Unsupervised Natural Language Processing

A Textual Analysis of Selected Articles from Grantland.com

~

Greg Condit – conditg@gmail.com

Find the code for this project at Github.com/conditg/nlp-grantland

What was Grantland.com?

Grantland was a long form blog owned by ESPN. Grantland was known for its award-winning writing, and it's contributors brilliantly mixed sports, popular culture, and data analytics & visualization into riveting stories and analysis.

Grantland was writing and media as it should be, and it was shut down because it's much harder to monetize than low-effort clickbait. This analysis is a lighthearted tribute to the site and it's contributors.

Corpus & Collection Process

126 Articles by selected contributors scraped from Grantland.com using *BeautifulSoup*

Contributors were chosen fairly arbitrarily by my memory of which writers I enjoyed most, which means the text will bias towards my interests, that is, mostly basketball and occasionally football.

I also added some contributors at random for variety.

The scraper pulled the most recent 10 posts for each contributor, then excluded podcast posts. Bill Simmons's most recent posts were almost all podcasts, so his were selected manually.

Contributor	Articles Scraped	Wordcount
Andrew Sharp	10	14,039
Bill Barnwell	8	23,783
Bill Simmons	11	62,242
Brian Phillips	10	17,841
Charles P. Pierce	10	12,764
Jonathan Abrams	10	41,791
Kirk Goldsberry	10	11,974
Mark Harris	10	15,744
Mark Titus	10	36,736
Sean McIndoe	10	20,946
Shea Serrano	10	14,811
Steven Hyden	10	17,857
Zach Lowe	7	24,857
TOTAL	**126**	**315,385**

Process Overview

- Clustering using Doc2Vec
  - K-Means vs Spectral Clustering vs Affinity Propagation
- Predicting authorship of each article
  - Features: Doc2vec
  - Features: Latent Semantic Analysis
- Just for fun: Most used words and most unique words per contributor over all selected articles

Bill Simmons was the site's creator

Article vectorization using Doc2vec

We're close. Four more days and there will be professional basketball on TV every night.

Text cleaned, contractions removed

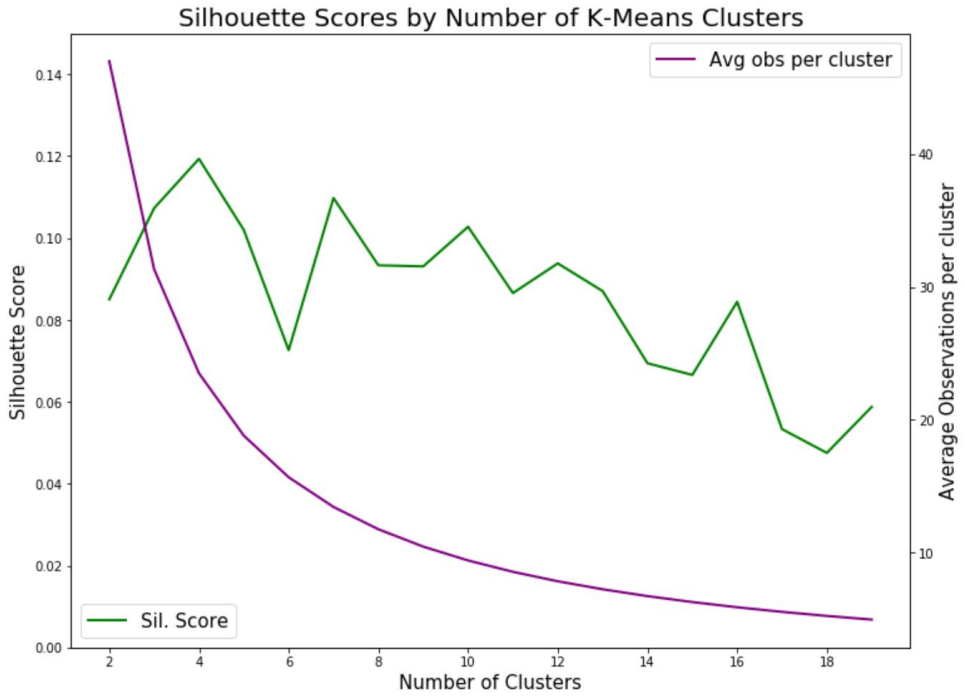
we are close. four more days and there will be professional basketball on tv every night.

Stop words & punctuation removed, remaining words changed to lemmas

['close']
['day', 'professional', 'basketball', 'tv', 'night']

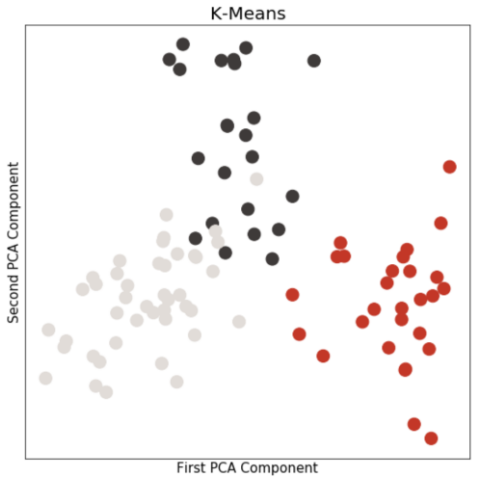
The doc2vec algorithm was used to create vector representations of 94 of the 126 articles, the rest were held out as a test set.

With the resulting vector set, silhouette scores were examined for various numbers of K-Means clusters. The highest was typically 3 or 4:

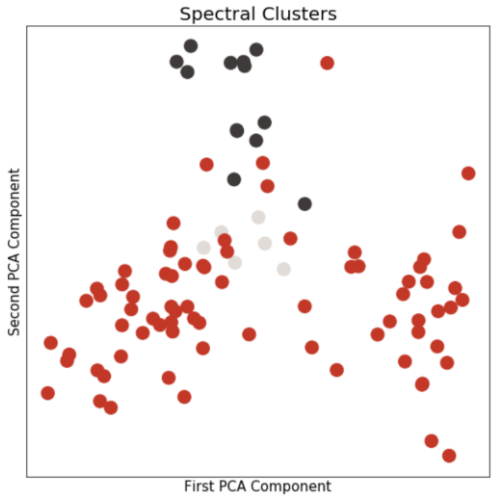


Exploration of 3-cluster methods

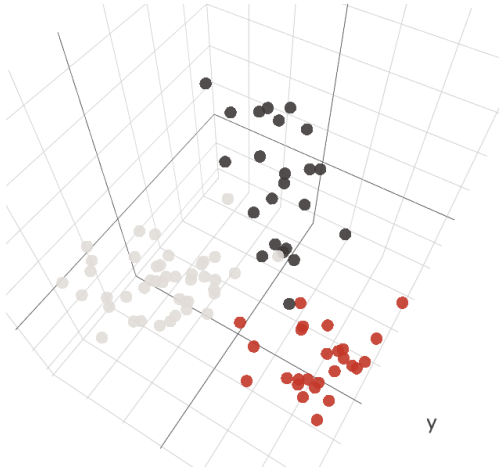
K-Means and Spectral Clustering were both used. The 65 dimensions created by doc2vec were reduced to 2 (or 3) dimensions for visualization using Primary Components Analysis.

