Raj Vardhan

student, passionate about computer science, coding and research

Stanford POS tagger in Eclipse

NOVEMBER 11, 2012   54 COMMENTS (HTTPS://RAJVARDHAN.WORDPRESS.COM/2012/11/11/STANFORD-POS-TAGGER-IN-ECLIPSE/#COMMENTS)

This post will get you started with **POS tagging** in **java** using **Eclipse**.

Why do it ?
Well, a Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word, such as noun, verb, adjective, etc. It is an extremely powerful and accurate tool. You can use it in any application that deals with natural language text to analyze words/tokens and classify them into categories (http://www.computing.dcu.ie/~acahill/tagset.html).

For pre-requisites, follow these simple steps:

1. Download and install java **JDK** and **JRE** on your system from here (http://www.oracle.com/technetwork/java/javase/downloads/index.html).
2. Edit **system environment variables** by right clicking on My Computer -> Properties -> Advance System Settings ->Environment Variables. Copy the path to the bin directory of your JDK installation to the beginning of your environment variable PATH. For default settings, this will look like this: "C:\Program Files\Java\jdk1.7.0_09\bin;" (without the quotes ofcourse).
3. Download **Eclipse IDE** from here (http://www.eclipse.org/downloads/) depending upon your system configuration. You can pick "Eclipse IDE for Java and DSL Developers" if you are not sure which one to chose.
4. Download **Stanford POS tagger** from here (http://nlp.stanford.edu/software/tagger.shtml#Download).

You're almost ready to go. Lets setup our work:

1. Open Eclipse and chose the location of your workspace. This is where all your projects will be stored.
2. Make a new project and name it anything you want. I'll go with the name "**practise**".
3. Add a new class to it. You can name it "**tagTex**t".
4. Go to the directory where your downloaded the Stanford POS tagger, and inside the folder "models". Copy a .tagger file and its corresponding .props file. I will assume these are: "**left3words-wsj-0-18.tagger**" and "left3words-wsj-0-18.props". In your workspace directory, inside your project folder make a new folder and name it "taggers". Go to this folder and paste the tagger and props files.

Alright people. Now lets start coding !

Add/write this code to the tagText.java file you created.

```
1   import java.io.IOException;
2   import edu.stanford.nlp.tagger.maxent.MaxentTagger;
3
4   public class tagText {
5   public static void main(String[] args) throws IOException,
6   ClassNotFoundException {
7
8   // Initialize the tagger
9   MaxentTagger tagger = new MaxentTagger("taggers/left3words-wsj-0-18.tagger");
10
11  // The sample string
12  String sample = "This is a sample text";
13
14  // The tagged string
15  String tagged = tagger.tagString(sample);
16
17  //output the tagged sample string onto your console
18  System.out.println("Input: " + sample);
19  System.out.println("Output: "+ tagged);
20  }
21  }
```

We are not done yet. We need to import the Stanford tagger library to eclipse. To do this:

Right click on your project "practise" -> Build Path -> Configure Build Path -> Click on Add External JARs -> Browse to the location of your download directory of the Stanford POS tagger and select the stanford-postagger.jar file -> Click OK.

 (https://rajvardhan.files.wordpress.com/2012/11/build.png)
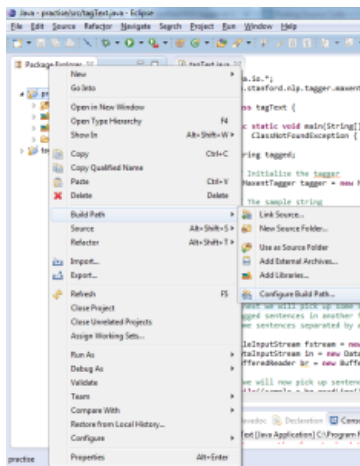
Import library to Eclipse

That's it guyz. Run your code and you should have this output:

Loading default properties from trained tagger taggers/left3words-wsj-0-18.tagger
Reading POS tagger model from taggers/left3words-wsj-0-18.tagger … done [2.1 sec].
i/FW can/MD man/VB the/DT controls/NNS of/IN this/DT machine/NN

 (https://rajvardhan.files.wordpress.com/2012/11/output1.png)
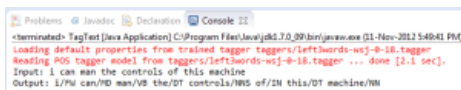
The output you will get

The "FW", "MD", "VB", etc next to each word are classes. For example, VB stands for Verb. The complete list of classes can be found here (http://www.computing.dcu.ie/~acahill/tagset.html).

