Speech Tagger

gumgum

Report File

DHRUVKUMAR PATEL

9379798797

Dhruvpatel401@yahoo.in

1) **Introduction:-** Idea is to develop a Speech Tagger which is used to tag the text present in the Web Pages using the Stanford part of Speech Tagger. Here this speech tagger that reads English text from HTML web pages and assigns part of speech to each word such as noun, verb, adjective etc. So it is generate output as tagged String. For example if the String is Hello World then Output is Hello_UH where UH stands for Interjection. This is used in Natural Language Processing.

2) **Specification:-** Here I used latest version 3.6.0 English Stanford Tagger library. One can download from http://nlp.stanford.edu/software/tagger.shtml this link. Also I used jsoup: Java HTML Parser to extract the text from HTML web page. One can refer https://jsoup.org/ this link for more details. Here following are the main classes and components.

   i) **PageTagger:-** this class contains following methods.

      a) **tagText() :-** this method is used to tag the text

      pre:- method takes String as an input. So, input is not null.

      Post:- generate tagged text as an output which is String. Output is not null.

      b) **getText():-** this method is to extract a text from HTML web pages body tags.

      Pre:- takes java.net.URL as an input. So, validate the URL.

      Post:- generate the text from given url html web page as a String. So, output is not null.

      c) **main():-** this is used to execute getText() and tagText() methods and write the output in output file.

3) **Design:-** following is the step by step design of Speech Tagger to fulfil the specific output requirement.

   ii) **PageTagger:-** this class contains following methods. That internally contains an object of MaxentTagger initialized with an english-left3words-distsim.tagger file.

a) **tagText() :-** this method is used to take String as an input and return tagged Text as an output.

Step1:- this method takes input as a String untagged text.

Step2:- initialize the instance of MaxentTagger

Step3:- use public String tagString(String toTag) method using an instance of MaxentTagger to get taggedString as an output.

b) **getText():-** this method takes URL as an input and return the extracted text in String as an output.

Step1:- this method takes java.net.URL as an input.

Step2:- validate the url using UrlValidator class from org.apache.commons.validator.

Step3:- connect the specific given url using Jsoup connect method which is a third party open source Html parser library to extract text from html web page.

Step4:- sometimes if the text containing html tag is right before that anchor tag<a> that not gives space between the content so I add space before the anchor tag using jsoup select() and before() methods.

Step5:- extract text from body of html page using jsoup document.body().text() method and return String as an output.

c) **main():-** this is used to execute getText() and tagText() method.

Step1:- initialize output file to store output.

Step2:- declare and define java.net.URL class and give specific url.

Step3:- call getText method and get the extracted text as a String from html web page.

Step4:- pass output of getText to input for tagText() method get the tagged String as an output and write into the ouput file.

4) **Implementation:-** I integrate all the libraries in eclipse build path and configured. Follow design to implement the source code. one can find into my source code.

