



# MiniStack

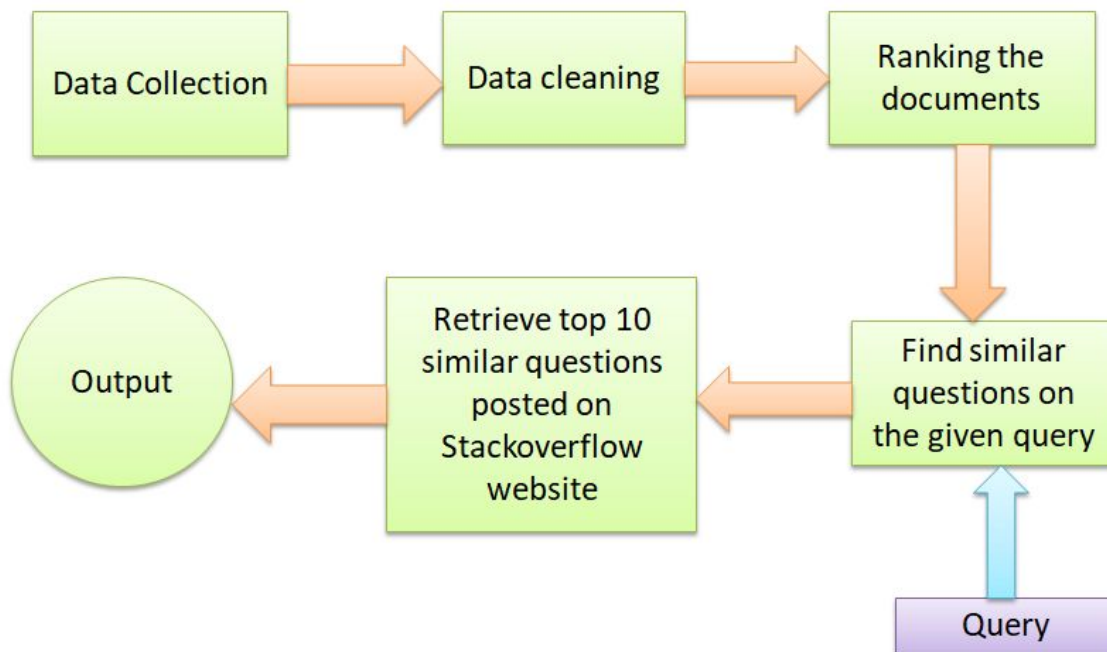
## ABOUT STACK OVERFLOW

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



## Dataset collection

I collected a dataset from [Stack Exchange Data Dump](#) . The URL contains a link to stack overflow corpus which contains a large number of questions along with that it had some other information like tags, URL ,up-vote. The stack overflow dump corpus contains data on several topics which is nearly 63GB but for the project I would consider a part of it.