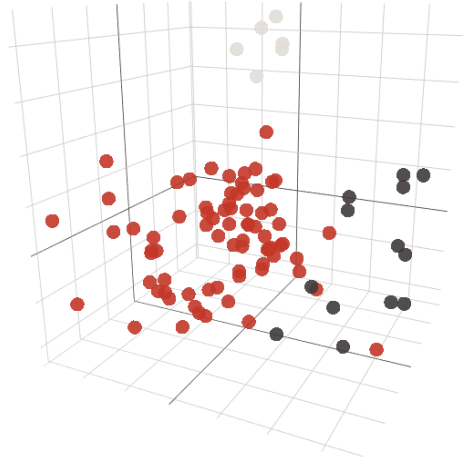


3 Components, Clustered using Kmeans



3 Components, Spectral Clustering





K-Means created 3 clusters that are very distinct in just 2 dimensions, and almost entirely linearly separable in 3 dimensions.

Spectral is not constrained to same-size clusters, and we see that it creates 2 small clusters and one huge one, and they are not quite as easy to separate visually.

Given that these are only 2D or 3D representations of entire articles, there's no clear-cut visual way to compare the two clustering methods comprehensively. K-Means has a better silhouette score (0.10) than Spectral (.05), which does match visual intuition.

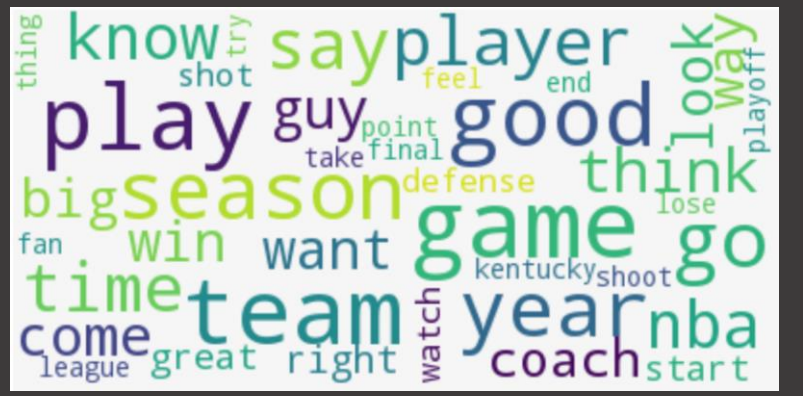
We also have no real "ground truth", i.e. we don't have 3 known classes for this data that we can compare these clusters to, so we can't use an Adjusted Rand Index.

What remains is to assess the clusters qualitatively based on which words are most frequent in each cluster's articles, which is done on the next page.

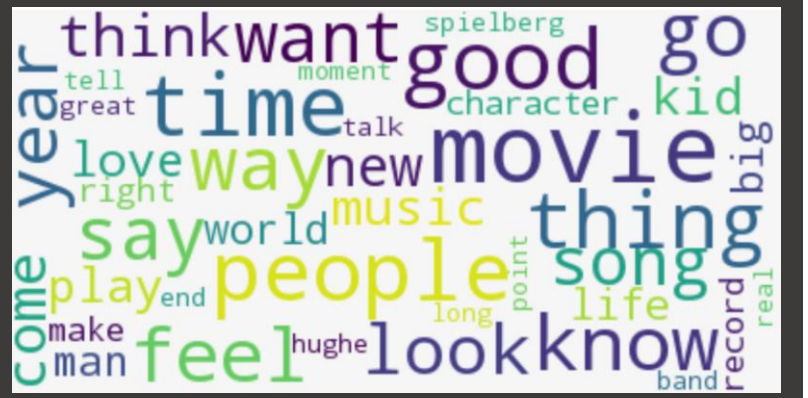
Most Common Words: 3 K-Means Clusters

Topic: Football

Topic: Basketball



Topic(s): Pop Culture/Media



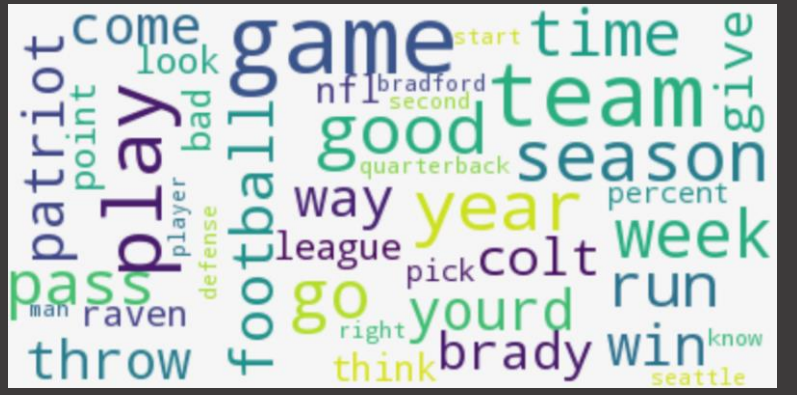
Basketball and Football are what made me love Grantland, so it's no surprise to see those consistently picked out.

The 3rd cluster is very informative. K-means assumes all clusters will be equal size, so it could never isolate a tiny minority population like the hockey articles. However, the Spectral Clustering seems to completely lose all the pop culture content that K-means captures.

This indicates that 3 was not a good choice. We need a clustering method that can decide the number of clusters for itself.

