% Accuracy

CountVectorizer and Multinomial Bayes

Modeling with spaCy

spaCy?

Industrial Strength NLP

Non-destructive tokenization

Named entity recognition

Part of speech tagging

Sentence Segmentation

Text classification

Our spaCy Model

Convolutional Neural Network

- Black Box
- Learns via:
- Gradient Errors
  - How incorrect was the prediction?
- Minibatches and Compounding

Dropout Rate

- How much data do we cycle out
- Learns and adjust
- 20%

Minibatches

- Max 64 tokens for Text Classification
- 4 (start), 32(stop)

How did it do?

With our lowest gradient, we finished at:

91.9%

13.7% increase!

Next steps

Go further into spaCy!

- Investigate the Adam solver
- Reevaluate
  - Minibatches and compounding
  - Gradients and loss function
  - Dropout rate and linear decay
- Learn to use L2 regularization in the model
- Get more data

Any Questions?

Courtesy of HBO