MS&E 231 HW3 Part I

Andrew (Foster) Docherty, Michael Spencer, Jorge Nam Song

Social science has for many years now been a part of the craze for "big data" solutions to modern questions and has seen numerous high-quality results because of it. With recent advances in storage and computation drastically reducing the time, effort, and cost associated with conducting rigorous analysis on massive datasets, we see old questions being answered with refreshed approaches and new questions emerging. Tackling these new questions, which are commonly of orders of magnitude greater scope and scale than traditional ones, open exciting new doors for academic research and productized applications. We will investigate a few of these questions new and old in the context of two example datasets, one a corpus of ~4% of all the books ever published (hosted by Google)[1] and the other a set of 2010 search queries on Google, Bing, and Yahoo! (collected by Microsoft Research).[2]

A potential research question to dive into with the Google book corpus data is understanding the rise and fall of colloquial language over time. Colloquial language typically refers to conversational or informal speech, but collecting data on the contextual usage of such language can be significantly challenging. Using this corpus of scanned books, we could at least take a crack at the written use of colloquial language. It is hard to imagine answering this question at the size and scale that is possible with such data: with over 5 million digitized books, this dataset is unique on several dimensions. For one, we would expect the breadth of literature that this dataset captures to represent variations in the ordering and usage of language far better than a traditional, smaller dataset. This is key to developing a representative sample of the full corpus written work, and will likely have a strong influence on the relative weighting of particular uses of colloquial language (e.g. "don't be chicken" may not be well-represented in British English where it is not as commonly used). In addition, this breadth goes back to works from the 1500s, offering valuable information on the timing, and thus trends, of written language. Moreover, this dataset is unique in that it retains all of the word data in-context. That is, we are not using summary statistics or other aggregations of book metadata that may leave out key grammatical and sentence-level context.

There are noteworthy downsides to approaching this question with such a massive dataset. The first is the learning curve to working with such data. Understanding of how to wrangle, clean, and pipe the data into a form that is ready for large-scale analysis is not a trivial task, and especially with inconsistencies in book formatting and encoding, these may present real challenges to handling the 5+ million books at one's disposal. Furthermore, while it is fantastic to have language presented in-context, this does present a challenge for knowing how to interpret the words themselves. However, a large body of research into how to encode grammatical relationships, conduct sentiment analysis, and understand natural language on a large scale has matured in the past few years in response to this challenge. With modern packages for running for example NLP analysis, questions such as understanding how colloquial language is used and has changed over time can be feasibly tackled.

Another question to ask is understanding and potentially predicting criminal behaviors such as gun shootings or harassment through the full log of search queries. People nowadays highly rely on online search. They use it for all kinds of reasons from checking the weather to

[1] Michel et al. (2011 Jan 14). Quantitative Analysis of Culture Using Millions of Digitized Books. *SCIENCE*, 331. Retrieved from https://5harad.com/mse231/papers/michel_et_al_books.pdf.

[2] White et al. (2013 May 1). Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association*, 20(3). Retrieved from https://5harad.com/mse231/papers/white_et_al_pharmacovigilance.pdf.

finding explanations for anything such as their emotions or sickness. Although people are different and their reasons behind their behaviors could vary, there could exist a pattern of behavior that could potentially lead to criminal acts. Similar to White et al. the prolific use of online search to gather information could resemble a pattern of behaviors which could be used to prevent criminal acts from happening. In addition, such patterns could also be used to analyze some of the reasons behind such behaviors and eliminate or reduce those factors.

Similarly to using the Google book corpus, parsing the data into a form to be efficiently analyzed could be a real challenge. According to some Google statistics, Google receives on average more than 60,000 searches per second, from which around 15% of them have never been searched before. Although the search logs might be stored in a similar format, the amount of data stored and to be analyzed would be enormous and expensive to compute on.

Modern day capabilities in data storage and computation have significantly widened the set of potential scientific questions we can answer. In the social sciences, this means that many problems of the last few decades demand a second look using comprehensive data and complex, data-hungry methods. Further, it means that we as a research community have much lower technical barriers to answering our most extraordinary questions. Hopefully, questions like the ones explored here and more will drive us to new, fascinating conclusions about ourselves and society as a whole.