Version 1.0

Natural Language Processing Investigation:

Data Science | Business Analytics

Graham Lim,
Sharmaine Cheung,
Stuart Daw

Interested in a career in data but confused?

We’re here to help!

Our goal:

Help YOU to identify -
i. Unique differences between data science and analytics
ii. Recommend tools based on your desired role

How? Investigate posts from 2 subreddits: **r/datascience** and **r/analytics**

Bonus! You get to observe the work of a data scientist.

Data process

1

Scraping

- Reddit api
- OOD Posts and Comments
- Key fields: title, selftext, subreddit
- Over 10,000 data points

2

Data Cleaning

- Null values
- Duplicate posts
- Binarize target variable
- Feature engineering: title + selftext

3

EDA

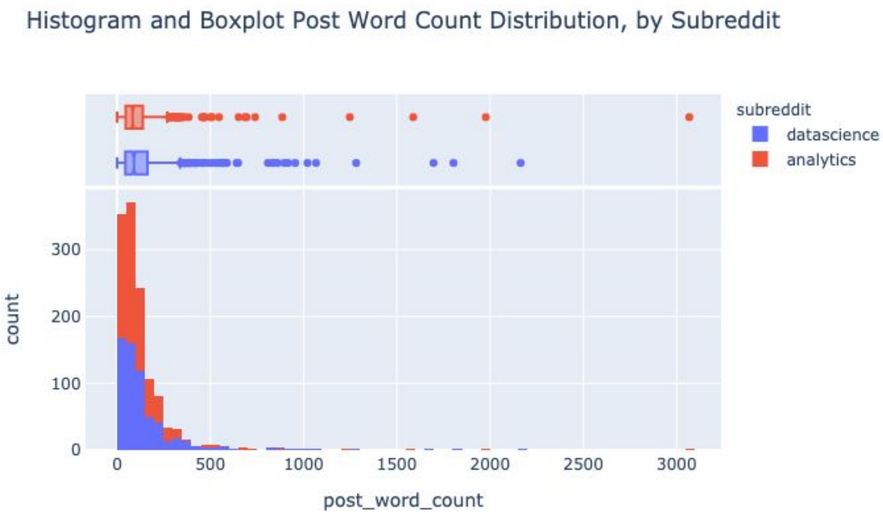
- Word count in posts, Comments
- Frequently occurring words

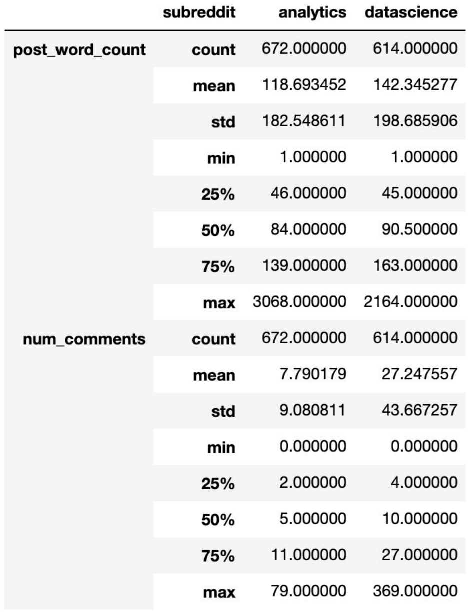
4

Modelling

- Lemmatizing, Stopwords
- CountVectorizer & TfidfVectorizer
- Logistic Regression, Multinomial NB, SVM