- Artificial intelligence
- Computer Science
- Computer graphics
- DataScience

The above mentioned are the part of stack overflow corpus which I have considered. The data in this link are in XML Files . This is the format of the XML files

```
<?xml version="1.0" encoding="UTF-8"?>
<posts>
  <row ContentLicense="CC BY-SA 4.0" FavoriteCount="1" CommentCount="0" AnswerCount="5" Tags="<neural-
    networks><backpropagation><terminology><definitions>" Title="What is "backprop"?"
    LastActivityDate="2020-04-22T00:49:46.747" LastEditDate="2019-11-16T17:56:22.093" LastEditorUserId="2444"
    OwnerUserId="8" Body="<p>What does "backprop" mean? Is the "backprop" term basically the same as
    "backpropagation" or does it have a different meaning?</p>" ViewCount="499" Score="8"
    CreationDate="2016-08-02T15:39:14.947" AcceptedAnswerId="3" PostTypeId="1" Id="1"/>
  <row ContentLicense="CC BY-SA 4.0" FavoriteCount="2" CommentCount="0" AnswerCount="3" Tags="<neural-
    networks><machine-learning><statistical-ai><generalization>" Title="How does noise affect
    generalization?" LastActivityDate="2019-02-23T22:36:37.133" LastEditDate="2019-02-23T22:36:19.090"
    LastEditorUserId="2444" OwnerUserId="8" Body="<p>Does increasing the noise in data help to improve the
    learning ability of a network? Does it make any difference or does it depend on the problem being solved?
    How is it affect the generalization process overall?</p>" ViewCount="649" Score="11" CreationDate="2016-
    08-02T15:40:20.623" AcceptedAnswerId="9" PostTypeId="1" Id="2"/>
  <row ContentLicense="CC BY-SA 3.0" CommentCount="0" LastActivityDate="2016-08-02T15:40:24.820"
    OwnerUserId="4" Body="<p>"Backprop" is the same as "backpropagation": it's just a shorter way to say it. It
    is sometimes abbreviated as "BP".</p>" Score="13" CreationDate="2016-08-02T15:40:24.820" PostTypeId="2"
    Id="3" ParentId="1"/>
  <row ContentLicense="CC BY-SA 3.0" FavoriteCount="11" CommentCount="0" AnswerCount="4" Tags="<deep-neural-
    networks><search><neurons>" Title="How to find the optimal number of neurons per layer?"
    LastActivityDate="2018-10-18T10:45:15.213" LastEditDate="2018-10-18T10:45:15.213"
    LastEditorUserId="10135" OwnerUserId="8" Body="<p>When you're writing your algorithm, how do you know
    how many neurons you need per single layer? Are there any methods for finding the optimal number of
    them, or is it a rule of thumb?</p>" ViewCount="955" Score="29" CreationDate="2016-08-02T15:41:22.020"
```

I should convert the XML files into 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 the csv files are available for each of the xml files, my next step would be merging all the csv files into one single file. Now this single csv file would look as shown

Next I removed posts which has no text (null values)

This single csv file has 161423 posts with 3 attributes (Id ,Text , Topic)

	Id	Text	Topic
0	0	<p>Besides being "one of the 7 meta questions ...	AlMeta
1	1	<p>I've clicked on <em>chat</em> link, but the...	AlMeta
2	2	<p>I think this will be a crucial thing to fig...	AlMeta
3	3	<p>Are all questions asked on stats and data s...	AlMeta
4	4	<p>I've seen several questions that use the <a...	AlMeta

## Data Cleaning

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



## Remove html tags

Input :

```
'<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'
```

This input contains html tags which has to be removed

Function :

```
def cleanhtml(raw_html):
```

```
    cleanr = re.compile('<.*?>')
```

```
    cleantext = re.sub(cleanr, "", raw_html)
```

```
    return cleantext.lower()
```

Output:

```
" my data set contains a number of numeric attributes and one categorical. \n\n say, , \n\n where
takes one of three possible values: , or . \n\n i'm using default k-means clustering algorithm
implementation for octave
https://blog.west.uni-koblenz.de/2012-07-14/a-working-k-means-code-for-octave/ .\nit works with
numeric data only. \n\n so my question: is it correct to split the categorical attribute  into three
numeric (binary) variables, like ? \n"
```

**Now the html tags are removed and everything is in lower case**



## Remove URLs

Input :

" my data set contains a number of numeric attributes and one categorical. \n\n say, , \n\n where takes one of three possible values: , or . \n\n i'm using default k-means clustering algorithm implementation for octave

<https://blog.west.uni-koblenz.de/2012-07-14/a-working-h-means-code-for-octave/> .\nit works with numeric data only. \n\n so my question: is it correct to split the categorical attribute into three numeric (binary) variables, like ? \n"

Function:

```
url_regex = 'http[s]?://(?:[a-zA-Z]|[0-9]|[$_-@.&+][!*\(\)\,\:]|(?%[0-9a-fA-F][0-9a-fA-F]))+';
```

```
for i in range(preprocessed_post_text.shape[0]):
    preprocessed_post_text[i] = re.sub(url_regex, "", preprocessed_post_text[i]);
```

