r/datascience
vs
r/machinelearning

By Dan Ovadia

Problem Statement

Problem: Data Science is a new field that is forming it's identity in the midst of many fields: Math/statistics, SWE, data analysis, business intelligence, predictive modeling.

Goal: Understand vernacular and topics in the DS and ML communities.

Inference: Examine the colloquial distinctions between the DS and ML communities.

Outcome: Created a variety of classifying models that can identify attributes in text that reflect DS related content vs machine learning related content and make classification predictions on some inputted text data.

Introduction

r/Datascience 

- 217,371 Subscribers (as of 04/24/2020)
- Founded Sat, August 6, 2011

- A place for data science practitioners and professionals to discuss and debate data science career questions.

r/MachineLearning 

- 1,017,245 Subscribers (as of 04/24/2020)
- Founded Wed, July 29, 2009

- MOOCs, Books, DL Resources
- Math Resources, Languages
- Datasets
- ML Research

Methodology

Scraped 40k subreddit posts including both the titles and selftext.

Parsed out URLs and examined domains commonly used by each community.

NLP model utilizing TFIDF Vectorizing and Logistic Regression to predict whether a sentence reflects the DS community or the ML community

Model

NLP Features:

- Merged Text (subreddit post's Title and Selftext)
- Title URLs
- Selftext URLs

TFIDF Vectorizer

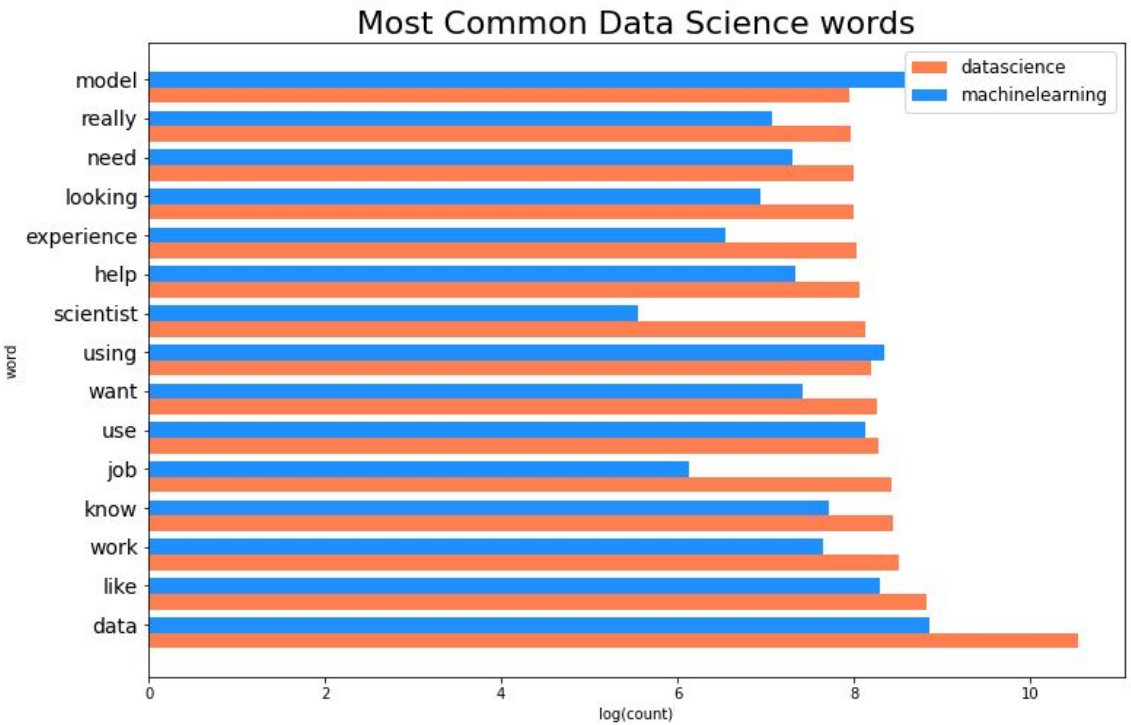
- Custom stopwords combining sklearns default, common terms, website terms
- Merged Text (173 features) | Title URLs (12 features) | Sefltext URLs (12 features)

Logistic Regression

- Default Parameters
  - Train accuracy score = 73.3%
  - Test accuracy score = 73.6%

