

Jacob Baskiewicz

CS429

12/01/25

Final Project Report

## **Abstract**

The objective of this project is to create a Flask-based search application for allowing text queries to return a list of scraped web pages and their correlating cosine similarity for the query provided. The list of the top 5 pages is sorted in descending order of their cosine similarity to provide the user with the webpage with the most relevant results. This project may be useful for individuals searching for several subpages or to have a local search engine environment. Next steps would be combining the system into one usable application with UI, such as providing a space for seed URLs, and entering in free-text queries and getting a result immediately. Another addition would be semantic searching and query expansion.

## **Overview**

To accomplish the application, the project is broken down into 3 major parts. Scrapy is used to build the html crawler, which can use any seed URL to return the first 20 pages with a max depth of 2. After scraping the page, Scikit-learn is used to build an index of all terms found. TF-IDF scores are calculated to find term relevance, and cosine similarity is found to compare with a provided query. Lastly, Flask is used as the user application. A user can provide a CSV file of queries, which will be searched in the inverted index to find the pages with the most relevance, and saved to a text document. The top-k results is set to be 5. In cases of spelling errors, NLTK is used to find edit distance between the term in the query, and the next closest in the inverted index. The suggested term is then shown to the user for the provided term.

## **Design**

The capabilities of the system are limited to scraping, building an inverted index, and providing a CSV file with queries to do a search. The interactions between the systems are as follows.

Firstly, the Scrapy crawler will crawl and save the webpages, which the Scikit-learn indexer will use to build a TF-IDF inverted index using a helper python function to run. The flask application will read the resulting JSON file or run the indexer if no JSON is found. The flask application then reads the resulting JSON file and will return a text file output based on the search.

## Architecture

- Scrapy crawler

The scrapy crawler named “html\_crawler”, takes a seed URL, and will crawl for 20 max pages with a max depth of 2. The default seed URL is a commonly used seed for testing Scrapy crawlers known as “quotes.toscrape.com”. The crawler will initialize and perform its crawl, saving all webpages in the project folder as “downloaded\_page\_x.html”.

- Scikit-learn Indexer

The Scikit-learn indexer uses BeautifulSoup 4 to parse the html pages and save all unique terms separated by spaces. It builds the TF-IDF matrix using the default list of English stop words, also making sure not to include duplicates within the index. The resulting TF-IDF matrix is organized as a JSONfile, with each term having resulted "doc\_id": "downloaded\_page\_x.html" and "tfidf": 0.x. The resulting index is saved as a JSON in the project folder with the name “inverted\_index.json”. To avoid needing console input, a python file called “run\_indexer.py” was created to run the indexer remotely.

- Flask Application

The Flask Application has two python files to run. The file “app.py” has the basic functionality of the application. The flask application first will check and see if an inverted index exists, and either use one if it exists or create one if not. The query processor function will check if a CSV file with queries exists and check to make sure that a column for such queries exists. If found, the application will build the output, first checking to see if the query has a suggestion if not found in the inverted index. If a word is found in the inverted index, the top 5 documents will be returned to the user. The output is formatted to read better for an average user. The entire application has many error checks responding with a status 400 if anything prompts an error when reading the inverted index.

## Operation

The following libraries are needed to operate the system. Scrapy, Flask, NLTK, Scikit-learn, BeautifulSoup 4.

When scraping, the following needs to be done to operate. Input a seed url into the HTML crawler python class, and in a terminal within the directory of the crawler, run the command “scrapy crawl html\_crawler” which will crawl the page based on the seed url and attributes described in the crawler initialization. All of the pages will then be saved in the project folder for use by the indexer.

When indexing, all that needs to be done is to run the python file “run\_indexer.py” which will run the python file “indexer.py”. This will save the resulting inverted index in the project folder.

When running the flask application, first, the “app.py” python file needs to be run from the terminal using “python app.py”. The user may input the queries they want to search in the “queries.csv” file in the project. Then, the user may run the queries by running the “test\_flask\_queries.py” file in a second terminal. From there, the results are saved as a text file and can be viewed.

