Subreddits:
DataIsBeautiful
vs
TodayILearned

—

Dmitriy Pavlov
12.21.2018
GA - Data Science Immersive

Overview

- Reddit Background
- Data Science Problem
- Exploratory Data Analysis
- Modeling
- Interpretation

Reddit is an American social news aggregation, web content rating, and discussion website

TodayILearned - You learn something new every day; what did you learn today? Submit interesting and specific facts about something that you just found out here.

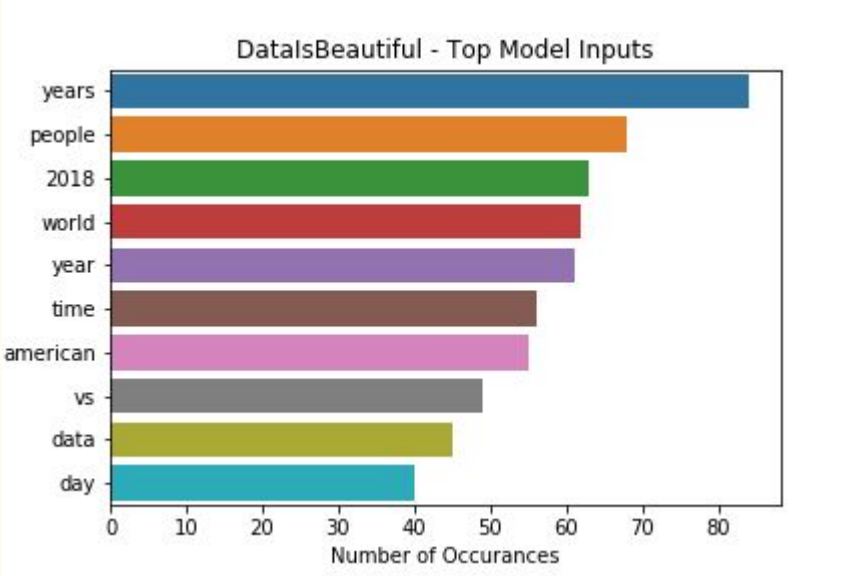
DataIsBeautiful - A place for visual representations of data: Graphs, charts, maps, etc. DataIsBeautiful is for visualizations that effectively convey information. ***Aesthetics are an important part of information visualization, but pretty pictures are not the aim of this subreddit.***

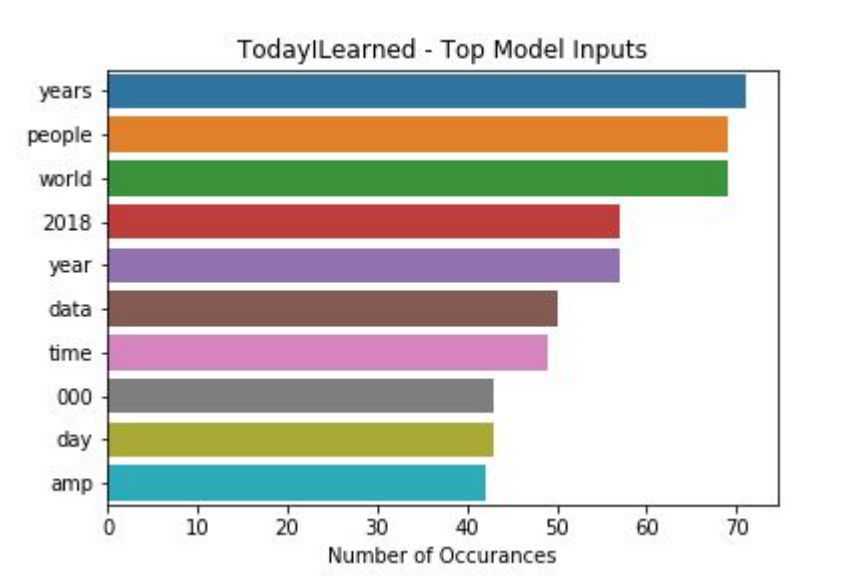
Data Science Problem:

Are we able to use Natural Language Processing to teach a model to distinguish between two popular subreddits, DataIsBeautiful and TodayILearned, only based on the text in the titles of the posts?

