CSC2552: Review 1, Paper 1

Due on January 24

Ashton Anderson

495 words

Thomas Hollis

Paper 1

This paper, by Michel, Shen et. al. in conjunction with some Google employees, aims to analyse culture and cultural evolution through ready-made big data provided by the Google Books corpus. The main result of this paper, which coined the term *culturomics*, is a series of independent hypotheses postulated to explain the cultural dynamics behind changes in prevalence of certain words or groups of words in books dating from the year 1800 to the year 2000.

The main weakness of this paper is its vulnerability to the fallacy of argument by selective observation. While each n-gram plot presented in the paper seems to be of academic value as it can be explained by a neat cultural hypothesis, this is not necessarily true for all n-grams. Indeed, it seems like some selective argumentation has been made, perhaps because of the high profile industrial partnership with Google. For example, if an n-gram becomes popular it is often given an acronym which results in messy n-grams that split across multiple graphs representing the same concept [1]. Limitations like these are almost never discussed in the paper as it is more of a demonstration of what is possible using Google-powered technology. Indeed, there is a trade-off between a consistent methodical but narrow set of analyses rather than a wide ranging but inconsistent palette of analyses. Although the latter has limitations, one can see how this is also a big advantage for the generation of future papers. Indeed, the main role of an exploratory paper such as this one is to reveal the breadth of possibilities of this new academic field, thus encouraging further publications. Other notable weaknesses include a dirty dataset, which relies on OCR methods with higher error rates than modern ML-based OCR, and irrelevant claims made for publicity purposes, such as multiple paragraphs of little relevance describing the vast size of the dataset used.

Conversely, a significant strength of this paper is the strong references back to alternative forms of cultural metrics. Indeed, by such comparisons with various dictionaries [2] [3] the paper is able to benchmark the performance with certain expected metrics. Other notable strengths include a highly ethical approach to research, significantly challenging in big data projects, as well as the avoidance of sensitive human data.

The implications of this paper are surprisingly important as they reveal a wide variety of possible methods for cultural analysis using simply the prevalence of words in books over time. Something that could have been done in this paper is a discussion of exactly which types of cultural analysis are possible using Google N-gram and which types of analysis are fallacious thus should be avoided. While more cautious language is ultimately a tradeoff with stupendousness, it seems that the scales could be balanced better toward a more rigorous methodology. It is worth noting that this paper has nonetheless inspired many others to complement this new field through more meticulous methodologies [4] [5].

[1] Google Books, Google N-gram Viewer for BFF, available at: https://books.google.com/ngrams/graph?content=BFF&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2CBFF%3B%2Cc0, Accessed: 21 Jan 2019

[2] Webster's Third New International Dictionary of the English Language, Unabridged, P. B. Gove, Ed. (Merriam-Webster, Springfield, MA, 1993).

[3] The American Heritage Dictionary of the English Language, Fourth Edition, J. P. Pickett, Ed. (Houghton Mifflin, Boston, 2000).

[4] Roth, S. (2016). Fashionable functions: A Google ngram view of trends in functional differentiation (1800-2000). In Pol. & Soc. Act.: Concepts, Methodologies, Tools, and App. (pp.177-203). IGI Global.

[5] Zeng, R., & Greenfield, P. M. (2015). Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values. International Journal of Psychology, 50(1), 47-55.