## **Conclusion**

Concluding the development of the process, there were a number of successes and failures during the development of the project. Setting up the Scrapy crawler became the first big roadblock, with a number of issues arising in its development such as infinite crawling and not stopping after reaching max depth, to not saving the pages correctly. These were overcome by changing a few of the project settings and debugging using Scrapy documentation and LLMs explaining the reason for the issue. The indexer was a large success, working as intended quite early in the process; however, it is only limited to the requirements of the project. The flask application is also a success, allowing free-text queries in a CSV format to search. It does, however, have the caveat of only running locally and requiring two terminals to operate. It also is less user friendly and it requires queries in a CSV format instead of console or some other type of input. It is also lacking spelling correction and query expansion.

A caution when developing a project like this is to plan the architecture of the project in advance, since managing several files and folders without a plan can lead to issues or inconveniences during development.

## **Data Sources**

[1] <https://quotes.toscrape.com>

## **Test Cases**

When testing during the development of the project, I first tested each individual section before testing the system as a whole. For the Scrapy crawler, I tested quotes.toscrape.com with different max pages and page depth until an efficient and effective parameter was found, that being 20 pages and 2 max depth. For testing the indexer, I ran it with groups of pages from 10-50, to see if the program correctly made an index for each group. During the development I also included a number of print statements to make sure TFIDF, Cosine similarity, and documents were calculated and read correctly. Saving the inverted index as a JSON file added clear visibility and organization to the parsed HTML pages. When testing the Flask application, I had to ensure that the queries the user passed existed, were correct, and had a correlating term in the inverted index. I also had to ensure the inverted index existed in the first place. If any issues arose, error status 400 was thrown from the flask application. When checking specific queries, I had to make sure that I included some that were misspelled to display suggested terms if one was not found within the inverted index. I also had to check for words that may show up frequently in a document, and

infrequently in many documents to see if the top 5 returned documents sorted correctly. Below is a test case example:

Query: quotes

Results:

downloaded\_page\_12.html -> 0.29662660165228943

downloaded\_page\_7.html -> 0.2546243690774928

downloaded\_page\_8.html -> 0.2515624369648701

downloaded\_page\_5.html -> 0.2510613501789357

downloaded\_page\_2.html -> 0.2443986794990816

Suggestions: None

Query: stupid

Results:

downloaded\_page\_12.html -> 0.20750141045778497

downloaded\_page\_10.html -> 0.15090770532077524

downloaded\_page\_11.html -> 0.10049849694766169

downloaded\_page\_1.html -> 0.05762524820913743

downloaded\_page\_17.html -> 0.050740502076156345

Suggestions: None

Query: stuplld

Results:

downloaded\_page\_1.html -> 0.0

downloaded\_page\_10.html -> 0.0

downloaded\_page\_11.html -> 0.0

downloaded\_page\_12.html -> 0.0

downloaded\_page\_13.html -> 0.0

Suggestions: stuplld -> stupid

## Source Code:

[1] Scrapy 2.13 documentation — Scrapy 2.1.3 documentation. (n.d.). Docs.scrapy.org.

<https://docs.scrapy.org/en/latest/>

[2] Examples. (2025). Scikit-Learn. [https://scikit-learn.org/stable/auto\\_examples/](https://scikit-learn.org/stable/auto_examples/)

[3] Flask. (2010). Welcome to Flask — Flask Documentation (3.0.x). Palletsprojects.com.

<https://flask.palletsprojects.com/en/stable/>

[4] NLTK :: nltk. (n.d.). Www.nltk.org. [https://www.nltk.org/\\_modules/nltk.html](https://www.nltk.org/_modules/nltk.html)

