MA5851 A3 - Web Crawler and NLP System

April 20, 2021

Document 1 - Overview

The phrase "if it bleeds, it leads" can allegedly be first attributed to journalist Eric Pooley as the author of a 1989 New York Times article titled "Grins, Gore and Videotape - The Trouble with Local TV News". At the time, he was angry about the volume of stories with grim and menacing storylines and the fact they were being prioritised over more thoughtful and optimistic pieces (Evaluating Conversations, n.d.). Historically, journalists and news producers aimed to report events as they actually happened, with an even balance between the depressing or otherwise negative stories and those showcasing the more positive side of life. Over time, however, it has evolved into a race to find the most sensationalist, stirring and spectacular stories to maintain ratings and profits (Serani, 2008). As result, the media has become less a source of information and more of another form of entertainment (Dmitrieva, 2017), where those working-class heroes quietly go about their life with nary a mention (Williams, 2005).

The topic of news sensationalism and an effort to understand why the human race is so attracted to it has been the object of many past studies. For example, Davis and McLeod (2003) found that thematic categorisation of the front-page news stories they had analysed thematically were remarkably stable over time and the contents of sensational news were unlikely to be socially constructed. More recently, after analysing the content of 14 television systems and 29 television stations, (Arbaoui et al., 2016) found that dependency on commercial revenue stimulates the use of vivid storytelling. Finally, Kilgo et al. (2016) investigated the relationship between the sensational treatment of news categories and values, and their associated interaction numbers on social media. They concluded that 'hard' news topics such as government affairs were sensationalised just as often as crime or lifestyle and society articles.

The intention of the broader project was to contribute to the research on sensationalism in the media by investigating the extent to which it is present in current times. The objective of this particular component was to develop and demonstrate a proof-of-concept product involving a single online news website that could then be used as a model for production-level development of a more general scraper[1] that could be used on multiple news websites. It consisted of three main parts: the first part involved the development of the web scraper that could automatically extract the desired content from a target website (along with some preliminary data cleaning and exploration), the second part involved a preliminary sentiment analysis using VADER and TextBlob, and the final part used non-negative matrix factorisation to investigate headline and summary topics.

[1] Whilst the terms web scraper and web crawler are often used interchangeably, they are actually related but different concepts. Web scraping refers to the targeted extraction of content from one or more websites whereas web crawling refers to the automated browsing of the internet for content indexing purposes (Patel, 2010). This project almost exclusively uses the term web scraper.

References

Arbaoui, B., De Swert, K., & van der Brug, W. (2016). Sensationalism in News Coverage: A Comparative Study in 14 Television Systems. Communication Research, 47(2), 299-320. https://doi.org/10.1177/0093650216663364

Davis, H., & McLeod, S. L. (2003). Why humans value sensational news: An evolutionary perspective. Evolution and Human Behavior, 24(3), 208-216. https://doi.org/https://doi.org/10.1016/S1090-5138(03)00012-6

Dmitrieva, K. I. (2017). Why are we fascinated with violence? An investigation of mass media's role in depicting violence as entertainment (Publication Number 574) [Senior Honors Projects, The University of Rhode Island]. DigitalCommons@URI. https://digitalcommons.uri.edu/srhonorsprog/574/

Evaluating Conversations. (n.d.). "If it bleeds, it leads". http://evaluatingconversations.weebly.com/if-it-bleeds-it-leads.html

Kilgo, D. K., Harlow, S., García-Perdomo, V., & Salaverría, R. (2016). A new sensation? An international exploration of sensationalism and social media recommendations in online news publications. Journalism, 19(11), 1497-1516. https://doi.org/10.1177/1464884916683549

Patel, H. (2010, March 16). Web scraping vs web crawling: What's the difference? DZone. https://dzone.com/articles/web-scraping-vs-web-crawling-whats-the-difference

Serani, D. (2008). If it bleeds, it leads: The clinical implications of fear-based programming in news media. Psychoanalysis and Psychotherapy, 24, 240-250. https://doi.org/10.3200/PSYC.24.4.240-250

Williams, A. (2005, December 6). Celebrating the everyday heroes. Townhall. https://townhall.com/columnists/armstrongwilliams/2005/12/06/celebrating-the-everyday-heroes-n1121673