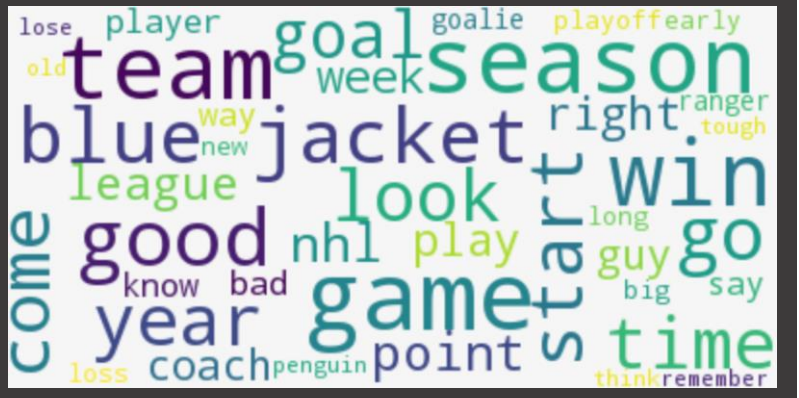
Most Common Words: 3 Spectral Clusters

Topic: Football



Topic: Basketball

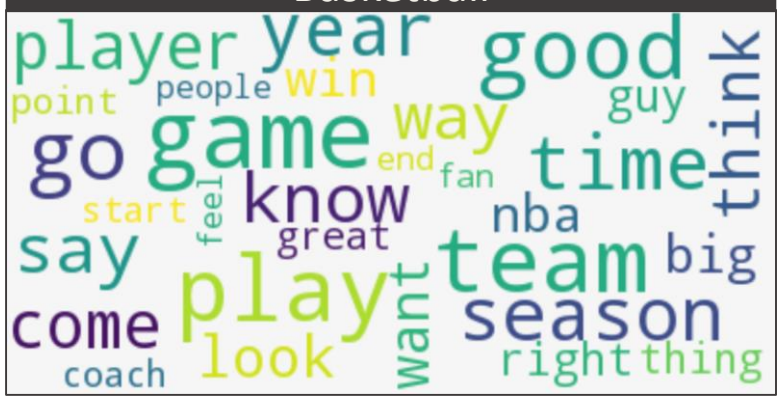
Topic: Hockey



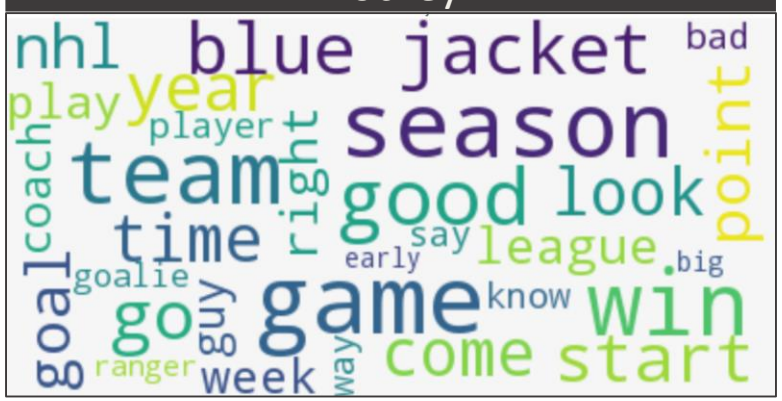
Affinity Propagation: Fitting data to a flexible # of clusters

Affinity Propagation selected 11 clusters, but upon review, 5 of them were duplicates with no topical differences. I merged those, and the below 6 are the best clustering method I found. With more data, I suspect the especially bizarre “Future” cluster would not remain consistent, it strikes me as a catch-all cluster for some irreverent outliers, whereas the other 5 are pretty defined and consistent.

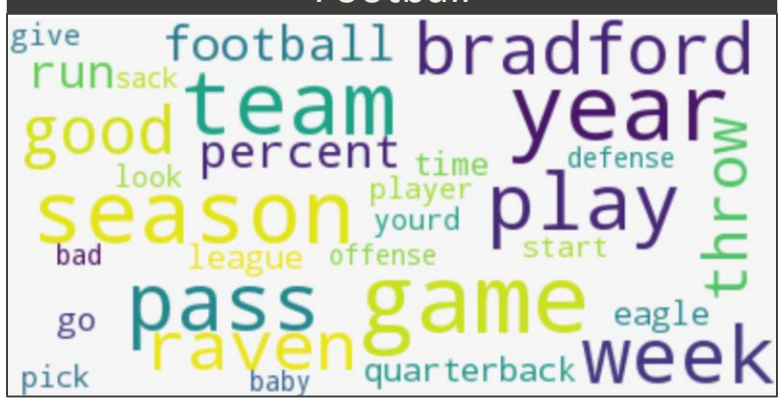
Basketball



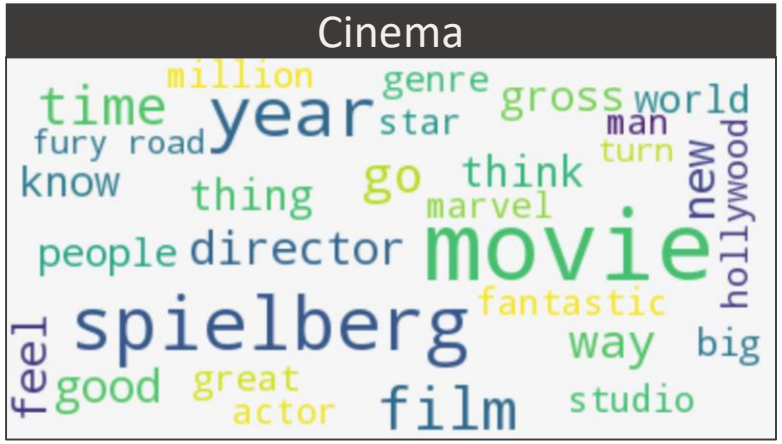
Hockey



Football

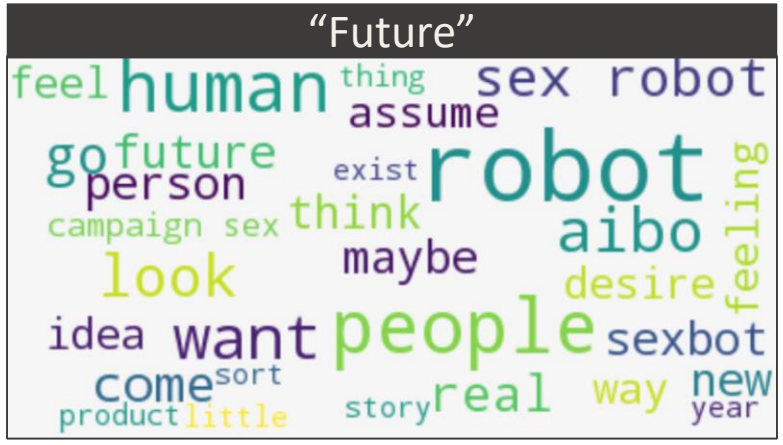


Cinema



Music

“Future”



