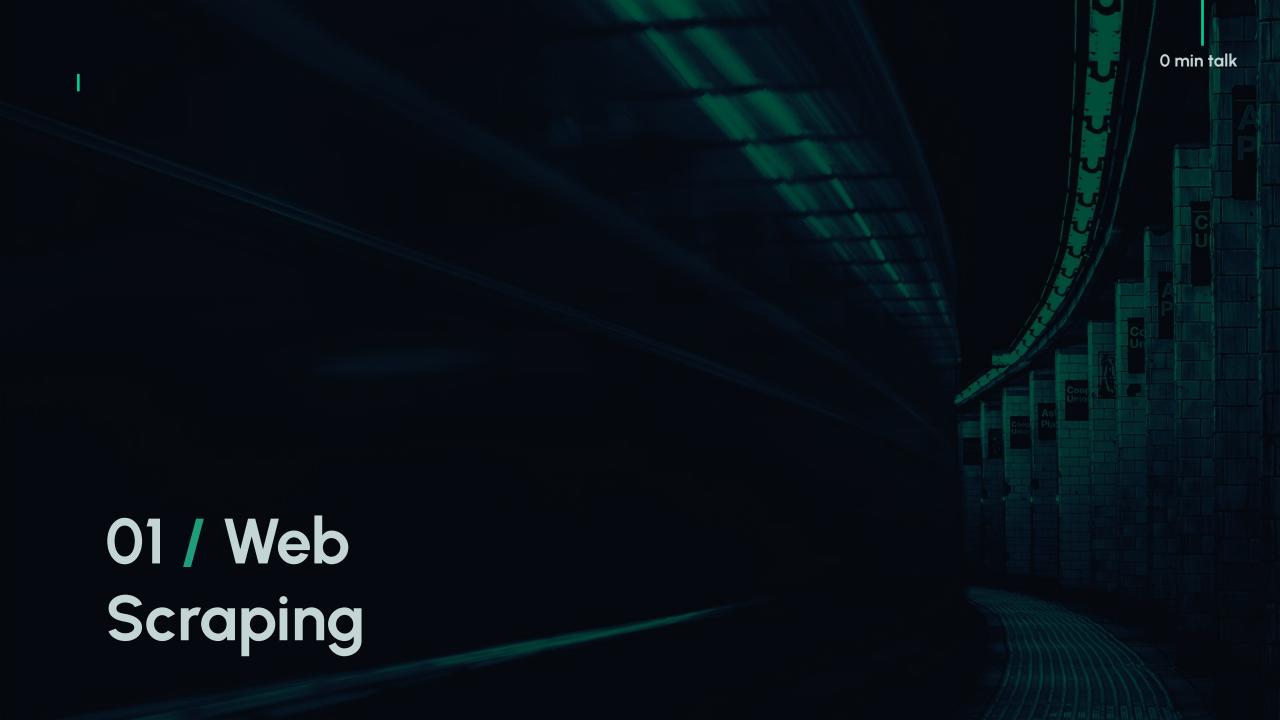
total talk: 15 min

DSI Project-3







Given a reddit post, determine which subreddit is it from?

Process

- Pushshift API,
- Removed empty posts
- Removed posts deleted by moderators
- 1000 posts from PS5 and 1100 posts from XboxSeriesX



Vectorizer

With logistic regression

CountVectorizer

- Max df: 0.7
- Min df: 2
- Max features 7000
- Ngram: (1, 2)
- Stop word: english
- Best score: 0.81

TF-IDF Vectorizer

- Max df: 0.7
- Mind df: 2
- Max features 7000
- Ngram: (1, 2)
- Stop word: english
- Best score: 0.81

I

Random Forest

Training Score: 0.866
Test Score: 0.837
N estomators: 100
Max Depth: 20
Min Samples Split 10:
Min Samples Leaf: 3

SVM

Training Score: 0.99 Test Score: 0.672 Kernel: C: KNN

Training Score: 0.99 Test Score: 0.55 N neighbors: 3 Weights: distance



By using random forest, gives us the best result to determine whether a post is from PS5 or XboxSeriesX subreddit.

Thank you!

