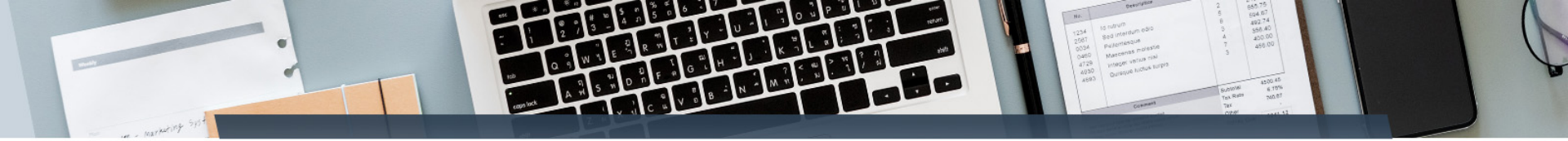
What Makes the News?

A Topic Modeling Project by Max Lan



The State of the Media

- US Newsroom employment declined 25% from 2008 to 2018
- 36% of young adults use social media for news; only 27% use traditional news websites
- What do news publications currently write about and where can they expand?

Methodology

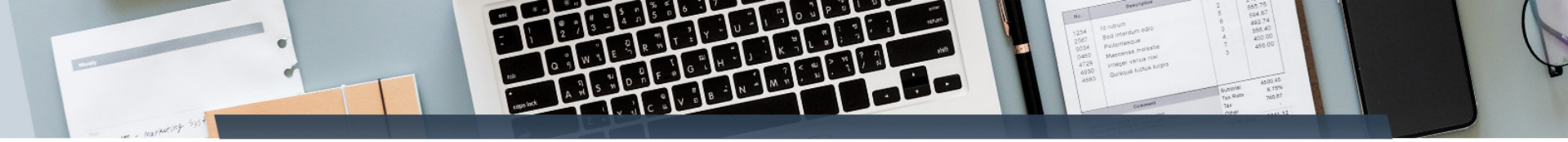
- Data: 143,000 articles from 15 news publications on Kaggle
- Topic Modeling: CountVectorizer, LDA, and A/B Testing
- Readability Analysis with spaCy readability
- Primary Packages used:

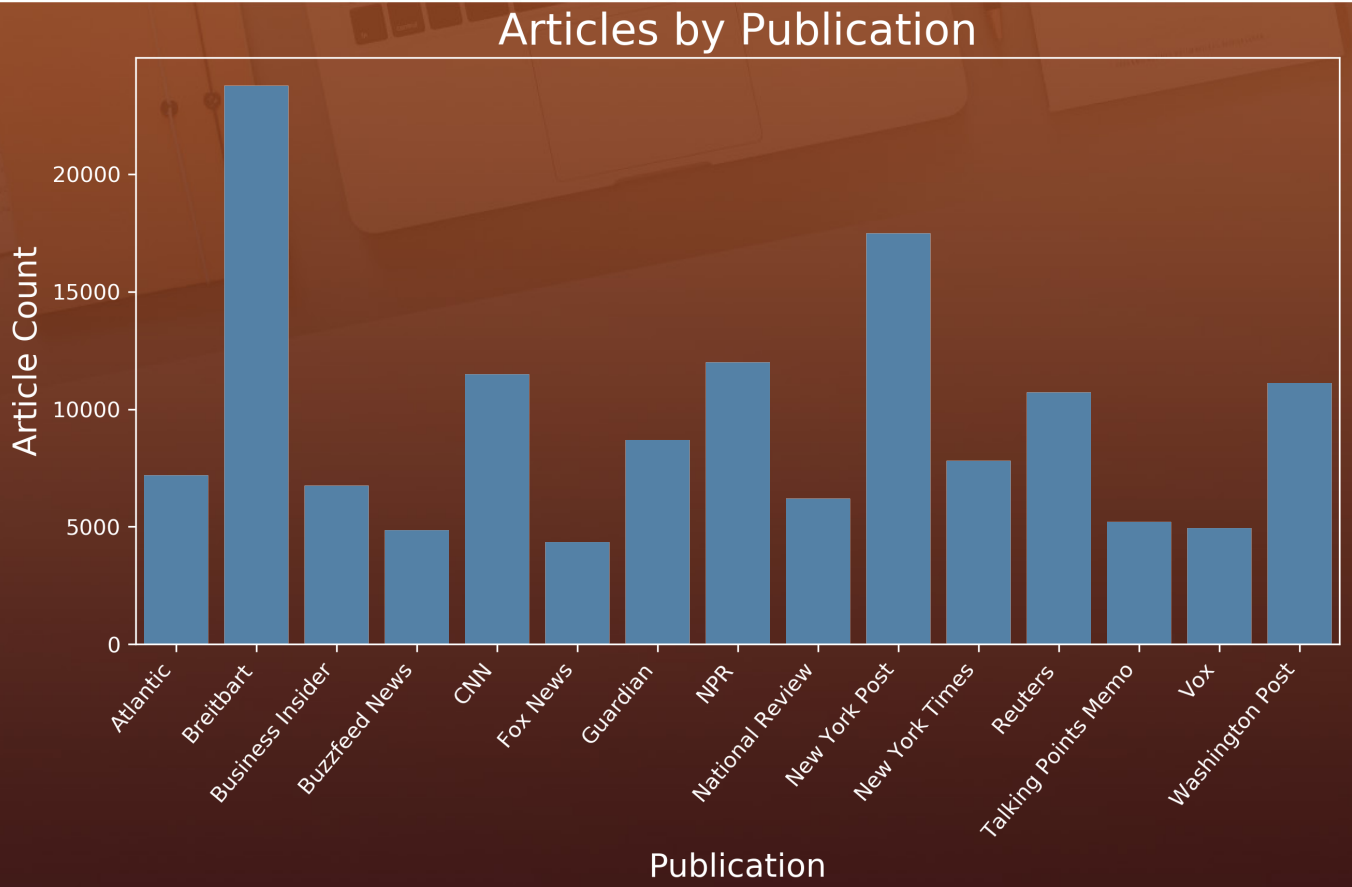
GENSIM topic modelling for humans

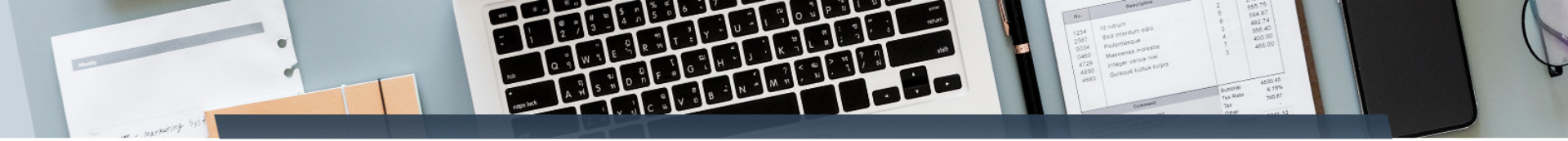
scikit learn

spaCy

SM StatsModels Statistics in Python



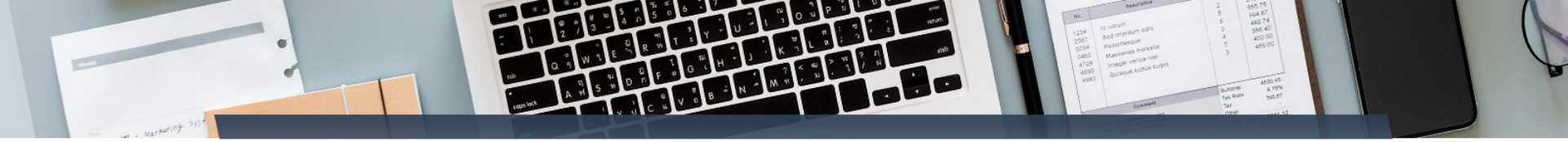

Articles by Publication
Article Count
0
5000
10000
15000
20000
Atlantic
Breitbart
Business Insider
Buzzfeed News
CNN
Fox News
Guardian
NPR
National Review
New York Post
New York Times
Reuters
Talking Points Memo
Vox
Washington Post
Publication