Following the problem I got during implementation:-

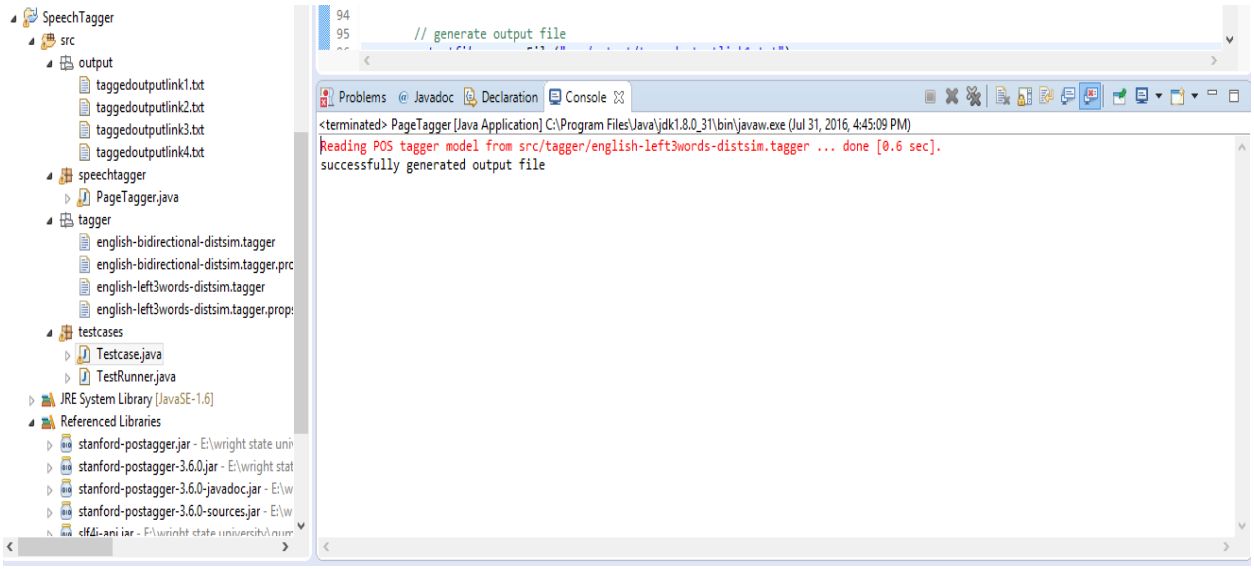
Step1: First I implemented with simple String and then implemented with URL.

Step2:- But, for given links in the PART #2 I got the output for all the given links.

Step3:- For link #3 I got Java heap space error because of running out of memory all the issues I explained in the PART#3 questions and answers.

Step4:- I used temporary file solution so, create file and write output of getText() method into that file. create the String with 4000 words let's say one chunk and then chunk by chunk I send the string to tagText() method and write the ouput into the output file. This solution works and I am getting the output of link #3 url.

Output generated and store into output file:-



5) **Testing:-** I write the test cases for tagText() and getText() method using java junit java testing framework which I used java assertion. Following are the test cases.

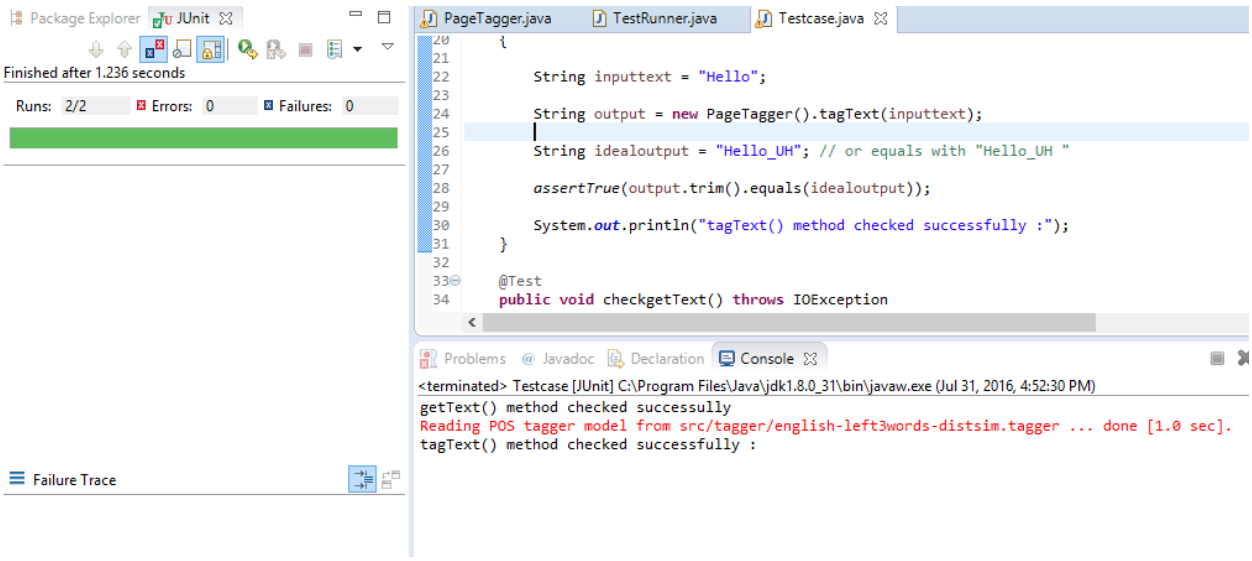
Step1:-test case for tagText() method

```
@Test
	public void checktagText()
	{

		String inputtext = "Hello";

		String output = new PageTagger().tagText(inputtext);

		String idealoutput = "Hello_UH"; // or equals with "Hello_UH "

		assertTrue(output.trim().equals(idealoutput));

		System.out.println("tagText() method checked successfully :");
	}
```

