SOCI 280 101 2023W1: Dear (My) Data

Jonas Stettner

Throughout the “Dear (My) Data” assignment, I explored my emoji usage on WhatsApp over the course of one week. In this report, I aim to provide an account of this process, detailing the decisions I made and the thought processes behind them. I will begin by sharing the motivations behind my choice of data source and the methods I employed for data collection. Subsequently, I will delve into the choices I made regarding the visualizations and the techniques I used to create them. Finally, I will offer a reflection on the outcomes and insights gained from this assignment.

Data Source

The task was to collect data on some aspect of our lives that upon analysis could potentially yield insights. Instead of actively recording data, I wanted to use data already collected by the digital services I use, as I believed it would provide me with an understanding for the knowledge these providers have about my live, sometimes in very personal areas. I considered using data from my Spotify listening history, contributions on GitHub, Google Maps Location History, and WhatsApp text messages. While this does not necessarily separate them from other comapnies that store digital traces for profit, the companies behind the services on the list provide ways to access and download personal data. Spotify and GitHub have convenient APIs for this purpose. Google offers a graphical user interface to access data collected by its services, including Google Maps. WhatsApp stores text messages locally and, if enabled, in the cloud. While the data file isn’t designed for direct user access and requires a decryption key, WhatsApp does allow manual export of messages from individual chats. However, this method is time-consuming if you want to access all your data at once.

Despite these limitations, I decided to use WhatsApp data. In my opinion, this is the most revealing and potentially most insightful data I could collect out of the options I mentioned beforehand. While a location history is also very personal and uncovers much of a persons life, sometimes the context of a location is not clear and, in terms of generating insights, I already know that there is a pattern to my movements. WhatsApp data is even more personal. This platform has witnessed intimate conversations and interactions. I have had disagreements, offered condolences, organized social gatherings, confided in others about my life’s challenges, and expressed affection for those close to me. The fact that all of this is potentially accessible to a company is worth exploring and reflecting on.

One limitation of this data source is that it is unstructured text data. While chats can indeed contain sensible information, information can only be extracted from raw text by humans that would have to read the chats one at a time. To generate spreadheet like data, texts need to be processed with predetermined goals. For this assignment, I wanted to find out if WhatsApp data could reveal something about my emotional state throughout the week. Emotional state in text messages is often intentionally revealed by using (or not using) emojis, so simply counted what emojis I used for a week. To do this, I systematically went through all my chats, starting from the first message on Monday, and recorded this data in a spreadsheet. Each day had its own row, and each emoji was represented as a column. Since emojis have various visual representations, I used standardized Unicode representations as column titles for consistency.

While there are more sophisticated tools available for detecting emotions in text, I decided not to use them for several reasons. Firstly, such methods appeared to be beyond the scope of this assignment. Secondly, given the personal nature of this data, it felt inappropriate to analyse it by utilizing automated algorithms. Some messages in these chats hold emotional significance for me and some I even avoid revisiting to prevent reliving the emotional states of their context. The thought of storing them in a folder for anonymous programs to process into numerical representations of reduced meaning felt uncomfortable. Additionally, the fact that the raw data includes messages from other people, even though my analysis would have focused only on messages I wrote, it seemed both unfair and unsafe to process this data without their explicit permission.