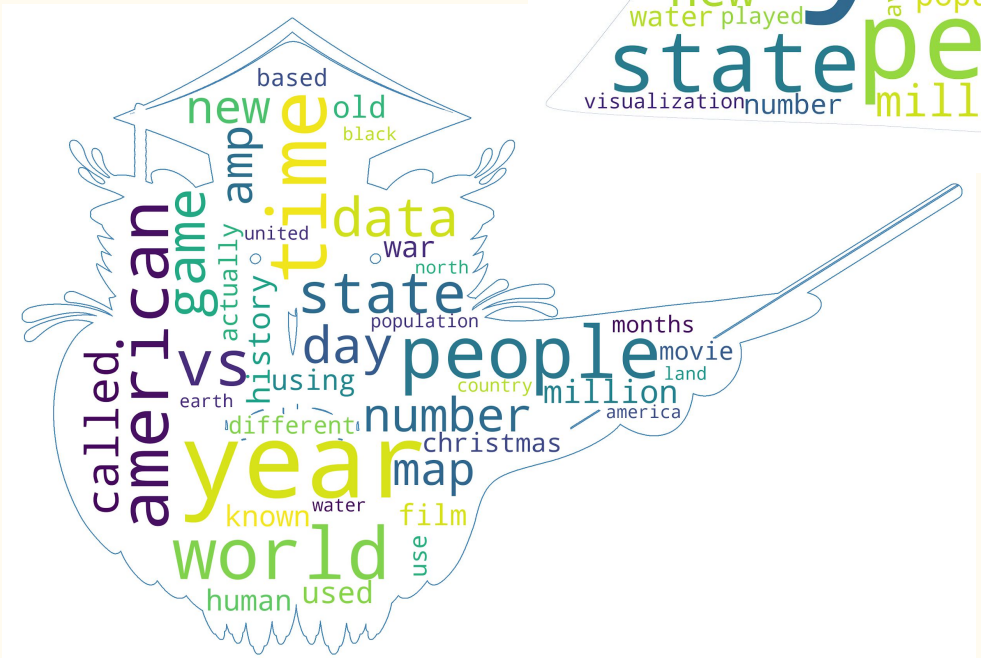
Exploratory Data Analysis

EDA - The two subreddits are very similar with most of the top features overlapping each other

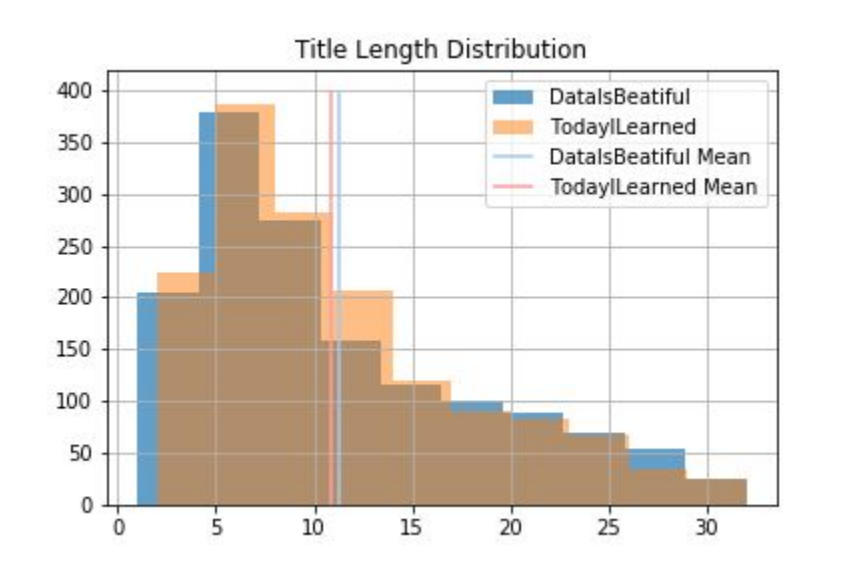




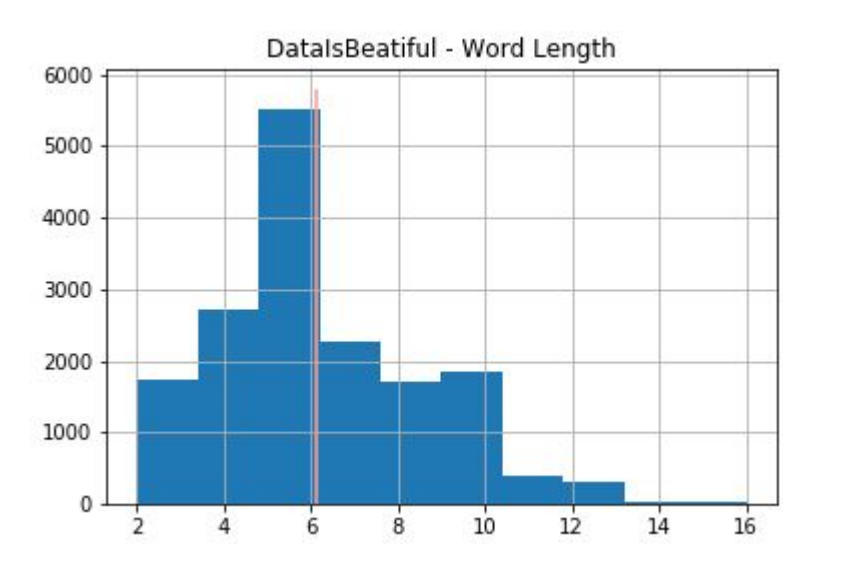
EDA - The overlap holds true as we expand our features

