Seyed Amin Fattahzadeh
ID: 531682

REPORT DCR Project

Search Engine using positional index and stop words

Introduction

This system is designed to efficiently index and search through various document types, including PDF, TXT, HTM and HTML files. The key features include using stop words, positional indexing, and calculating term frequency-inverse document frequency (TF-IDF) scores.

Stop Words

Stop words are often excluded from indexing and search operations to improve performance and relevance. In my system, I use a predefined set of English stop words to filter out terms that do not contribute significantly to the meaning of the documents.

List of Common Stop Words:

```
# List of common English stop words
STOP_WORDS = set([
    'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your',
    'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it',
    "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this',
    'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had',
    'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while',
    'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above',
    'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once',
    'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
    'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just',
    'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',
    "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn',
    "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't",
    'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"
])
```

By excluding these words, we focus on indexing and searching for more meaningful content within the documents.

Positional Indexing

Positional indexing is a powerful technique used in my system to keep track of the positions of words within documents. This allows for advanced search capabilities, including phrase searches and proximity queries.

How Positional Indexing Works:

1. Tokenisation: Text is tokenised into individual words, excluding stop words.
2. Index Creation: For each document, the positions of each token are recorded.
3. Posting Lists: These positions are stored in posting lists, which map each token to its occurrences in the documents.

For example :

Document 1: "The quick brown fox jumps over the lazy dog"

Token: "quick" -> Positions: [1]

Token: "fox" -> Positions: [3]

This indexing method ensures that we can quickly locate the exact occurrences of terms within the documents.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). My system uses TF-IDF to rank documents based on their relevance to the search query.

TF-IDF Calculation:

1. **Term Frequency (TF):** Measures how frequently a term appears in a document.

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

2. **Inverse Document Frequency (IDF):** Measures how important a term is in the entire corpus.

3. **TF-IDF Score:** Combines TF and IDF to give the final score.

Example: Term: "fox"

Document Frequency (DF): 1 (appears in 1 out of 10 documents)

IDF: log(10 / 1) = 1

TF in Document 1: 1 / 9

TF-IDF: (1/9) * 1 ≈ 0.111

Using TF-IDF, effectively ranks documents based on their relevance, improving the accuracy of search results.

Results (snapshots)

Searching the word “music”

1. As one word which is not a phrase just 1 word:

The output shows us Document IDs, the file addresses, total sequences and the relevance score 🙂