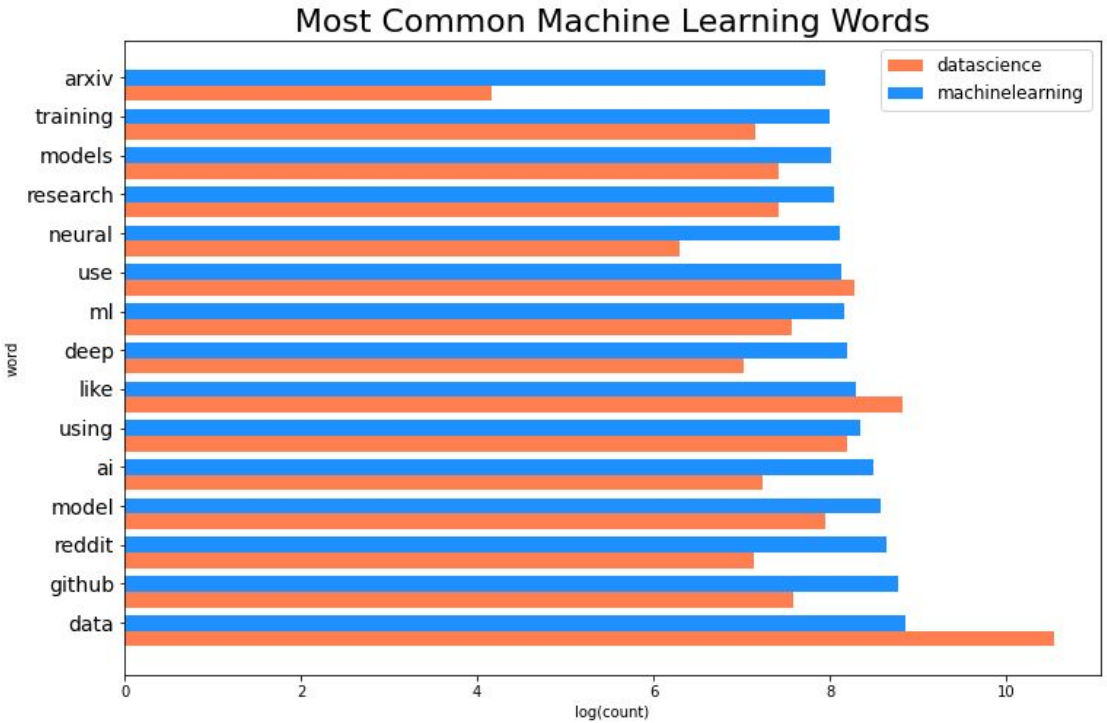
Word Frequencies - Data Science

- Aspiring Field
  - Job / career/ work
- Concepts
  - Data / model
- Requests
  - Looking
  - Work,
  - Like
  - Really
  - Want
  - Experience,
  - Need
  - Help?