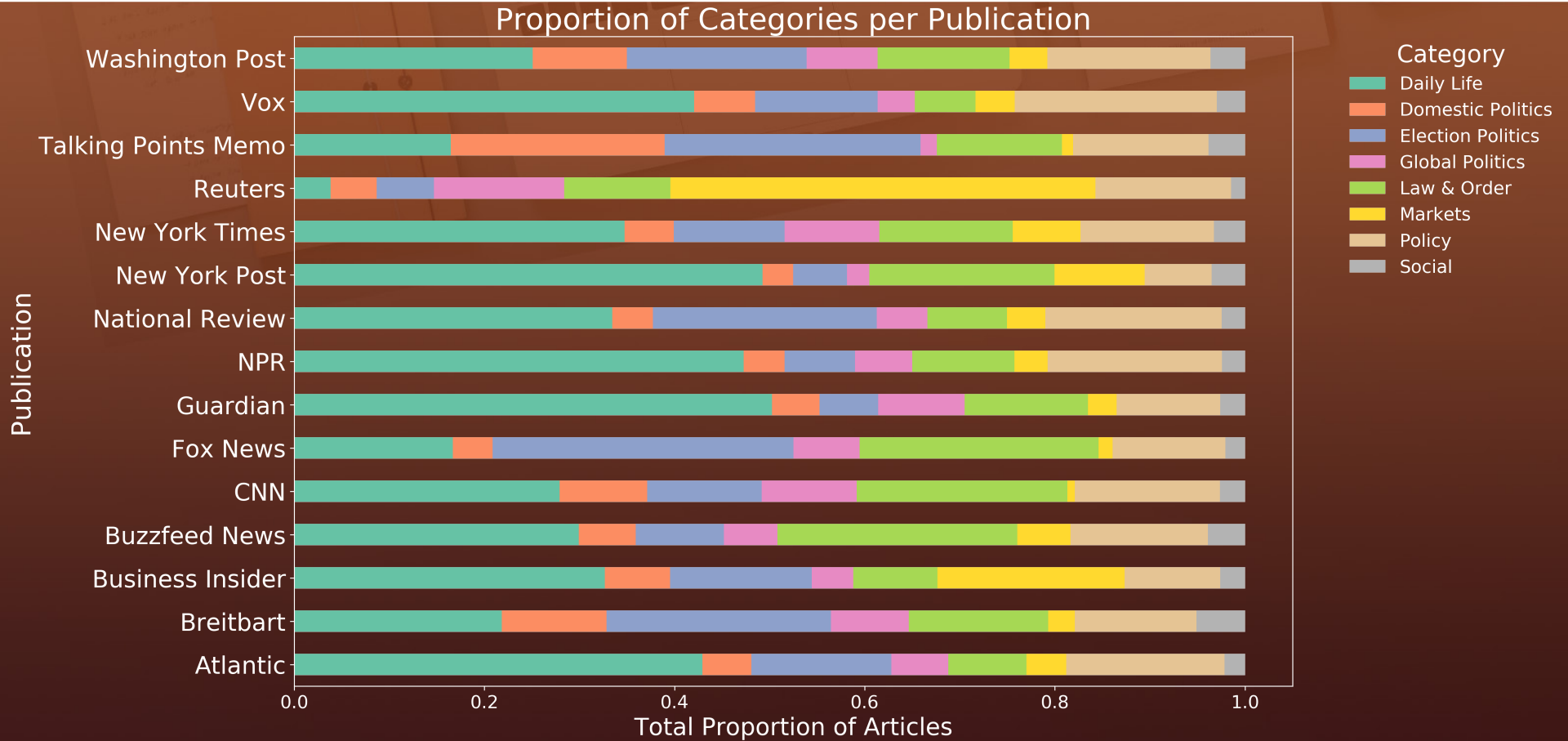
Topics and Categories

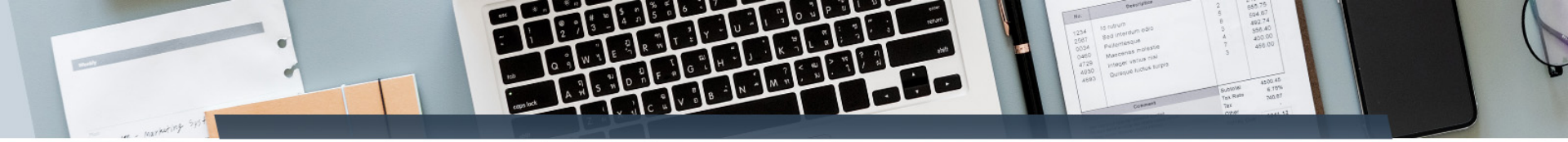
- 24 topics from LDA
  - Most relevant words used
  - pyLDAvis
- 8 primary categories from 24 topics
  - Cosine similarity