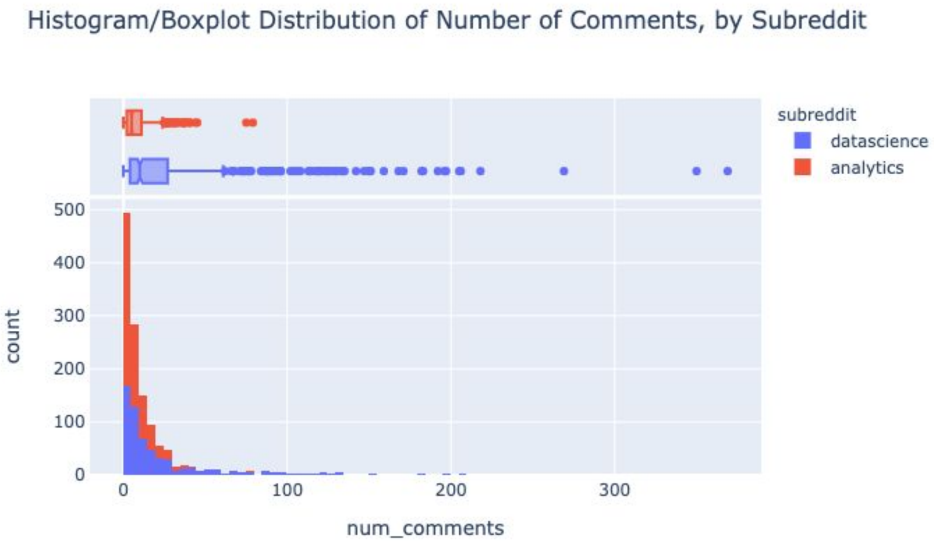
Summary Stats: Word Count in Posts



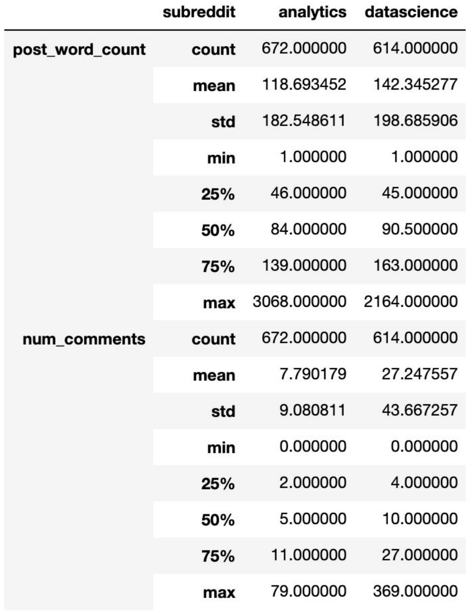


	subreddit	analytics	datascience
post_word_count	count	672.000000	614.000000
	mean	118.693452	142.345277
	std	182.548611	198.685906
	min	1.000000	1.000000
	25%	46.000000	45.000000
	50%	84.000000	90.500000
	75%	139.000000	163.000000
	max	3068.000000	2164.000000
num_comments	count	672.000000	614.000000
	mean	7.790179	27.247557
	std	9.080811	43.667257
	min	0.000000	0.000000
	25%	2.000000	4.000000
	50%	5.000000	10.000000
	75%	11.000000	27.000000
	max	79.000000	369.000000