Output:

"my data set contains a number of numeric attributes and one categorical. \n\n say, , \n\n where takes one of three possible values: , or . \n\n i'm using default k-means clustering algorithm implementation for octave . \n\n it works with numeric data only. \n\n so my question: is it correct to split the categorical attribute into three numeric (binary) variables, like ? \n"

**URLS are removed**

## Remove Punctuations

Input:

" my data set contains a number of numeric attributes and one categorical. \n\n say, , \n\n where takes one of three possible values: , or . \n\n i'm using default k-means clustering algorithm implementation for octave . \n\n it works with numeric data only. \n\n so my question: is it correct to split the categorical attribute into three numeric (binary) variables, like ? \n"

Function:

```
def cleanpunc(sentence):
    cleaned = re.sub(r'[?!|"|#|:|=|+|_|{|}|[]|!|$|%|^|&|]',r'',sentence)
    cleaned = re.sub(r'[\.,!)(\|\\|/|~|'|>|<|*|$|@|;|→|]',r'',cleaned)
    return cleaned
```

### Output:

" my data set contains a number of numeric attributes and one categorical \n\n say \n\n where takes one of three possible values or \n\n i'm using default kmeans clustering algorithm implementation for octave \nit works with numeric data only \n\n so my question is it correct to split the categorical attribute into three numeric binary variables like \n"

### **Removed the punctuations**

### **Remove Stopwords**

#### Input:

" my data set contains a number of numeric attributes and one categorical \n\n say \n\n where takes one of three possible values or \n\n i'm using default kmeans clustering algorithm implementation for octave \nit works with numeric data only \n\n so my question is it correct to split the categorical attribute into three numeric binary variables like \n"

#### Function:

```
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
print('list of stop words:', stop_words)

def nlp_preprocessing(total_text):
    """Removes stop words and alpha numeric values"""
    if type(total_text) is not int: # Numbers doesn't make any sense in searching them
        string = ""
        for words in total_text.split():
            # remove the special chars in review like "$@!%^&*()_+~?>< etc.
            word = "".join(e for e in words if e.isalnum())
            # stop-word removal
            if not word in stop_words:
                string += word + " "
        return string
```

#### Output:

'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 '

Data is cleaned and stored in the a separate column as shown

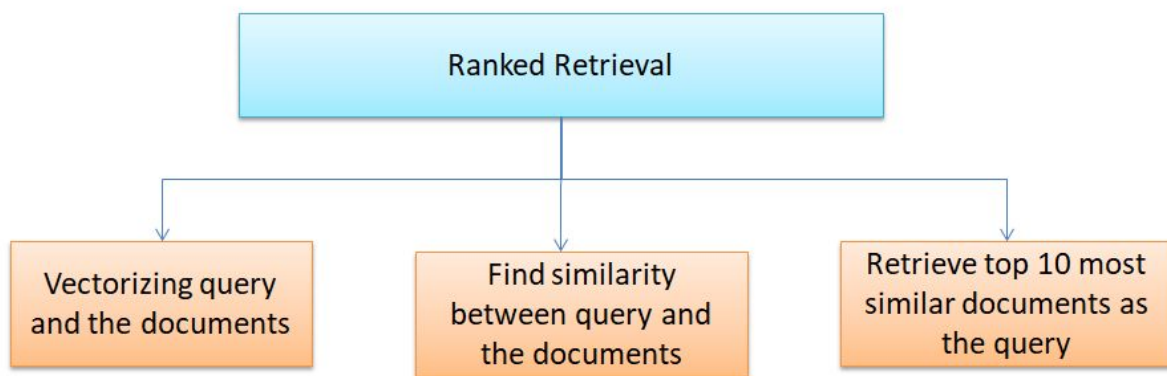
Shape (161423, 5)