- Examples:
- Topic:
  - Law: court, case, law, justice
  - Police: police officer, shooting
- Categories:
  - Daily Life: sports, lifestyle, social media

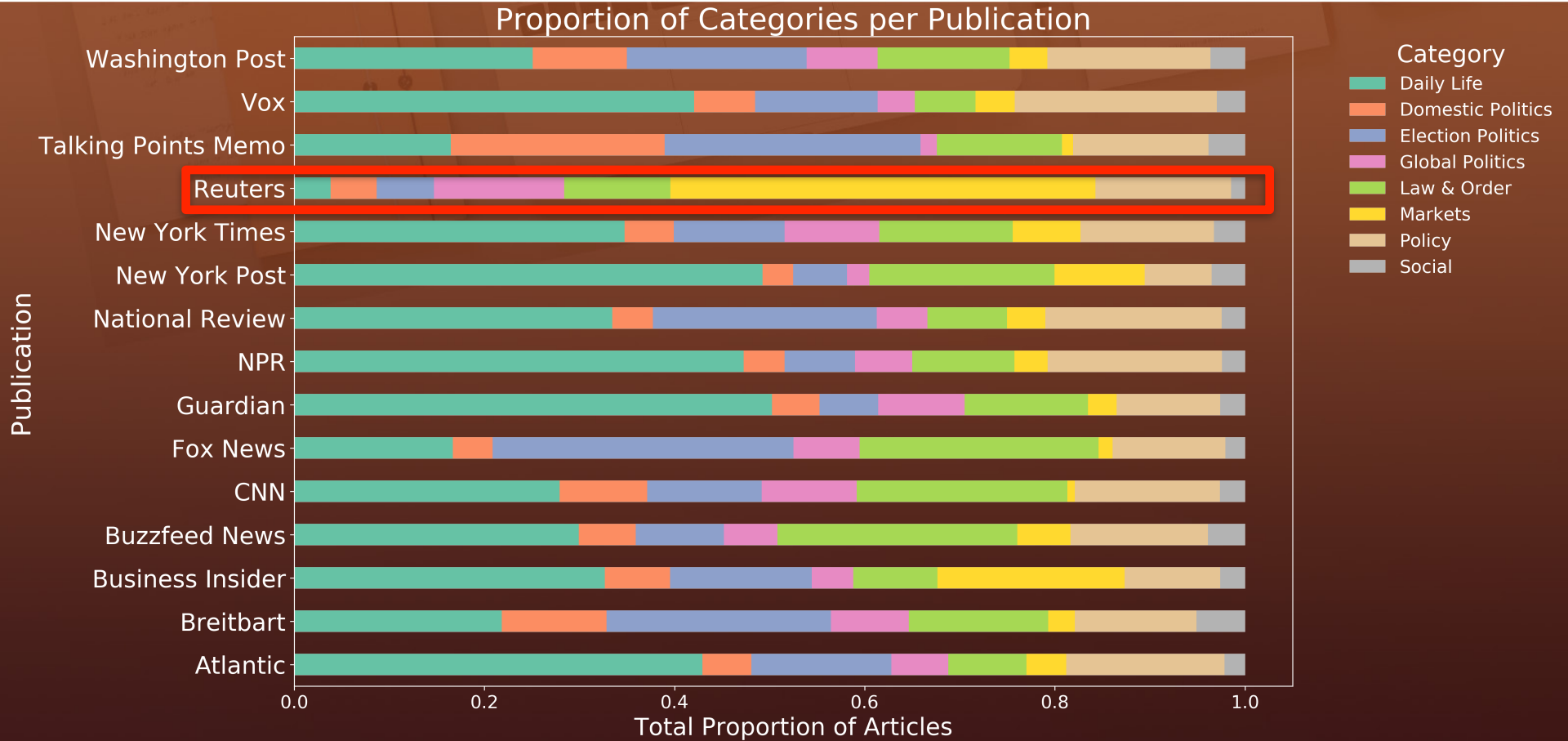


Proportion of Categories per Publication





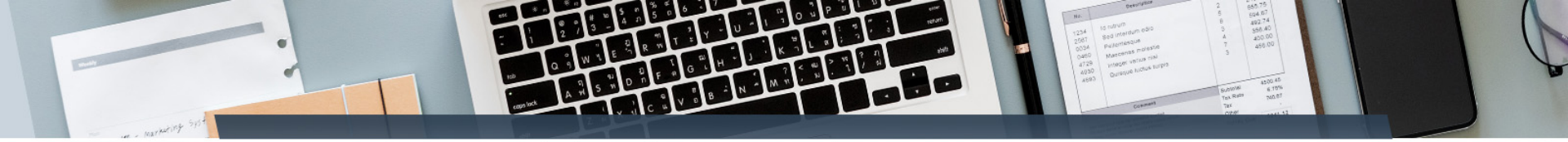
Proportion of Categories per Publication



