LGBTQ+ Identity Terms, Dataset “Cleanliness” Filters and Community-Led Artificial Intelligence

Lani Wang

Terms and Definitions

Artificial Intelligence (A.I.): The simulation of human intelligence (usually done through computers).

Natural Language Processing (NLP): A.I. that focuses on the manipulation, comprehension and interpretation of human language.

Large Language Model (LLM): A program that can perform NLP tasks. It deals with vast amounts of text-based data.

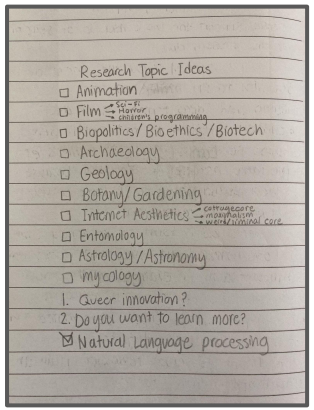
Open Source: Any program whose source code is made available for use or modification to the public.

Common Crawl (CC) Dataset: A 6.4 petabyte (~27,000 times larger than my computer!) open source collection of raw web page data, text extracts and metadata.

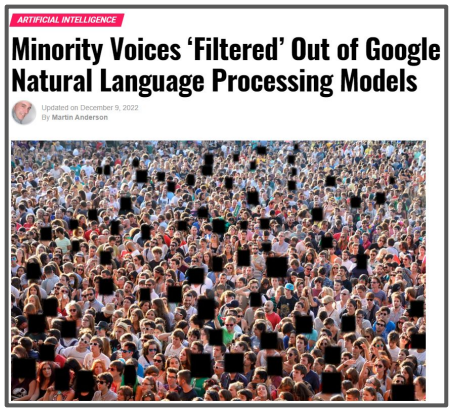
Colossal Cleaned Crawled Corpus (C4) Dataset: A 750 gigabyte (~3 times larger than my computer) algorithmically filtered version of the Common Crawl.

Focusing on a Research Topic

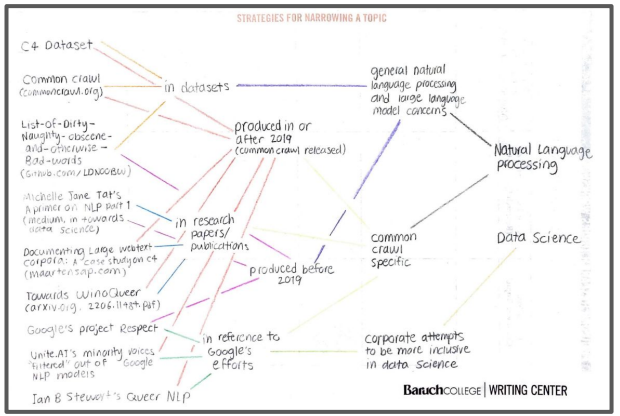
Personal Interests



Preliminary Research



Broadening Research Objectives



Research Paper Diagram

First Half:

How has the Common Crawl Dataset been “cleaned” of LGBTQ+ identity terms and related content? How do A.I. models built with such datasets vastly underserve those that identify with those terms?

Second Half:

What tools are being developed in community-led efforts to redesign A.I.? How does it differ from corporate, capitalized A.I. projects?

C4 v.s. “Common Crawl” Pre-Trained A.I. Experiment

Google/t5-v1_1-small (Filtered Data)

- Only trained on the C4 dataset
- When asked about LGBTQ+ terms, it would usually refuse to elaborate
- Stopped outputting responses after 2+ texts

Two input/output examples are depicted below:

```
>> You:tell me about queer history and context
T5 0:
T5 1: No!!.
T5 2:
T5 3:
T5 4:
```

```
>> You:tell me about queer history
T5 0: ?? OK I say. Definitely not!
T5 1: Thanks.
T5 2: Help?tial predictions?tial predictions??formedformed.
T5 3: Helpful?d most.
T5 4:
```

Google/flan-t5-small (Unfiltered Data)

- Trained on datasets that contained unpublished books, K-8 math word problems and explanations
- When asked about LGBTQ+ terms, it would give some response containing the term but it was usually in some harmful or irrelevant manner
- It was difficult to find unfiltered text2text models as well as models that did not include C4

An input/output example is depicted below:

```
>> You:tell me about queer history and context
T5 0: one queer woman was a womanizer who slept alone with a group of people.
T5 1: sandes
T5 2: killed in a car accident in the wake of the events on the street,
T5 3: the queer community?
T5 4:
```

Sources

Datasets/Data

- C4 Dataset
- Common Crawl Dataset
- List-Of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words Repository

Research Papers

- Queer People are People First
- Michelle Jane Tat's A Primer on NLP Part 1
- A Case Study on C4
- Towards WinoQueer
- "I'm Fully Who I Am"

Corporate Content

- Unite.AI's minority voices "filtered" out of Google NLP models
- Google's Project Respect
- Ian B Stewart's Queer NLP
- Google Deepmind Advocacy Interview

Model Cards

- t5-v1_1-small Model Card
- flan-t5-small Model Card

Community Content

- Queer in A.I.'s statement to Google
- WideningNLP

Problems & Questions

Problems

1. Very few models have been trained on the original Common Crawl.
2. There is much more content on the critique of current LLMs rather than the improvement of them.
3. AI research done to advance science instead of social change is far more prevalent.
4. There are quite a few community-led AI projects out there, but there is a lack of LGBTQ+-focused ones.
5. There is a lack of extensive LGBTQ+-specific datasets as well as models that use them.

Questions

1. Do browsers encourage models made by the company that operates them? (Ex: Bing for DialoGPT)
2. How can I make my source search as unbiased as possible?
3. What are common factors in content filters for LLMs?
4. What are the main differences between community-led efforts, community-led efforts backed by corporations, and corporate-led efforts?
5. Why do community-led AI projects often struggle to achieve the same level of success and recognition as corporate-led projects?