Sorting reddit's subreddits!

This-or-That Consultants

Agenda

First:	**Review the Problem**
Second:	**How we can Help**
Third:	**Intro to Our Data**
Fourth:	**Our Process**
Fifth:	**What we Found / Takeaways**
Sixth:	**Recommendations**
Seventh:	**Conclusion / Questions**

First: **Review the Problem**

- Nefarious and disgruntled ex-Employee
- Replaced all subreddit fields with ¯_(ツ)_/¯
- Subreddit links essentially down
- Posts **require sorting**!

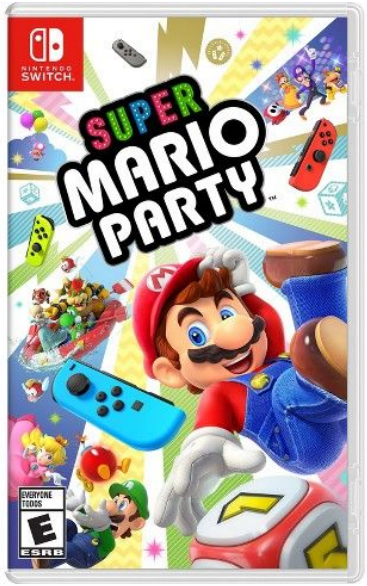
Second: **How we can Help**

Scrape reddit's API to gather post attributes

Build a Classification Model with these attributes!

Third: **Intro to Our Data**

r/MARIOPARTY



Post Traits Scraped
title
selftext (post_paragraph)
clicked
ups
downs
likes
num_comments

r/SuperSmashUltimate

Third: **Intro to Our Data** ctd..

Why we chose these subreddits:

- Expected content-overlap
- Recently Released
- Holiday Season

Fourth: **Our Process**

First: **Gathered the Data** (scraped reddit's API)

Second: **Cleaned, Formatted, and "Vectorized"**

Third: **Selected our Features/Target variables**

Fourth: **Train-Test-Split**

Fifth: **Ran through Four Models/"Gridsearched"**

Sixth **Analyzed different models/parameters' performance**

Fifth: **What we Found / Takeaways**

Model:	Logistic Regression, Gridsearched
Vectorization:	CountVectorize
CV Folds:	5
Parameter_1:	C = 1.0
Parameter_2:	penalty = l2
Parameter_3:	tol = 0.0001
Train Set Score:	0.999
Test Set Score:	0.950

Model:	Logistic Regression, Gridsearched
Vectorization:	TF-IDF
CV Folds:	5
Parameter_1:	C = 1.0
Parameter_2:	penalty = l2
Parameter_3:	tol = 0.0001
Train Set Score:	0.998
Test Set Score:	0.945

Model:	Extra Trees
Vectorization:	CountVectorize
CV Folds:	5
Parameter_1:	criterion = gini
Parameter_2:	n_estimators = 10
Parameter_3:	min_samp._split = 2
Train Set Score:	0.890
Test Set Score:	0.948

Fifth: **What we Found / Takeaways** ctd..

- **Not as much overlap as expected**
  - Results when model run on different pairs of subreddits?
- **Seems to be small list of important features**
  - Extra Trees only had 10 for its number of trees
- **Removed authors & words found in title**

Potential Pitfalls

- Older subreddits
- More diverse subreddits
- Let’s find out!

Sixth: **Recommendations**

1. **Use our model** on **r/MARIOPARTY** and **r/SuperSmashUltimate !**

2. **“Misplaced Post”** button

3. **Try the model** on **different pairs**

Questions?