To play around more with this, you can have lots of English sentences stored in a file, say "input.txt" and we can run the tagger and store all tagged sentences in another file, say "output.txt".

To accomplish this, add a new class named "tagTextToFile" to your project with the following code:

```java
1    import java.io.*;
2    import edu.stanford.nlp.tagger.maxent.MaxentTagger;
3
4    public class tagTextToFile {
5
6      public static void main(String[] args) throws IOException,
7      ClassNotFoundException {
8
9      String tagged;
10
11     // Initialize the tagger
12     MaxentTagger tagger = new MaxentTagger("taggers/left3words-wsj-0-18.tagger");
13
14     // The sample string
15     String sample = "i can man the controls of this machine";
16
17     //The tagged string
18     tagged = tagger.tagString(sample);
19
20     //output the tagged sample string onto your console
21     System.out.println(tagged);
22
23     /* next we will pick up some sentences from a file input.txt and store the output of
24     tagged sentences in another file output.txt. So make a file input.txt and write down
25     some sentences separated by a new line */
26
27     FileInputStream fstream = new FileInputStream("input.txt");
28     DataInputStream in = new DataInputStream(fstream);
29     BufferedReader br = new BufferedReader(new InputStreamReader(in));
30
31     //we will now pick up sentences line by line from the file input.txt and store it in the s
32     while((sample = br.readLine())!=null)
33     {
34     //tag the string
35     tagged = tagger.tagString(sample);
36     FileWriter q = new FileWriter("output.txt",true);
37     BufferedWriter out =new BufferedWriter(q);
38     //write it to the file output.txt
39     out.write(tagged);
40     out.newLine();
41     out.close();
42     }
43
44     }
45
46     }
```

You can throw the exceptions (if any) given by Eclipse.

You can continue to play around as much. Happy coding !

References:

1. http://www.galalaly.me/index.php/2011/05/tagging-text-with-stanford-pos-tagger-in-java-applications/(http://www.computir
   /~acahill/tagset.html)
2. http://nlp.stanford.edu/software/tagger.shtml (http://nlp.stanford.edu/software/tagger.shtml)

FILED UNDER <u>NATURAL LANGUAGE PROCESSING</u>   TAGGED WITH <u>ECLIPSE</u>, <u>STANFORD POS TAGGER</u>, <u>TAGGING</u>

About Raj Vardhan
NLP and Machine Learning enthusiast working at McAfee, Part of Intel Security, India.

54 Responses to *Stanford POS tagger in Eclipse*

Abdullah says:
<u>March 16, 2013 at 9:17 pm</u>
Hi Raj

Thank you for this great tutorial. And I want to ask if I want build POS tagger , will be the Standford POS tagger useful ?

I am confusing actually , because I want to implement HMM and try to get best result for word tag.

If you can help me or guide me to do that I will appreciate that.

Thank you

 Reply
<u>**Raj Vardhan**</u> **says:**
<u>March 17, 2013 at 5:03 am</u>
Hi Abdullah.

Building your own POS tagger through Hidden Markov Models is different from using a ready-made POS tagger like that provided by Stanford's NLP group. Stanford POS tagger will provide you direct results. It will function as a black box. You simply pass an input sentence to it and it returns you a tagged output.

 Reply
Balachandar says:
<u>April 8, 2013 at 1:21 am</u>
Hi ,

Thanks a lot for the tutorial. Is there a function to extract the tags from the tagged sentence ? e.g from This/DT is/VBZ a/DT great/JJ place/NN , I have to extract the POS tags alone.

