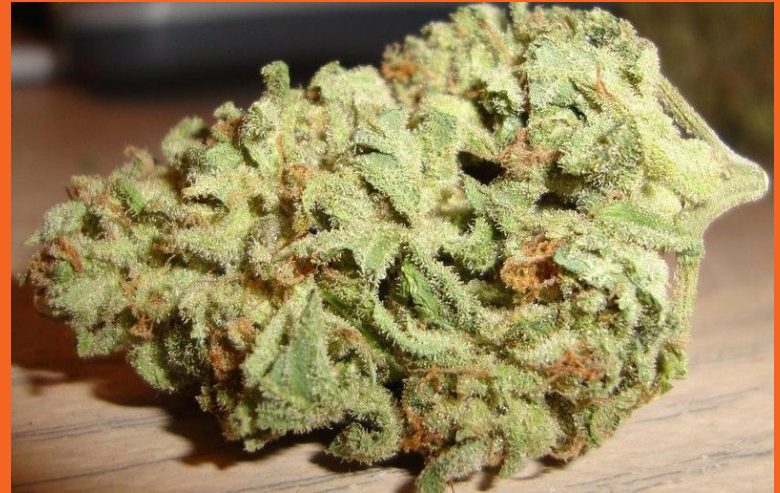# Subreddits: Cheese vs Weed

By Michael Knight

(Fried Cheese Curds)

(Blue Cheese Strain)

Scraping The Reddit Bowl

Using Reddit's API, I gathered thousands of posts from the subreddit threads regarding cheese (r/cheese) and marijuana (r/weed) to see if I could train a model to determine whether a post was from one or the other, and the specific words that distinguished the two posts from either subreddit.

Why Grass & Cheese?

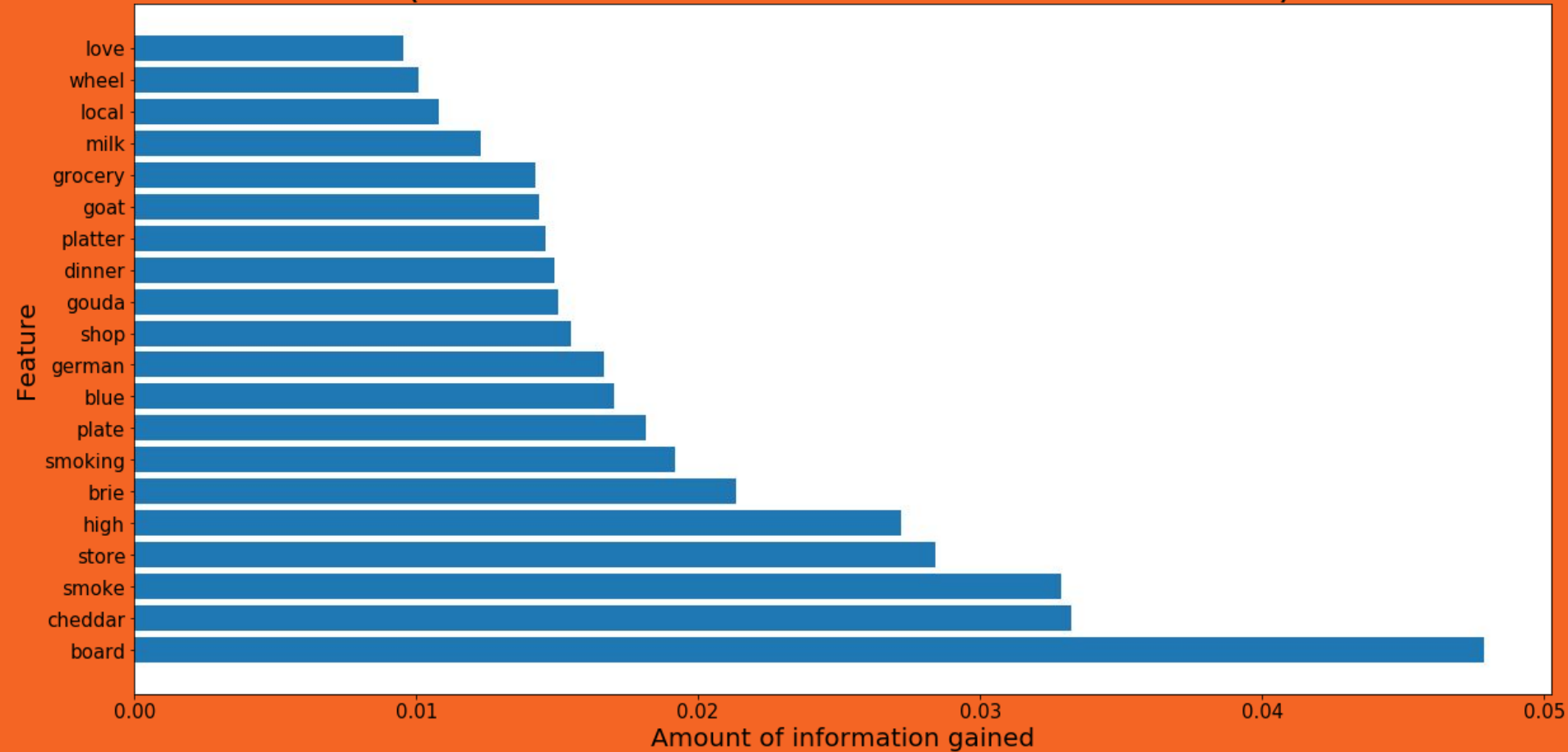Marijuana and cheese are two instances where smelly odor implies quality