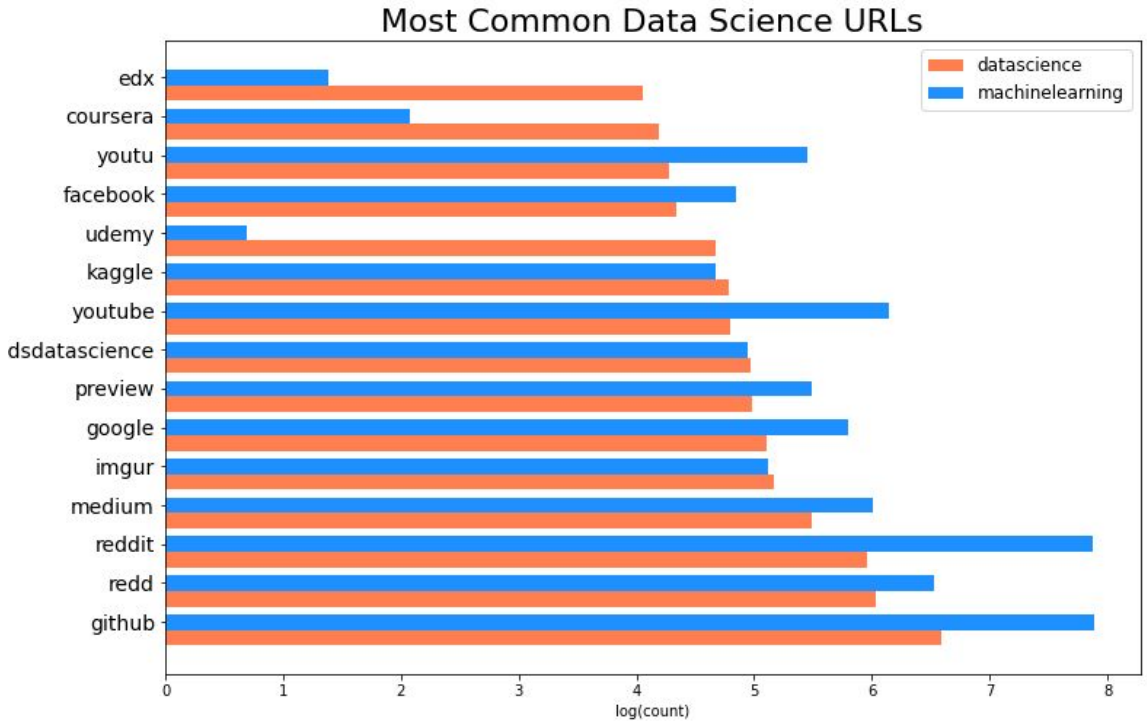
Word Frequencies - Machine Learning

- Focused on tools
  - Github / arXiv / reddit
- Resources
  - Website links
- Concepts
  - Neural
  - ML
  - Deep Learning
  - Training / Modeling



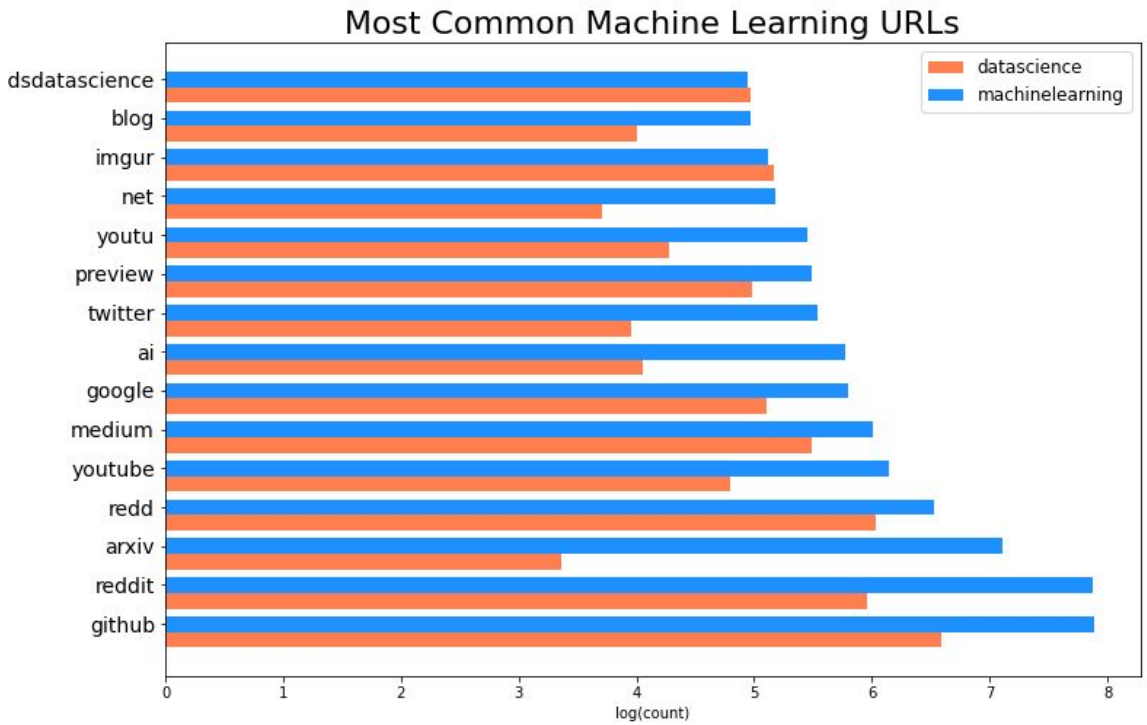
URL Frequencies - r/datascience

- Resources
  - Coursera
  - Edx
  - Udemy
  - Kaggle
  - Github
  - Reddit
- Other
  - Facebook
  - Youtube