 Reply
   <u>**Raj Vardhan**</u> **says:**
   <u>June 7, 2013 at 12:05 pm</u>
   Hi Balachandar,

   Yes, you can do easily do this by making a simple function where you will be doing some string processing on the tagged output. One way would be to tokenize the tagged sentence using 'space character' as delimiter. This way you would get tokens such as {This/DT , is/VBZ,….,place/NN}. Then using '/' as delimiter you can get tokens each for the word and its corresponding tag.

   Hope it helps.

   Reply
Hamdi says:
<u>May 29, 2013 at 10:56 am</u>
Thanx alot Raj.

 Reply
priyanka says:
<u>July 10, 2013 at 12:55 pm</u>
hi,

I'm new in the field of natural language processing and is doing a project in which i will be using a Hindi POS tagger for input text and then i will be assigning a POS tag to unknown words (MWE). My query is how my system will pick the value for the unknown words from the database that i have created and how will it show that result in the given input text.

Reply

Raj Vardhan says:
July 12, 2013 at 6:45 am
Hi Priyanka

Are you using any open-source Hindi POS tagger or developing one yourself? Also, to be clear on your question first, do you want to assign POS tags to Multiword Expressions (MWE) in Hindi?

Reply

priyanka says:
July 12, 2013 at 8:27 am
yes, i'm using open-source Hindi POS tagger and with that i want to assign POS tags to MWE(idioms) in Hindi.
For example: if a user input a Hindi text having an idiom then the POS tagger will tag each word individually whereas i want that idiom part to be tagged as a single unit.
for this i have assigned POS tags to the list of idioms and want my system to pick these POS tags for the idioms in the text.

I would like to know how to implement this task ….plz help me out.

Raj Vardhan says:
July 13, 2013 at 3:55 pm
If you have 'n' number of idioms in your database, you can simply scan the input text to find the occurrence of any of those idioms in it. A pattern matching or template matching algorithm can be used for this.

Have you started off with any such approach? It will be easier to help you out if you let me know where exactly are you stuck.

MnT says:
July 13, 2013 at 5:22 am
I would like to know how POS tagger can be used for implementing human interactions in meetings?

Reply

Raj Vardhan says:
July 13, 2013 at 3:58 pm
Hi
Can you please elaborate on what you're trying to achieve with POS tagging? If you are trying to model human interactions that occur in meetings, where do you need POS tagging to be done?

Reply

Venkata Ramana says:
July 31, 2013 at 12:53 pm
Hi Raj,

Can we do a spell check using teh Stanford NLP libraries? If yes, how?

Reply

Raj Vardhan says:
July 31, 2013 at 3:40 pm
Hi Venkata,

There is no standard provision for spell check in Stanford NLP libraries, as far as I know.

You can give a look at this article by Peter Norvig on writing your own spell checker.

Hope it helps.

Reply

vibha says:
October 22, 2013 at 6:54 am
hello sir,
can we use eclipse and pos tagger for the implementation of so-pmi algorithm.

Reply

vibha says:
October 22, 2013 at 6:59 am
Ge sir,
can we use eclipse and pos tagger in the implementation of so-pmi algorithm.or can we extracts those words that matchs with pos pattern provided by turney

Reply

Raj Vardhan says:
November 14, 2013 at 2:39 pm
Hi Vibha,

Can you tell me what do you intend to do with the POS tagger in so-pmi algorithm. I am not quite familiar with the algorithm, so you'll have to frame a problem statement. I can then see whether POS tagger can be put to use or not.

Reply

vibha says:
November 16, 2013 at 7:17 am
thank u sir for reply.
Sir,firstly I can't use Eclipse. I had download "eclipse-jee-helios-win32-x86_64" and I follow your steps defined in above but it can't work.I am using stanford-postagger tagger. Also which type of project we select in eclipse as there is lot of options.whenever i try this code ,eclipse hangs out. I want to store the output of the pos tagger in separate file which can be use for further implementation .