2. now search for a phrase with random words related to social science, try to add some words related to it and see the output (second, third and fourth images)

```
import pickle ···

Term/phrase 'music' found in the following documents:
Document ID: 53, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Music.htm, Relevance Score: 998.2640
Document ID: 39, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Elements_of_music.htm, Relevance Score: 118.8409
Document ID: 72, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Albert_Einstein.htm, Relevance S
Document ID: 30, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\Werner_Meyer-Eppler.htm, Relevance Score: 16.7775
Document ID: 64, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Classical_Greece.htm, Relevance Score
Document ID: 78, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Classical_Greece.htm, Relevance
Document ID: 47, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Religion.htm, Relevance Score: 4.1944
Document ID: 57, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Number_theory.htm, Relevance Score: 4.1944
Document ID: 59, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Architecture.htm, Relevance Score: 4.
Document ID: 73, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Architecture.htm, Relevance Scor
Document ID: 25, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\Social_science.htm, Relevance Score: 2.7963
Document ID: 40, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Erikson's_stages_of_psychosocial_development.htm, Rel
Document ID: 52, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Mathematics.htm, Relevance Score: 2.7963
Document ID: 70, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Acoustic_wave.htm, Relevance Sco
Document ID: 34, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Creator_deity.htm, Relevance Score: 1.3981
Document ID: 42, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Francis_Scott_Key_Bridge_collapse.htm, Relevance Scor
Document ID: 51, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Marriage.htm, Relevance Score: 1.3981
Document ID: 55, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\New_Kingdom_of_Egypt.htm, Relevance Score: 1.
Document ID: 58, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Outdoor_literature.htm, Relevance Score: 1.39
Document ID: 65, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Patapsco_River.htm, Relevance Score:
Document ID: 80, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\entrepreunership course\ENT2_Business Model Innovation.pdf, Relevance Score: 1.3981
Total documents found: 21
```

```
import pickle ···

Term/phrase 'social science' found in the following documents:
Document ID: 25, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\Social_science.htm, Relevance Score: 375.4775
Document ID: 36, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Economics.htm, Relevance Score: 67.0150
Document ID: 37, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Society.htm, Relevance Score: 66.6492
Document ID: 47, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Religion.htm, Relevance Score: 74.5969
Document ID: 57, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Number_theory.htm, Relevance Score: 10.2215
Document ID: 62, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Chronic_condition.htm, Relevance Scor
Document ID: 63, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Civilization.htm, Relevance Score: 46
Document ID: 69, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Abrahamic_religions.htm, Relevar
Document ID: 72, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Albert_Einstein.htm, Relevance S
Document ID: 76, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Chronic_condition.htm, Relevance
Document ID: 77, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Civilization.htm, Relevance Scor
Total documents found: 11
```

```
al science began' found in the following documents:
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\Social_science.htm, Relevance Score: 378.7036
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Religion.htm, Relevance Score: 75.6723
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Mathematics.htm, Relevance Score: 70.9786
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Marriage.htm, Relevance Score: 68.9638
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Society.htm, Relevance Score: 67.7246
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Economics.htm, Relevance Score: 67.0150
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Civilization.htm, Relevance Score: 46.1305
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Civilization.htm, Relevance Score: 46.1305
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Albert_Einstein.htm, Relevance Score: 36.5082
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\Tax.htm, Relevance Score: 30.1350
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Architecture.htm, Relevance Score: 24.8972

ath: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\computer vision\project of CV\Computer Vision Projects\4_5879903974475173794 (1).pdf, Relevance Score: 0.7291
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Outdoor_literature.htm, Relevance Score: 0.7291
Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Acoustic_wave.htm, Relevance Score: 0.7291
ound: 51
'iew as a scrollable element or open in a text editor. Adjust cell output settings...
```

```
Term/phrase 'social sciences developed' found in the following documents:
Document ID: 25, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\Social_science.htm, Relevance Score: 400.6287
Document ID: 51, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Marriage.htm, Relevance Score: 72.8443
Document ID: 37, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Society.htm, Relevance Score: 71.1435
Document ID: 52, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Mathematics.htm, Relevance Score: 66.2710
Document ID: 36, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Economics.htm, Relevance Score: 65.0956
Document ID: 47, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Religion.htm, Relevance Score: 56.7304
Document ID: 63, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Civilization.htm, Relevance Score: 4
Document ID: 77, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Civilization.htm, Relevance Sco
Document ID: 72, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Albert_Einstein.htm, Relevance
Document ID: 27, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\Tax.htm, Relevance Score: 33.4426
Document ID: 53, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Music.htm, Relevance Score: 20.9096
Document ID: 59, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Architecture.htm, Relevance Score: 2
Document ID: 73, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Architecture.htm, Relevance Sco
Document ID: 61, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Axiom.htm, Relevance Score: 16.1568
Document ID: 75, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Axiom.htm, Relevance Score: 16.
Document ID: 40, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Erikson's_stages_of_psychosocial_development.htm, Re
Document ID: 62, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\Chronic_condition.htm, Relevance Sco
Document ID: 76, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Chronic_condition.htm, Relevanc
Document ID: 57, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Number_theory.htm, Relevance Score: 13.8305
Document ID: 49, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\James_Fenimore_Cooper.htm, Relevance Score:
Document ID: 43, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\Glucose.htm, Relevance Score: 9.9835
Document ID: 56, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\Nobel_Prize_in_Physics.htm, Relevance Score:
Document ID: 69, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\level 2\level 3\level 4\DCRB\Abrahamic_religions.htm, Releva
Document ID: 35, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Diabetes.htm, Relevance Score: 7.6941
---
Document ID: 31, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Cooperstown,_New_York.htm, Relevance Score: 0.9163
Document ID: 33, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Cost-of-living_index_-_Wikipedia.html, Relevance Score: 0.91
Document ID: 34, Path: E:\University\Digital Content with Matlab for Digital Content\resutls_of_DCRB_2\my directory\DC Project\wikipedia\level 1\Creator_deity.htm, Relevance Score: 0.9163
Total documents found: 45
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Every project in the world has strengths and weaknesses as do my project.

Strengths:

Stop Words Removal:

- Implementing stop words removal helps in reducing the noise and improving the relevance of the search results.

Positional Indexing:

- Keeping track of the positions of words within documents allows for advanced search capabilities like phrase searches and proximity queries.

TF-IDF Calculation:

- Using TF-IDF for scoring search results is a robust method for evaluating the importance of a word in a document relative to the entire corpus.

Index Persistence:

- Saving the index to disk using `pickle` allows for efficient reloading and avoids the need to reprocess documents each time.

Weaknesses:

Error Handling:

- There's limited error handling in the code. For instance, handling cases where a file cannot be opened or read, or when the input directory is invalid.

User Interface:

- The search functionality is interactive through the console, which might not be user-friendly for non-technical users. Consider creating a simple web interface using Flask or Django.

Sincerely yours.