Can I differentiate r/AskCulinary from r/cookingforbeginners

By Michael Winder
Data Scientist
mhwinder@gmail.com

12 /4/2020

Overview

- Data Collection
- Similarities / Differences
- Model Comparison
- Best Model
- Conclusion / Recommendations

Data Source

- Subreddits of www.reddit.com
- AskCulinary and cookingforbeginners
- AskCulinary 3093 posts
- cookingforbeginners 3056 posts
- Title, Selftext, author, much more metadata

https://www.reddit.com/r/AskCulinary/

https://www.reddit.com/r/cookingforbeginners/

Data Example

Title: What to do with leftover white rice?

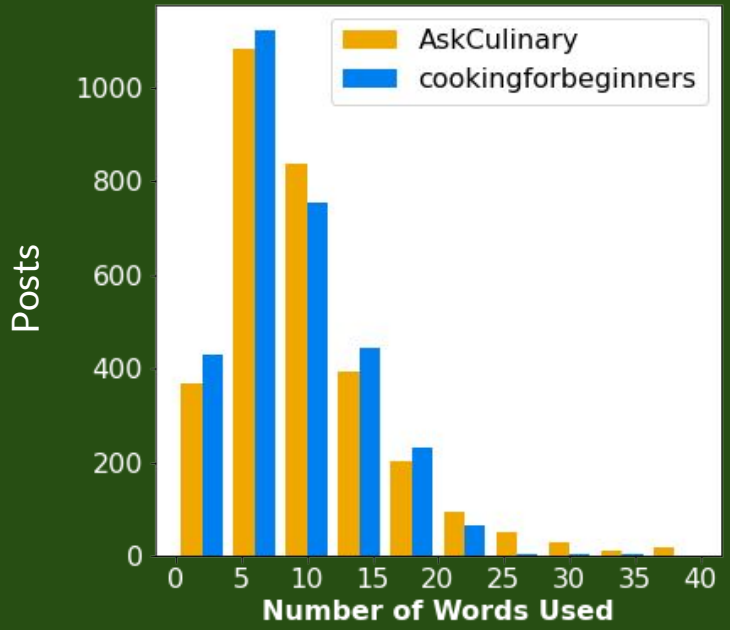
Author: Ascatman

Selftext: I made fried rice yesterday, but I made way too much white rice for the recipe. While I love fried rice, I don't want to have it two days in a row. Does anyone have any suggestions for what I could do with the leftover white rice so it doesn't go to waste?

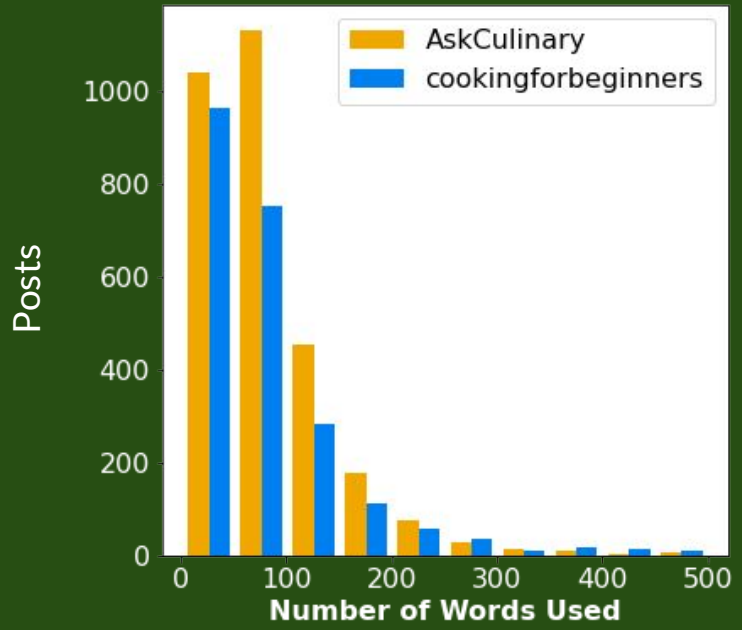
https://www.reddit.com/r/cookingforbeginners/

