WikiMood

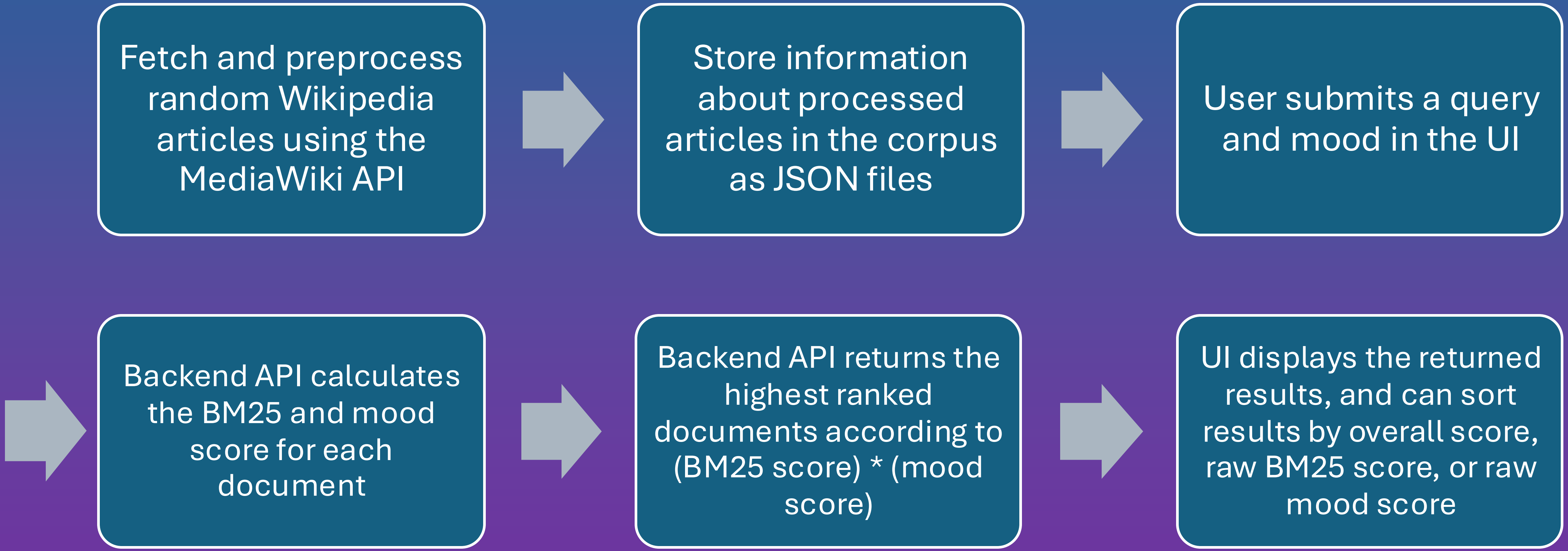
A tool to find Wikipedia articles

Caleb Frye

Overview of the WikiMood Tool

- Consists of:
  - A backend API which fetches articles, performs document processing, and scores documents based on BM25 and mood score
  - A frontend UI where the user can input a search query and mood, and see articles which match the query and mood
- Motivation:
  - I like browsing random Wikipedia articles when I’m bored
  - I looked online but surprisingly didn’t find any website like this
  - I also wanted to add something that made it more advanced than just a simple BM25 search, so I added the mood score

System Flow



Fetch and preprocess random Wikipedia articles using the MediaWiki API

Store information about processed articles in the corpus as JSON files

User submits a query and mood in the UI

Backend API calculates the BM25 and mood score for each document

Backend API returns the highest ranked documents according to (BM25 score) * (mood score)

UI displays the returned results, and can sort results by overall score, raw BM25 score, or raw mood score

Algorithm Choice: BM25

- Does a decent job of scoring documents while allowing for fast performance and efficient document storage (with the preprocessing that I’ve done)
- See the file ‘bm25.py’ for the functions I use to calculate the BM25 score
- But how to combine it with a mood score?