Raj Vardhan says:
November 17, 2013 at 7:18 am
What is your system-configuration?
Make sure you have downloaded the corresponding eclipse version. You can download this version if you aren't sure which one to pick.
After installation, create a new "Java Project" and try to run a simple hello-world java program. Let me know if that's working.
We'll then try to add the library and begin coding.

Regards

priyanka says:
October 25, 2013 at 10:42 am
hey i want to use tagger in dynamic web project . When i m creating simple java project im able to use tagger but in web project im getting no classdeffound exception .. can you help me to solve it

Reply

Raj Vardhan says:
November 14, 2013 at 12:53 pm
Hi
Sorry for the late reply. Have you been able to resolve the issue yet?

Reply

brandon says:
October 29, 2013 at 5:48 pm
Hi Raj,

may i ask you just one question?
in that code, every sentence in the text file should be in each line.
But, my sentences are very long. I mean some sentence starts from line 1 and finishes at line 3.

for example, in the text file "input.txt"
line 1 : hello my name is brandon …………………………………………………………………………..and i was l
line 2 : ooking for something……………………………………………………………………………….toda
line 3 : y, i am going to visit school.

in this case, what can i do?
because i think it will tag terms like "ooking" and "toda" in the wrong way.

i'm looking forward to your reply.
thank you,Raj.

Reply

brandon says:
October 30, 2013 at 5:58 am
Hi Raj,

my apology…
it works well. but i don't know how it worked.

it tagged POS correctly on that problem that i mentioned above comment.

thank you very much.

Reply
Raj Vardhan says:
November 14, 2013 at 2:48 pm
Hi Brandon,

Consider a text file with the lines of text that you mentioned above. Try to think what would happen if you reduce the font by say 4 points. Do you think that line2 will still end with 'toda' and line 3 would still begin with 'y, i am...' ?

The text is simply overflowing to the next line, and there is no 'return character' between letters of the same word.

A File reader will not read a file based on how the text is overflowing, as the text contained is independent of that (font, size, etc)

I hope that helps.

Regards

Reply
sonam.gupta11 says:
November 1, 2013 at 8:48 pm
Hello Raj,

I am very new in the field of NLP. I want to implement the Brill POS tagger algorithm. Can you guide me a bit.

Reply
Raj Vardhan says:
November 14, 2013 at 2:42 pm
Hi Sonam,

Brill POS tagger algorithm uses supervised machine learning algorithm to assign tags based on error-driven transformations. Before I can suggest any methods to implement it, can you tell me your background knowledge in the concerned areas (machine learning, theory-of-computation, A.I, NLP)?

Regards

Reply
Jesudas says:
January 19, 2014 at 7:09 pm
Hi I want to code on my own pos tagger in c++ but I m new to this field. Can u guide me how can I start from scratch.
Thanks!
Jesudas

Reply
Dough ng says:
January 25, 2014 at 7:41 am
Wow! thank you very much..you simplified the whole process for me.
If I wish to annotate the output..say Ravi New Delhi

How can I do that?
My objective is to find and annotate persons and locations from a given text.

Reply
Raj Vardhan says:
January 26, 2014 at 12:58 pm
Hi

Annotating names and persons in a given text is done through what is called "Named-entity recognition".
Have a look here and here

Hope it helps

Reply
Dough ng says:

February 2, 2014 at 11:19 pm
Thank you for the links Raj,
I read from the link that Stanford NER model trained for CoNLL and MUC, which recognizes Person, location, date etc.

But how can I retrain it to tag entities like phones. For example, if I have a sentence like this:
the new Iphone 5s seems bigger than Nexus S

then I get this result: Theiphone 5sseems bigger thanNexus S
Do you any example for a beginner?
Thanks

 Reply
Dough ng says:
February 2, 2014 at 11:24 pm
I meant a result like this…(the post hid the xml format I wrote before)
The iphone 5s/PHONE seems bigger than Nexus S/PHONE

