Information Retrieval Project -REPORT
Search Engine on Stack Overflow corpus and a web Crawler on stackoverflow website

Korupolu Saideepthi S2180010087

ABSTRACT

A search engine on stackoverflow corpus, using Term frequency inverse term frequency (TF-IDF) and cosine similarity for retrieving top 10 similar documents as the given query, next a web crawler on stackoverflow website and finds most popular technologies by getting the tags information of the newest questions asked on the website

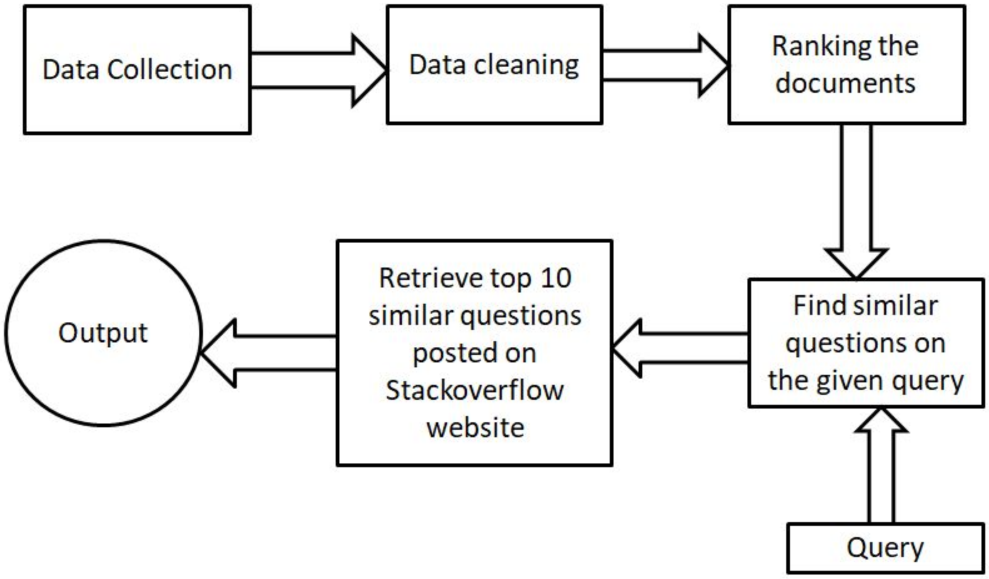
About Stack Overflow

Stack Overflow is the largest, most trusted online community for developers to learn and share their knowledge

Task-1 Problem Statement

Build a Search Engine on Stack Overflow corpus

Overview of the approach


Data Collection
Data cleaning
Ranking the documents
Find similar questions on the given query
Query
Retrieve top 10 similar questions posted on Stackoverflow website
Output

Data Collection

Collected data from Stack Exchange Data Dump , for the project I download posts on AI, DataScience, Computer Science and Computer graphics. The files in this link are in XML-files

- First I convert the xml files to csv files However all the tags are not useful hence I extracted body(questions) and topic from XML by using XML parser and converted the extracted tag’s into a data frame and then stored each of the data frame in a csv file.
- Once all the csv files are available for the xml files, I merged all the csv files into one file
- Next I removed posts which has no text (null values)
- This single csv file has 161423 posts with 3 attributes (Id ,Text , Topic)

Data Cleaning

Since all the posts are not merely text it is a html components, some preprocessing is required

1. Removal of html tags and converting text to lower case
2. Removal of urls
3. Removal of punctuations
4. Removal of stopwords

Text before cleaning

```
'<p>My data set contains a number of numeric attributes and one categorical.</p>\n\n<p>Say,
<code>NumericAttr1, NumericAttr2, ..., NumericAttrN, CategoricalAttr</code>, </p>\n\n<p>where
<code>CategoricalAttr</code> takes one of three possible values:
<code>CategoricalAttrValue1</code>, <code>CategoricalAttrValue2</code> or
<code>CategoricalAttrValue3</code>.</p>\n\n<p>I\'m using default k-means clustering algorithm
implementation for Octave <a
href="https://blog.west.uni-koblenz.de/2012-07-14/a-working-k-means-code-for-octave/">https://blo
g.west.uni-koblenz.de/2012-07-14/a-working-k-means-code-for-octave/</a>.\nIt works with numeric
data only.</p>\n\n<p>So my question: is it correct to split the categorical attribute
<code>CategoricalAttr</code> into three numeric (binary) variables, like
<code>IsCategoricalAttrValue1, IsCategoricalAttrValue2, IsCategoricalAttrValue3</code> ?</p>\n'
```

Text after cleaning

```
'data set contains number numeric attributes one categorical say takes one three possible values
using default kmeans clustering algorithm implementation octave works numeric data question
correct split categorical attribute three numeric binary variables like’
```

