Homework5 report

(steps & result analysis)

Name: CHEN XU

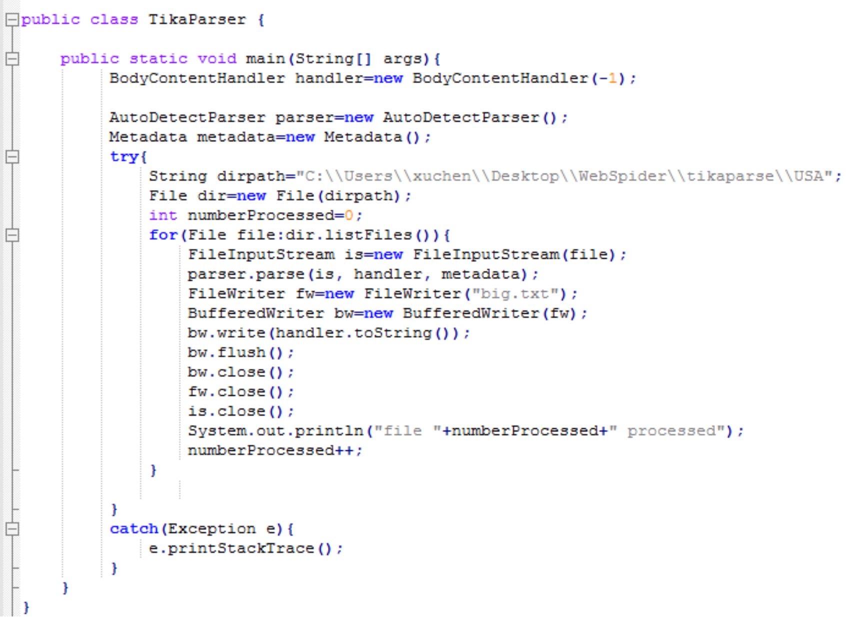
Student ID: 9012039827

Email:cxu264@usc.edu

目录

Steps for homework5:

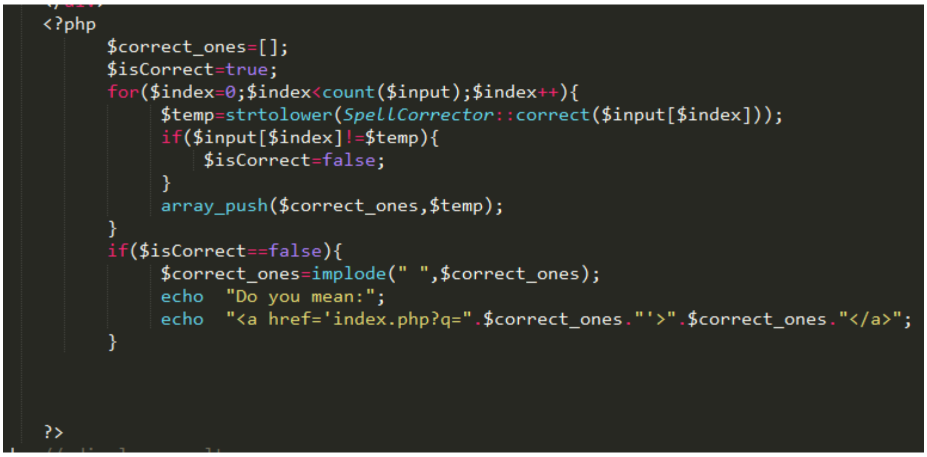
1. generate the big.txt using tika:
   write a program in java and generate big.txt(running about 30 minutes)



```java
public class TikaParser {

    public static void main(String[] args){
        BodyContentHandler handler=new BodyContentHandler(-1);

        AutoDetectParser parser=new AutoDetectParser();
        Metadata metadata=new Metadata();
        try{
            String dirpath="C:\\Users\\xuchen\\Desktop\\WebSpider\\tikaparse\\USA";
            File dir=new File(dirpath);
            int numberProcessed=0;
            for(File file:dir.listFiles()){
                FileInputStream is=new FileInputStream(file);
                parser.parse(is, handler, metadata);
                FileWriter fw=new FileWriter("big.txt");
                BufferedWriter bw=new BufferedWriter(fw);
                bw.write(handler.toString());
                bw.flush();
                bw.close();
                fw.close();
                is.close();
                System.out.println("file "+numberProcessed+" processed");
                numberProcessed++;
            }

        }
        catch(Exception e){
            e.printStackTrace();
        }
    }
}
```