URL Frequencies - r/machinelearning

- Resources
  - Dsdatascience
  - Medium
  - Reddit
  - Github
  - arXiv
- Others
  - Twitter
  - Youtube



T-tests across common top 100 terms

- Our T-tests revealed some statistically significant values.
- Machine learning was more unique
- Data science was more colloquial

r/datascience

	pvalue	statistic
really	1.210827e-113	-22.729723
use	5.681134e-72	-17.977118
using	4.739231e-46	-14.264408
think	4.061204e-39	-13.098248
way	5.379948e-34	-12.166627
computer	4.547270e-33	-11.990457
know	6.311123e-24	-10.093454
help	1.270140e-12	-7.099815
thanks	1.160639e-10	-6.446117
want	2.512089e-10	-6.327871
work	4.949131e-08	-5.454183
training	2.346760e-07	-5.170437
lot	6.992263e-05	-3.976969
data	1.945658e-03	-3.098611
questions	1.272819e-01	-1.524941

r/machinelearning

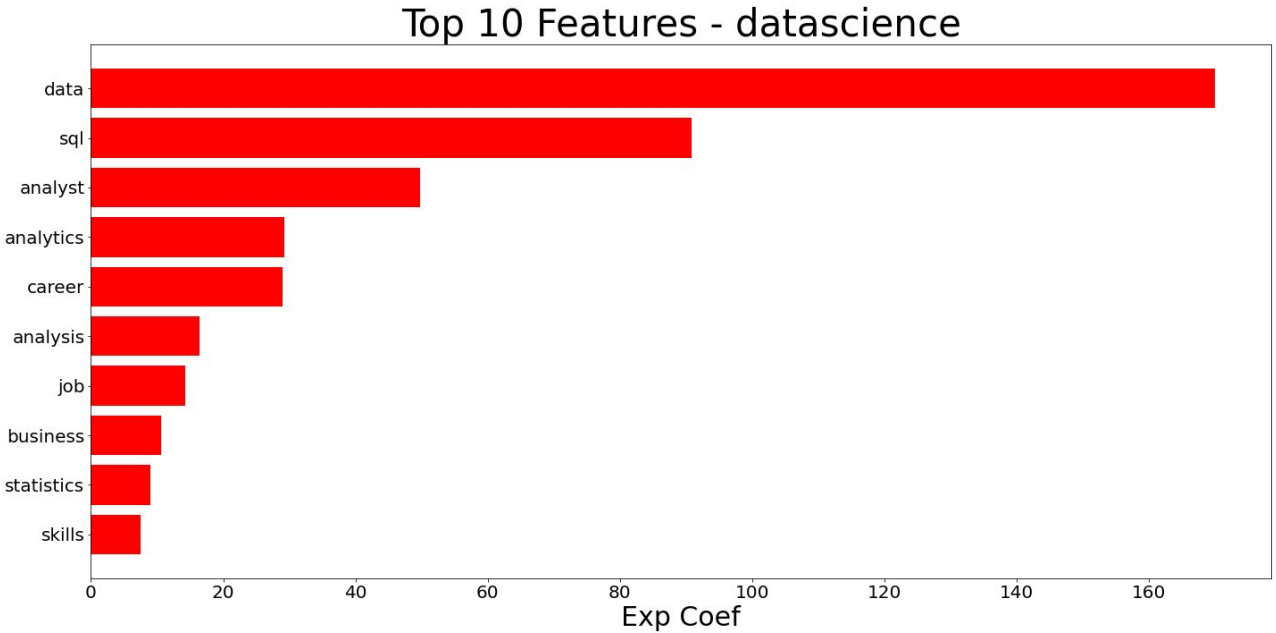
	pvalue	statistic
ai	0.000000e+00	65.393146
dataset	1.268217e-143	25.622278
ml	5.504446e-119	23.270656
model	3.448299e-115	22.887482
research	4.546729e-111	22.464524
different	3.336559e-108	22.165730
reddit	3.424985e-91	20.304278
based	1.874666e-87	19.872502
github	2.377845e-85	19.625502
code	7.184177e-81	19.088973
google	4.242441e-69	17.603456
better	5.772906e-49	14.727590
set	2.227273e-45	14.155487
make	1.116283e-33	12.106664
problem	7.662089e-32	11.753371

Beta Coefficients - Data Science

- Data and SQL
- Business, Analytics
- Job and Career
- Courses and Degrees

Theme revolving data analysis and business data solutions.