Predicting who wrote each article

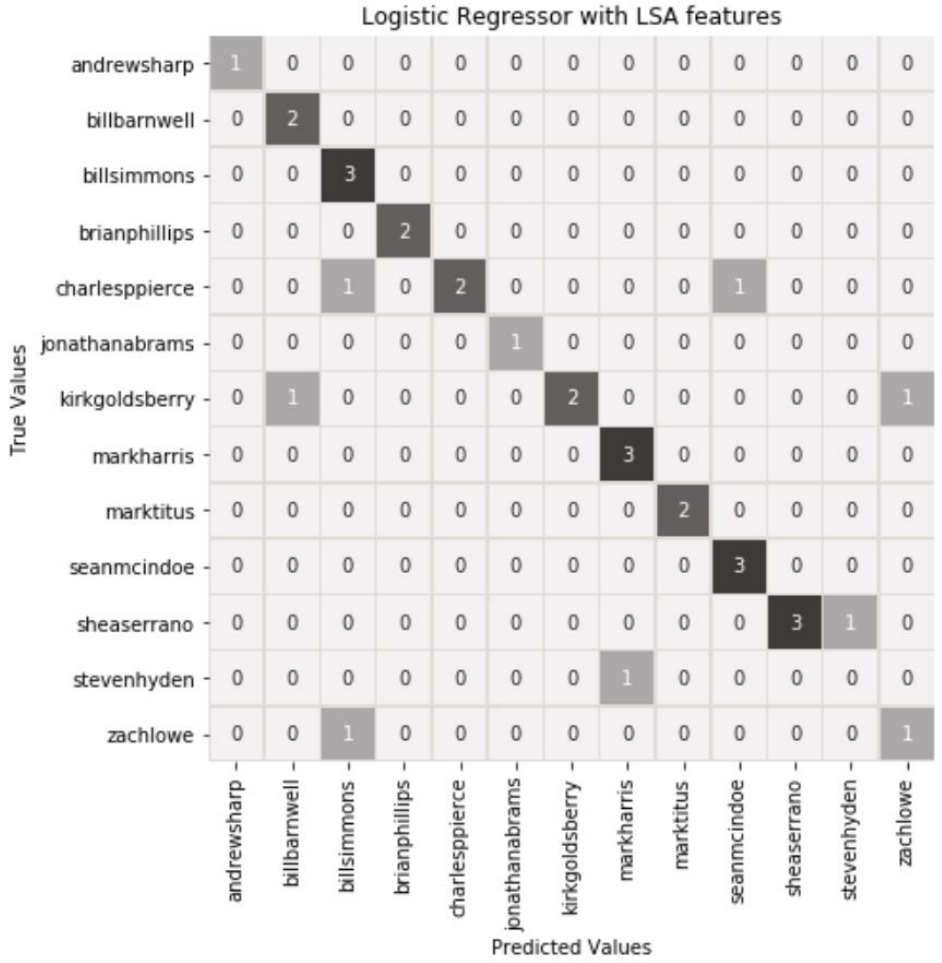
Random Forest, K-Nearest Neighbors, and Multi-class Logistic Regression models were applied to two different feature sets:

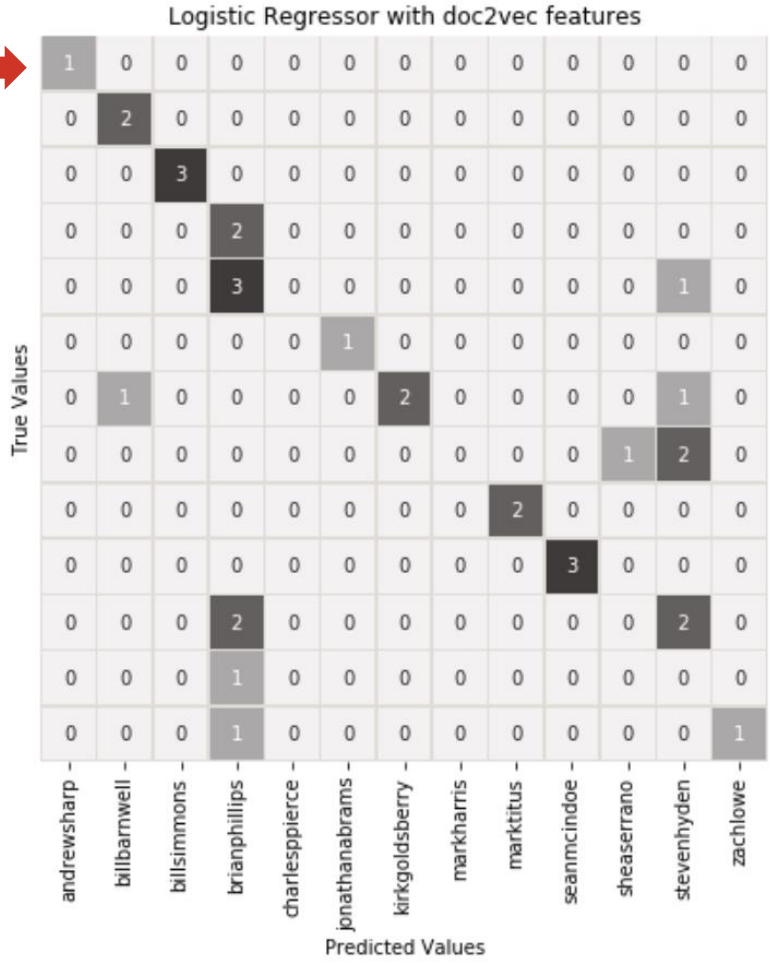
- Document vectors created via Latent Semantic Analysis
- Document vectors created via doc2vec

Across all models, the LSA features were better for classification than the doc2vec features were.

For both feature sets, Logistic Regression was most effective as a contributor classifier.

Test Data F1 Scores	LSA Features	Doc2vec Features
Random Forest	0.69	.16
KNN	0.75	.25
Logistic Regression	0.78	.53