Combining BM25 and Mood Score

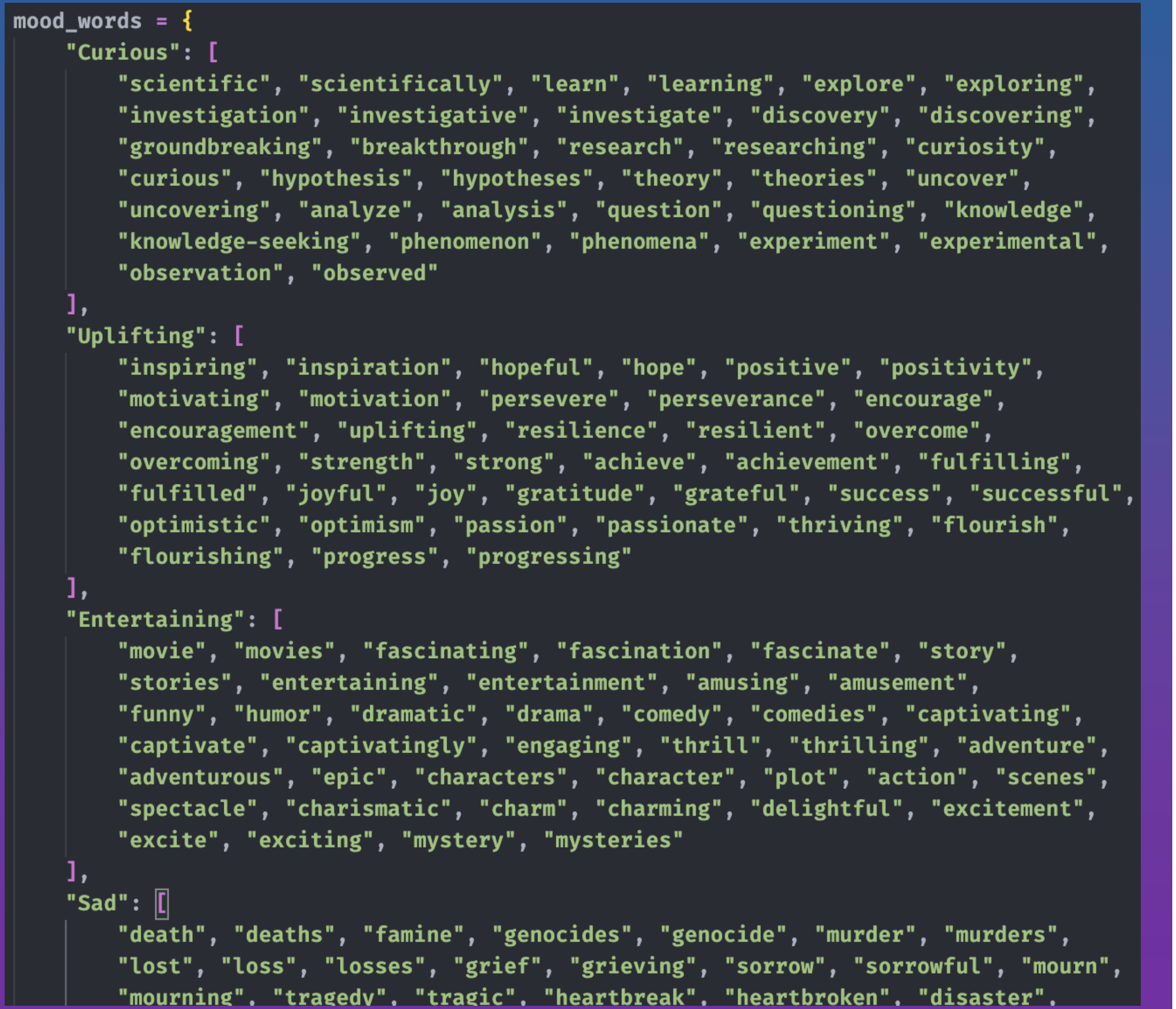
- First, I defined lists of mood words for each mood in ‘constants.py’
- When fetching articles, the system checks each term to see if it is in one of the mood word lists
- If so, it adds +1 to that article’s ‘[mood]_frequency’ value
- A document’s mood multiplier is then:

  ([mood]_frequency) ÷ (number of mood words in [mood])

- <u>Overall score </u>= (BM25 score) × (mood multiplier)

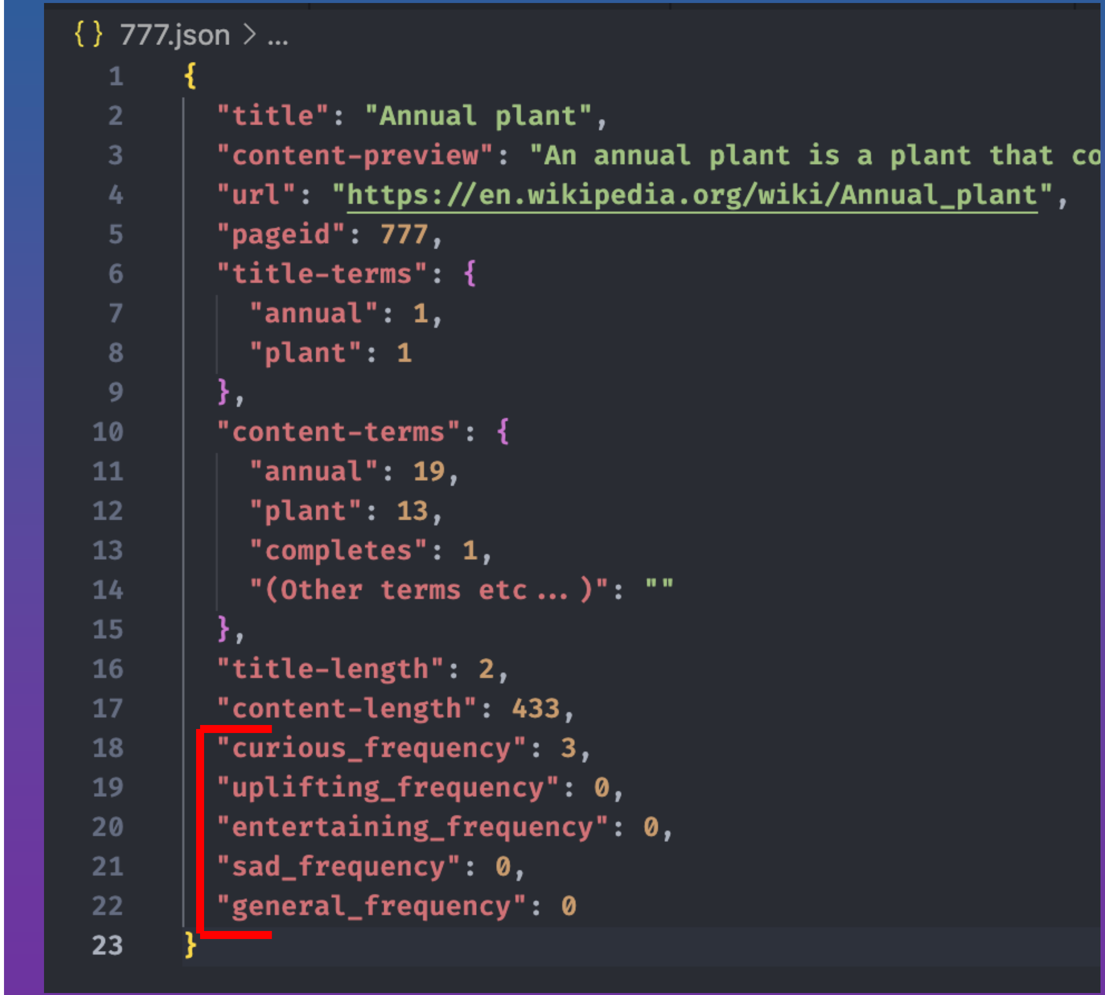
Combining BM25 and Mood Score cont.