Summary Stats: Number of Comments



Histogram/Boxplot Distribution of Number of Comments, by Subreddit



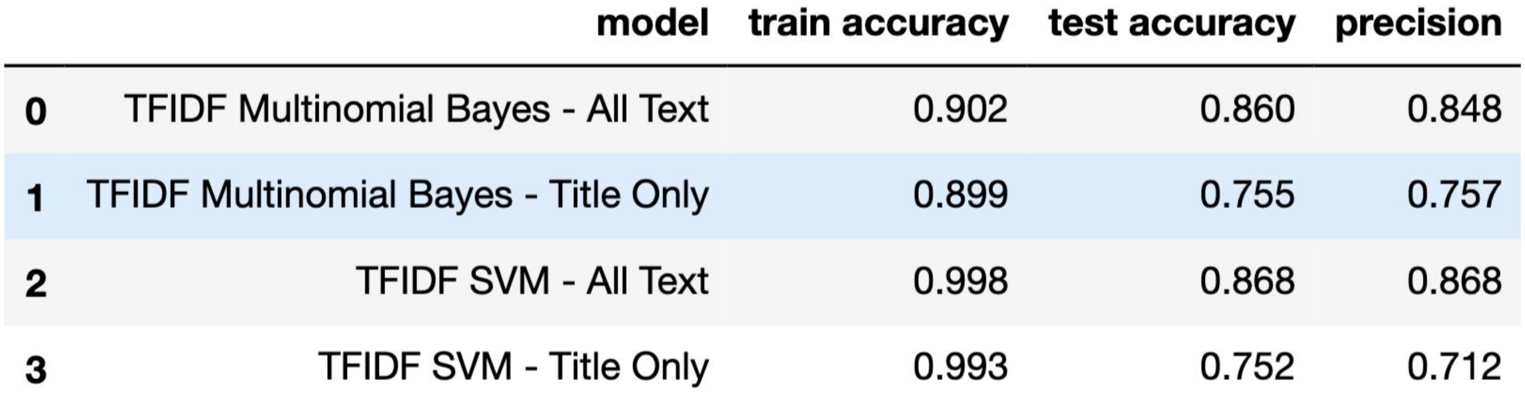
	subreddit	analytics	datascience
post_word_count	count	672.000000	614.000000
	mean	118.693452	142.345277
	std	182.548611	198.685906
	min	1.000000	1.000000
	25%	46.000000	45.000000
	50%	84.000000	90.500000
	75%	139.000000	163.000000
	max	3068.000000	2164.000000
num_comments	count	672.000000	614.000000
	mean	7.790179	27.247557
	std	9.080811	43.667257
	min	0.000000	0.000000
	25%	2.000000	4.000000
	50%	5.000000	10.000000
	75%	11.000000	27.000000
	max	79.000000	369.000000

Initial Model for Accuracy

- **Baseline Score - Multinomial Bayes - All Text - Accuracy of 83%**
- Our Production Model Tests

Initial Model for Accuracy

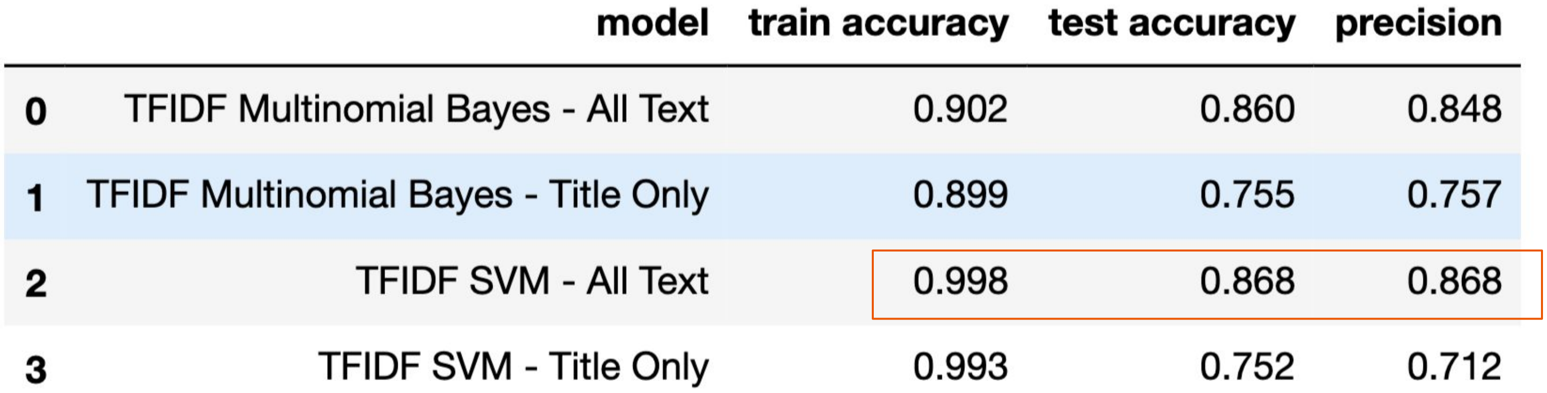
- **Baseline Score - Multinomial Bayes - All Text - Accuracy of 83%**
- Our Production Model Tests



	model	train accuracy	test accuracy	precision
0	TFIDF Multinomial Bayes - All Text	0.902	0.860	0.848
1	TFIDF Multinomial Bayes - Title Only	0.899	0.755	0.757
2	TFIDF SVM - All Text	0.998	0.868	0.868
3	TFIDF SVM - Title Only	0.993	0.752	0.712