After finishing the above 4 steps , I stored this preprocessed text in separate column in the dataframe

TI-IDF Vectorization

For query search I am doing ranked retrieval , so as a weighting factor I am using TF-IDF (term frequency-inverse document frequency) this is used to understand how important a word is to a document in a collection or corpus

I did TF-IDF vectorization for each post in the collection and also for the given query

Cosine Similarity

After vectorizing all the posts and the query , to understand which posts are more similar for the given query , I did cosine similarity

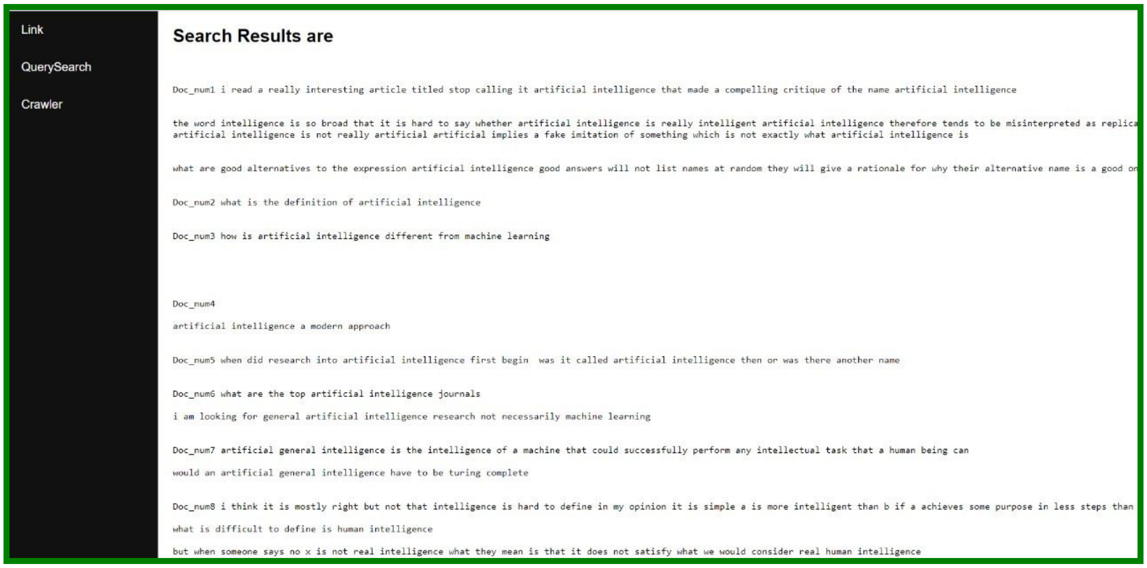
After finding the cosine similarity scores between the query and the each post i stored the values in a dictionary

Top 10 retrieval

Next I sorted the dictionary to get the top 10 most similar questions for the given query that are posted on the Stack overflow website