	Id	Text	Topic	non_stopword_removed_preprocessed_text	preprocessed_text
0	0	<p>Besides being "one of the 7 meta questions ...	/Almeta	besides being one of the 7 meta questions ever...	besides one 7 meta questions every site ask pl...
1	1	<p>I've clicked on <em>chat</em> link, but the...	/Almeta	i have clicked on chat link but the list is em...	clicked chat link list empty also tried create...
2	2	<p>I think this will be a crucial thing to fig...	/Almeta	i think this will be a crucial thing to figure...	think crucial thing figure one hand think impo...
3	3	<p>Are all questions asked on stats and data s...	/Almeta	are all questions asked on stats and data scie...	questions asked stats data science se also top...
4	4	<p>I've seen several questions that use the <a...	/Almeta	i have seen several questions that use the art...	seen several questions use artificialintellige...

Non\_stopword\_removed\_preprocessed\_text is the data after removing html tags, urls and punctuations

Preprocessed\_text is the data after removing stop words also

## Ranked Retrieval



### TF-IDF Vectorization

For vectorizing the data and the query I did TF-IDF vectorization (Term frequency inverse term frequency)

Function:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(total_df['non_stopword_removed_preprocessed_text'].values)
```

As the size of the data is very large I used in-built function for vectorization



For example let's consider a query **Query = "What is artificial intelligence"**

Code:

```
Query_Bow = vectorizer.transform([Query])
```

## Cosine Similarity

For finding the similarity between two vectors i.e each sentence and the query I used cosine similarity

Function:

```
doc_dict = dict()
```

```
for i in range(X.shape[0]):  
    doc_dict[i] = cosine_similarity(X[i], Query_Bow)
```

After finding similarity scores between sentences and the query, I stored them in the dictionary

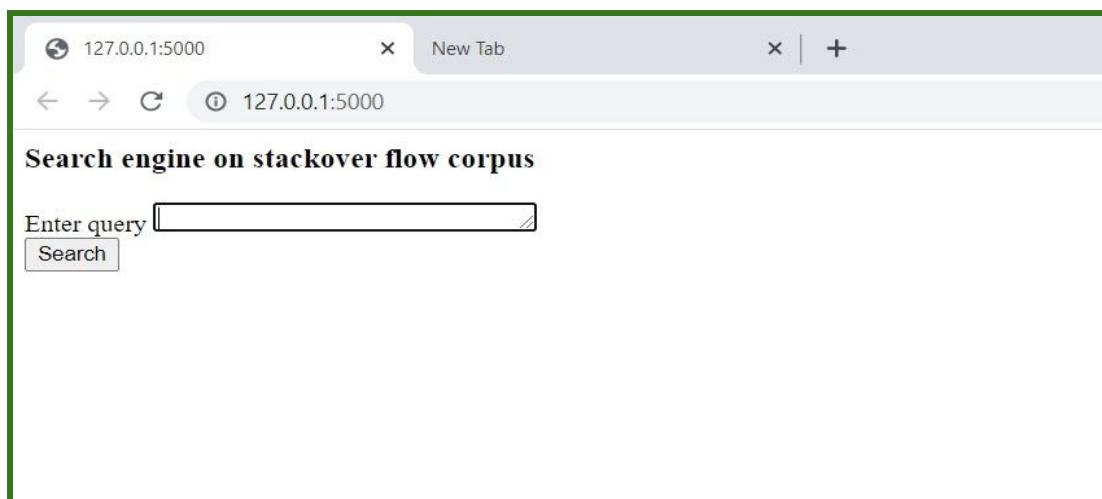
## Get Top 10

After storing the similarity scores in the dictionary I sorted them and retrieved top 10 documents

Code:

