YANR.IN – A Project Proposal

Introduction

My course project is an online web application called yanr.in (Yet Another News Reader) which is basically an Objective News Summarizer and Aggregator. Yanrin means “sand” or “silica” in Yoruba language.

Team

Jaideep Singh – jaideep2@illinois.edu

Background

In recent years, we have seen a seismic shift in how news is reported and consumed. Whereas earlier there was a clear distinction between what was news and what was an opinion, the lines have blurred. This can be attributed to various factors, but the big three are as follows:

1. Rise of social media: The social media bubble is real and creates an echo chamber for extreme opinions that propagate quickly. This leads to highly polarized discussion when the subject does come out in the open.
2. Click-bait driven analytics: Because controversial twist on topics generate more revenue, media online resort to presenting provocative headlines and content. Up and coming online news outlets like Buzzfeed and Breitbart are challenging the big media in share for the public eyes. Even established media like Washington Post and New York times have started doing this.
3. Established media downturn: In recent years, old media has suffered a massive loss of revenue. Giants like New York Times and Wall Street Journals have managed to survive because of their size, but smaller/regional players are involved in record mergers and bankruptcies.

Objective

Yanr.in aims to remove bias in news articles. Instead of giving subjective takes on a subject, it aims to get straight up factual news, instead of clickbait headlines and content customized for maximum revenue. It will aggregate news using different techniques and algorithms and present user with simple outlines of the articles that is straight to the point. This is an everyday news aggregator that can be used by anyone, so it will benefit the general public. News aggregators and document summarizers exist but this kind of project does not exist as of now.

Details

This project will be completed by doing the following steps:

1. **Writing a web scraper**
   This scraper will collect news articles based on the same topic. The websites will range from established media (like Reuters, BBC etc.) to click bait media (BuzzFeed, TMZ etc.). It will also scrape based on affiliation, i.e. websites that lean left (New York Times, HuffPost etc.) and right (Fox News, Breitbart etc.)

2. **Topic modeling** [1]
   Next step would be to use topic modeling to group the articles into topic "themes". An example of a theme can be "Russian Invasion of Crimea" that can contain articles from New York Times and Russia Times, among other news sources.
   Algorithms that will be used in this would be Latent Dirichlet Allocation (LDA) [1] or recent ones using Word2Vec modeling [2] [4].
3. **Multi-Document Summarizer** [3]
   Next step would be to summarize each topic into an outline. There are many ways to do this [6] [7] but for starting out, we can use features such as word frequency, document weights based on type and source of article etc. Last step would be to aggregate all outlines into the web application [8] where user can view and interact with various topics.

Timeline

This is a rough timeline of how the project will be developed. The weeks are based on course week schedule:
Week 7 (current week) – Submit this project proposal.
Week 8 – Build scraper, test on various news website and store articles in database.
Week 9 – Test different topic modeling techniques various news datasets.
Week 10 – Test finalized topic modeling technique on scraped articles database.
Week 11 – Test different multi document summarization techniques on collected topics.
Week 12 – Test finalized summarization technique on scraped articles database.
Week 13 – Project progress report + Building yanr.in web application to view news outlines.
Week 14 – Technology review submission + Work on finalizing code and presentation.
Week 15 – Submit course project and presentation.
After course is finished – Make the scraper, topic generation and summarization real-time.

Assumptions

1. Words in articles that are most common would be more objective in nature.
2. Articles that have higher weightage due to source, author, country or origin etc. will be more objective.
3. Topic modeling would create buckets with distinct and varied topics.
4. More data/articles there are, more accurate this project.
5. The scraper will follow the policies of the news website and the news article will always list the sources.

Bibliography

[1] Various, "Topic Model," [Online]. Available: https://en.wikipedia.org/wiki/Topic_model.

[2] Various, "Latent Dirichlet Allocation," [Online]. Available: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation.

[3] Various, "Word2vec," [Online]. Available: https://en.wikipedia.org/wiki/Word2vec.

[4] M. Z. C. D. Rajarshi Das, "Gaussian LDA for Topic Models with Word Embeddings," [Online]. Available: http://rajarshd.github.io/papers/acl2015.pdf.

[5] Various, "Multi document summarization," [Online]. Available: https://en.wikipedia.org/wiki/Multi-document_summarization.

[6] S. K. S. C. J. S. Souneil Park, "NewsCube," [Online]. Available: https://pdfs.semanticscholar.org/e86a/6d68cb928850e33819372a2112abc0e00600.pdf.

[7] N. M. B. G. Felix Hamborg, "Matrix-based News Aggregation: Exploring Di!erent News Perspectives," [Online]. Available: https://www.gipp.com/wp-content/papercite-data/pdf/hamborg2017b.pdf.

[8] Various, "News Aggregator," [Online]. Available: https://en.wikipedia.org/wiki/News_aggregator.