

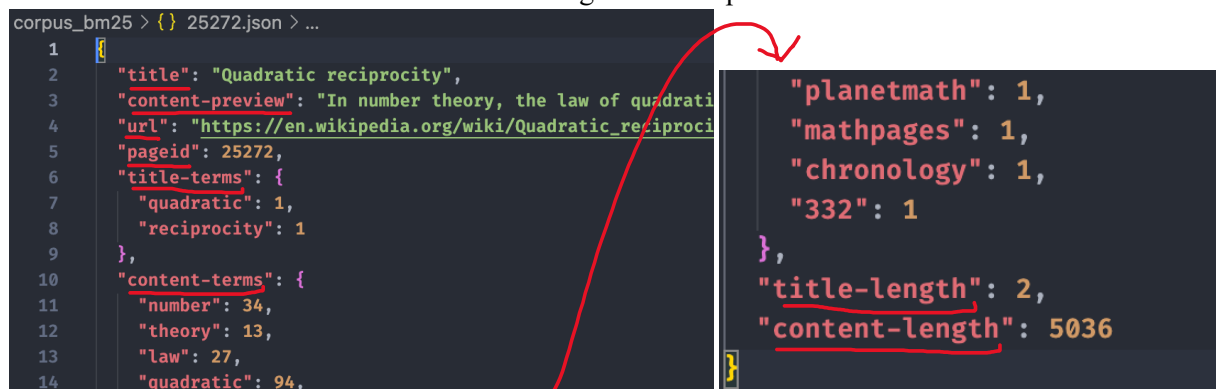
WikiMood Checkpoint 1: Data

My project uses BM25 to allow a user to find Wikipedia articles based off mood, and they can also use a query to search for specific kinds of documents. I am storing my documents as JSON files in a folder called 'corpus.' Each json file has the following attributes:

- title – title of the article
- content-preview – a short preview of the article's content to show in the UI
- url – the article's URL, so my UI can link to the article
- pageid – the article's pageid, used for backend article-fetching purposes
- title-terms – term frequencies of each word in the title of the given article
- content-terms – term frequencies of each word in the body of the given article
- title-length – number of terms in the title after stopwords are removed, used for BM25
- content-length – number of terms in the after stopwords are removed, used for BM25

Collecting the Data

I am using the Wikimedia API to get articles to include in my corpus. I fetch the title, content, and pageid from the API, and then calculate the other JSON fields in my backend. After using the entire document's content to create its term frequency vectors, I only store a small preview of the content. This allows me to store many documents without using much space. I currently have almost 1000 documents stored as JSON files, which are taking up less than 6MB. The files are named "*pageid*.json." I will greatly expand this corpus before I deploy my project; I hope to have at least 50,000 articles in the collection, but I will have to test to ensure runtime is not too long. See a sample of one of the article's JSON files below:



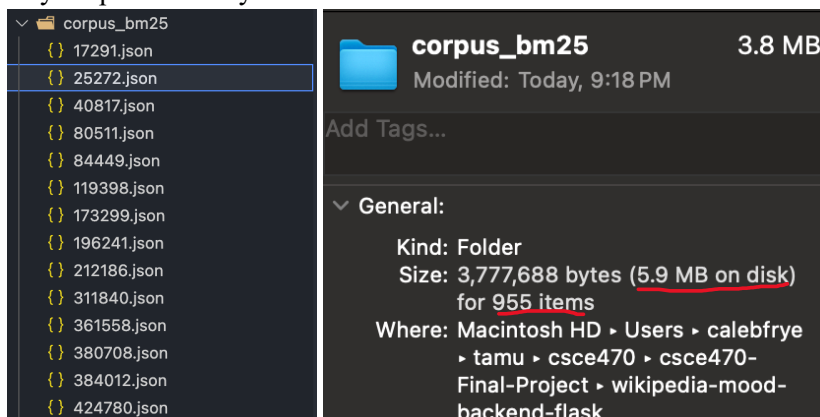
The image shows a code editor with two panels. The left panel displays a JSON file named '25272.json' from a directory 'corpus_bm25'. The JSON content is as follows:

```
{
  "title": "Quadratic reciprocity",
  "content-preview": "In number theory, the law of quadratic reciprocity...",
  "url": "https://en.wikipedia.org/wiki/Quadratic_reciprocity",
  "pageid": 25272,
  "title-terms": {
    "quadratic": 1,
    "reciprocity": 1
  },
  "content-terms": {
    "number": 34,
    "theory": 13,
    "law": 27,
    "quadratic": 94,
    "planetmath": 1,
    "mathpages": 1,
    "chronology": 1,
    "332": 1
  },
  "title-length": 2,
  "content-length": 5036
}
```

The right panel shows a JavaScript object representation of the JSON data, with a red arrow pointing from the 'content-terms' object in the JSON to the object in the code:

```
{
  title: "Quadratic reciprocity",
  "content-preview": "In number theory, the law of quadratic reciprocity...",
  url: "https://en.wikipedia.org/wiki/Quadratic_reciprocity",
  pageid: 25272,
  titleTerms: {
    quadratic: 1,
    reciprocity: 1
  },
  contentTerms: {
    number: 34,
    theory: 13,
    law: 27,
    quadratic: 94,
    planetmath: 1,
    mathpages: 1,
    chronology: 1,
    332: 1
  },
  titleLength: 2,
  contentLength: 5036
}
```

My corpus currently looks like this:



The image shows a file explorer window. On the left, a list of files in the 'corpus_bm25' folder is shown, with '25272.json' selected. On the right, the details for the 'corpus_bm25' folder are displayed:

- Folder name: corpus_bm25
- Size: 3.8 MB
- Modified: Today, 9:18 PM
- General information:
 - Kind: Folder
 - Size: 3,777,688 bytes (5.9 MB on disk) for 955 items
 - Where: Macintosh HD > Users > calebfrye > tamu > csce470 > csce470-Final-Project > wikipedia-mood-backend-flask