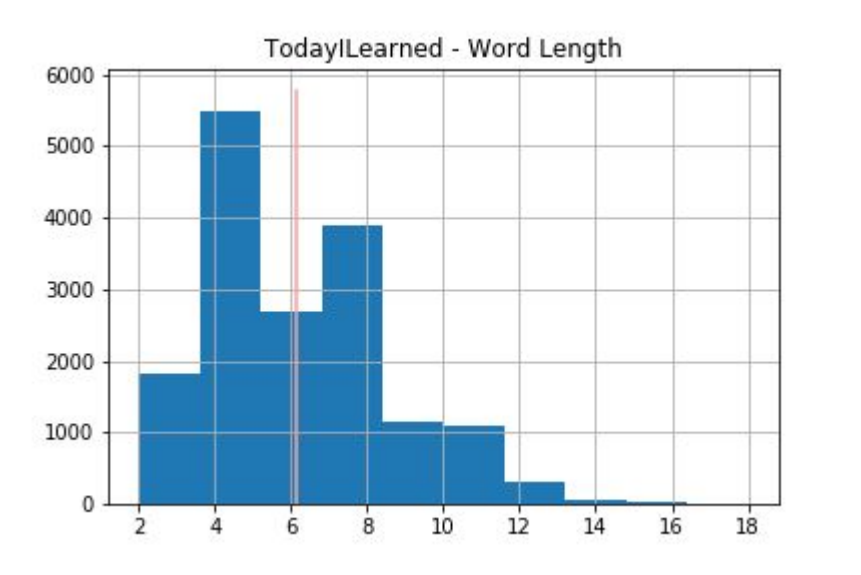


EDA - The lengths of the titles are similar to each other, with DisB having higher deviation

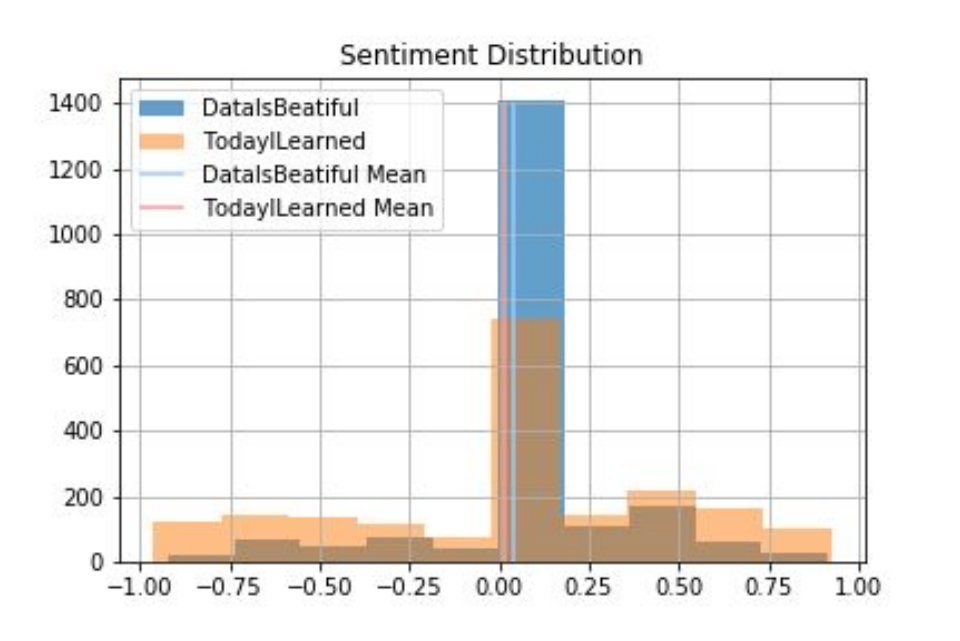


EDA - The lengths of words are similar between the two subreddits, DisB has a more normal dist