 Reply
aNq zHrA (@aNq_zHrA) says:
February 26, 2014 at 11:49 am
hi raj,
please help me out with this problem when ever i run the project after completing the step i get this in console Usage: java
GrammaticalStructure [options]* [-sentFile | -treeFile | -conllxFile file] [-testGraph]
options: -basic, -collapsed, -CCprocessed [the default], -collapsedTree, -parseTree, -test, -parserFile file, -conllx, -keepPunct,
-altprinter -altreader -altreaderfile
nsubj(died-2, Sam-1)
root(ROOT-0, died-2)
tmod(died-2, today-3)
rather than tagged words
kindly help me out

 Reply
　Raj Vardhan says:
　March 4, 2014 at 5:58 pm
　Hi,
　This is the kind of output you get when you use Stanford Dependencies.
　The relations you have got such as nsubj(died-2, Sam-1) represent a dependency tree.
　Are you looking for POS tagging or getting such dependencies?
　Let me know.

　 Reply
　　armaan jain says:
　　March 22, 2015 at 11:19 am
　　hey how do i get the dependencies if i want them ?

Payal says:
March 10, 2014 at 12:23 pm
hello Raj,
firstly thank you for this tutorial. It helped me a lot. But i have a question. I am not able to find the .props file inside my POS
tagger folder. Please help me with it. I have the .tagger file.

 Reply
Pingback: Finding date range in natural language text | Raj Vardhan

anupamjamatia says:
June 19, 2014 at 8:45 am
I am working on Hindi Language and I have taken Standford POS tagger as my base. I would like to know how to create .props
file for specific language to train? In the Standford model have chinese, english , german language but how to create model and
its props file?

 Reply
　Raj Vardhan says:
　June 20, 2014 at 6:23 am
　I would recommend asking this question in the "java-nlp-user@lists.stanford.edu" mailing list. The creators of this tool might

Reply
Dwijen says:
June 19, 2014 at 9:29 am
Hello Raj,
Your tutorial helped me a lot. Now i am facing a problem to train stanford POS tagger by my tagger file. I don't have .props file for this tagger. I want to train by the .tagger file and tag my text. Please help me.

Reply
poornima mehta says:
June 27, 2014 at 11:48 am
I am getting this error….please help

Exception in thread "main" edu.stanford.nlp.io.RuntimeIOException: Unrecoverable error while loading a tagger model
at edu.stanford.nlp.tagger.maxent.MaxentTagger.readModelAndInit(MaxentTagger.java:763)
at edu.stanford.nlp.tagger.maxent.MaxentTagger.(MaxentTagger.java:294)
at edu.stanford.nlp.tagger.maxent.MaxentTagger.(MaxentTagger.java:259)
at tagText.main(tagText.java:9)
Caused by: java.io.IOException: Unable to resolve "taggers/left3words-wsj-0-18.tagger" as either class path, filename or URL
at edu.stanford.nlp.io.IOUtils.getInputStreamFromURLOrClasspathOrFileSystem(IOUtils.java:446)
at edu.stanford.nlp.
tagger.maxent.MaxentTagger.readModelAndInit(MaxentTagger.java:758)
… 3 more

Reply
aeesha says:
September 1, 2014 at 8:07 am
i give this input
input : i can man the controls of this machine .
i want that output
output : i/FW can/MD man/VB the/DT controls/NNS of/IN this/DT machine/NN.
but my output is this , which i dont know what is it ?
i want the tragger file from which i get the required output , plzzzz send me the required tragger file .
Freq Norm Add1 Dir1 Dir2 GT SGT
——— ——— ——— ——— ——— ——— ———
1000 0.11837121 0.08041452 0.08041450 0.08041450 0.11673059 0.11798246
500 0.05918561 0.04024743 0.04024741 0.04024741 0.05836529 0.05893220
333 0.03941761 0.02683162 0.02683160 0.02683160 0.03887128 0.03920949
250 0.02959280 0.02016388 0.02016386 0.02016386 0.02918265 0.02940724
——— ——— ——— ——— ——— ——— ———
1 0.00011837 0.00016067 0.00016065 0.00016065 0.00004659 0.00004713
——— ——— ——— ——— ——— ——— ———
UNK 0.00000000 0.00008033 0.00008031 0.00008031 0.00007885 0.00007885
——— ——— ——— ——— ——— ——— ———
RESERVE 0.00000000 0.16074871 0.00000000 0.00000000 0.15778883 0.15778883
——— ——— ——— ——— ——— ——— ———
Total 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000

Reply
Reena says:
September 5, 2014 at 4:38 pm
Hi,
I am trying to implement the example in a dynamic web project using servlet.
I have an issue: java.io.FileNotFoundException: \.\models\left3words-wsj-0-18.tagger (The system cannot find the path specified).
Please help me.