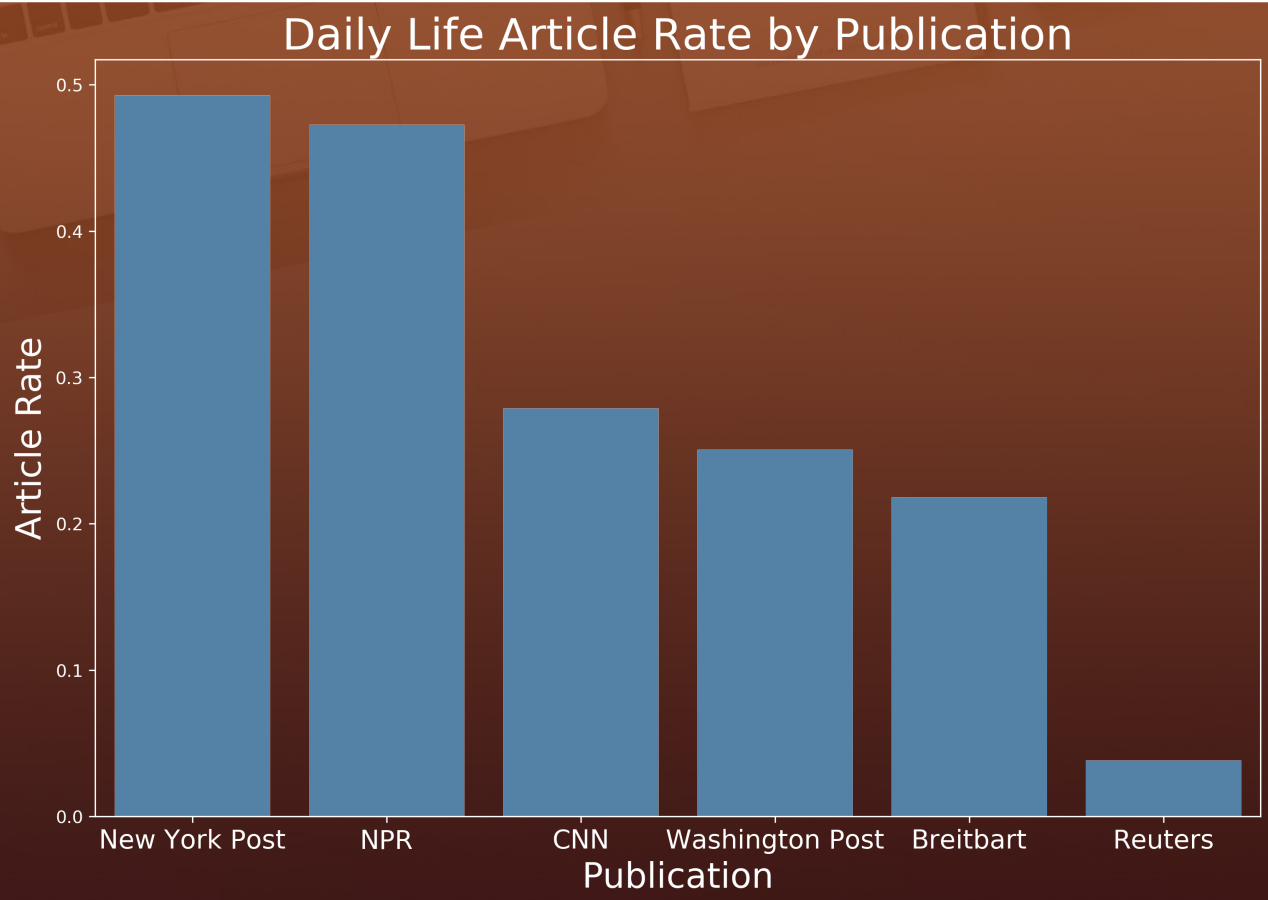
Categories - A Statistical Approach

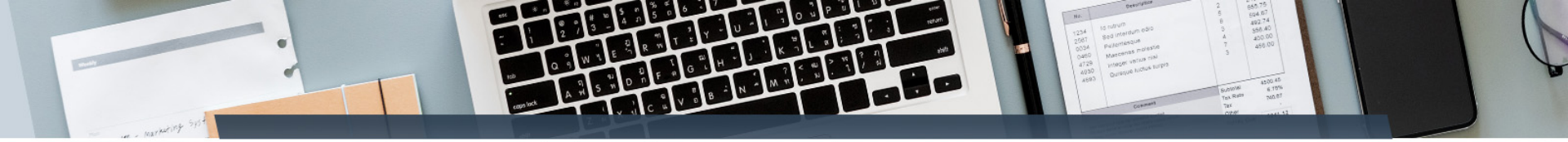
- Multiple hypothesis A/B testing by category
  - Top 6 publications by article count
  - Holm-Bonferroni Correction

- Determine statistically significant differences in article proportion



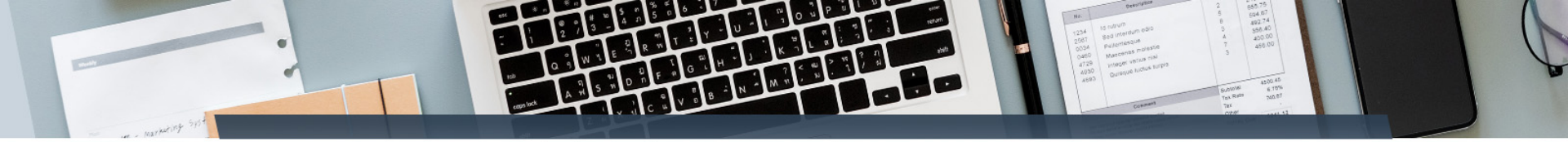
Category	Test Result
Daily Life	Pass
Domestic Politics	Fail
Election Politics	Fail
Global Politics	Fail
Law & Order	Fail
Markets	Fail
Policy	Fail
Social	Fail