Work experiences.



Predictions

"Describe generalized linear models. Fit Poisson and Gamma regression models in statsmodels. Interpret coefficients from Poisson and Gamma regression models. Describe iteratively reweighted least squares."

- DSI-11 Lecture 6.06

Predictions

"Describe generalized linear models. Fit Poisson and Gamma regression models in statsmodels. Interpret coefficients from Poisson and Gamma regression models. Describe iteratively reweighted least squares."

- DSI-11 Lecture 6.06 **(Predicted (55%): r/datascience)** Tim Book

Predictions

"Describe generalized linear models. Fit Poisson and Gamma regression models in statsmodels. Interpret coefficients from Poisson and Gamma regression models. Describe iteratively reweighted least squares."

- DSI-11 Lecture 6.06 **(Predicted (55%): r/datascience)** Tim Book

“Understand the intuition behind the KNN algorithm Describe the Bias / Variance tradeoff using hyper-parameters of KNN Implement KNN with sklearn”

- DSI-11 Lecture 4.02

Predictions

"Describe generalized linear models. Fit Poisson and Gamma regression models in statsmodels. Interpret coefficients from Poisson and Gamma regression models. Describe iteratively reweighted least squares."

- DSI-11 Lecture 6.06 **(Predicted (55%): r/datascience)** Tim Book

“Understand the intuition behind the KNN algorithm Describe the Bias / Variance tradeoff using hyper-parameters of KNN Implement KNN with sklearn”

- DSI-11 Lecture 4.02 **(Predicted (67%): r/machinelearning)** Riley Dallas

Predictions

"Describe generalized linear models. Fit Poisson and Gamma regression models in statsmodels. Interpret coefficients from Poisson and Gamma regression models. Describe iteratively reweighted least squares."

- DSI-11 Lecture 6.06 **(Predicted (55%): r/datascience)** Tim Book

“Understand the intuition behind the KNN algorithm Describe the Bias / Variance tradeoff using hyper-parameters of KNN Implement KNN with sklearn”

- DSI-11 Lecture 4.02 **(Predicted (67%): r/machinelearning)** Riley Dallas

"After the program, you'll want to continue developing your skills. Being comfortable with documentation and being confident in your ability to read something new and decide whether or not it is an appropriate method for the problem you're trying to solve is incredibly valuable."

- DSI-11 Lab 5.02

Predictions

"Describe generalized linear models. Fit Poisson and Gamma regression models in statsmodels. Interpret coefficients from Poisson and Gamma regression models. Describe iteratively reweighted least squares."

- DSI-11 Lecture 6.06 **(Predicted (55%): r/datascience)** Tim Book

“Understand the intuition behind the KNN algorithm Describe the Bias / Variance tradeoff using hyper-parameters of KNN Implement KNN with sklearn”

- DSI-11 Lecture 4.02 **(Predicted (67%): r/machinelearning)** Riley Dallas

"After the program, you'll want to continue developing your skills. Being comfortable with documentation and being confident in your ability to read something new and decide whether or not it is an appropriate method for the problem you're trying to solve is incredibly valuable."

- DSI-11 Lab 5.02 **(Predicted (84%): r/datascience)** Matt Breams

Aside

- My Brand Statement
  - “Data engineer with a passion for putting data in context and building robust data workflows for meaningful interpretation. I am fascinated by the insurance industry's methods of carving out and assessing risk, and excited by the data science community's ability to manage and analyze data to provide insights in responding to and mitigating risk.

Aside

- My Brand Statement
  - “Data engineer with a passion for putting data in context and building robust data workflows for meaningful interpretation. I am fascinated by the insurance industry's methods of carving out and assessing risk, and excited by the data science community's ability to manage and analyze data to provide insights in responding to and mitigating risk.
  - **(Predicted (98.3%): r/datascience)**

Conclusions

r/Datascience

- Still figuring out it’s fundamentals and foundations, at least colloquially.
- Business oriented and analysis oriented
- They have strict rules and regulations, such as no self promotion, and no videos.

r/MachineLearning

- An established community with resources for any contextual goals
- Very research oriented
- The about me, focuses on “beginners” and provides many resources