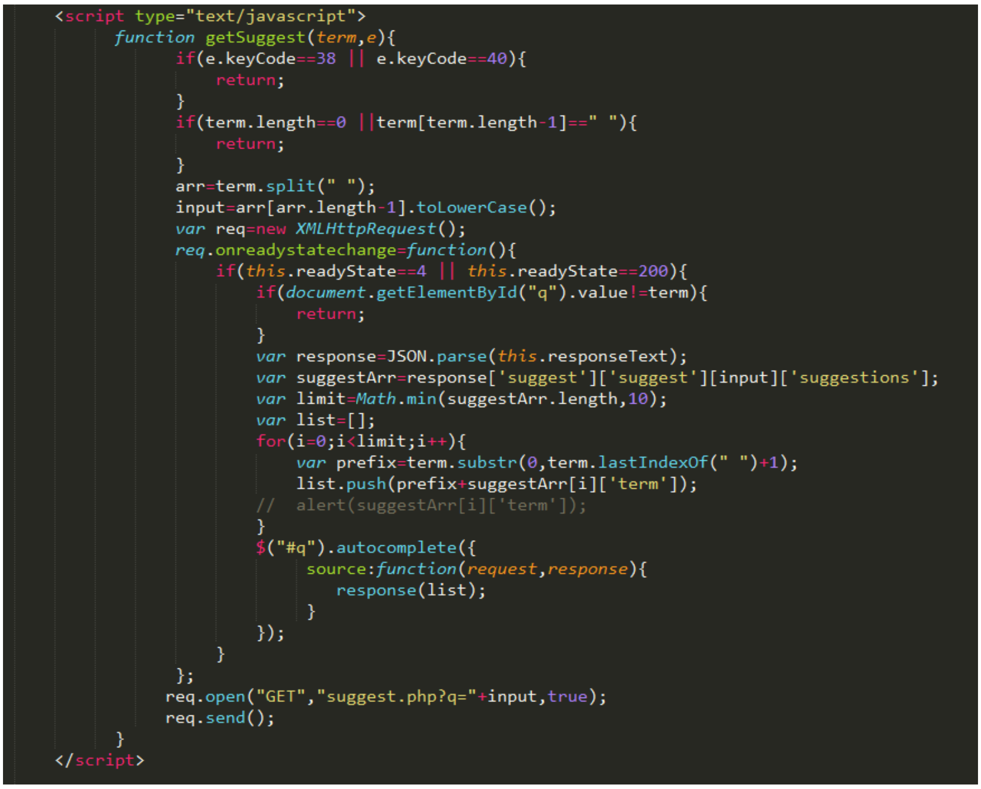
   After that, I change the solrconfig.xml like the instructions of home and add suggestions to the solr in order to get the suggestions list, here below are the preparations for the homework

2. use Peter Norvig’s SpellCorrector.php for the first time to generate serialized_dictionary.txt (running about 15 minutes) you need to set ini_set(‘memory_limit’,-1) in order to make it accept the large memory
3. write a piece of code of correction:



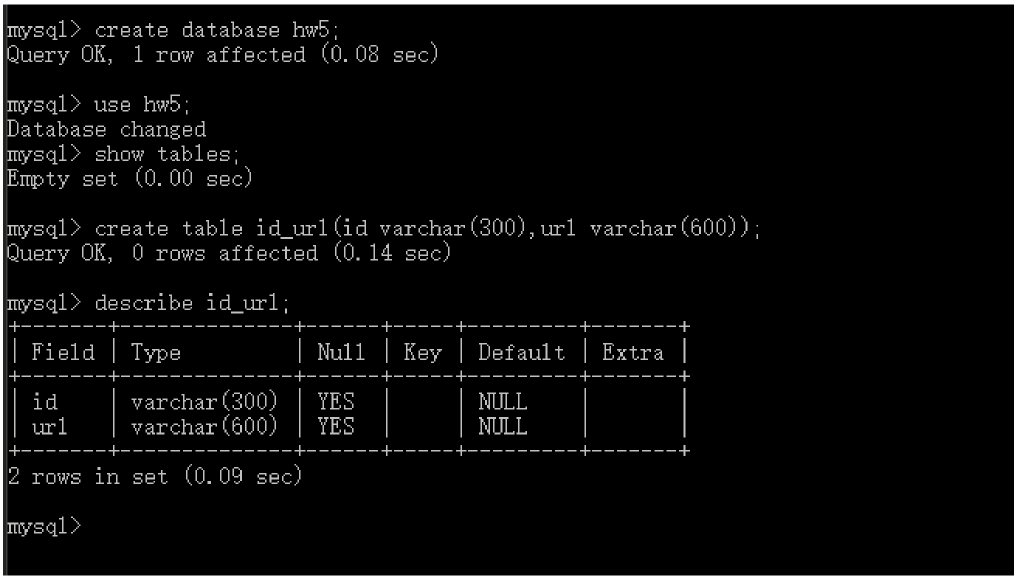
```php
<?php
        $correct_ones=[];
        $isCorrect=true;
        for($index=0;$index<count($input);$index++){
            $temp=strtolower(SpellCorrector::correct($input[$index]));
            if($input[$index]!=$temp){
                $isCorrect=false;
            }
            array_push($correct_ones,$temp);
        }
        if($isCorrect==false){
            $correct_ones=implode(" ",$correct_ones);
            echo  "Do you mean:";
            echo  "<a href='index.php?q=".$correct_ones."'>".$correct_ones."</a>";
        }


?>
```

4. write the code of auto complete : use JQUERY functions for that: add class=’ui-widget’ to the body and the input part add onkeyup event to the input and write javascript code which calls suggest.php:



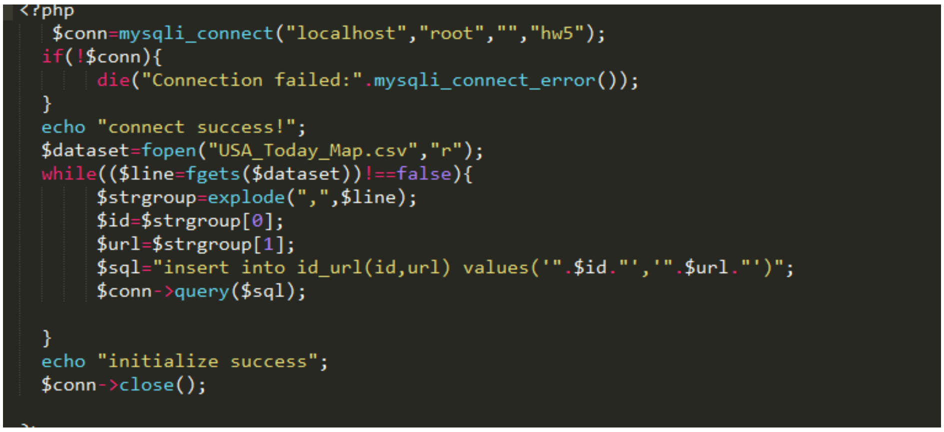
```
<script type="text/javascript">
      function getSuggest(term,e){
            if(e.keyCode==38 || e.keyCode==40){
                return;
            }
            if(term.length==0 ||term[term.length-1]==" "){
                return;
            }
            arr=term.split(" ");
            input=arr[arr.length-1].toLowerCase();
            var req=new XMLHttpRequest();
            req.onreadystatechange=function(){
                if(this.readyState==4 || this.readyState==200){
                    if(document.getElementById("q").value!=term){
                        return;
                    }
                    var response=JSON.parse(this.responseText);
                    var suggestArr=response['suggest']['suggest'][input]['suggestions'];
                    var limit=Math.min(suggestArr.length,10);
                    var list=[];
                    for(i=0;i<limit;i++){
                        var prefix=term.substr(0,term.lastIndexOf(" ")+1);
                        list.push(prefix+suggestArr[i]['term']);
                    //  alert(suggestArr[i]['term']);
                    }
                    $("#q").autocomplete({
                         source:function(request,response){
                            response(list);
                         }
                    });
                }
            };
           req.open("GET","suggest.php?q="+input,true);
           req.send();
      }
</script>
```