UI Demo

Search Results for the **Query = "What is artificial intelligence"**



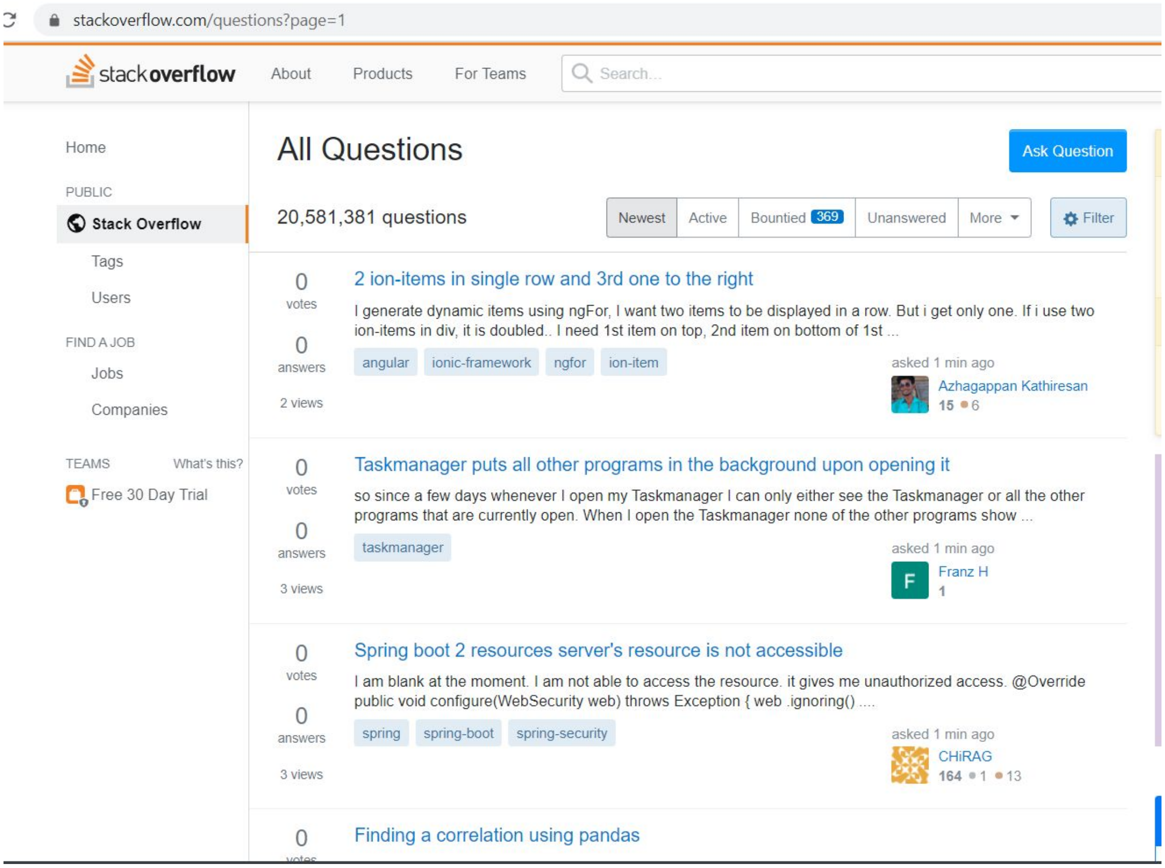
Search Results are

```
Doc_num1 i read a really interesting article titled stop calling it artificial intelligence that made a compelling critique of the name artificial intelligence

the word intelligence is so broad that it is hard to say whether artificial intelligence is really intelligent artificial intelligence therefore tends to be misinterpreted as replica
artificial intelligence is not really artificial artificial implies a fake imitation of something which is not exactly what artificial intelligence is

what are good alternatives to the expression artificial intelligence good answers will not list names at random they will give a rationale for why their alternative name is a good on

Doc_num2 what is the definition of artificial intelligence

Doc_num3 how is artificial intelligence different from machine learning



Doc_num4

artificial intelligence a modern approach

Doc_num5 when did research into artificial intelligence first begin  was it called artificial intelligence then or was there another name

Doc_num6 what are the top artificial intelligence journals

i am looking for general artificial intelligence research not necessarily machine learning

Doc_num7 artificial general intelligence is the intelligence of a machine that could successfully perform any intellectual task that a human being can

would an artificial general intelligence have to be turing complete

Doc_num8 i think it is mostly right but not that intelligence is hard to define in my opinion it is simple a is more intelligent than b if a achieves some purpose in less steps than

what is difficult to define is human intelligence

but when someone says no x is not real intelligence what they mean is that it does not satisfy what we would consider real human intelligence
```

TASK_2 Problem statement

Web crawler on Stack overflow website to find most popular technologies used by the developers

Stack Overflow website page



All Questions

20,581,381 questions

0 votes 0 answers 2 views

2 ion-items in single row and 3rd one to the right

I generate dynamic items using ngFor, I want two items to be displayed in a row. But i get only one. If i use two ion-items in div, it is doubled.. I need 1st item on top, 2nd item on bottom of 1st ...

angular ionic-framework ngfor ion-item

asked 1 min ago Azhagappan Kathiresan 15 ● 6

0 votes 0 answers 3 views

Taskmanager puts all other programs in the background upon opening it

so since a few days whenever I open my Taskmanager I can only either see the Taskmanager or all the other programs that are currently open. When I open the Taskmanager none of the other programs show ...

taskmanager

asked 1 min ago Franz H 1

0 votes 0 answers 3 views

Spring boot 2 resources server's resource is not accessible

I am blank at the moment. I am not able to access the resource. it gives me unauthorized access. @Override public void configure(WebSecurity web) throws Exception { web .ignoring() ....

spring spring-boot spring-security

asked 1 min ago CHiRAG 164 ● 1 ● 13

0 votes

Finding a correlation using pandas

Preprocessing steps in crawler

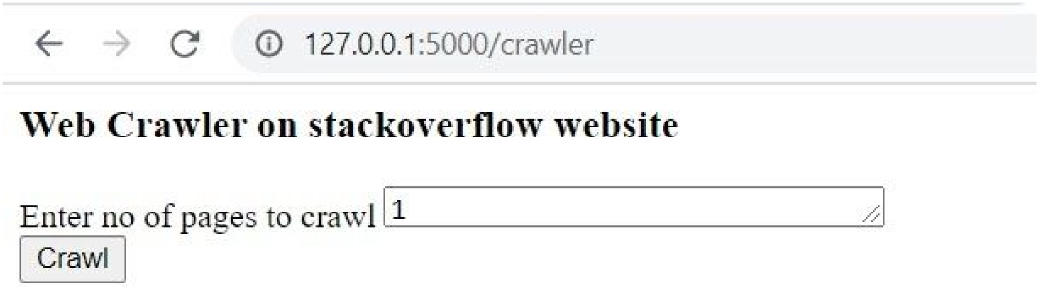
1. Fetched this Stack overflow Website URI to know what are the latest questions asked on the website
2. Parsed the Url to extract links from it to get all the questions
3. Extracted each question url
4. Found the tags of the questions