Example of mood words:



```
mood_words = {
    "Curious": [
        "scientific", "scientifically", "learn", "learning", "explore", "exploring",
        "investigation", "investigative", "investigate", "discovery", "discovering",
        "groundbreaking", "breakthrough", "research", "researching", "curiosity",
        "curious", "hypothesis", "hypotheses", "theory", "theories", "uncover",
        "uncovering", "analyze", "analysis", "question", "questioning", "knowledge",
        "knowledge-seeking", "phenomenon", "phenomena", "experiment", "experimental",
        "observation", "observed"
    ],
    "Uplifting": [
        "inspiring", "inspiration", "hopeful", "hope", "positive", "positivity",
        "motivating", "motivation", "persevere", "perseverance", "encourage",
        "encouragement", "uplifting", "resilience", "resilient", "overcome",
        "overcoming", "strength", "strong", "achieve", "achievement", "fulfilling",
        "fulfilled", "joyful", "joy", "gratitude", "grateful", "success", "successful",
        "optimistic", "optimism", "passion", "passionate", "thriving", "flourish",
        "flourishing", "progress", "progressing"
    ],
    "Entertaining": [
        "movie", "movies", "fascinating", "fascination", "fascinate", "story",
        "stories", "entertaining", "entertainment", "amusing", "amusement",
        "funny", "humor", "dramatic", "drama", "comedy", "comedies", "captivating",
        "captivate", "captivatingly", "engaging", "thrill", "thrilling", "adventure",
        "adventurous", "epic", "characters", "character", "plot", "action", "scenes",
        "spectacle", "charismatic", "charm", "charming", "delightful", "excitement",
        "excite", "exciting", "mystery", "mysteries"
    ],
    "Sad": [
        "death", "deaths", "famine", "genocides", "genocide", "murder", "murders",
        "lost", "loss", "losses", "grief", "grieving", "sorrow", "sorrowful", "mourn",
        "mourning", "tragedy", "tragic", "heartbreak", "heartbroken", "disaster",
```

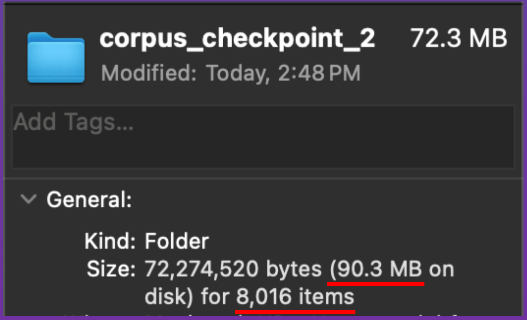
Mood term frequency information in the JSON:



```
{
  "title": "Annual plant",
  "content-preview": "An annual plant is a plant that co
  "url": "https://en.wikipedia.org/wiki/Annual_plant",
  "pageid": 777,
  "title-terms": {
    "annual": 1,
    "plant": 1
  },
  "content-terms": {
    "annual": 19,
    "plant": 13,
    "completes": 1,
    "(Other terms etc ... )": ""
  },
  "title-length": 2,
  "content-length": 433,
  "curious_frequency": 3,
  "uplifting_frequency": 0,
  "entertaining_frequency": 0,
  "sad_frequency": 0,
  "general_frequency": 0
}
```

The Corpus

- The system stores document information as JSON after stripping out stopwords
- This allows for very efficient document storage and fast search performance
  - At the time of making this, the corpus is 8,000 documents and takes up only 90MB:



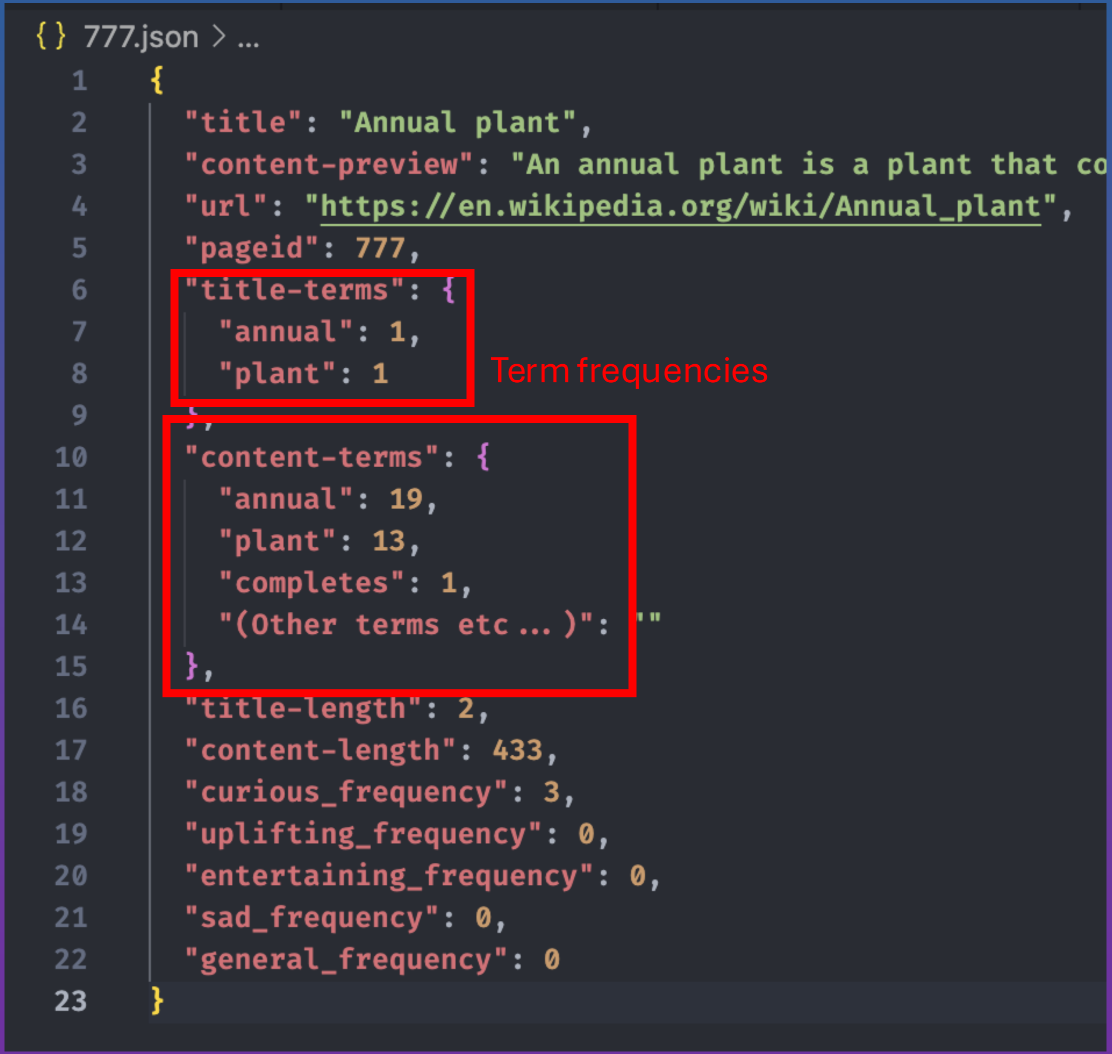
corpus_checkpoint_2 72.3 MB
Modified: Today, 2:48 PM

Add Tags...

General:

Kind: Folder
Size: 72,274,520 bytes (90.3 MB on disk) for 8,016 items

Example document JSON file:



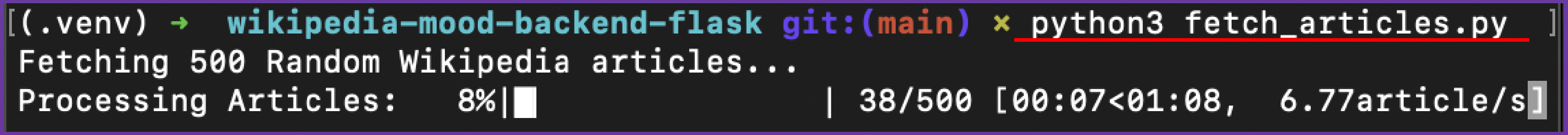
{ } 777.json > ...

```
{
  "title": "Annual plant",
  "content-preview": "An annual plant is a plant that co
  "url": "https://en.wikipedia.org/wiki/Annual_plant",
  "pageid": 777,
  "title-terms": {
    "annual": 1,
    "plant": 1
  },
  "content-terms": {
    "annual": 19,
    "plant": 13,
    "completes": 1,
    "(Other terms etc ... )": ""
  },
  "title-length": 2,
  "content-length": 433,
  "curious_frequency": 3,
  "uplifting_frequency": 0,
  "entertaining_frequency": 0,
  "sad_frequency": 0,
  "general_frequency": 0
}
```