Post Length

Words in Title



Words in Selftext

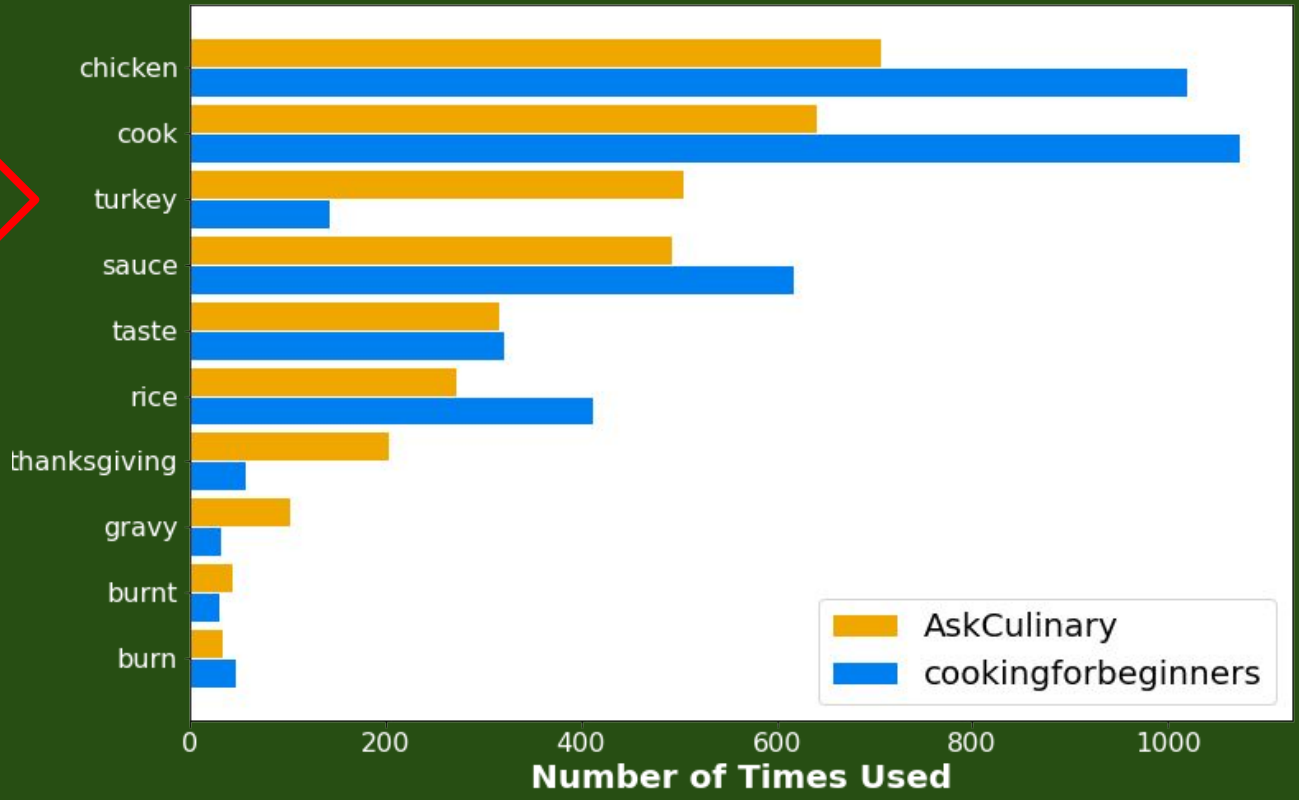


Authors

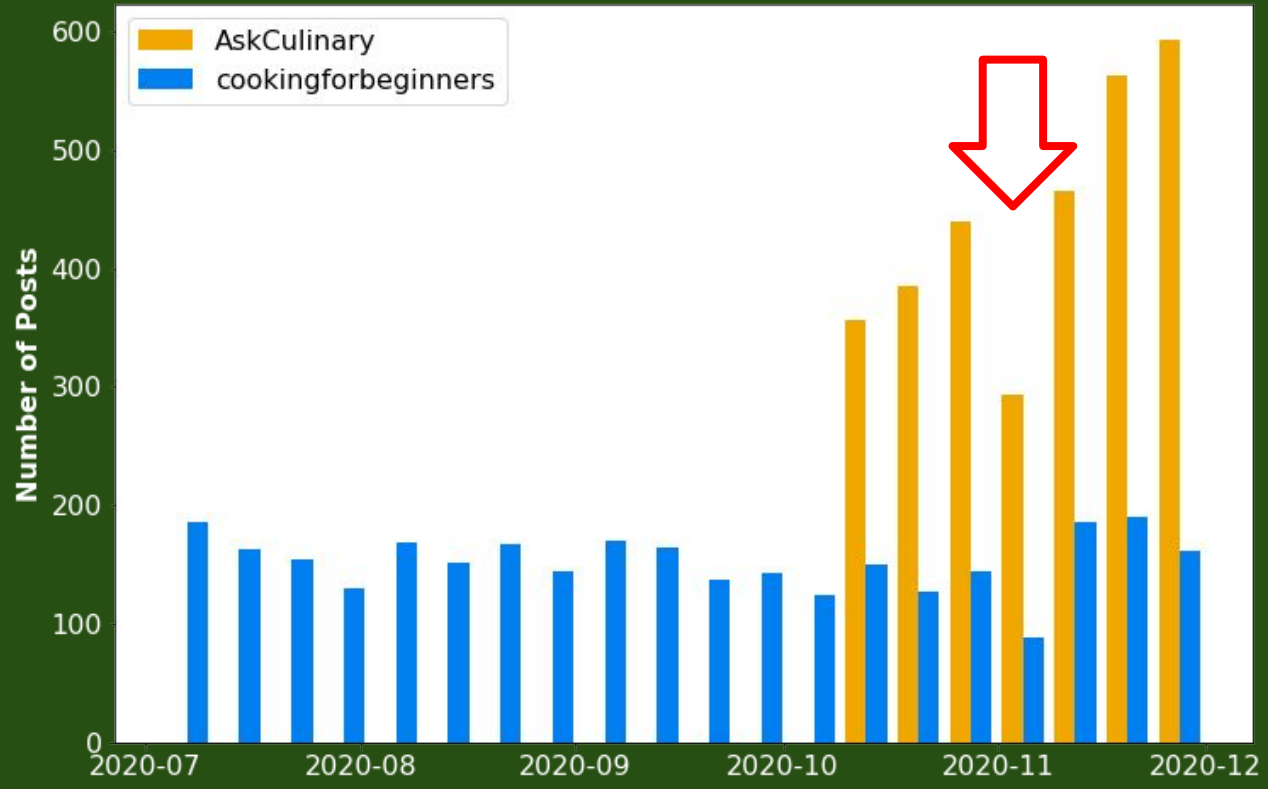
	AskCulinary	cookingforbeginners
Unique Users	2630	2263
Deleted Posts	12	37
Most Active	Kitchenrehab162563	hurrianawaz
User Posts	8	53

There were 57 users out of 4837 who posted in both subreddits. southerngentelman90 had 17 posts, only 1 from AskCulinary

Word Counts


chicken
cook
turkey
sauce
taste
rice
thanksgiving
gravy
burnt
burn
0
200
400
600
800
1000
Number of Times Used
AskCulinary
cookingforbeginners

Post Dates


AskCulinary
cookingforbeginners
Number of Posts
0
100
200
300
400
500
600
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12

Model Parameters

- Cross Validated GridSearch
- Document = Title + Selftext
- Stemmed, Unstemmed and Lemmatized tokens
- Ngram ranges 1 and 2
- Count Vectorizer and TfidfT
- With or without english stop words
- Hyperparameters specific to each model

Model Comparison

Models Used:	Best Accuracy:
Baseline Score	0.503
Logistic Regression	0.715 lemmatized tokens
K Nearest Neighbors	0.675 stemmed tokens
Random Forest	0.706 stemmed tokens
Bagging Classifier	0.714 lemmatized tokens
Support Vector Classifier	0.703 lemmatized tokens

Best Model

- Logistic Regression Model
- Lemmatized Tokens
- No stop words
- TfidfTransformer
- Ngrams = 1
- Specificity = 0.68
- Sensitivity = 0.75

Conclusion/Recommendations

- Forums were diffcult to differentiate
- All models outperformed baseline model
- Results were inconclusive 0.715

Further research:

Regular Expressions/Selecting Data with identical calendar ranges/more CPU/explore comments

Additional models needed in the future to increase accuracy rate

QUESTIONS!

By Michael Winder
Data Scientist
mhwinder@gmail.com