Most Used & Most Unique Words by Contributor

Bill Simmons

Most Used

Most Unique

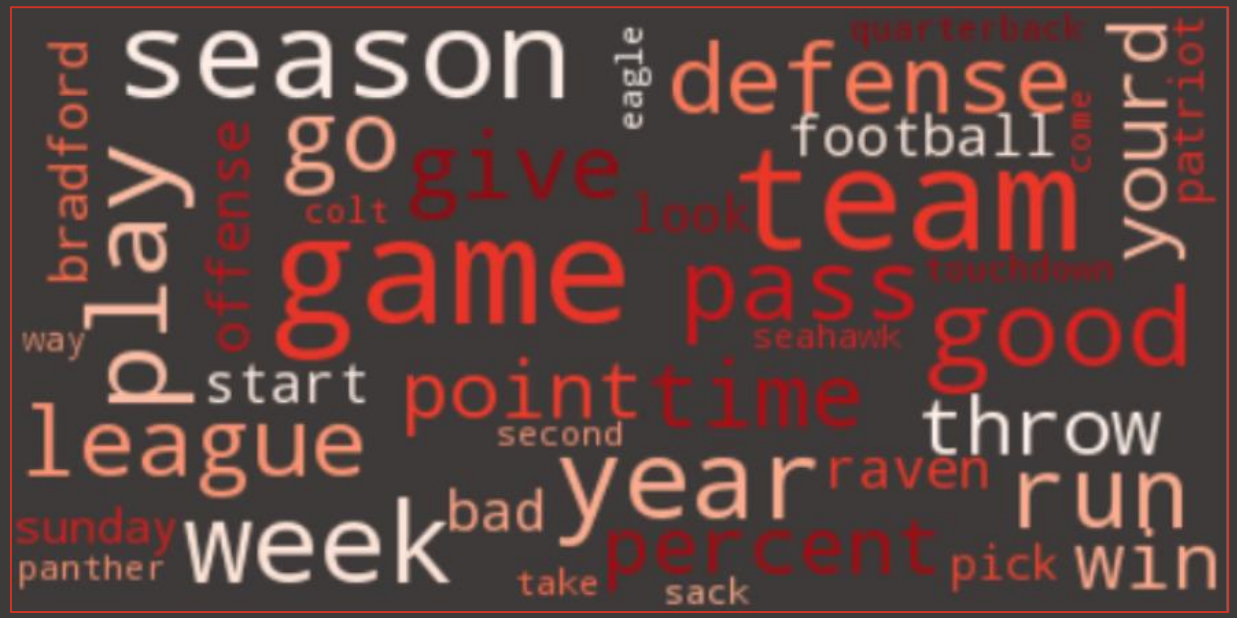
Andrew Sharp

Most Used

Most Unique

Bill Barnwell

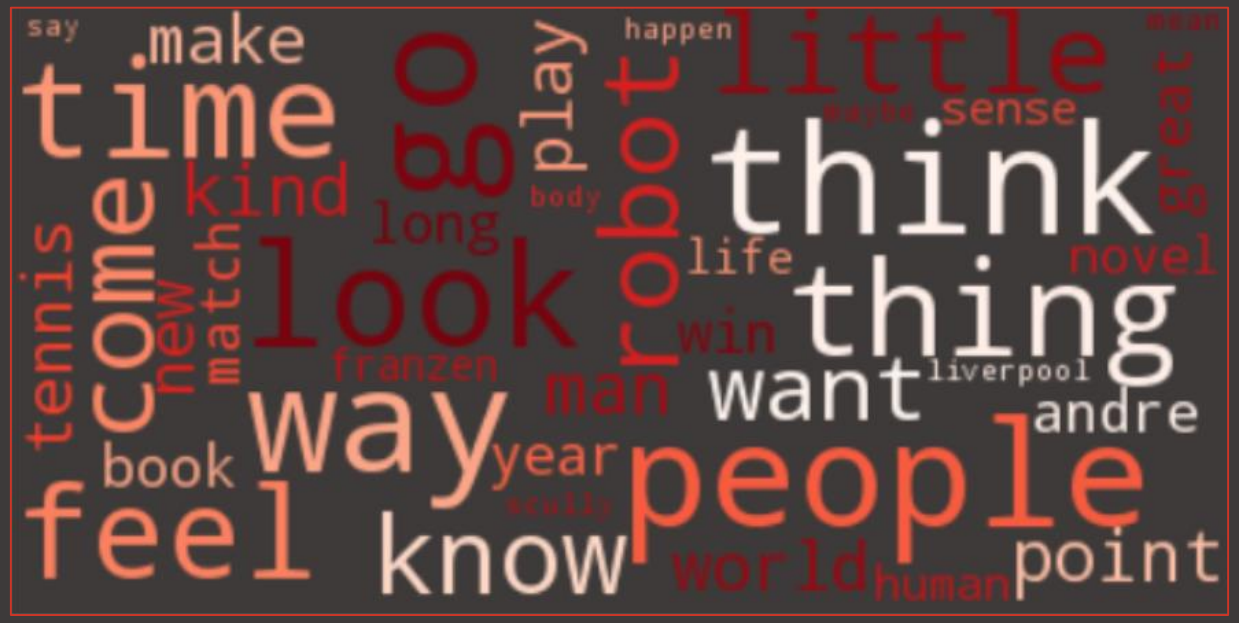
Most Used



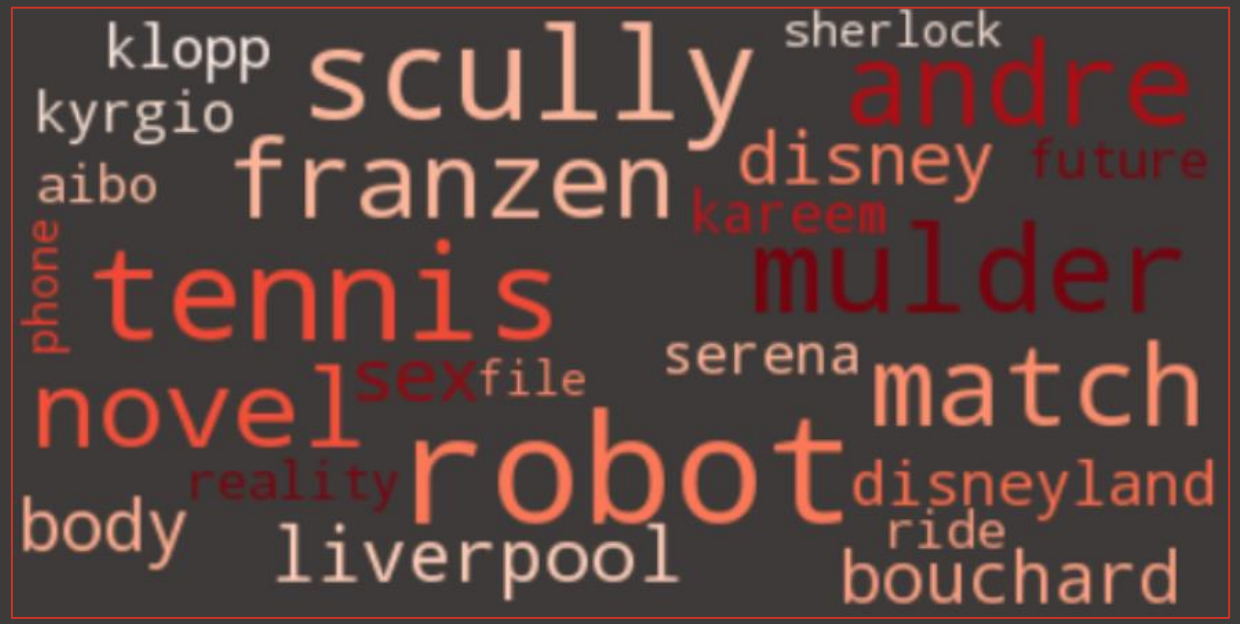