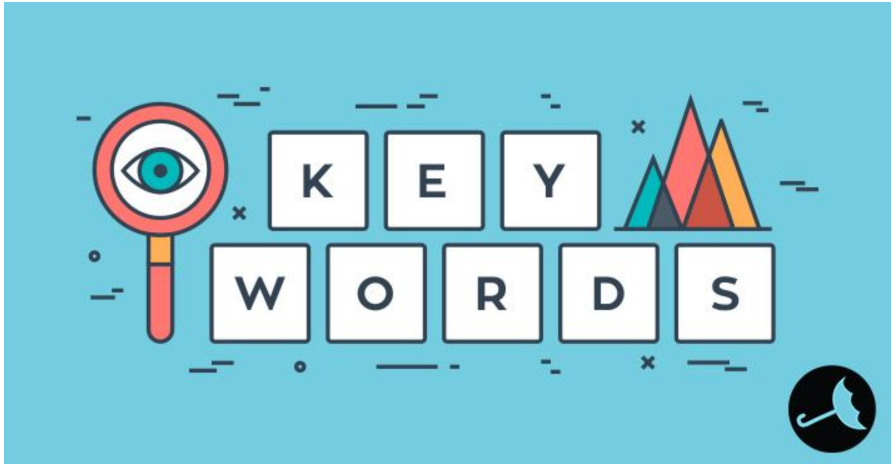
Initial Model for Accuracy: SVM

- Baseline Score - Multinomial Bayes - Accuracy of 83%
- Our Production Model Tests



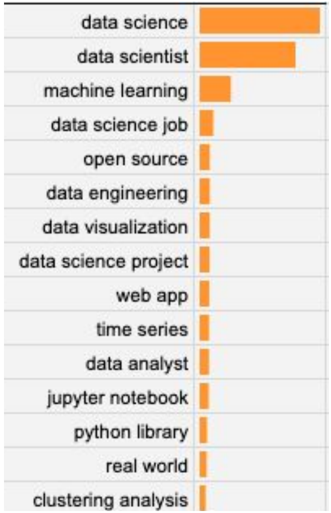
	model	train accuracy	test accuracy	precision
0	TFIDF Multinomial Bayes - All Text	0.902	0.860	0.848
1	TFIDF Multinomial Bayes - Title Only	0.899	0.755	0.757
2	TFIDF SVM - All Text	0.998	0.868	0.868
3	TFIDF SVM - Title Only	0.993	0.752	0.712

Accurate Prediction and Finding MORE

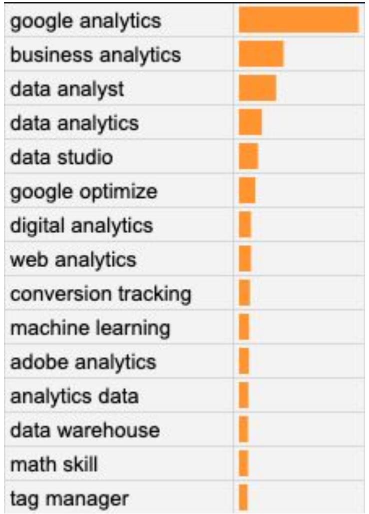


Key Words

Data Science



Analytics



Recommendations and Insights

Data Science

01 | Interest in building and creating

02 | Data Engineering

03 | Web Applications

04 | Technical skills: Python, Jupyter

Analytics

01 | Interest in using software to quickly gain actionable insights

02 | Google Analytics, Optimise, Data Studio

03 | Data Warehouse

04 | Technical skills : Online tools

Next Steps

More

Subreddits

deeplearning
dataisbeautiful
dataanalysis

Other

Forums

Towardsdatascience.com
Quora

Scrape

LinkedIn

Job posts
Profile and skills

Questions?