[5] Richardson, L. (2019). Beautiful Soup Documentation — Beautiful Soup 4.4.0

documentation. Crummy.com. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

## Bibliography

- [1] Scrapy Tutorial — Scrapy 2.3.0 documentation. (n.d.). Docs.scrapy.org.  
<https://docs.scrapy.org/en/latest/intro/tutorial.html>
- [2] OpenAI. (2025, September 12). ChatGPT. Chatgpt.com; OpenAI. <https://chatgpt.com>
- [3] Working With Text Data. (2024). Scikit-Learn. [https://scikit-learn.org/1.4/tutorial/text\\_analytics/working\\_with\\_text\\_data.html](https://scikit-learn.org/1.4/tutorial/text_analytics/working_with_text_data.html)
- [4] GeeksforGeeks. (2017, October 23). Flask (Creating first simple application).  
GeeksforGeeks. <https://www.geeksforgeeks.org/python/flask-creating-first-simple-application/>

```
In [ ]: import scrapy, os
from scrapy.linkextractors import LinkExtractor
from urllib.parse import urlparse

class HtmlCrawler(scrapy.Spider):
    name = "html_crawler"
    # initialize crawler
    def __init__(self, seed="https://quotes.toscrape.com/", max_pages=20, max_depth=2, *args, **kwargs):
        super(HtmlCrawler, self).__init__(*args, **kwargs)
        self.start_urls = [seed]
        self.allowed_domain = urlparse(seed).netloc
        self.max_pages = int(max_pages)
        self.max_depth = int(max_depth)
        self.page_count = 0
        self.link_extractor = LinkExtractor(allow_domains=[self.allowed_domain])

    def parse(self, response):
        # stop if reached max pages
        if self.page_count >= self.max_pages:
            self.crawler.engine.close_spider(self, reason="max pages reached")
            return

        # increment and get page url and text
        self.page_count += 1
        pageURL = response.url
        pageHTML = response.text

        # save content in file
        filename = f"downloaded_page_{self.page_count}.html"
        path = os.path.join(".", filename)

        with open(path, "w", encoding="utf-8") as f:
            f.write(pageHTML)
            self.logger.info(f"Saved {pageURL} as {filename}")

        # get next link if within depth
        curDepth = response.meta.get('depth', 0)
        if curDepth < self.max_depth:
            for link in self.link_extractor.extract_links(response):
                yield scrapy.Request(link.url, callback=self.parse, meta={'depth': curDepth + 1})
```

2: Scikit-learn Indexer  
Done with assistance from ChatGPT (bug fixing and explanation)  
Examples. (2025). Scikit-Learn. [https://scikit-learn.org/stable/auto\\_examples/](https://scikit-learn.org/stable/auto_examples/)  
Working With Text Data. (2024). Scikit-Learn. [https://scikit-learn.org/1.4/tutorial/text\\_analytics/working\\_with\\_text\\_data.html](https://scikit-learn.org/1.4/tutorial/text_analytics/working_with_text_data.html)

```
In [ ]: # Project assisted and partially generated using the assistance of LLM ChatGPT
import os
import json
import glob
import re
from test import BeautifulSoup
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

class SklearnIndexer:
    def __init__(self, input_folder=".", output_file="./inverted_index.json"):
        self.input_folder = input_folder
        self.output_file = output_file
        self.documents = []
        self.doc_ids = []
        self.vectorizer = None
        self.tfidf_matrix = None
        self.feature_names = None

    # html to plaintext
    def html_to_text(self, html):
        soup = BeautifulSoup(html, "html.parser")
        text = soup.get_text(separator=" ", strip=True)
        splitText = re.sub(r"<\/?a>", " ", text)
        return splitText

    # load the downloaded files
    def load_documents(self):
        files = glob.glob(os.path.join(self.input_folder, "downloaded_page_*.html"))

        for path in files:
            with open(path, "r", encoding="utf-8", errors="ignore") as f:
                html = f.read()

                text = self.html_to_text(html)

                self.documents.append(text)
                self.doc_ids.append(os.path.basename(path))

        print(f"Loaded {len(self.documents)} documents.")

    # build tfidf matrix
    def build_tfidf(self):
        self.vectorizer = TfidfVectorizer(
            stop_words="english",
            lowercase=True,
        )

        self.tfidf_matrix = self.vectorizer.fit_transform(self.documents)
        self.feature_names = self.vectorizer.get_feature_names_out()

    # build the inverted index
    def build_inverted_index(self):
        inverted = {}

        rows, cols = self.tfidf_matrix.nonzero()

        for row, col in zip(rows, cols):
            term = self.feature_names[col]
            score = float(self.tfidf_matrix[row, col])
            docID = self.doc_ids[row]

            if term not in inverted:
                inverted[term] = []

            inverted[term].append({
                "doc_id": docID,
                "tfidf": score
            })

        return inverted

    # save index to a json file
    def save_index(self, inverted_index):
        os.makedirs(os.path.dirname(self.output_file), exist_ok=True)

        with open(self.output_file, "w", encoding="utf-8") as f:
            json.dump(inverted_index, f, indent=4)

        print(f"Saved index at {self.output_file}")

    # load docs, build tfidf, build index
    def run(self):
        self.load_documents()
        self.build_tfidf()
        inverted = self.build_inverted_index()
        self.save_index(inverted)

    # find based on cosine similarity
    def search(self, query, top_k=5):
        # create query vec and find cosine similarity
        queryVec = self.vectorizer.transform([query])
        scores = cosine_similarity(queryVec, self.tfidf_matrix)[0]

        # return sorted by rank
        ranked = sorted(zip(self.doc_ids, scores), key=lambda x: x[1], reverse=True)

        return ranked[:top_k]
```