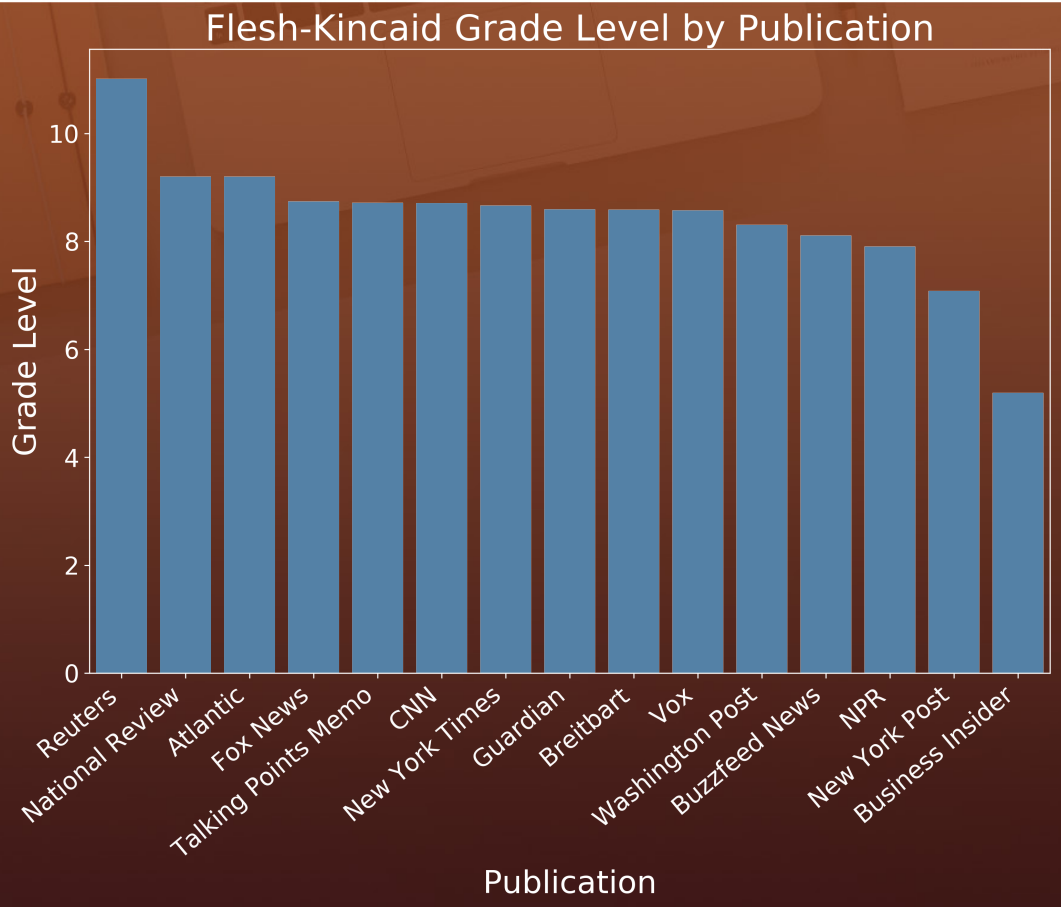


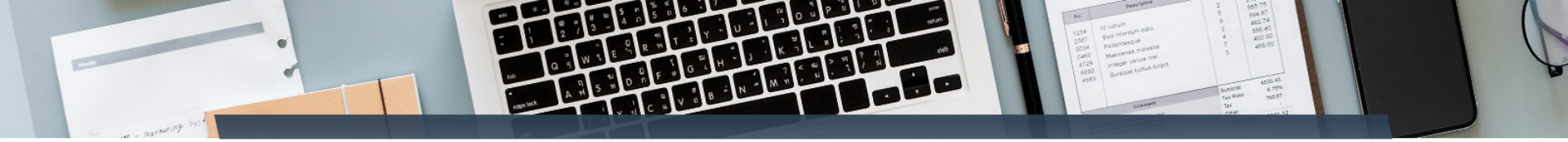


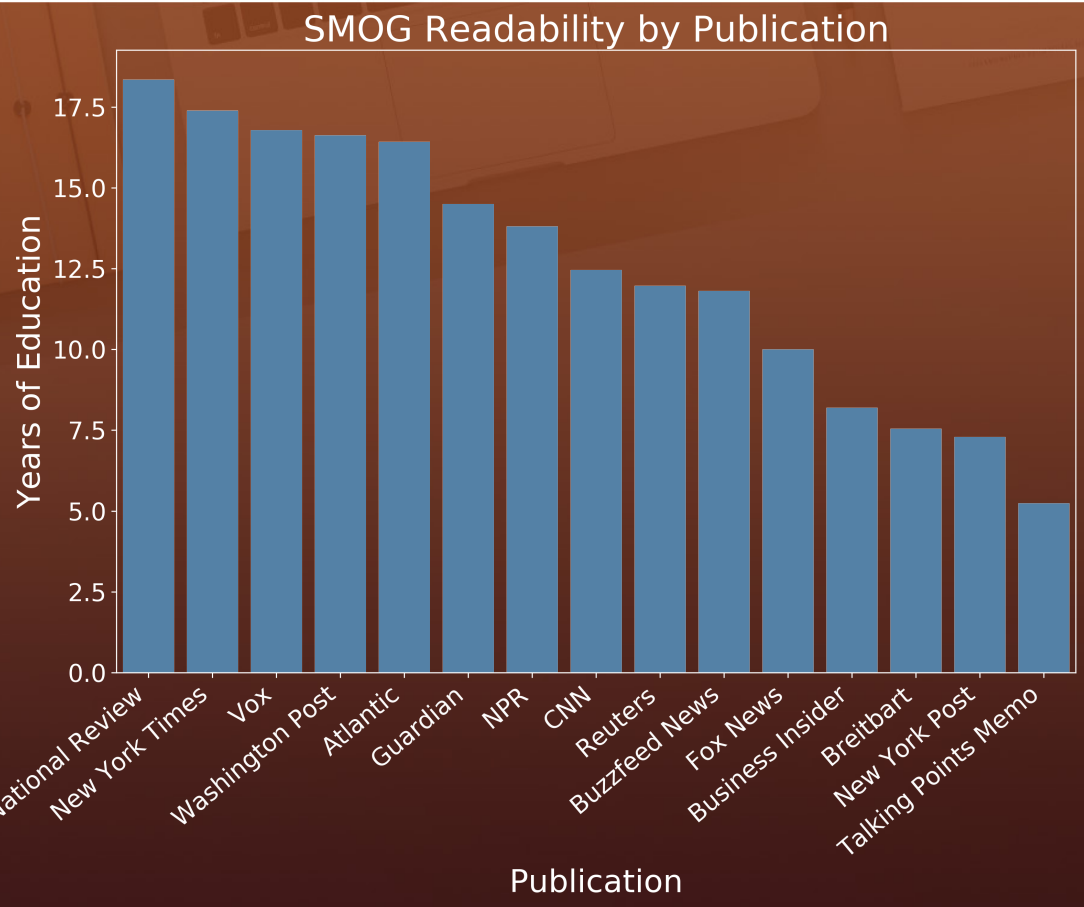
Readability Analysis

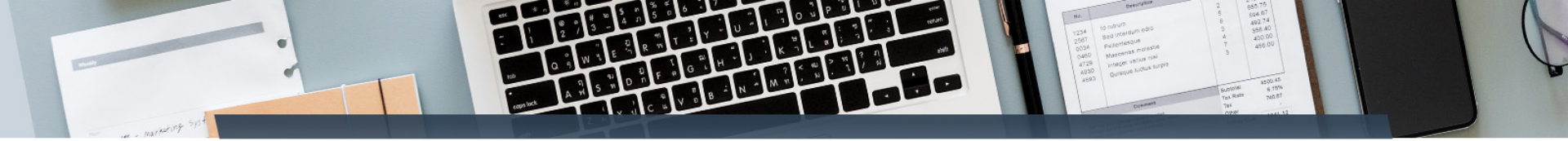
- Examine publications by multiple readability metrics
- Determine if publications focus on different writing levels




Flesh-Kincaid Grade Level by Publication
Grade Level
0
2
4
6
8
10
Reuters
National Review
Atlantic
Fox News
Talking Points Memo
CNN
New York Times
Guardian
Breitbart
Vox
Washington Post
Buzzfeed News
NPR
