Chunge Wang
NetId: chungew2
Team Name: CW
CS410 Text Information System

Instagram Text Content Friendliness Analysis for Underage Users Project Proposal

Choice of Theme

- I have chosen *Theme 5: Free Topics*.

Section 1

- What are the names and NetIDs of all your team members?
    - Our team consists of 1 member
        - Name: Chunge Wang
        - NetID: chungew2
- Who is the captain? The captain will have more administrative duties than team members.
    - Captain: Chunge Wang

Section 2

- What is your free topic? Please give a detailed description.
    - My free topic aims to develop a Python app and scoring algorithms that collates Instagram text from a particular public Instagram profile and provide standardized topical scores that inform my users of how friendly an Instagram profile is for an underage viewer. The ideal user group of this app will be parents who have underage children. Those scores will reflect educationalness, general emotional positivity, violence, and general emotional negativity. To use this Python app, users input a public Instagram username, and the app will output numerical information that describes the friendliness of this profile for an underage viewer.
- What is the task?
    - The task is to utilize Python Instagram scraping frameworks to systematically extract text data from a public Instagram profile, algorithmically compute topical scores text against representative buckets of keywords (where each bucket represents a categorical score) for the text data, and present the results in an informative and digestive manner for the end user of this Python app.
- Why is it important or interesting?
    - This topic is important because children consume much social media content nowadays without quick informatics that inform children and parents of the harmfulness of the social media content for their cognitive development. Recent research points to that “individuals who are involved in social media, games, texts, mobile phones, etc. are more likely to experience depression” (Karim). It is evident that the cognitive development of children is crucial for future humanity. Providing easily digestible and standardized age appropriateness metrics with text

information analytical techniques based on the text content of social media content is an essential step to inform underage viewers and their guardians to make fast decisions that facilitate healthy cognitive development for children. Also, it is interesting because algorithmically computed standardized psychometrics for social media content appropriateness for underage viewers is both technically and definitively challenging. For example, assigning a score to indicate whether an Instagram post is educational is still an underexplored technical task.

- What is your planned approach?
    - My planned approach is to use a Python Instagram scraping framework, develop a numerical method that algorithmically calculates the similarity between a topical keyword bucket (e.g. educationalness) and the text data from the Instagram profile, and use Python to implement the similarity score calculation.
- What tools, systems or datasets are involved?
    - Tools
        - Python
        - Instaloader (python library for scraping Instagram posts of a given profile)
        - Gensim (python library for topic modeling)
    - Dataset
        - Google News Corpus (dataset for creating topical buckets)
- What is the expected outcome?
    - The expected outcome is when a user inputs a valid Instagram profile username (e.g. *joebiden*), the Python program will automatically download Instagram text data of this profile, calculate underage friendliness scores and present them in a readable and digestive manner for the user.
- How are you going to evaluate your work?
    - A small (due to limited resources for this one-person team) human-evaluated test dataset will be manually created for evaluating the efficacy of this Python program. Once the numerical scores have been generated from this Python program, they will be compared against the test dataset and their human evaluated scores and topical assignments.

Section 3

- Which programming language do you plan to use?
    - I plan on using Python as the programming language.

Section 4

- Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.
    - Our team has 1 person, so the expected workload is at least 20 hours.

        - Write code that scrapes and preprocesses Instagram profile text data (4 hours)
        - Design metrics that define underage appropriateness of Instagram text content (3 hours)
        - Write code that computes topical scores of underage appropriateness of Instagram text content (5 hours)
        - Write evaluation code that shows the efficacy of this project’s Python program (1 hour)
        - Write code that displays the underage friendliness scores and evaluation results (1 hour)
        - Create small test dataset for evaluation (3 hours)
        - Progress report (3 hours)
        - Software code submission with documentation (2 hours)
        - Create usage tutorial presentation video (4 hours)
    - Total estimated hours: 26