5. For the reason I use WAMP SERVER for developing, then I store the URL and id into MYSQL and here below are the codes for creating database and table:



```
mysql> create database hw5;
Query OK, 1 row affected (0.08 sec)

mysql> use hw5;
Database changed
mysql> show tables;
Empty set (0.00 sec)

mysql> create table id_url(id varchar(300),url varchar(600));
Query OK, 0 rows affected (0.14 sec)

mysql> describe id_url;
+-------+--------------+------+-----+---------+-------+
| Field | Type         | Null | Key | Default | Extra |
+-------+--------------+------+-----+---------+-------+
| id    | varchar(300) | YES  |     | NULL    |       |
| url   | varchar(600) | YES  |     | NULL    |       |
+-------+--------------+------+-----+---------+-------+
2 rows in set (0.09 sec)

mysql>
```

6. then I write the code of initializing database and run it on the server to store all the information in excel file into MYSQL:which is used for generating clickable title like google



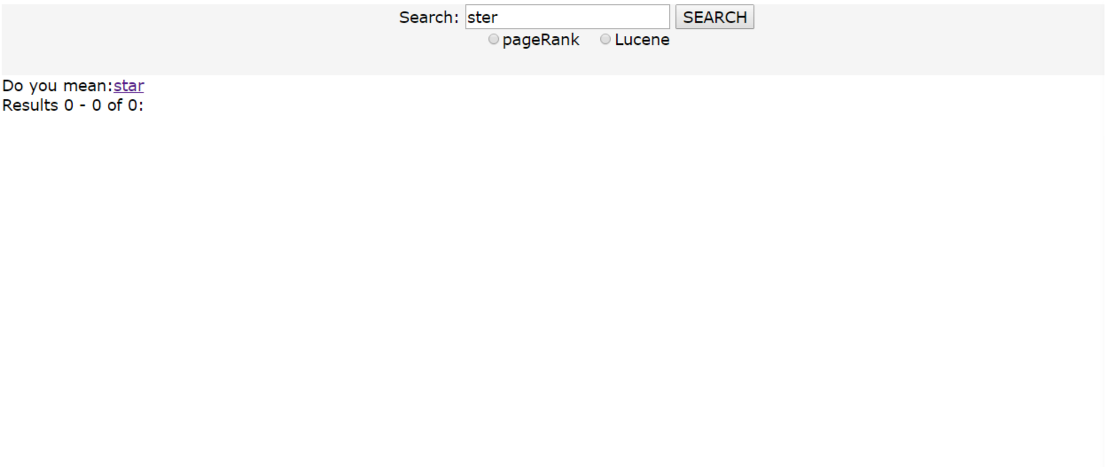
```php
<?php
   $conn=mysqli_connect("localhost","root","","hw5");
  if(!$conn){
       die("Connection failed:".mysqli_connect_error());
  }
  echo "connect success!";
  $dataset=fopen("USA_Today_Map.csv","r");
  while(($line=fgets($dataset))!==false){
       $strgroup=explode(",",$line);
       $id=$strgroup[0];
       $url=$strgroup[1];
       $sql="insert into id_url(id,url) values('".$id."','".$url."')";
       $conn->query($sql);

  }
  echo "initialize success";
  $conn->close();
?>
```

7. then I use php preg_match and domdocument to get the snippet and show it on the page(more details see GetSnippets.php function getsnippet())

result:

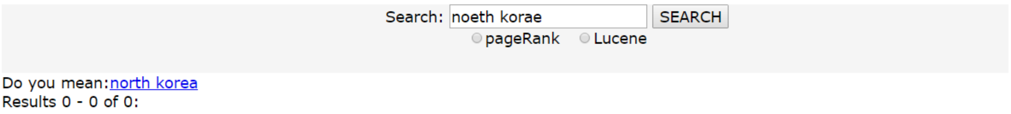
correction:

1.ster ------star



Do you mean:star
Results 0 - 0 of 0:

2.noeth korae-------north korea

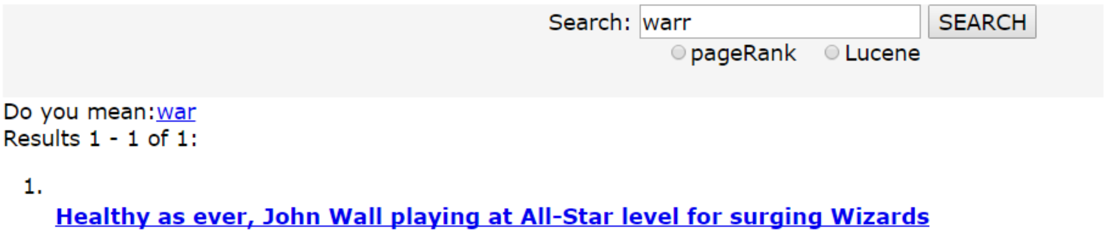

Search: noeth korae SEARCH

○ pageRank ○ Lucene

Do you mean:north korea

Results 0 - 0 of 0:

3.War-------war


Search: warr SEARCH

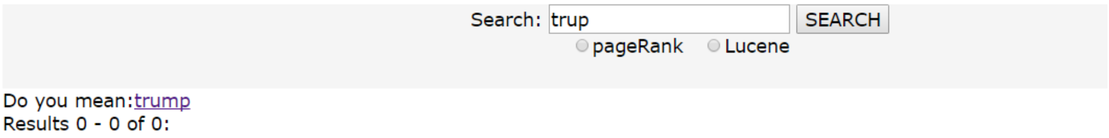
○ pageRank ○ Lucene

Do you mean:war

Results 1 - 1 of 1:

1. Healthy as ever, John Wall playing at All-Star level for surging Wizards

4.trup------trump

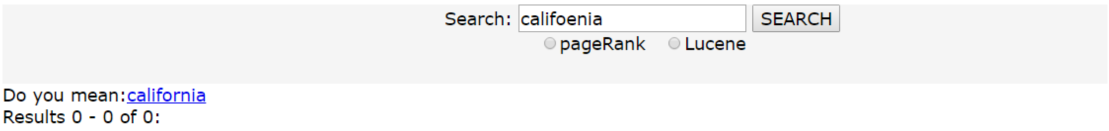

Search: trup SEARCH

○ pageRank ○ Lucene

Do you mean:trump

Results 0 - 0 of 0:

5.califoenia----california


Search: califoenia SEARCH

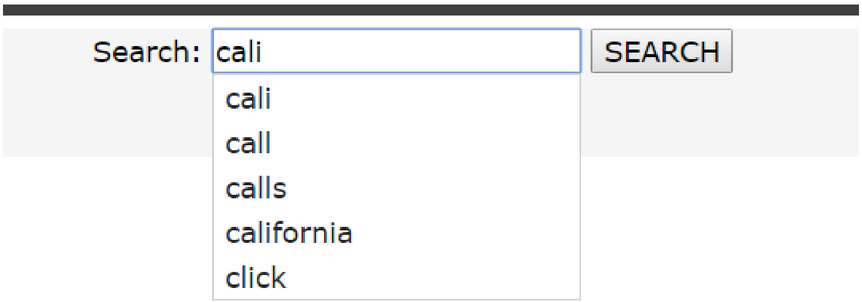
○ pageRank ○ Lucene

Do you mean:california

Results 0 - 0 of 0:

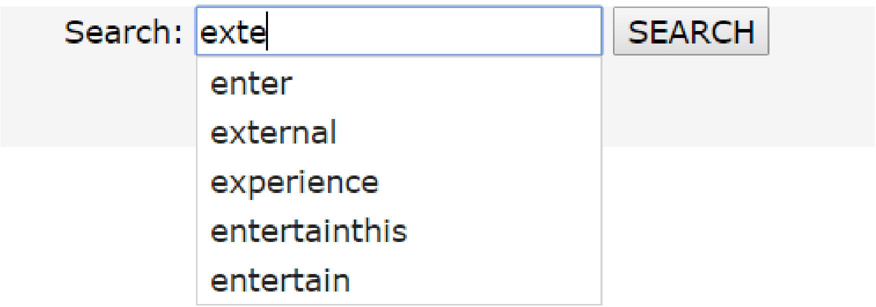
Autocomplete:

1.input:cali


Search: cali

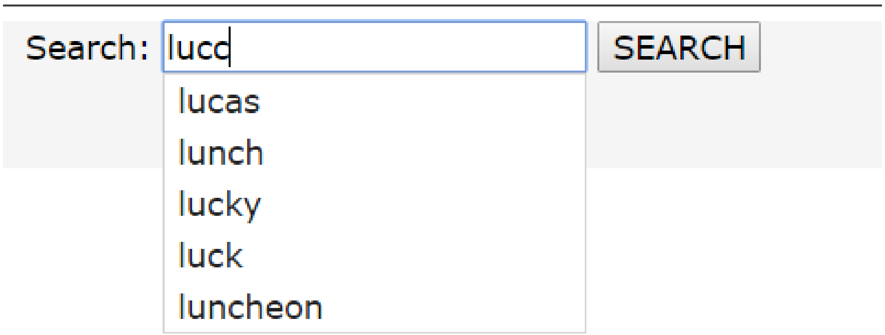
```
cali
call
calls
california
click
```

2.input:exte


Search: exte

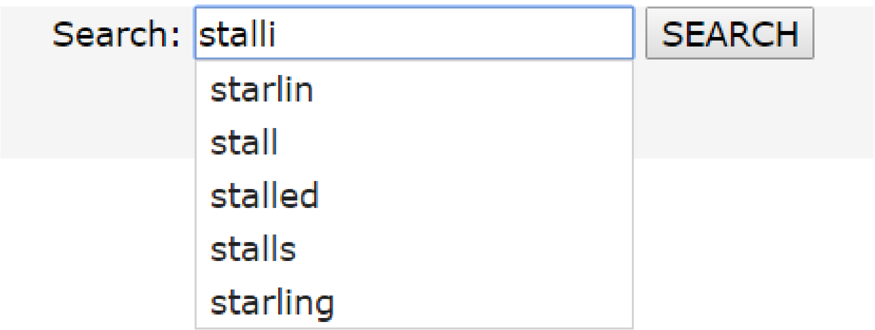
```
enter
external
experience
entertainthis
entertain
```

3.input:lucc


Search: lucc

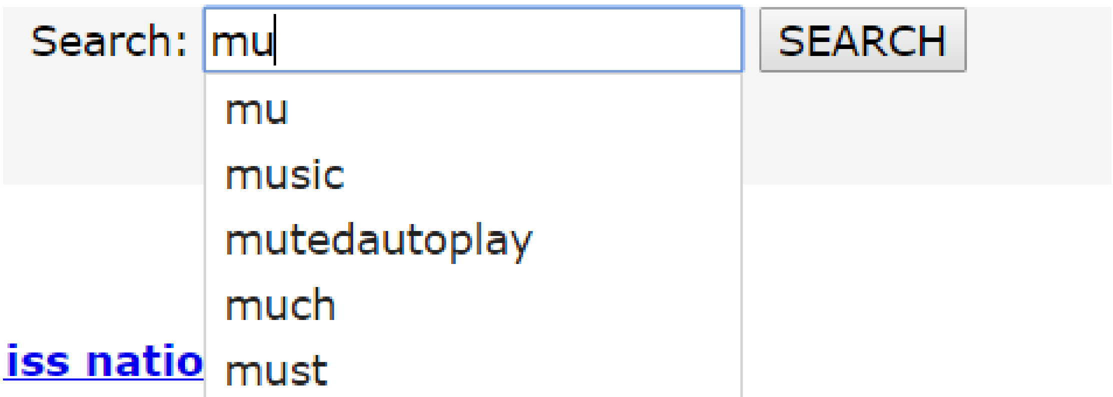
```
lucas
lunch
lucky
luck
luncheon
```

4.input:stalli


Search: stalli

```
starlin
stall
stalled
stalls
starling
```

5.input:mu



Search: mu