Unsupervised Natural Language Processing

A Textual Analysis of Selected Articles from Grantland.com

 \sim

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

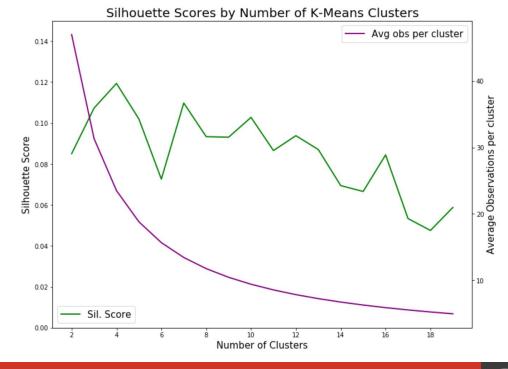
we are close. four more days and there will be professional basketball on tv every night.

Stop words & punctuation removed, remaining words changed to lemmas

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

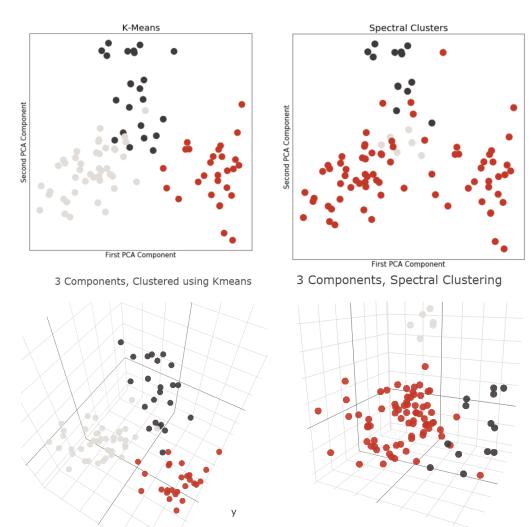
The doc2vec algorithm was used to create vector representations of 94 of the 126 articles, the rest were held out as a test set.

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 Principal Components Analysis.



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

way yearsecondbad football start football season time player season player season player season player season to play game week of the player bowl think to be super bowl to be season to patriot start start seattle football season to patriot start start seattle football season to patriot season to pa

Topic: Basketball

play guypoint good Think bigseason game good time team by yearnba

Iopic(s): Pop ulture/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

opic: Football

Took Same start time of the look of the lo

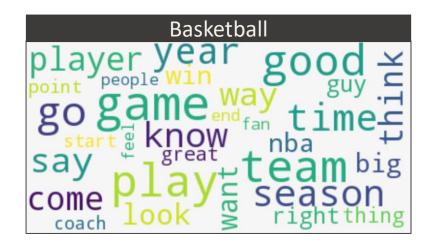
Topic: Basketball

opic: Hockey

bluenew jacket right tought league look long good nhl play to guy go year game year good nhl play to guy go year game thinkremembe

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.

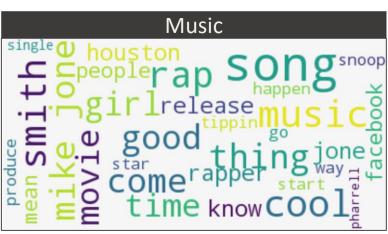


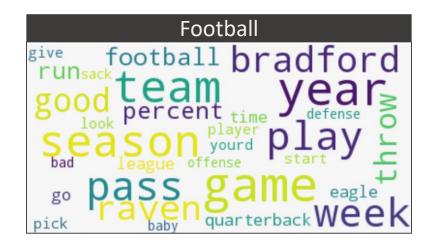
Cinema

studio

know







```
feel human thing sex robot
gofuture exist robot
gerson exist robot
campaign sex think aibo desire look
idea want people sexbot
comesort comesort story real way new year
```