Both are things one ingests to feel a type of euphoria

Both synonymous with California and Vermont

Seeing The Forrest for the trees

Using the Random Forrest model over 100 n_estimators, through a gridsearch testing over parameters of 8, 10 and 12 for 'min_samples_split' and 50, 75, and 100 for 'max_depth', the training score was about 89% and the testing score was about 82%
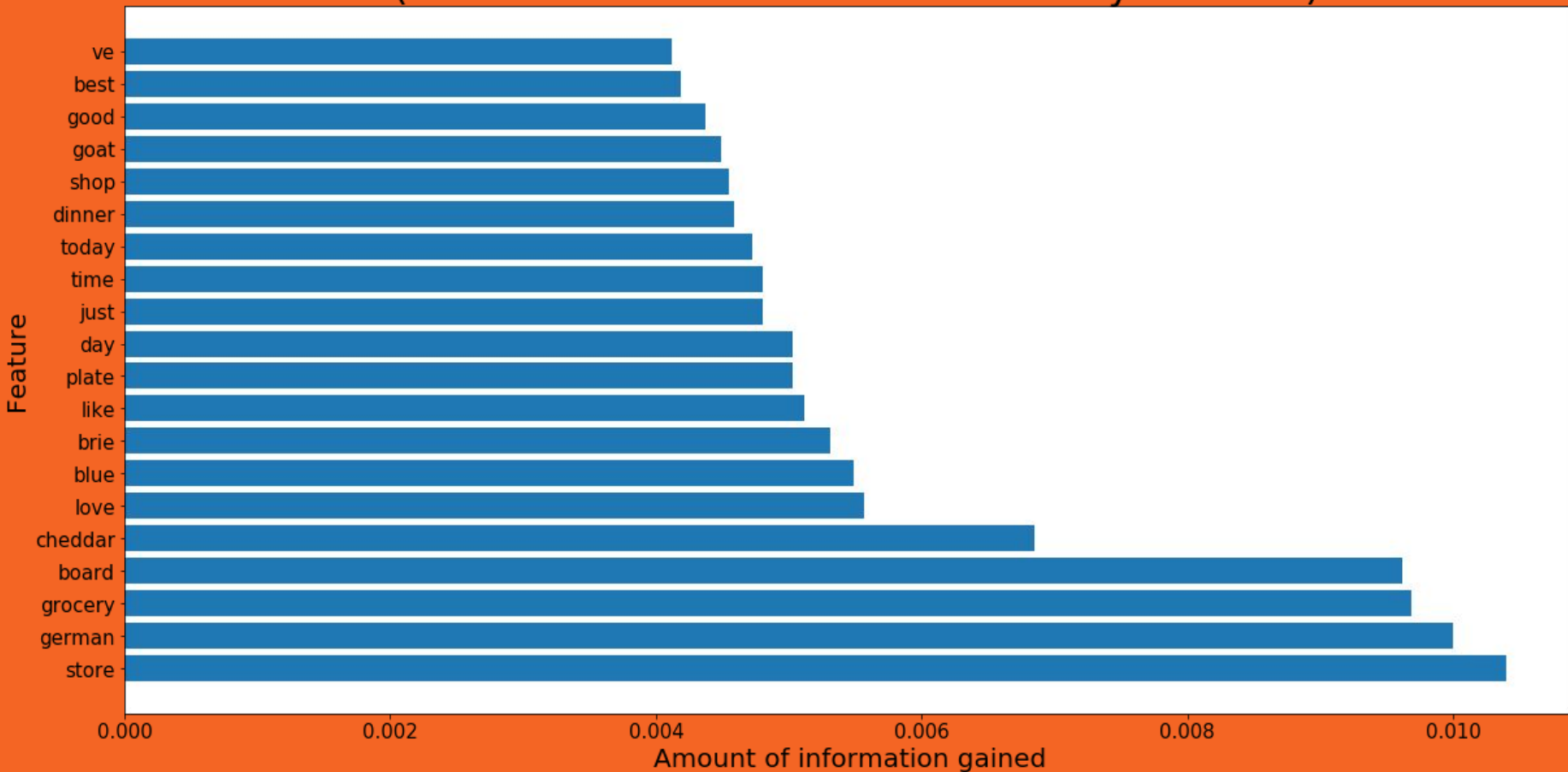
Top 20 Features
(Based On The RandomForestClassifier Model)

Multinomial Naive Bayes Model

Using the Multinomial Naive Bayes model with the built in parameters, the training score was 89% and the testing score was 84%. Both models worked fairly well, and almost equally accurately, with the Random Forrest Model being just a hair overfit and weaker than the Multinomial Naive Bayes model.

Top 20 Features
(Based On The Multinomial Naive Bayes Model)

Comparing subreddits is like comparing chocolate to cheese, but 84% accuracy a'int too shabby!

Michael Knight

Army of One
For Cheese & Weed