Run Indexer

```
In [ ]: from indexer import SklearnIndexer

if __name__ == "__main__":
    indexer = SklearnIndexer(input_folder=".", output_file="./inverted_index.json")
    indexer.run()
```

Example Output

```
In [ ]: "tea": [
    {
        "doc_id": "downloaded_page_1.html",
        "tfidf": 0.06385495787246827
    },
    {
        "doc_id": "downloaded_page_10.html",
        "tfidf": 0.10246253140246371
    },
    {
        "doc_id": "downloaded_page_15.html",
        "tfidf": 0.1406372739365191
    },
    {
        "doc_id": "downloaded_page_3.html",
        "tfidf": 0.09926167934880873
    },
    {
        "doc_id": "downloaded_page_9.html",
        "tfidf": 0.1322218782641707
    },
],
"bag": [
    {
        "doc_id": "downloaded_page_1.html",
        "tfidf": 0.083923151450086
    },
    {
        "doc_id": "downloaded_page_15.html",
        "tfidf": 0.18390961606661183
    },
],
"know": [
    {
        "doc_id": "downloaded_page_1.html",
        "tfidf": 0.09875103357053636
    },
    {
        "doc_id": "downloaded_page_11.html",
        "tfidf": 0.0426386553148255
    },
    {
        "doc_id": "downloaded_page_12.html",
        "tfidf": 0.18689371662278184
    },
    {
        "doc_id": "downloaded_page_13.html",
        "tfidf": 0.12017676879799435
    },
    {
        "doc_id": "downloaded_page_14.html",
        "tfidf": 0.03229951675649862
    },
    {
        "doc_id": "downloaded_page_15.html",
        "tfidf": 0.10074784660765194
    },
    {
        "doc_id": "downloaded_page_2.html",
        "tfidf": 0.12017676879799435
    },
    {
        "doc_id": "downloaded_page_7.html",
        "tfidf": 0.060310442047078756
    },
    {
        "doc_id": "downloaded_page_9.html",
        "tfidf": 0.10223988530801692
    },
],
```

3: Flask Application and Query Handling  
Done with assistance from ChatGPT (bug fixing and explanation) Flask. (2010). Welcome to Flask — Flask Documentation (3.0.x). Palletsprojects.com. <https://flask.palletsprojects.com/en/stable/> NLTK: nltk. (n.d.). Www.nltk.org. <https://www.nltk.org/modules/nltk.html>

```
In [ ]: # Project assisted and partially generated using the assistance of LLM ChatGPT
from flask import Flask, request, Response
import csv
from indexer import SklearnIndexer
import os
from nltk.metrics import edit_distance

app = Flask(__name__)
INDEX_FILE = os.path.join(".", "inverted_index.json")
DOC_FOLDER = "."

indexer = SklearnIndexer(input_folder=DOC_FOLDER, output_file=INDEX_FILE)
VOCAB = set()

# run indexer if not existant
if not os.path.exists(INDEX_FILE):
    print("Inverted index not found, running indexer")
    indexer.run()
    VOCAB = set(indexer.vectorizer.get_feature_names_out())
else:
    print("Loading existing index")
    indexer.load_documents()
    indexer.build_tfidf()
    VOCAB = set(indexer.vectorizer.get_feature_names_out())

# spelling correction using edit distance
def spelling_check(query, vocab, max_distance=2):
    tokens = query.lower().split()
    suggestions = []

    # go through each token and find closest word using edit distance
    for t in tokens:
        if len(t) <= 2:
            continue

        closestWord = None
        closestDist = 999

        for v in vocab:
            d = edit_distance(t, v)
            if d < closestDist:
                closestDist = d
                closestWord = v
            if closestDist == 0:
                break

        if closestDist <= max_distance and closestWord != t:
            suggestions[t] = closestWord

    return suggestions if suggestions else None

# format output
def format_output(query, data):
    lines = []
    lines.append(f"<strong>Query: {query}</strong>")
    lines.append("Results:")
    for item in data.get("results", []):
        lines.append(f"<span>{item['document']}</span> -> {item['score']}")

    suggestions = data.get("suggestions")
    if suggestions:
        lines.append("Suggestions:")
        for wrong, correct in suggestions.items():
            lines.append(f"<span>{wrong}</span> -> {correct}")
    else:
        lines.append("Suggestions: None")

    return "\n".join(lines)

@app.route("/query", methods=["POST"])
def query_processor():
    if 'file' not in request.files:
        return Response("ERROR: No file provided", status=400)

    file = request.files['file']

    if not file.filename.endswith(".csv"):
        return Response("ERROR: Only CSV files allowed", status=400)

    decoded = file.read().decode("utf-8").splitlines()
    reader = csv.DictReader(decoded)

    if 'query' not in reader.fieldnames:
        return Response("ERROR: CSV must contain a 'query' column", status=400)

    queries = []
    for row in reader:
        text = row['query'].strip()
        if text:
            queries.append(text)

    if not queries:
        return Response("ERROR: No valid queries found", status=400)

    # build text output
    output_text = ""
    for q in queries:
        suggestion = spelling_check(q, VOCAB)
        raw_results = indexer.search(q, top_k=5)

        results = [
            {"document": doc, "score": float(score)}
            for doc, score in raw_results
        ]

        format_results = format_output(q, {"results": results, "suggestions": suggestion})
        output_text += format_results + "\n\n"

    return output_text

if __name__ == "__main__":
    app.run(host="0.0.0.0", port=5000, debug=True)
```

