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

Project Progress Report

Introduction

This is a project progress report for my team of 1 member. My project is called *Instagram Text Content Friendliness Analysis for Underage Users*.

(1) Progress made thus far

Currently, I have made progress in the following tasks:

- I have identified tools and frameworks that assist in the creation of the crawling logic and data preprocessing of the Instagram data that the Python app will retrieve. Now, users can input an Instagram profile name and download the top ten posts and photos from an Instagram profile, given that there is no rate limiting applied from Instagram's server.
- Writing crawler logic in Python for retrieving the text information and photos from Instagram posts of an Instagram profile
- I have written python code for preprocessing the text, including stop word removal and tokenization.
- I am in the process of finding ways to calculate how much a tokenized text dataset has closeness to a central theme word like "education".

(2) Remaining tasks

- Design metrics that define underage appropriateness of Instagram text content
- Write code that computes topical scores of underage appropriateness of Instagram text content
- Write evaluation code that shows the efficacy of this project's Python program

- Write code that displays the underage friendliness scores and evaluation results
- Create small test dataset for evaluation
- Software code submission with documentation
- Create usage tutorial presentation video

(3)Any challenges/issues being faced

The main challenge I have encountered is the rate limiting of Instagram APIs. This poses the challenge where when I scrape photos and text from Instagram with Python code, the GET API requests would fail because Instagram is limiting the number of requests my Python code could make. I tried to overcome this challenge by implementing Instagram login functionality and using different cookies from browsers. Usually, the solution is to wait out the API blocking period, but this has seriously impaired the crawling capabilities of the app.

Also, designing a way to score the age appropriateness based on a tokenized text dataset is complex as well. I may resort to a basic set of metrics for this purpose to assign scores that describe how age appropriate an Instagram profile is based on its closeness to a topic such as “emotional negativity”.