Term frequencies

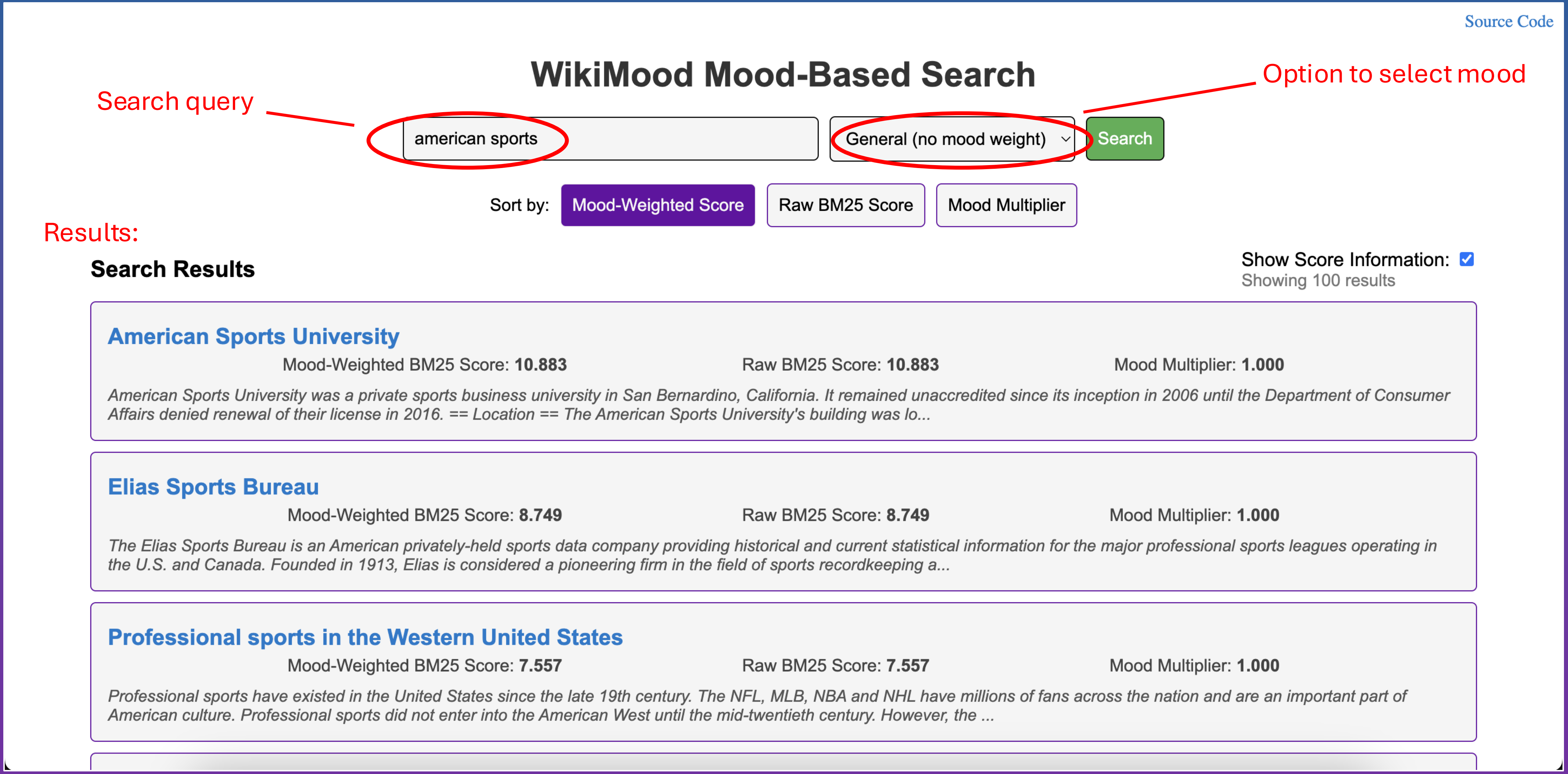
Adding Documents to the Corpus

- I wrote a python script which automatically fetches, preprocesses, and adds new articles to the corpus
- Parameters are configurable in ‘constants.py’ and explained in the README on Github
- Simply use the script as seen below:



```
[(.venv) ➜  wikipedia-mood-backend-flask git:(main) ✗ python3 fetch_articles.py  ]
Fetching 500 Random Wikipedia articles...
Processing Articles:   8%|█                    | 38/500 [00:07<01:08,  6.77article/s]
```

Using the UI



Source Code

WikiMood Mood-Based Search

Search query

Option to select mood

american sports | General (no mood weight) | Search

Sort by: Mood-Weighted Score | Raw BM25 Score | Mood Multiplier

Results:

Search Results

Show Score Information: ☑
Showing 100 results

American Sports University

Mood-Weighted BM25 Score: **10.883** Raw BM25 Score: **10.883** Mood Multiplier: **1.000**

American Sports University was a private sports business university in San Bernardino, California. It remained unaccredited since its inception in 2006 until the Department of Consumer Affairs denied renewal of their license in 2016. == Location == The American Sports University's building was lo...