Most Unique

Brian Phillips

Most Used

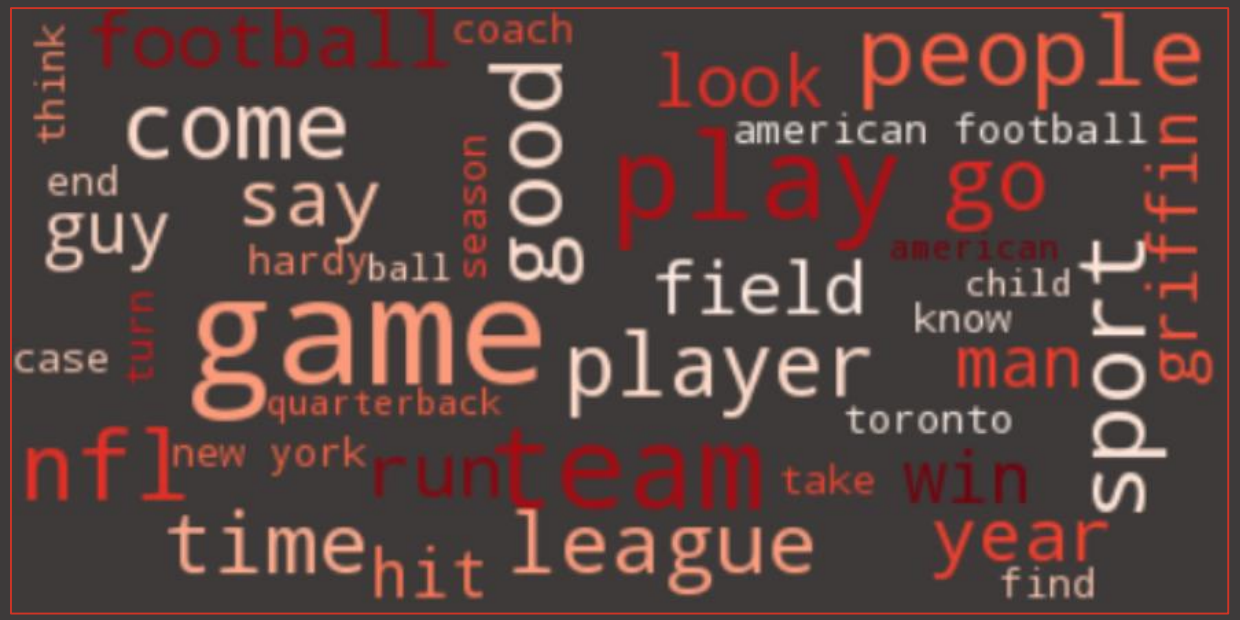


Most Unique



Charles P. Pierce

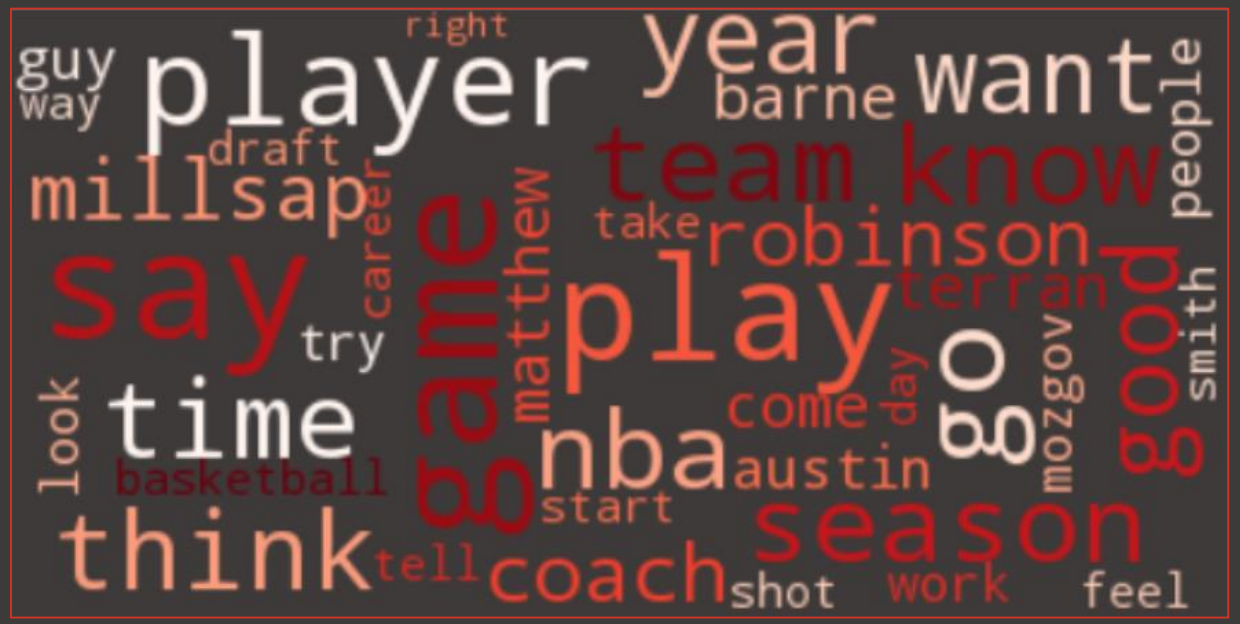
Most Used



Most Unique

Jonathan Abrams

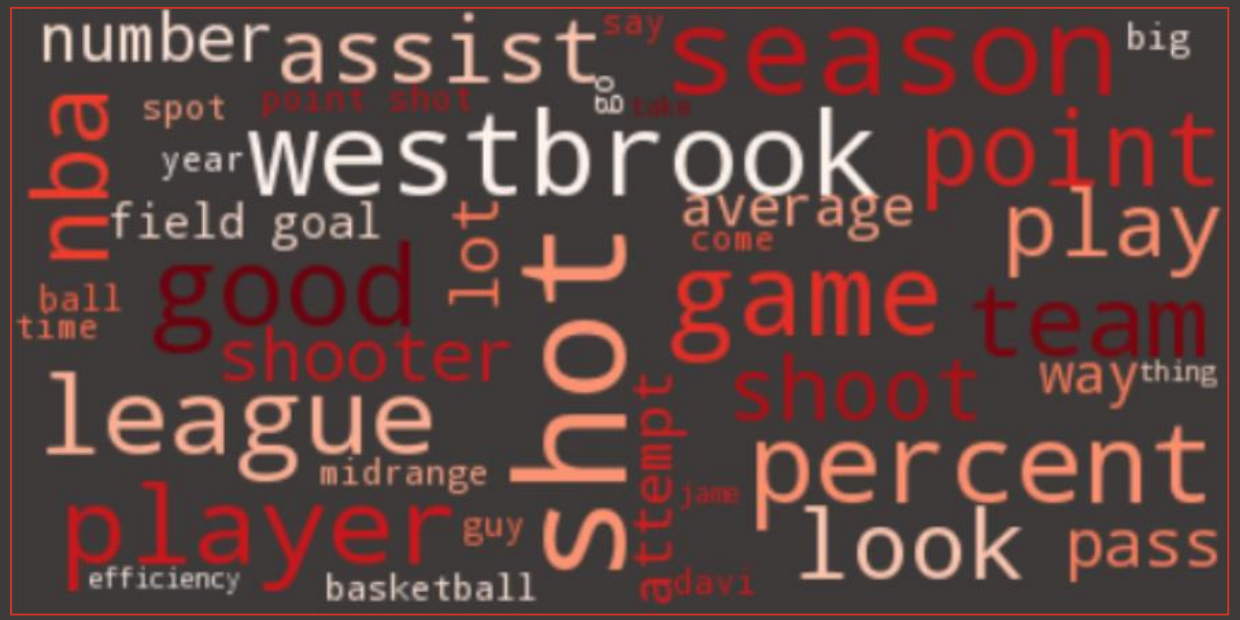