```
a = sorted(doc_dict.items(), key=lambda x: x[1], reverse=True) [:10]
```

## UI-Application





Query enter

[←](#) [→](#) [×](#) 127.0.0.1:5000

### Search engine on stackover flow corpus

Enter query

Similar questions as the given query

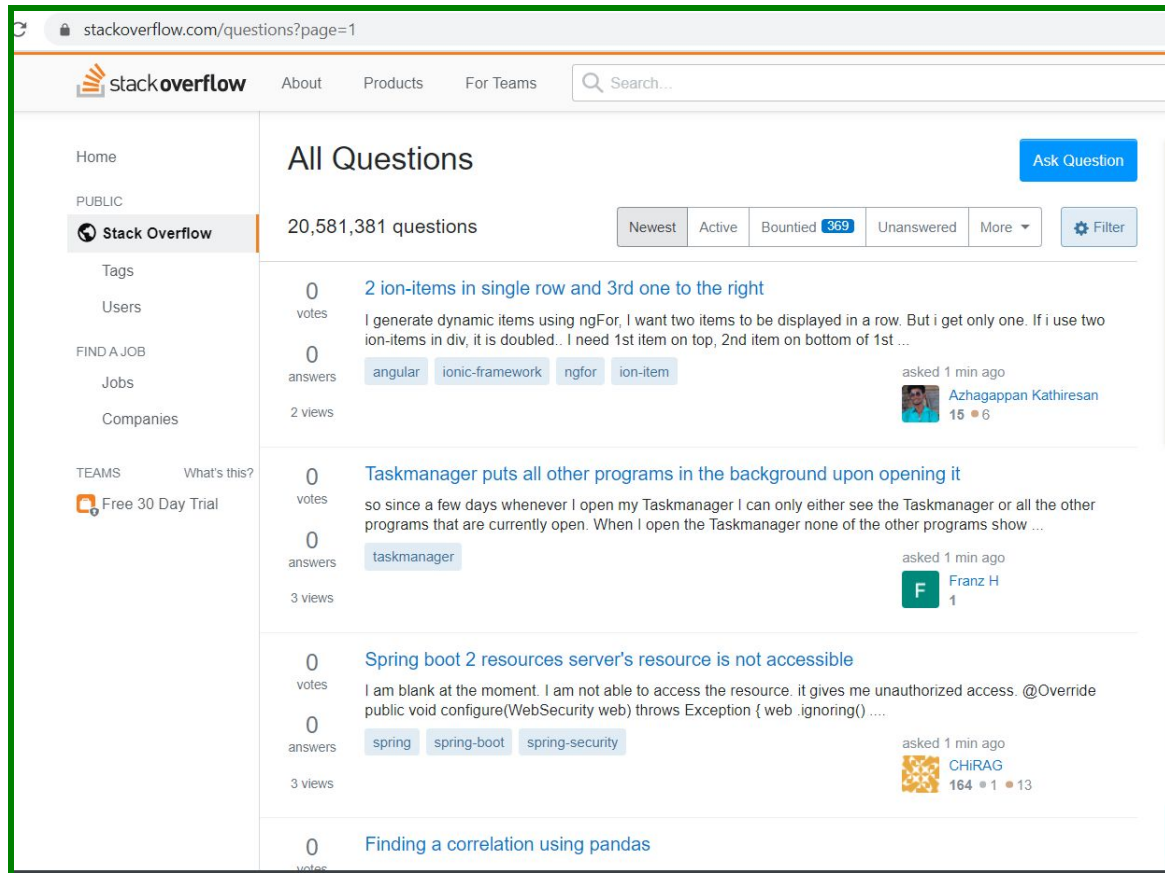
Link	Search Results are
QuerySearch	
Crawler	<p>Doc_num1 i read a really interesting article titled stop calling it artificial intelligence that made a compelling critique of the name artificial intelligence</p> <p>the word intelligence is so broad that it is hard to say whether artificial intelligence 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</p> <p>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 one</p> <p>Doc_num2 what is the definition of artificial intelligence</p> <p>Doc_num3 how is artificial intelligence different from machine learning</p> <p>Doc_num4</p> <p>artificial intelligence a modern approach</p> <p>Doc_num5 when did research into artificial intelligence first begin was it called artificial intelligence then or was there another name</p> <p>Doc_num6 what are the top artificial intelligence journals</p> <p>i am looking for general artificial intelligence research not necessarily machine learning</p> <p>Doc_num7 artificial general intelligence is the intelligence of a machine that could successfully perform any intellectual task that a human being can</p> <p>would an artificial general intelligence have to be turing complete</p> <p>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</p> <p>what is difficult to define is human intelligence</p> <p>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</p>

