Week 2

VALERIE LOPEZ

Background

Computer Science Major

University of Central Florida – Senior

Interests: AI in Education, Ethical AI



Computer Science

- Algorithms for Machine Learning
- Robot Vision
- Ethics in Science & Tech

Mathematics

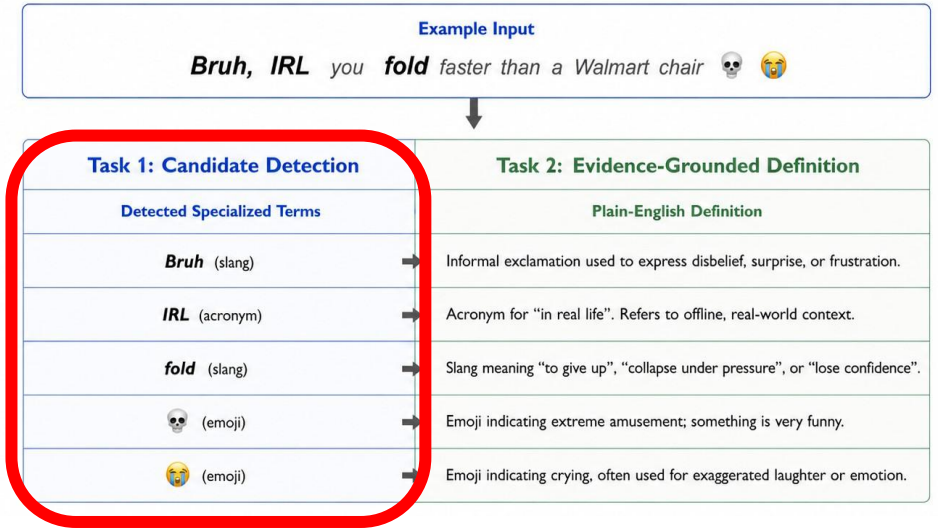
- Calculus I, II, III
- Differential Equations
- Discrete Math

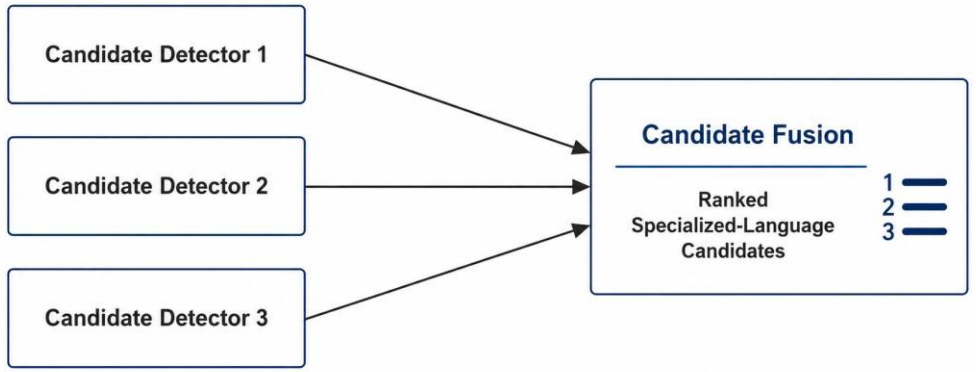
Fun Fact: I teach Latin dance classes!

Project: Detecting and Defining Specialized Language in Text

Overview

- **Given a text input:**
- **Detect** specialized terms, aka **candidates**
- Build **evidence bundles** to **define candidates** alongside confidence levels





Updated scope:



Defining the Problem:

Standard NLP tools fail with specialized terms

- Domain specific jargon, acronyms, emojis, slang, code words

Difficult to overcome rarity, context, community

How can we improve the foundational pipeline?

Applications:

Benefits non-native speakers

Understanding social media

Breaking coded phrases

Removing cultural barriers

Week 1

Background research: NLP, LRM, term extraction

Selecting datasets

Planning Module 1

Meta-Llama Inference

Discuss narrowing research scope

Week 2-5

Begin building Module 1 candidate detection

Apply detection methods:

Language surprise, machine translation, model agreement

Week 6-7

Stabilize Module 1 detector

Expand to multilingual dataset

Compare with direct prompting baselines

Week 7-8

Continue stabilizing Module 1

Error analysis

Final report

Poster Presentation

Final Expectation:

Build a module that can effectively identify specialized language with confidence scores