Tags-Frequency

1. Found number of questions asked on each tag

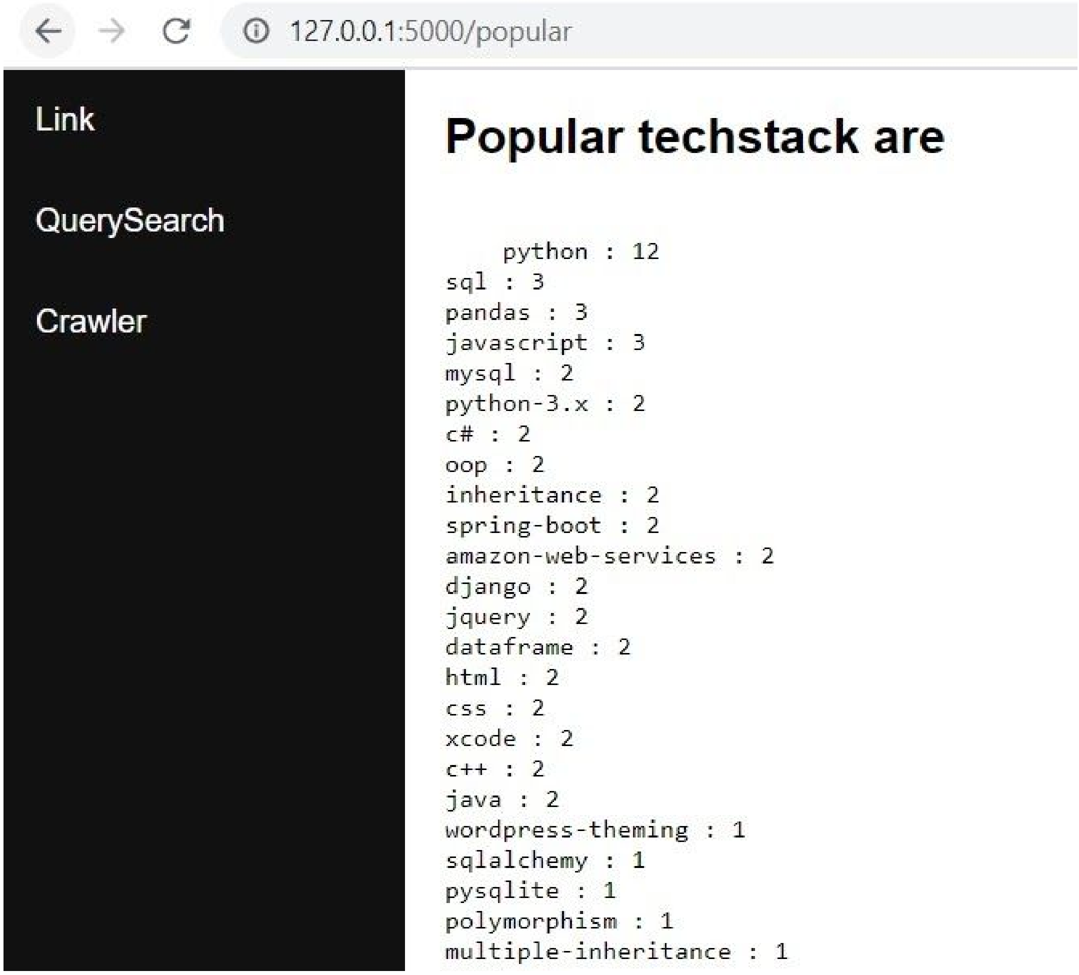
2. Sorted them and saved them in a Tags_frequency.txt

UI Demo



Web Crawler on stackoverflow website

Enter no of pages to crawl 1



Popular techstack are

```
    python : 12
sql : 3
pandas : 3
javascript : 3
mysql : 2
python-3.x : 2
c# : 2
oop : 2
inheritance : 2
spring-boot : 2
amazon-web-services : 2
django : 2
jquery : 2
dataframe : 2
html : 2
css : 2
xcode : 2
c++ : 2
java : 2
wordpress-theming : 1
sqlalchemy : 1
pysqlite : 1
polymorphism : 1
multiple-inheritance : 1
```

Conclusion

By crawling the 1st page , more questions are asked on python , so as of now most popular technology is python.