Elias Sports Bureau

Mood-Weighted BM25 Score: **8.749** Raw BM25 Score: **8.749** Mood Multiplier: **1.000**

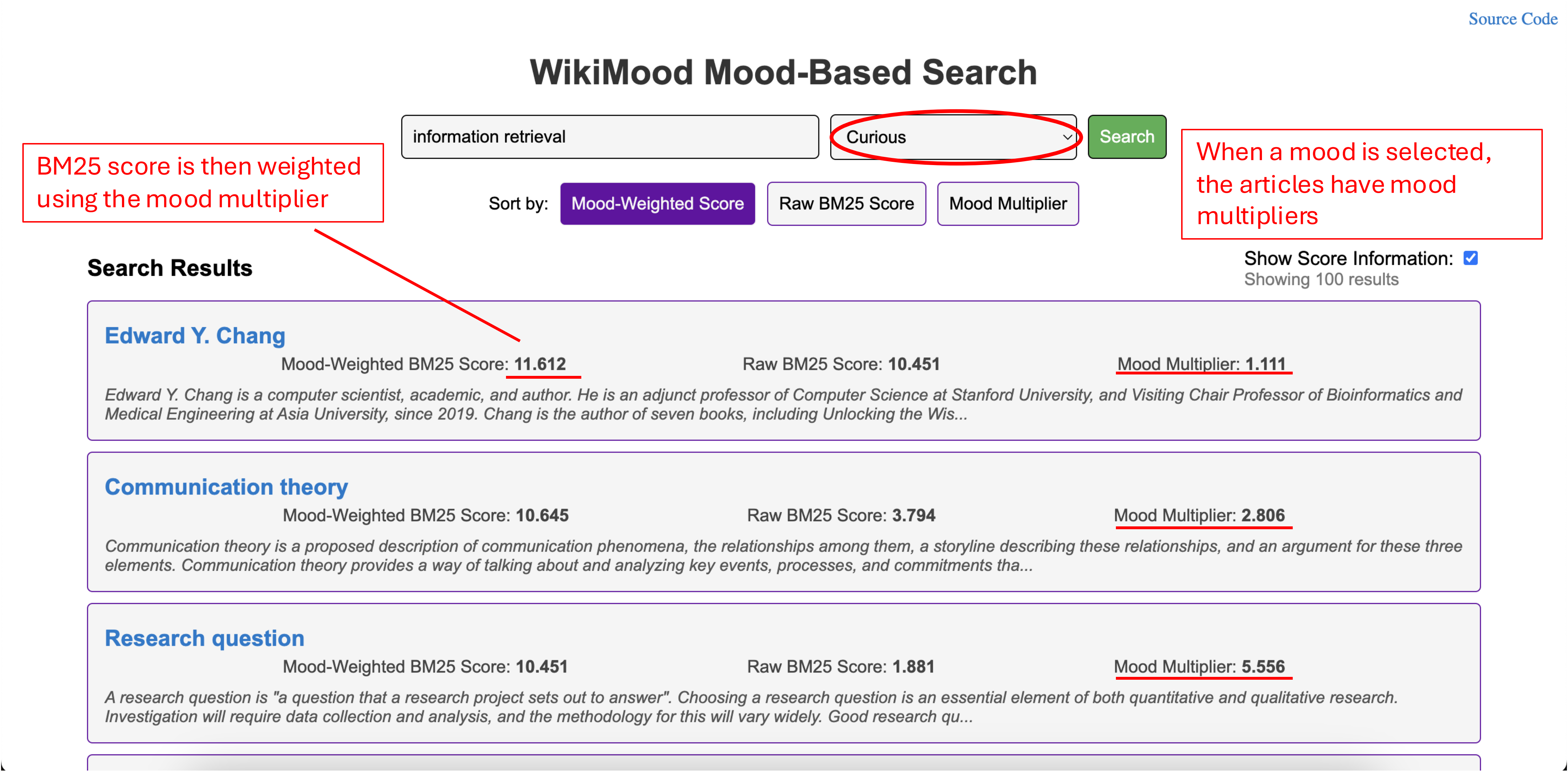
The Elias Sports Bureau is an American privately-held sports data company providing historical and current statistical information for the major professional sports leagues operating in the U.S. and Canada. Founded in 1913, Elias is considered a pioneering firm in the field of sports recordkeeping a...

Professional sports in the Western United States

Mood-Weighted BM25 Score: **7.557** Raw BM25 Score: **7.557** Mood Multiplier: **1.000**

Professional sports have existed in the United States since the late 19th century. The NFL, MLB, NBA and NHL have millions of fans across the nation and are an important part of American culture. Professional sports did not enter into the American West until the mid-twentieth century. However, the ...

Using the UI cont.



Source Code

WikiMood Mood-Based Search

Sort by: Mood-Weighted Score | Raw BM25 Score | Mood Multiplier

BM25 score is then weighted using the mood multiplier

When a mood is selected, the articles have mood multipliers

Show Score Information: ☑
Showing 100 results

Search Results

Edward Y. Chang

Mood-Weighted BM25 Score: **11.612** Raw BM25 Score: **10.451** Mood Multiplier: **1.111**

Edward Y. Chang is a computer scientist, academic, and author. He is an adjunct professor of Computer Science at Stanford University, and Visiting Chair Professor of Bioinformatics and Medical Engineering at Asia University, since 2019. Chang is the author of seven books, including Unlocking the Wis...

Communication theory

Mood-Weighted BM25 Score: **10.645** Raw BM25 Score: **3.794** Mood Multiplier: **2.806**

Communication theory is a proposed description of communication phenomena, the relationships among them, a storyline describing these relationships, and an argument for these three elements. Communication theory provides a way of talking about and analyzing key events, processes, and commitments tha...