Test Flask Queries

```
In [ ]: import requests

csv_file_path = "queries.csv"
url = "http://127.0.0.1:5000/query"

with open(csv_file_path, "rb") as f:
    files = {"file": f}
    try:
        response = requests.post(url, files=files)
        response.raise_for_status()
    except requests.exceptions.RequestException as e:
        print(f"Request failed: {e}")
    else:
        # Save returned text file
        filename = "results.txt"
        with open(filename, "wb") as out:
            out.write(response.content)

        print(f"Saved output to {filename}")
```

Query Search Example:

```
In [ ]: Query: quotes
Results:
downloaded_page_12.html -> 0.2962660165228943
downloaded_page_7.html -> 0.2560243690774018
downloaded_page_8.html -> 0.06355193616720467
downloaded_page_5.html -> 0.2510613501789357
downloaded_page_2.html -> 0.2443980794990816
Suggestions: None

Query: stupid
Results:
downloaded_page_12.html -> 0.20750141045778497
downloaded_page_10.html -> 0.15090778520977524
downloaded_page_11.html -> 0.10849849604766169
downloaded_page_1.html -> 0.05762524820913743
downloaded_page_17.html -> 0.050740502076156345
Suggestions: None

Query: world
Results:
downloaded_page_1.html -> 0.12371606314979401
downloaded_page_16.html -> 0.10103151203536786
downloaded_page_13.html -> 0.06355193616720467
downloaded_page_19.html -> 0.047534820618943
downloaded_page_9.html -> 0.03420312693054097
Suggestions: None

Query: stupid1d
Results:
downloaded_page_1.html -> 0.0
downloaded_page_10.html -> 0.0
downloaded_page_11.html -> 0.0
downloaded_page_12.html -> 0.0
downloaded_page_13.html -> 0.0
Suggestions:
stupid1d -> stupid

Query: lady
Results:
downloaded_page_12.html -> 0.10420842272174035
downloaded_page_10.html -> 0.1433056855338832
downloaded_page_19.html -> 0.0829290060000552
downloaded_page_1.html -> 0.0539586447997341
downloaded_page_17.html -> 0.04751196309343366
Suggestions: None

Query: ladyy
Results:
downloaded_page_1.html -> 0.0
downloaded_page_10.html -> 0.0
downloaded_page_12.html -> 0.0
downloaded_page_13.html -> 0.0
Suggestions:
ladyy -> lady

Query: love
Results:
downloaded_page_9.html -> 0.36624082650321055
downloaded_page_14.html -> 0.233642004676565
downloaded_page_4.html -> 0.1206697354160547
downloaded_page_15.html -> 0.10309506870464176
downloaded_page_7.html -> 0.0890158461540505
Suggestions: None
```