Predicting who wrote each article

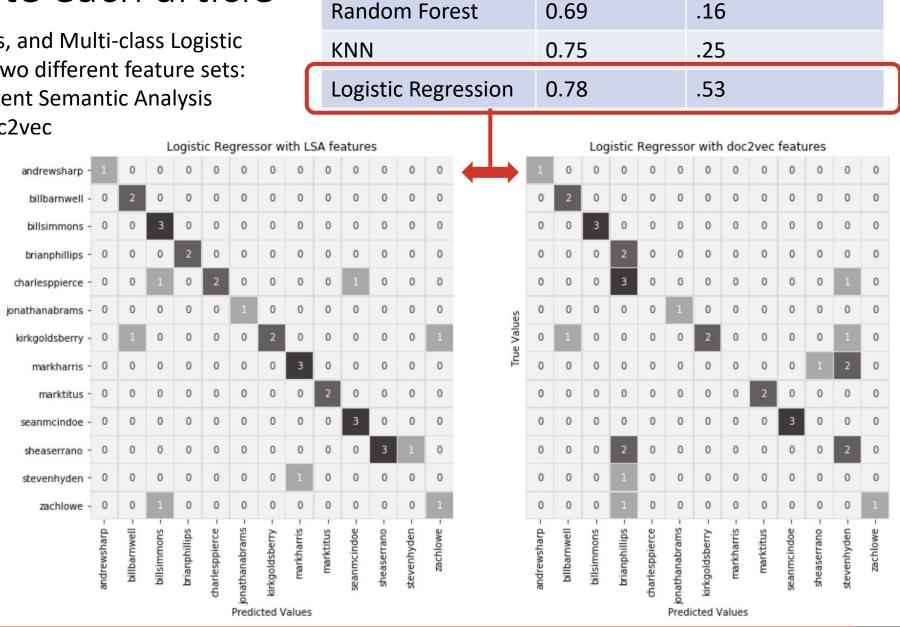
Random Forest, K-Nearest Neighbors, and Multi-class Logistic Regression models were applied to two different feature sets:

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

Most Used & Most Unique Words by Contributor

Bill Simmons

```
granpick year bowlend love year bowlend love play season completed by the player go play good winder good winder good winder good winder good watchtime bottome watchtime bottome by start ight brady
```

```
wilson Sfallrussell
Eduncan

kawhi Sfine duncan

boogie god mailbag protect

god mailbag celtic
```

Andrew Sharp

```
look point play year basketball player make time big to good win team players wantthing players wantthing sayseason nbasketball knick happen and watch week tell knick happen and sayseason make time to be players.
```

```
concern lawsonsteph pacer wade knick of the skile knick perfect september sharp stretch answer dragic
```

Bill Barnwell

```
Season defense in team of throw league sunday week take peace throw take peace throw the part with the pick win
```

```
dvoapanther seahawk dolphin raider seahawk approach raider seahawk approach some brown Tessunday interception Sunday titan cardinal
```

Brian Phillips

```
time of thinksense time of thinksense thinks
```

```
klopp SCULLY andre aibo franzen disney future tennis mulder movelsexfile serena match body liverpool bouchard
```

Charles P. Pierce

```
Tootball coach look people look people look people look people american football guy say say field shift child chi
```

```
peyton owenfantasy tulowitzki younkee arrieta younkee arrieta
```

Jonathan Abrams

```
guy player year wanted millsape team knowed takerobinson of ta
```

```
mile josh family have mile josh phawkportland phawkportland mozgov mozgov mozgov austin recall recall doc dad robinson by doc
```

Kirk Goldsberry

```
numberassist Season big spot pull season big year west brook point of special goal of game team league Shooter Officiency basketball of the look pass
```

```
midrangetarget point and midrange pickup davitare promorrow paint paint rank jump bryount paint range ibaka assist convertefficiency left
```

Mark Harris

```
action colbert box drama decade billion akermangreenlight gross festival audience Cruse jurassic couplefury office clooney
```

Mark Titus

```
year Songame bigguy
time
wisconsinduke
wisconsinduke
kentuckywin
good
think
playcollege basketball
think
arizona tournament
```



Sean McIndoe

```
game team of start want Win season guy keld league Bookey big far Bookey big far Bookey
```

```
hockey division division ranger bruin duck ranger bruin penguin schultz penguin sabre cupgoalie cupghighlightpuck flame to schedule columbus division ranger bruin duck ranger bruin cupgoalie schedule cupghighlightpuck flame to schedule cupgoalie cupgoalie schedule cupgoalie schedule cupgoalie schedule cupgoalie cupgoalie schedule cupgoalie cupgoalie schedule cupgoalie cupgoalie schedule cupgoalie cupgoa
```

Shea Serrano

```
michaelastronaut
prerequisite
special presequisite
vinship
dieselfettyrap
dieselfettygirl
wap
wap
mill
bond
pharrell
conversation
situation
```

Steven Hyden

```
Time new come hughe hughe albumdamon Say find know people world evil dead poople song song pockbandfeel want
```

```
Chrucelisten evil pop seach slang alexhughe cooper
```

Zach Lowe

```
playneed playoff
shoth
come nba good
strade deal know look
guy player look
minute feel thing
win seasonend hard ball
bigmet defense game
offense defense game
good
work
work
work
season
to
go
work
year
to
go
work
to
go
wor
to
go
work
to
go
wor
```

```
preseason martin
stepfit ton position johnson
lavine 3s wing
pauphilly Cap lineup agency
contract met to minnesota draw dribble
```