Research question

Mood-Weighted BM25 Score: **10.451** Raw BM25 Score: **1.881** Mood Multiplier: **5.556**

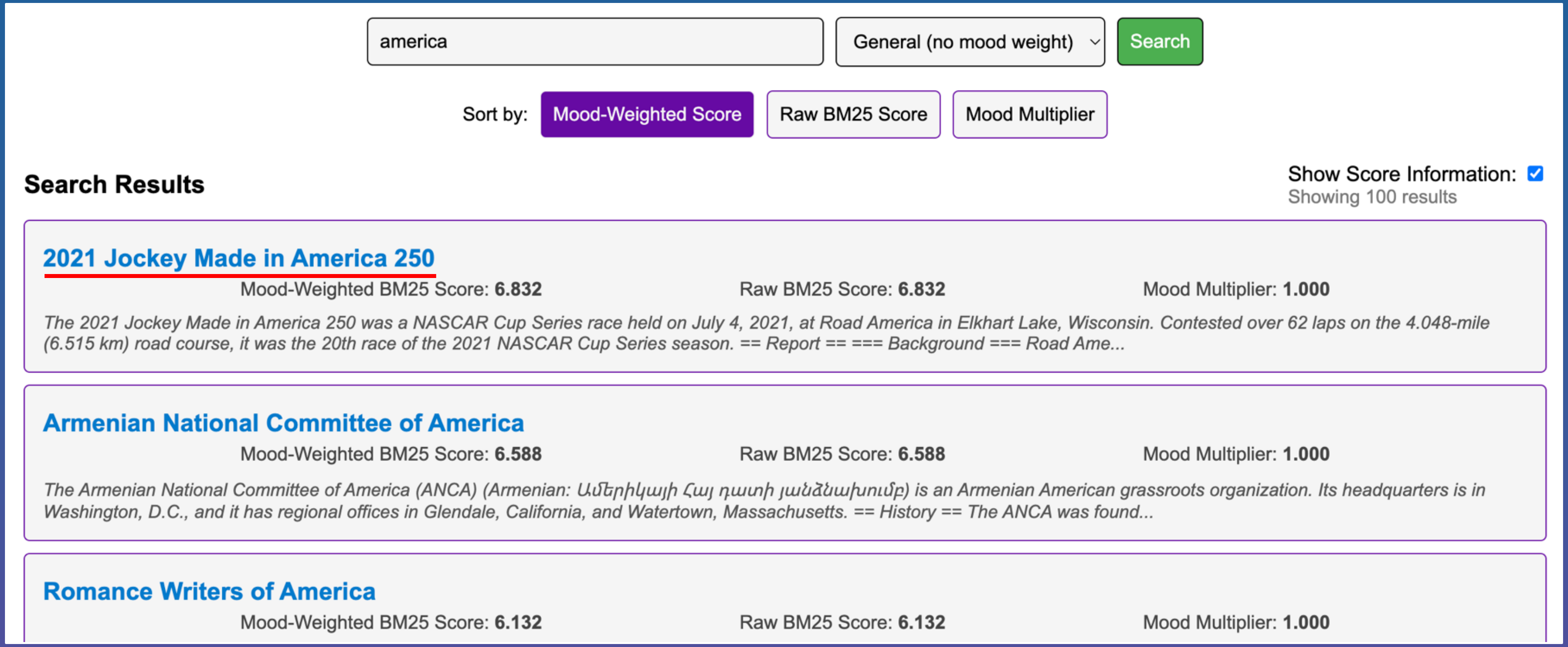
A research question is "a question that a research project sets out to answer". Choosing a research question is an essential element of both quantitative and qualitative research. Investigation will require data collection and analysis, and the methodology for this will vary widely. Good research qu...

Evaluation

- The BM25 algorithm returns relevant articles decently well, but…
- Since I don't take term positions into account the results can be slightly unrelated
  - For example, searching "tennis player" with any mood other than 'general' returns a lot of articles about video games, since the word "player" appears in them frequently, and the articles that are actually about tennis players don't have high mood scores
- However, I think this is ok! The goal of the system is to help you find articles to read, so if it exposes you to more topics, you're more likely to find something interesting to read
- That said, if I made the system again I would definitely take term positions into account

Evaluation cont.

- The mood multiplier does affect which documents are ranked higher, especially noticeable for the ‘sad’ and ‘entertaining’ moods
- For example, see the top result for searching ‘America’ with no mood vs. ‘sad’ mood:
- It’s not perfect, but the mood does make a difference