Most Used



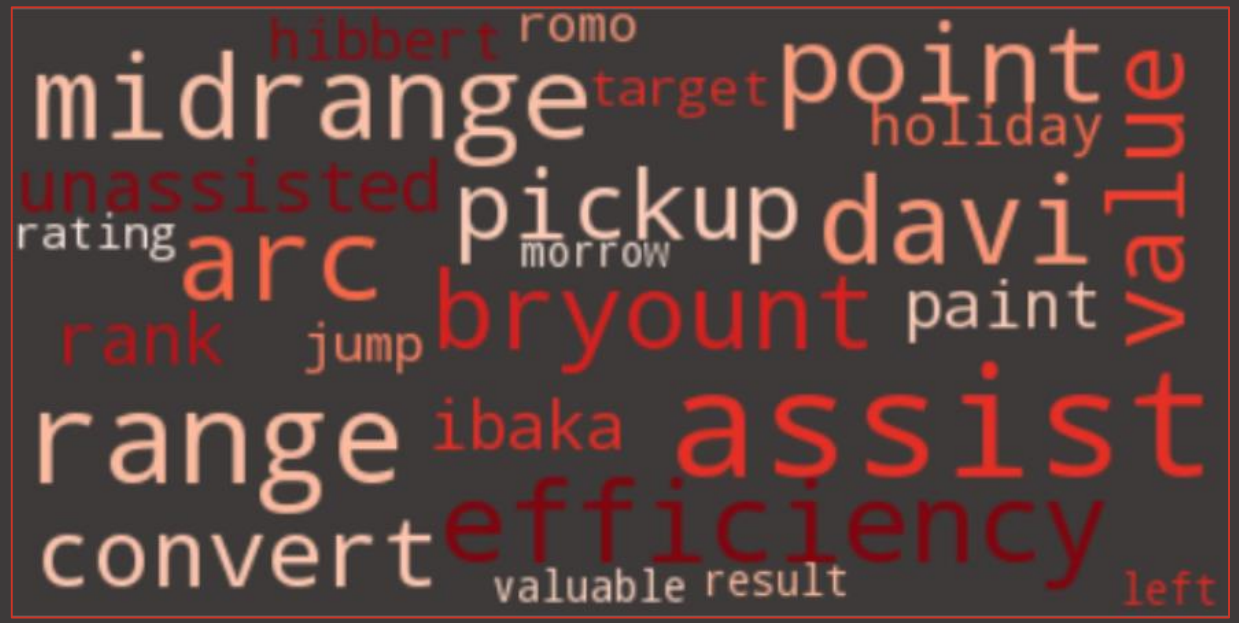
Most Unique

Kirk Goldsberry

Most Used

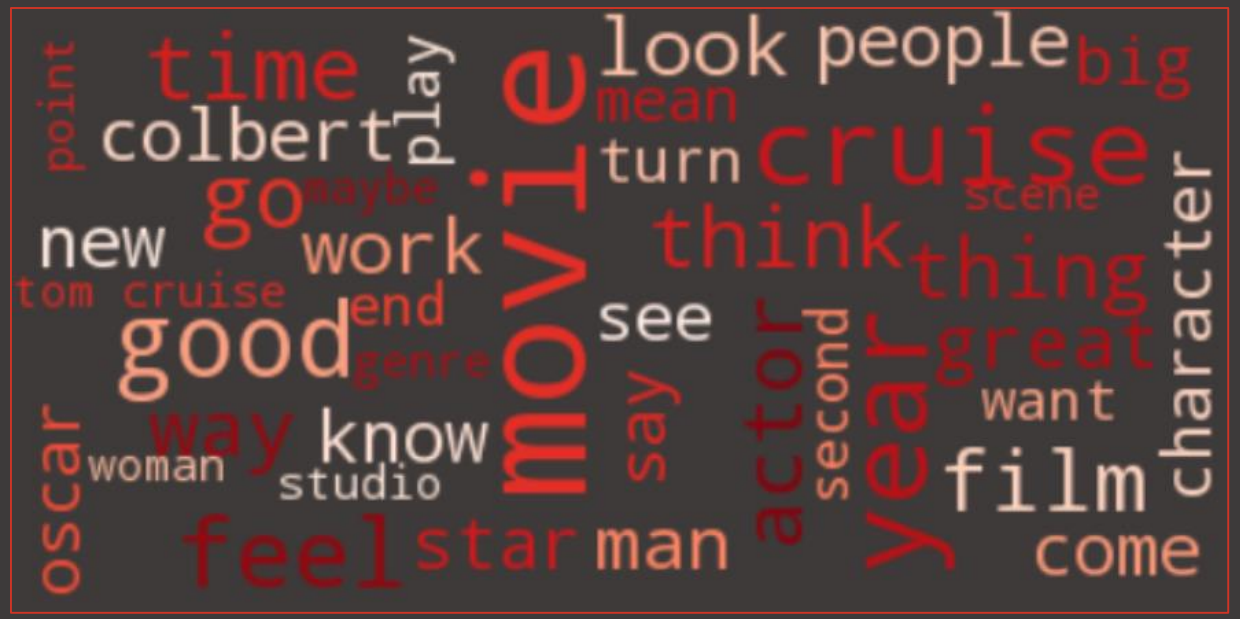


Most Unique



Mark Harris

Most Used



Most Unique

Mark Titus

Most Used

Most Unique

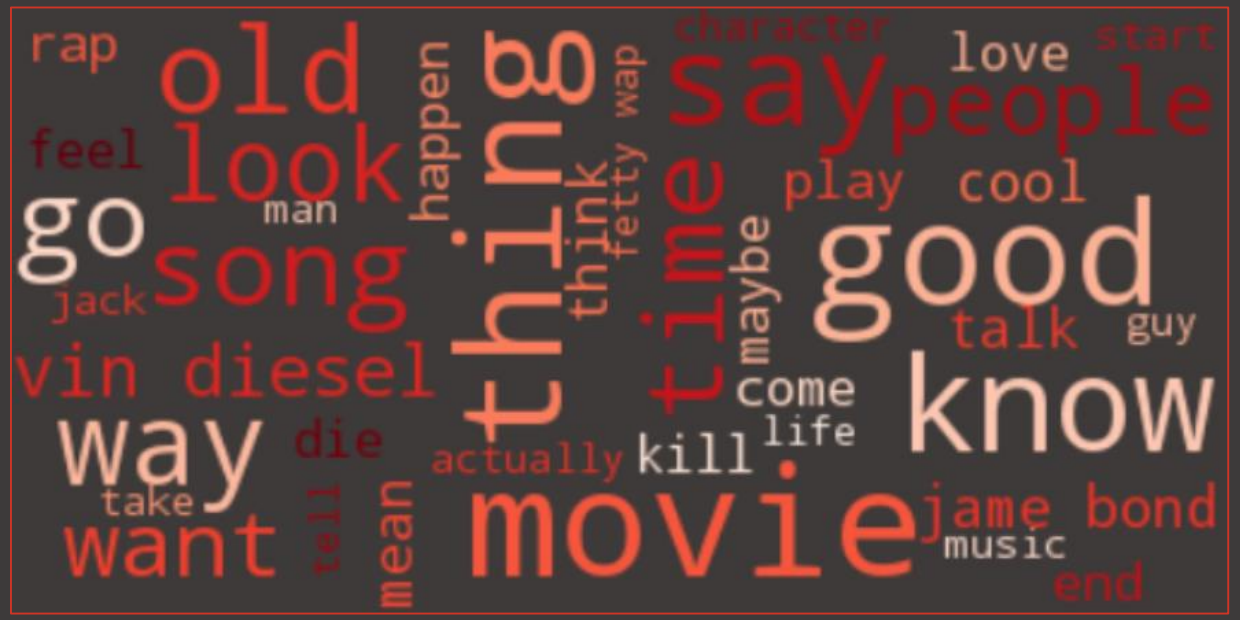
