Final Projects

CS 485, Spring 2024

Applications of Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs485_s24/

[Slides by Laure Thompson]

Final Projects

https://people.cs.umass.edu/~brenocon/cs485_s24/project.html

Project Overview

Investigate, analyze, and come to research findings about new methods, or insights on previously existing methods.

In groups of 2 - 3, you will either *build* a natural language processing system or *apply* them to some task.

Your project must: (1) use or develop a dataset, and
(2) report empirical results/analyses with this dataset

Project Components

Proposal: A 2 page document outlining the problem, your approach, possible dataset(s) and/or software systems to use.

Progress Report: A 4 - 8 page document that describes your preliminary work and results

Presentation: An opportunity to present your near- complete project to the class.

Final Report: An 8 - 12 page document that describes your project and final results.

Where to start

- What *core question(s)* are you trying to answer?
- How will you *operationalize* this question?
- What work are you building off of? What has been done before?
- What experiments will you run?
- How will you measure the success of these experiments?
  e.g., held - out accuracy, error analysis, manual evaluation, etc.

Where to look for related work?

NLP research papers:

- The ACL Anthology is a good place to start
- Some Resources:
    - On how to read research papers
    - On navigating the NLP research space

How to search for papers

- Search keywords in the ACL anthology, Google Scholar, Semantic Scholar
- Look at the papers that a paper references and those that cite it
- Examine other papers by a given author and their lab

Where to look for related work?

A standard web search can also be useful for finding…

- Research blog posts
- Datasets
- Related codebases
- Recorded Talks
- …and more!

Choice of emphasis

- Implementing and developing algorithms and features
- Defining a new linguistic / text analysis task, and tackling it with off - the- shelf NLP software
- Collect and explore a new textual dataset to address research hypotheses about it

A large variety of tasks

Detection Tasks

Classification Tasks

Prediction Tasks

- Predict external information from text (e.g. movie revenue, post popularity, stock volatility, etc.)

Structured Linguistic Prediction

- Relation, event extraction
- Narrative chain extraction
- Parsing

Text Generation Tasks

- Machine Translation
- Summarization & Normalization
- Poetry / Lyric generation

End - to - End Systems

- Question Answering
- Conversational dialogue systems

Visualization & Exploration

- Temporal analysis of events
- Topic modeling & clustering

For more dataset and task ideas

- Shared task websites
  - SemEval: Series of semantic evaluation tasks.
    - SemEval 2023 tasks, 2022, 2021, etc.
      There may be access to data!
  - CoNLL shared tasks
- HuggingFace datasets website

Some projects from recent years

Text Classification

- Song genre classification using lyrics
- Comparing models for multi - labeled classification of book genres
- Distinguishing between 19th and 20th century literature
- Predicting political slant in news comments
- Classification of political views on Reddit
- Classifying BBC news articles into their section/category types
- Language classification

Some projects from recent years

Detection Tasks

· Paraphrase detection

· Toxicity level detection in social media posts

Prediction Tasks

· Estimating stock volatility from news articles

· r/ AmITheAsshole verdict prediction

· Predicting tweet popularity

Text Generation Tasks

· Text summarization for lectures

End - to - End Systems

· FAQ answering

· Medical diagnosis chatbot

Visualization & Exploration

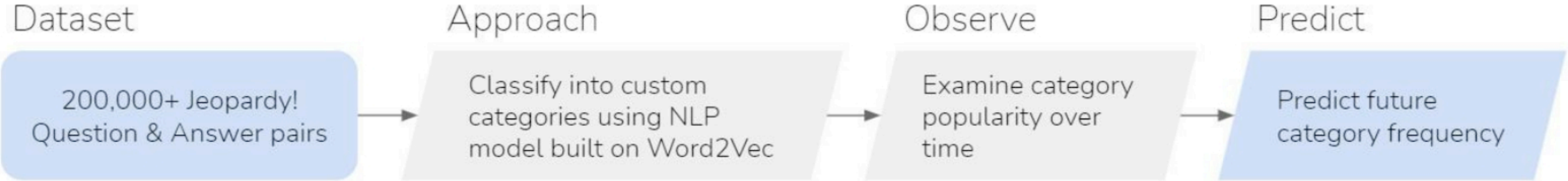
· Sentiment analysis of songs throughout time

· Sentiment analysis of r/ wallstreetbets

JEOPARDY! Category Analysis

Evan Risas & Alisa Kotliarova

Task: Analyze each question-answer pair to determine which broad category it most closely fits, then predict category frequency for future Jeopardy games.



Dataset

200,000+ Jeopardy! Question & Answer pairs

Approach

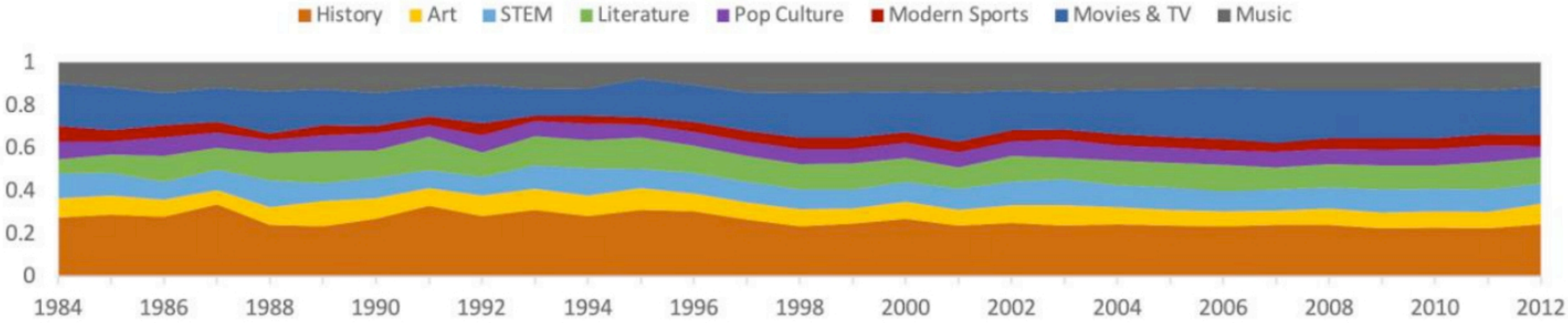
Classify into custom categories using NLP model built on Word2Vec

Observe

Examine category popularity over time

Predict

Predict future category frequency



Brainstorming Session