Search Results

Show Score Information: ☑
Showing 100 results

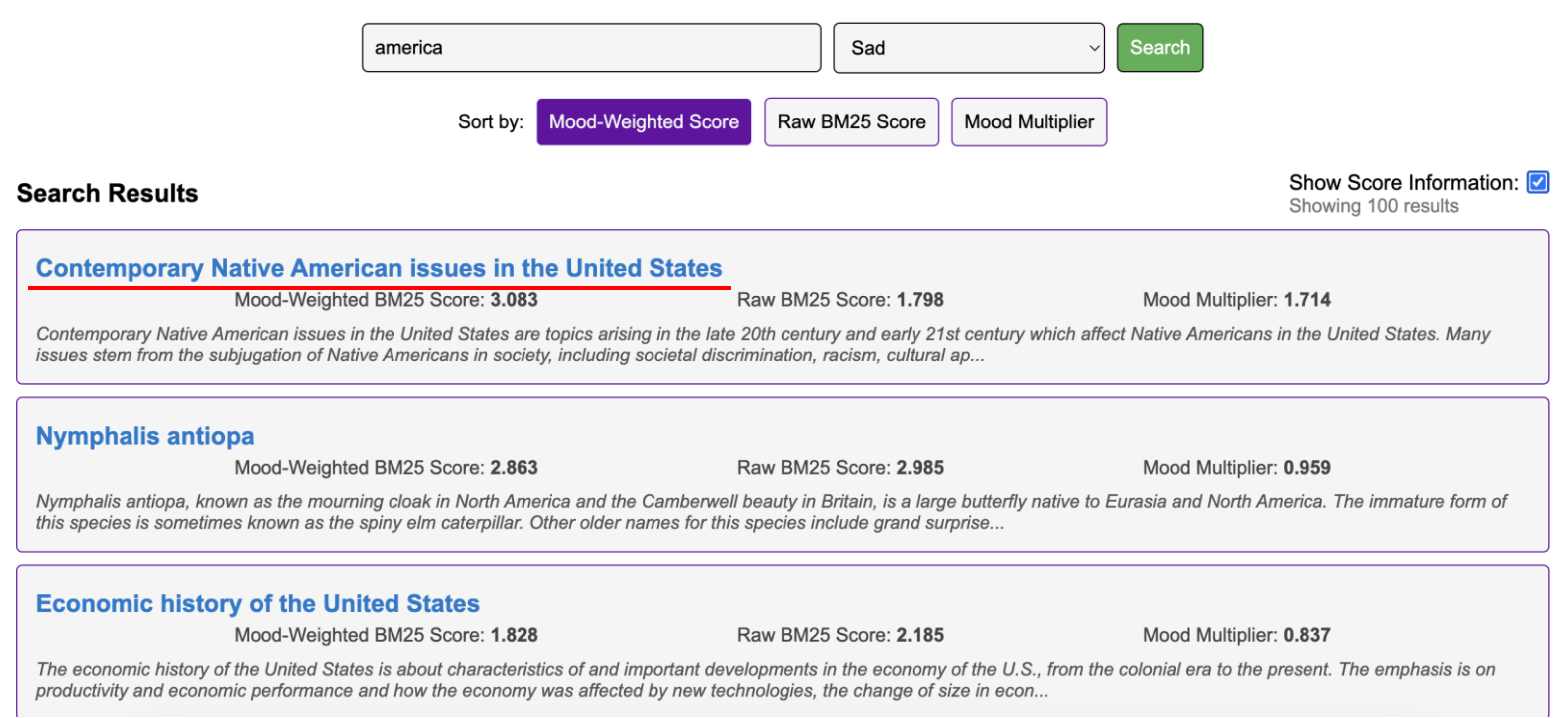
2021 Jockey Made in America 250
Mood-Weighted BM25 Score: **6.832** Raw BM25 Score: **6.832** Mood Multiplier: **1.000**

The 2021 Jockey Made in America 250 was a NASCAR Cup Series race held on July 4, 2021, at Road America in Elkhart Lake, Wisconsin. Contested over 62 laps on the 4.048-mile (6.515 km) road course, it was the 20th race of the 2021 NASCAR Cup Series season. == Report == === Background === Road Ame...

Armenian National Committee of America
Mood-Weighted BM25 Score: **6.588** Raw BM25 Score: **6.588** Mood Multiplier: **1.000**

The Armenian National Committee of America (ANCA) (Armenian: Ամերիկայի Հայ դատի յանձնախումբ) is an Armenian American grassroots organization. Its headquarters is in Washington, D.C., and it has regional offices in Glendale, California, and Watertown, Massachusetts. == History == The ANCA was found...

Romance Writers of America
Mood-Weighted BM25 Score: **6.132** Raw BM25 Score: **6.132** Mood Multiplier: **1.000**



Search Results

Show Score Information: ☑
Showing 100 results

Contemporary Native American issues in the United States
Mood-Weighted BM25 Score: **3.083** Raw BM25 Score: **1.798** Mood Multiplier: **1.714**

Contemporary Native American issues in the United States are topics arising in the late 20th century and early 21st century which affect Native Americans in the United States. Many issues stem from the subjugation of Native Americans in society, including societal discrimination, racism, cultural ap...

Nymphalis antiopa
Mood-Weighted BM25 Score: **2.863** Raw BM25 Score: **2.985** Mood Multiplier: **0.959**

Nymphalis antiopa, known as the mourning cloak in North America and the Camberwell beauty in Britain, is a large butterfly native to Eurasia and North America. The immature form of this species is sometimes known as the spiny elm caterpillar. Other older names for this species include grand surprise...

Economic history of the United States
Mood-Weighted BM25 Score: **1.828** Raw BM25 Score: **2.185** Mood Multiplier: **0.837**

The economic history of the United States is about characteristics of and important developments in the economy of the U.S., from the colonial era to the present. The emphasis is on productivity and economic performance and how the economy was affected by new technologies, the change of size in econ...

Conclusion

- Even though it’s not perfect, the system does return interesting Wikipedia articles based on your query and mood selection
- The considerations I made in document preprocessing allow for fast searches and efficient document storage
- I know the tool works, because there were many times while I was working on it that I got distracted and started reading the articles returned by the tool

Thank you!