MSiA 421 Data Mining
Group J: Zanesville

Content Clustering

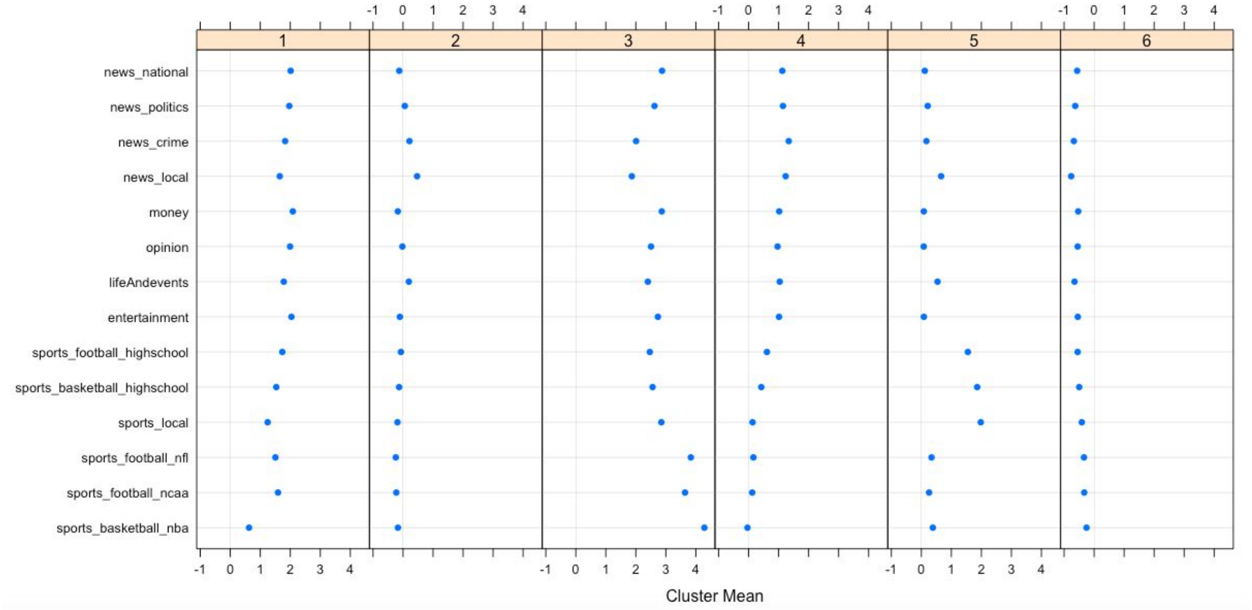
Content Variable Cleaning:
The Zanesville newspaper data initially had 3 sections indicating content, **Section**, **Sub_section,** and **Topic.** We dropped rows with section labels that were non-content related such as 'error pages' or 'contact form' as they are not informative helpful for this analysis. Within the **Section** column, we noticed that there were redundant labels as well as differing levels of specificity. Some columns said 'sports', while others said 'baseball', and still others said 'highschool sports'. We decided to reorganize all of the section, sub_section, and topic values to be more understandably structured.

This required aggregating the different **section** values into large categories such as 'sports', 'news', 'Life&Events', etc. Then we made **sub_section** indicate a subcategory of the larger section, such as 'news:local', 'sports:basketball', etc. Finally, **Topic** was a 3rd level of specificity, allowing for granularity such as 'sports:basketball:nba' or 'sports:football:highschool'.

We then looked at which of these new **section**, **sub_section**, and **topic** values showed up most and decided to focus on those.

Cluster Creation Process:
After identifying our content features based on the number of clicks and subscribers per content tag, we performed log transformations on the counts for each individual and then standardized and scaled our data frame to account for different standard deviations across features. Our initial clustering attempts identified a couple significant outlier observations with orders of magnitude more page views than all other users, which caused significant issues with the interpretability of our solutions, so we dropped those observations (fire_fly_id 2438551 & 5010813). Once we dropped those observations, we were able to fit solutions of 6-12 clusters. The fit statistics (pseudo-F statistic, R-squared) indicated only small benefits to adding more than 6 clusters and we did not find particularly more informative interpretations with more clusters, so we decided on the 6 cluster solution as the best.

Clustering Solution:



Cluster 1 (n = 212, 6.1%):
Medium engagement readers. These readers display a mild engagement with most content types, although have a little bit less local focus within those content types.

Cluster 2 (n = 722, 20.7%):
Light Local news consumers. These readers consume relatively little of every type of news, but consume local news more so than the other categories.

Cluster 3 (n = 100, 2.9%):
Heavy readers, Nationally Oriented. These readers read the most articles between clusters for many categories. Noticeable is that they are less interested in local sports and news than they are in other types of articles. For instance their reading of NBA and NFL articles are much higher than of corresponding sports in Zanesville high schools.

Cluster 4 (n = 386, 10.0%):
Medium readers. This group is engaged with the news across a wide variety of topics with less of an interest in sports.

Cluster 5 (n = 159, 4.6%):
Sports readers. The people who care most about local sports, especially high school and keep up with local news.

Cluster 6 (n = 1902, 54.6 %): Non-readers. Minimal engagement across all categories.