Step2:-test case for getText() method

```
@Test
	public void checkgetText() throws IOException
	{

		URL inputurl = new URL("http://www.example.com/");

		String output = new PageTagger().getText(inputurl);

		String idealoutput = "Example Domain This domain is
established to be used for illustrative examples in documents. You may use
this domain in examples without prior coordination or asking for
permission.  More information...";

		assertTrue(output.equals(idealoutput));

		System.out.println("getText() method checked successully");


	}
```

Step3:-output screenshot which indicates all test cases are successful.



6) Question & Answers:- (part #3)

1) What class and method is responsible for using too much memory?

Answer:- PageTagger class , tagText() method is responsible for using too much memory. Because when the volume of text is high String object inside tagText() method cannot handle all the text as an output from MaxentTagger tagString method.

2) Is this considered a memory leak? Can you elaborate either way?

Answer:- when I run link#3 URL as an input I got java.lang.OutOfMemoryError java heap space error. Yes this is considered as a memory leak because it causes because of usage of large data volume and memory leak. Java application memory is separated in Heap space and permanent generation. this is occurred application try to add more memory into heap space but there is not enough space. So, leaking functionality leaves some objects into the java heap space so after some time leaked objects consume all the available java heap space and give

this type of error. Memory leaks is situation that garbage collection fails to identify unused objects in application.

3) Can you adjust the JVM to provide enough memory to tag this page? How?

Answer:- yes we can adjust the JVM to provide enough memory to tag link#3 URL web page. Heap size is set during the JVM launch and customized by JVM parameters –Xmx and –XX:MaxPermSize. Otherwise platform specific default will be used. I set the parameter in eclipse run conffiguartion arguments using following two parameters: -Xms512M -Xmx1024M which describes 1024 MB of heap space. But, after did this step still I had this problem so I used temporary file solution which I mentioned in implementation part to solve this problem.

4) What tool(s) and techniques can you use to find out why there is a memory problem?

Answer:- I used Eclipse memory analyzer tool to generate report and find out the memory leak problem. I used eclipse plugin of Eclipse Memory Analyzer. When java.lang.OutOfMemory error throws at the same time .hprof file will be generate in your workspace and open that file and select Leak suspect Report and it generates whole report with amount of memory leak and all the other details.

5) Can this problem be solved by changing the garbage collection algorithm?

Answer:- Garbage collector is a mechanism in java used to delete memory automatically. Finalize() method used to cleanup before delete an unused objects. I do not know exact solution using garbage collection but we can use garbage collector System.gc() method to call garbage collection functionalities and delete unused objects and release a memory.

6) Suppose we are using this class to study the distribution of particular parts of speech on web pages in our publisher network. What changes would you suggest in order to

continue to process pages without crashing on pages such as this?

Answer:- yes one can use class to execute the particular part of speech on web pages just need to make change in getText() method in jsoup html parser. For example, if one want to extract text from only specific tag suppose <p> then use parse, manipulation functionality of jsoup parser.

For example, public Element select(String query) where you passes p tag so it will return all the elements containing p tag and use text() method to extract the text which is only inside paragraph tag.

7) Journal & feedback:- journal represent the time wise my efforts to complete this task.

1) part#1 for introduction and understanding it took me only 1 hour to integrate and implement source code of this project.

2) After getting errors in part#2 took me one more hour to redesign the code, write test cases.

3) 1 hour to prepare report and write the questions and answers.

4) overall it took me 4 hours of work and overall it was nice experience for me. I learned all the third party library implementation what I used while implementing this task.