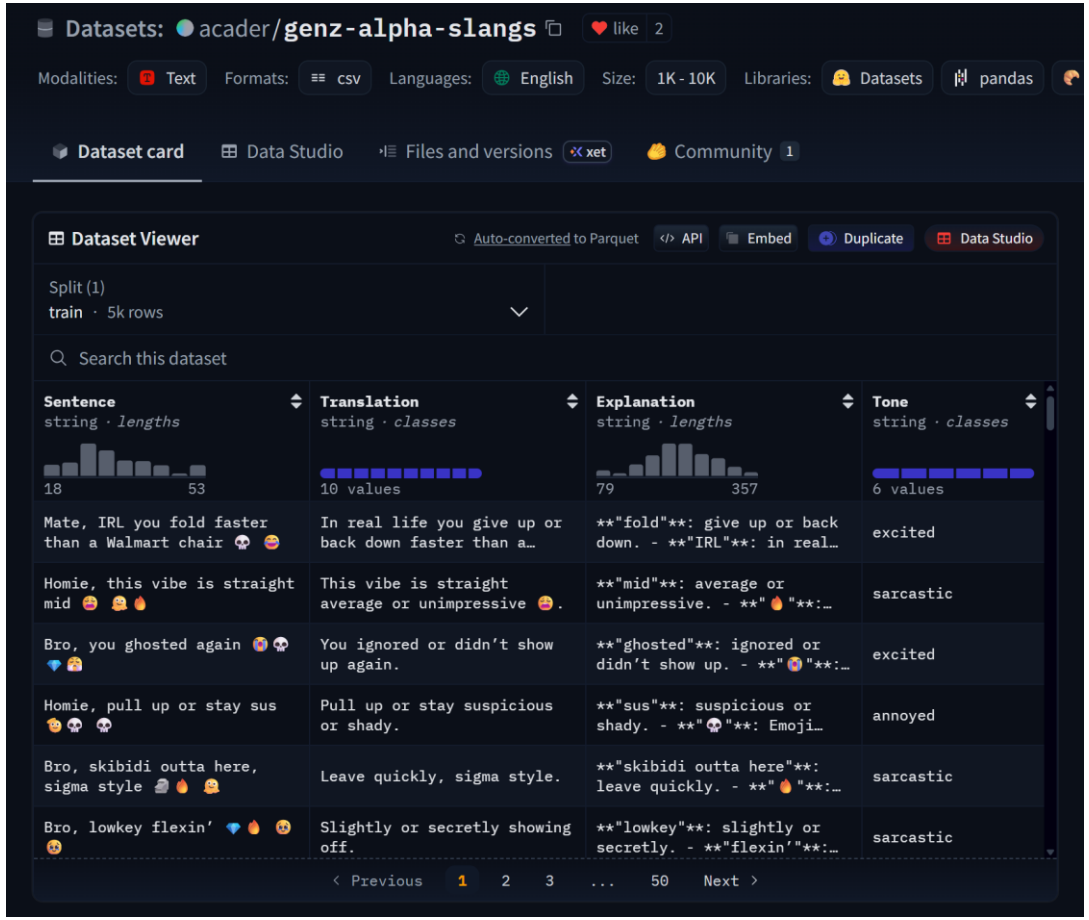
WEEK 1 PROGRESS

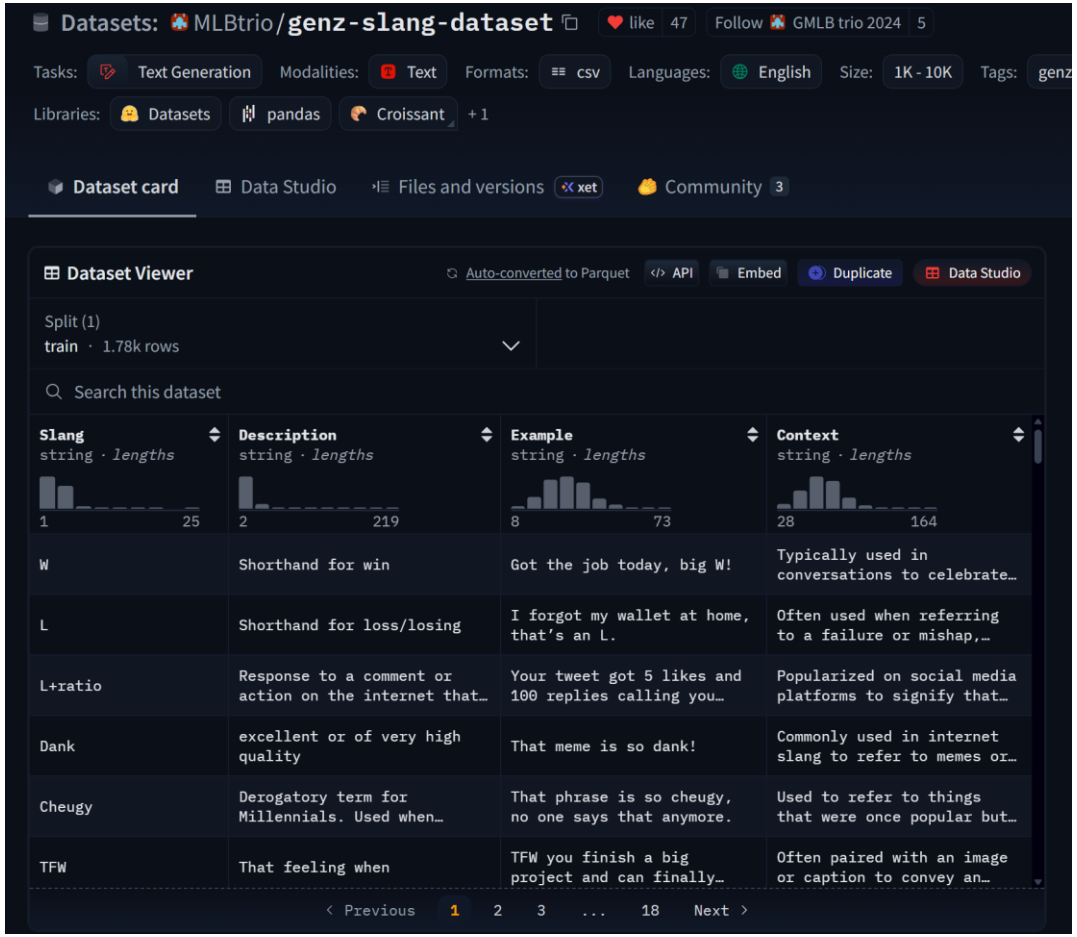
Research – “Attention is All You Need”, “Towards Large Reasoning Models: A Survey of Reinforced Reasoning with LLM”

Set Up – Python environment, dataset collections

Meta-Llama Inference – familiarizing with detection process

DATASET OVERVIEW





DATASET STATISTICS

```
================================================
DATASET LOADED SUCCESSFULLY
DATASET: acader/genz-alpha-slangs
================================================
Sentence       object
Translation    object
Explanation    object
Tone           object
dtype: object

Missing values:
Sentence 0
Translation 0
Explanation 0
Tone 0

Duplicated rows:
243

Total rows:
5000

================================================
DATASET LOADED SUCCESSFULLY
DATASET: MLBtrio/genz-slang-dataset
================================================
Slang          object
Description    object
Example        object
Context        object
dtype: object

Missing values:
Slang 0
Description 0
Example 0
Context 0

Duplicated rows:
0

Total rows:
1779
```

```
================================================
DATASET LOADED SUCCESSFULLY
DATASET: nikesh66/Slang-Dataset
================================================
Tweet            object
Slang (yes/no)   object
dtype: object

Missing values:
Tweet 0
Slang (yes/no) 0

Duplicated rows:
4744

Total rows:
5000

================================================
DATASET LOADED SUCCESSFULLY
DATASET: LM-Lexicon/Slang
================================================
term           object
pos            object
context        object
definition     object
source         object
synonyms       object
instruction    object
dtype: object

Missing values:
term 0
pos 0
context 0
definition 0
source 0
synonyms 0
instruction 0

Duplicated rows:
127973

Total rows:
411382
```

Recommendation:

- **arcader/genz-alpha-slangs**

Plan to discuss further during weekly team meeting

WEEK 1 PLAN

Research – “Towards Large Reasoning Models: A Survey of Reinforced Reasoning with LLM,” “DeepSeekR1”

Weekly Meeting – Discuss updated project scope, select early datasets

Module 1 – Plan approach on filters to be used

Meta-Llama – Run inference

Thank You!

ANY QUESTIONS?