Sean McIndoe

Most Used

Most Unique

Shea Serrano

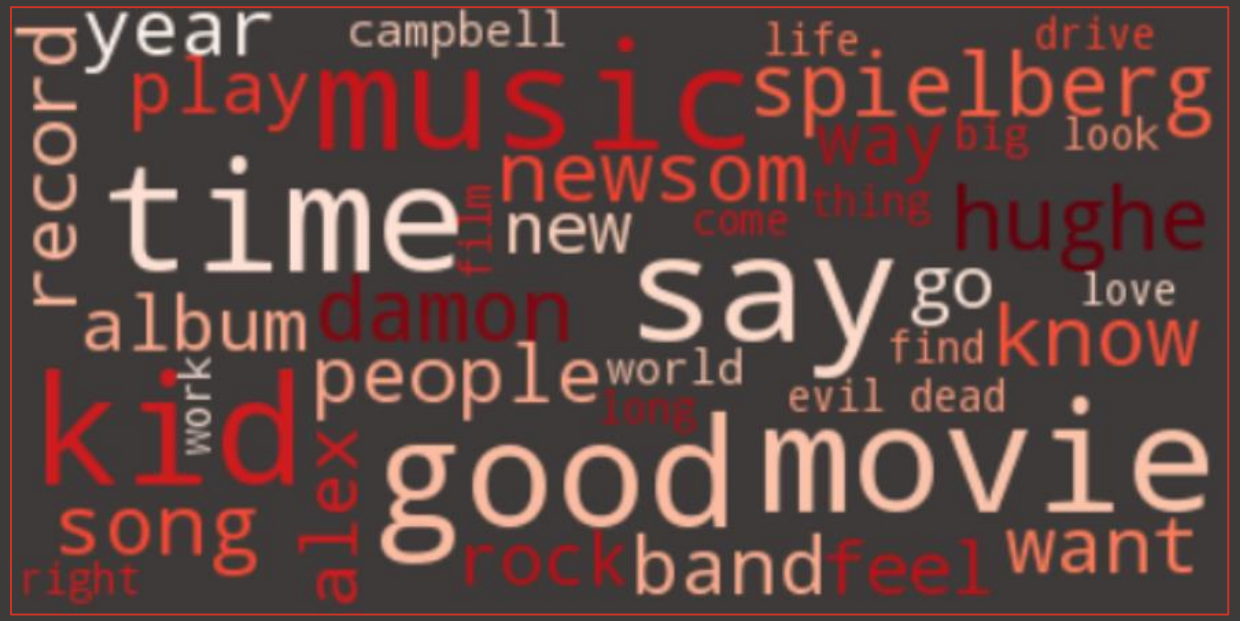
Most Used



Most Unique

Steven Hyden

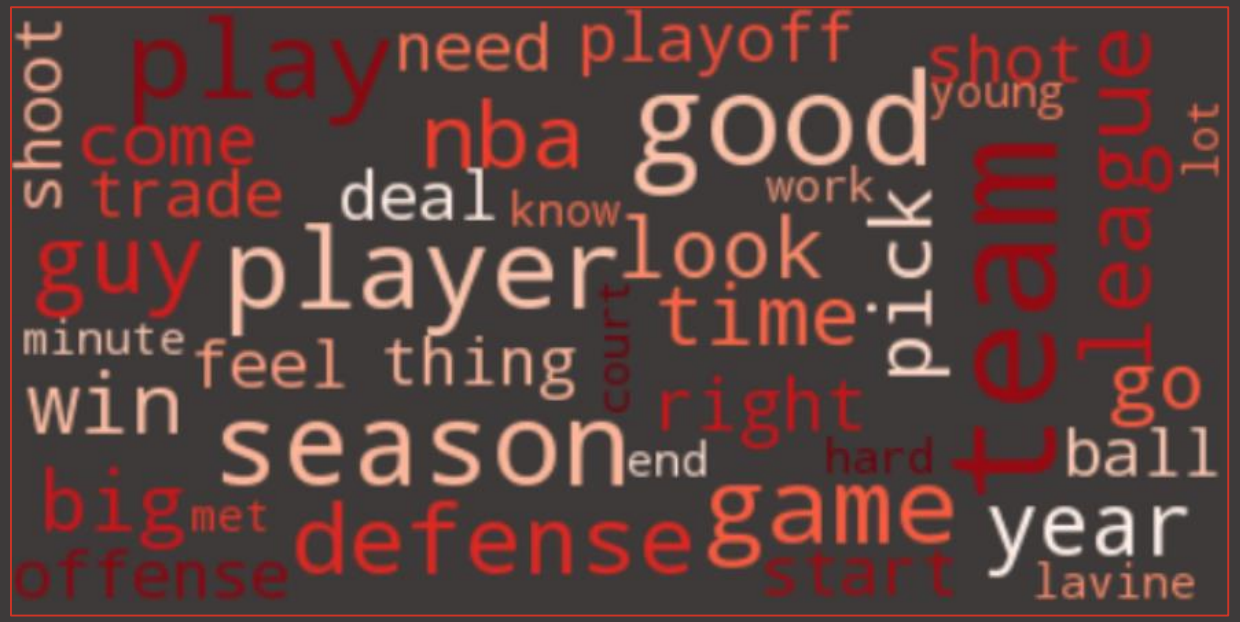
Most Used



Most Unique

Zach Lowe

Most Used



Most Unique