EDA - TodayILearned uses more emotional text and on average are more negative



Modeling and Interpretation

Modeling - Logistic Regression performed best with the data and is my top choice

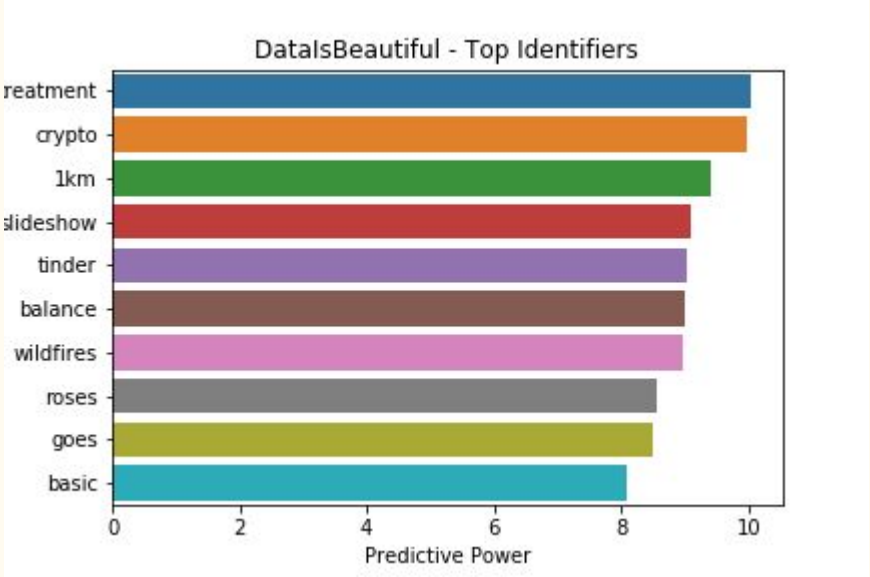
Preprocessing - CountVectorizer and removed StopWords + ('oc', 'til', 'OC', 'TIL')
Model Optimization - GridSearch through parameters

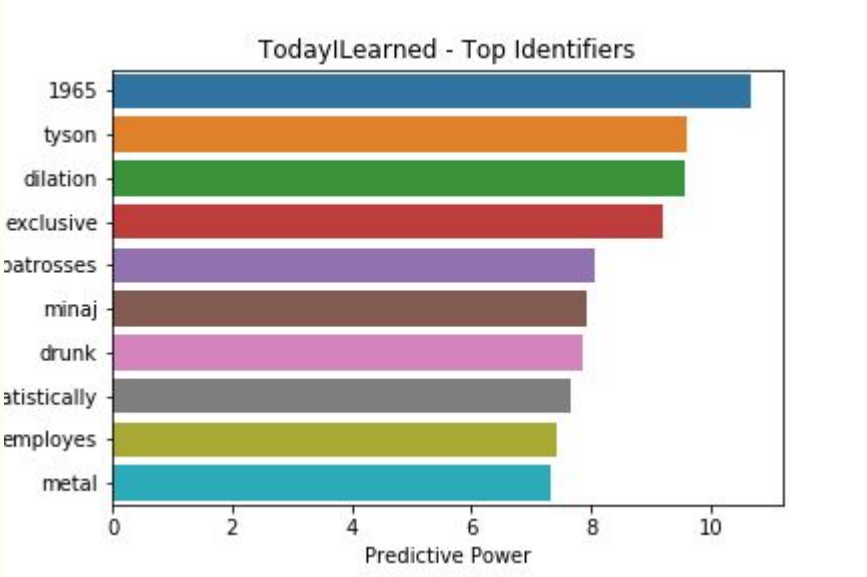
Model	Cross Val Score	Test Data Score
LogisticRegression	0.9632	0.9819
RandomForestClassifier	0.9491	0.9880
AdaBoostClassifier	0.8324	0.8923

Interpretation - The model predicts very accurately, but is prone to Type I error

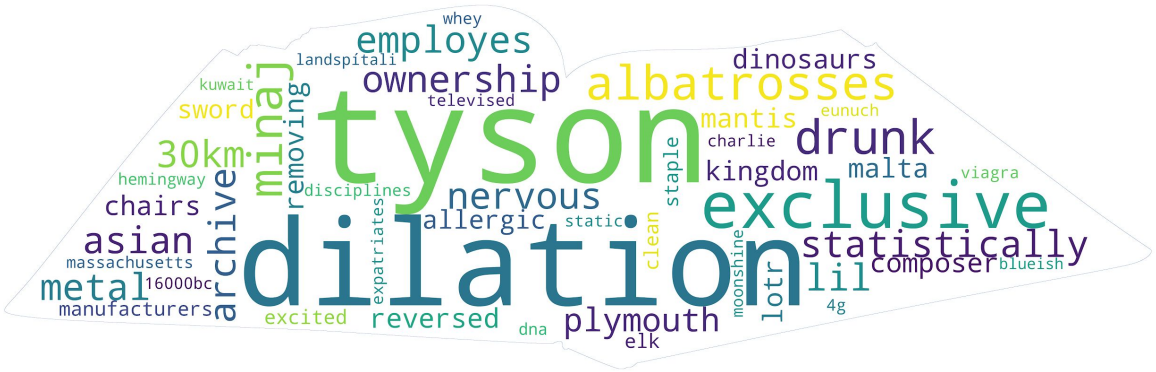
Matrix Category	Count
True Positive	507
True Negative	475
False Positive	15
False Negative	0

Interpretation - The model identified the less popular features, but with more predictive power





Interpretation - Looking at the word cloud of top features shows the distinction



Questions?