Retrieved top 10 similar questions as the given query that are posted on the Stackoverflow website

## TASK\_2 Problem statement

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

### Stack Overflow website page



### Preprocessing steps in crawler

1. Fetched this [Stack overflow Website](https://stackoverflow.com/questions?page=1) URL 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

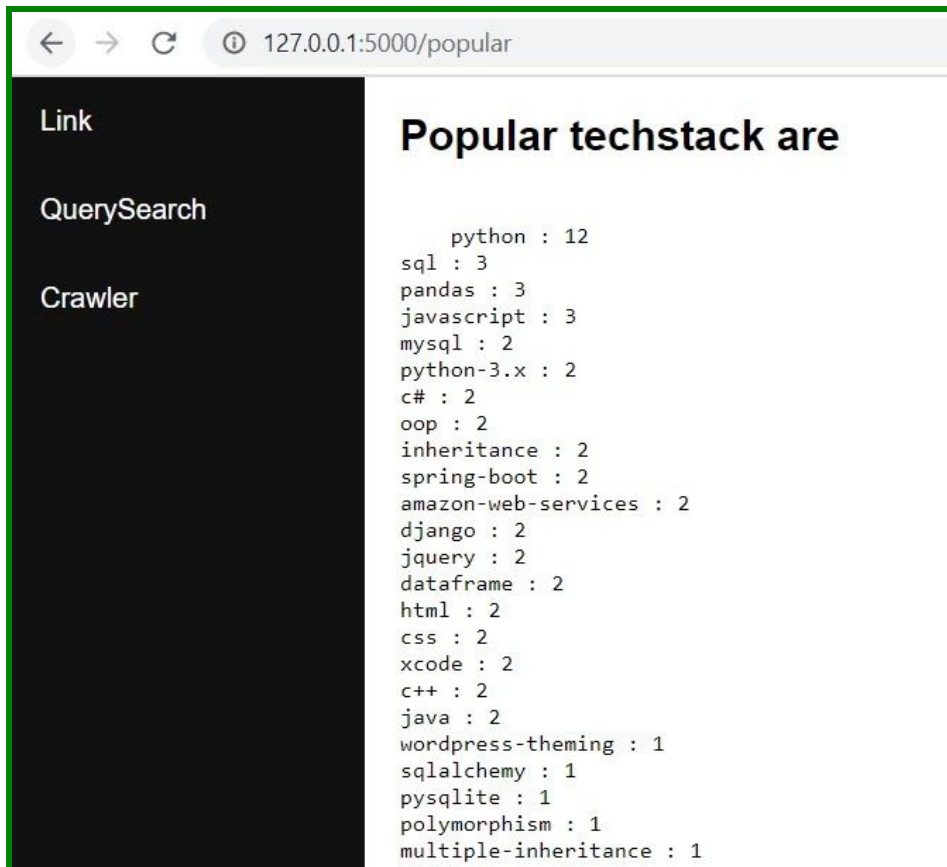
### Tags-Frequency

1. Found number of questions asked on each tag
2. Sorted them and saved them in a Tags\_frequency.txt

## UI Demo



A screenshot of a web browser showing the URL `127.0.0.1:5000/crawler`. The page title is "Web Crawler on stackoverflow website". Below the title, there is a text input field labeled "Enter no of pages to crawl" with the value "1" entered. To the right of the input field is a "Crawl" button.



A screenshot of a web browser showing the URL `127.0.0.1:5000/popular`. The page title is "Popular techstack are". On the left side, there is a dark sidebar with three links: "Link", "QuerySearch", and "Crawler". The main content area displays a list of techstacks and their frequencies.

Link	QuerySearch	Crawler
<b>Popular techstack are</b>		
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.