New York Post
Business Insider
Publication



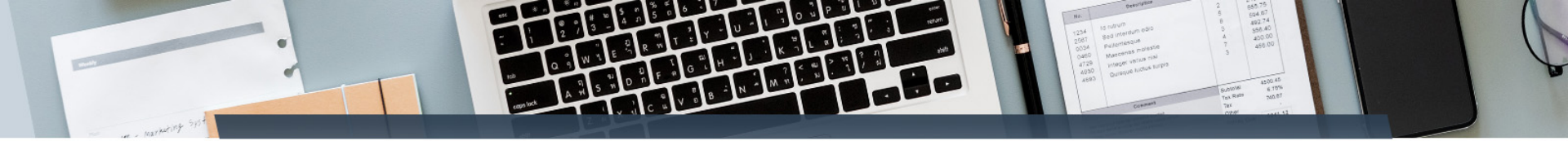

SMOG Readability by Publication
Years of Education
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
National Review
New York Times
Vox
Washington Post
Atlantic
Guardian
NPR
CNN
Reuters
Buzzfeed News
Fox News
Business Insider
Breitbart
New York Post
Talking Points Memo
Publication



Recommendations

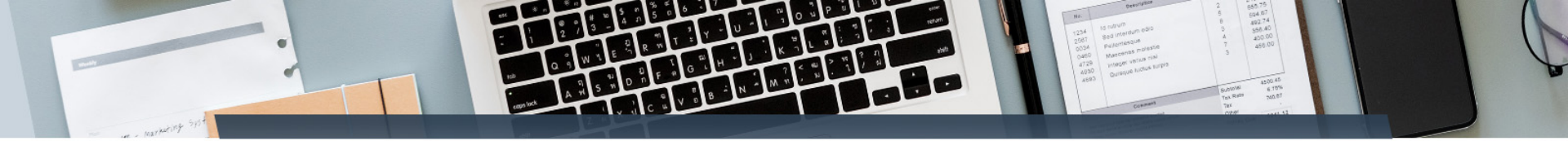
- Daily life is a differentiator:
  - Reuters: more reader-friendly publication with daily life articles
  - New York Post, NPR: consider separate focus on other topics

- Reading level: cater to your audience
  - Breitbart and Talking Points Memo: consider longer/more sophisticated articles or pieces to expand



Additional Topics

- Obtaining more articles across more publications
- Creating a similar article recommender
- Identifying plagiarism between news publications



Thank You!

- Github: https://github.com/mlan93/project_04_NLP
- LinkedIn: https://www.linkedin.com/in/maxlan93/
-