Computation and Discrete Structures III

SOCIAL MEDIA CONTENT MODERATOR

Introduction

Social media platform needs automated moderation of the posts that users create.
This project uses:

- Regex: Extract Mentions, Hashtags and URLs.
- DFA: Classify posts as safe, review and violation.
- FST: Censor offensive words and hate speech
- CFG: Validate structure and format the post with html.

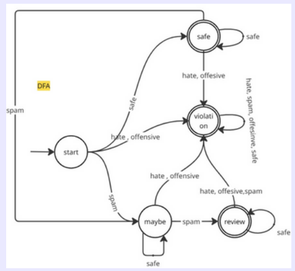
Objective

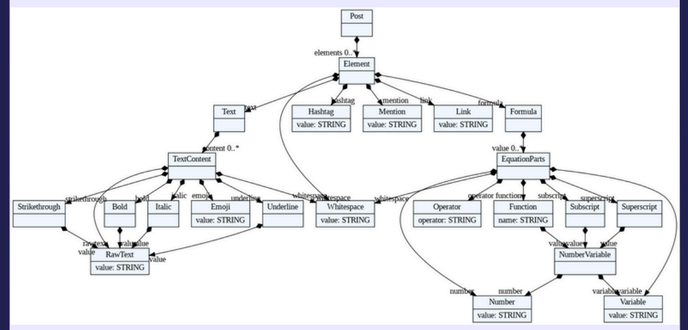
Develop a post moderator that detects policy violations, classifies posts and transform's texts using formal language theory

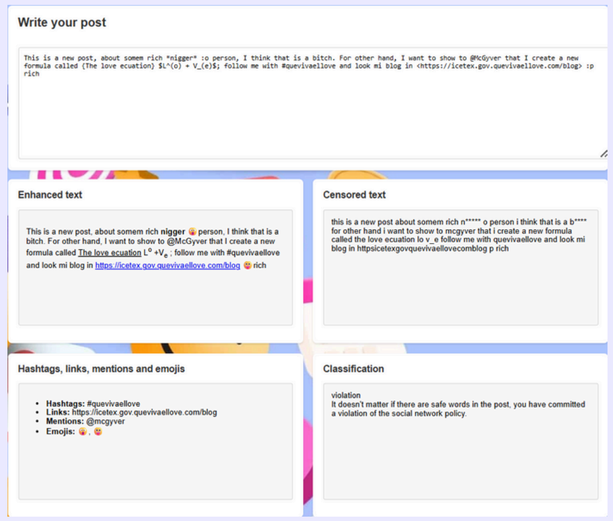
Methodology

1. Regex preprocessing considering common hashtags, mentions and URLs using python's module re.
2. Post classification using pyformlang DFA:
   a. Determine list of classifiable words
   b. Filter known and unknown words
   c. Replace spam words with generic classifiable words
   d. Every known word changes the DFA state
3. Transform offensive words with pyformlang FST
   a. Import all considered offensive and hate words
   b. Store censored version of the word within the FST
   c. Filter unknown words so the FST words appropriately
4. CFG post enhancement using textX library
   a. Define the general structure of a post using textX
   b. Implement a compiler to structure an HTML file

Results







Limitations

- DFA relies on predefined word lists, slang or subtle changes in words are not detected.
- FST reduces the expressiveness
- CFG considers only basic formulas with no recursive structures

What we learned

- Formal languages are a powerful tool but they can be restrictive and need constant updates.
- Automata are useful to classify strings (DFA) and to transform them (FST)
- Python is a great language to work the backend using re and pyformlang and is easily integrated with FastAPI.

Conclusions

- Formal Language theory is a good starting point for written content moderation.
- Regex + DFA + FST + CFG was a great combination to make the post moderator.
- We learned the basics of a new framework called FastAPI to do the web app with UI and manage the communication accross the web.

Daniela Castaño (A00401805), Johan Suarez (A00404253), Simón García (A00371828)