Reply
ghani says:
September 26, 2014 at 3:52 pm
great work thinks Raj, you lot of student and research in NLP

Reply
priya says:
October 4, 2014 at 12:49 am
I have followed the steps as per your tutorial. Still I'm facing the problem.

Exception in thread "main" java.lang.RuntimeException: 'model' parameter must be specified
at edu.stanford.nlp.tagger.maxent.TaggerConfig.setProperties(TaggerConfig.java:196)
at edu.stanford.nlp.tagger.maxent.TaggerConfig.(TaggerConfig.java:155)
at edu.stanford.nlp.tagger.maxent.TaggerConfig.(TaggerConfig.java:128)
at edu.stanford.nlp.tagger.maxent.MaxentTagger.main(MaxentTagger.java:1837)

I could not understand why im getting this error message. Please help me out?

Thanks

 Reply
Nirali says:
December 3, 2014 at 11:11 am
Hi.. I need to do stemming and pos tagging for wsd in java. is there any pos tagger or stemmer for hindi in java?

 Reply
Sulabh says:
December 4, 2014 at 7:17 pm
Hi,

I am going to build a POS tagger for Sanskrit language.
I have time limit around 3-4 months for this.
So, is it better to create a new one or to use one that exists for other language?

What should we do to use one language's POS tagger for other language?
(However, I will have tagged corpus for training my tagger.)

Thanks

 Reply
uma says:
January 7, 2015 at 6:27 am
thankyou,your tutorial helped me a lot,tha examole specifies only one sentence what to do if we want to include hundreds of sentences,do we need to connect a database please xplain me the procedure

 Reply
atikah says:
January 8, 2015 at 7:29 pm
hai raj,

i need the answer in xml format. can you help me ?

 Reply
Musheer Ahamed says:
February 13, 2015 at 7:14 pm
thank you very much

 Reply
Akash D Souza says:
February 18, 2015 at 11:45 pm
How can I print parse tree of the senence using stanford parser?

 Reply
manisha says:
March 5, 2015 at 9:09 pm
i have followed the steps told by u but i am getting the following error
Exception in thread "main" java.lang.UnsupportedClassVersionError: edu/stanford/nlp/tagger/maxent/MaxentTagger :
Unsupported major.minor version 52.0
at java.lang.ClassLoader.defineClass1(Native Method)
at java.lang.ClassLoader.defineClass(Unknown Source)
at java.security.SecureClassLoader.defineClass(Unknown Source)
at java.net.URLClassLoader.defineClass(Unknown Source)
at java.net.URLClassLoader.access$000(Unknown Source)
at java.net.URLClassLoader$1.run(Unknown Source)
at java.security.AccessController.doPrivileged(Native Method)
at java.net.URLClassLoader.findClass(Unknown Source)

at java.lang.ClassLoader.loadClass(Unknown Source)
at sun.misc.Launcher$AppClassLoader.loadClass(Unknown Source)
at java.lang.ClassLoader.loadClass(Unknown Source)
at java.lang.ClassLoader.loadClassInternal(Unknown Source)
at tagText.main(tagText.java:10)
plz help ????
thanks in advance

Reply
Sanju says:
March 17, 2015 at 6:41 am
hello Raj
I need some help in NLP for preprocessing of data but I m facing some problems.

Reply
sanju says:
March 26, 2015 at 4:33 pm
I have followed the steps as per your tutorial. Still I'm facing the problem.

Exception in thread "main" java.lang.RuntimeException: 'model' parameter must be specified
at edu.stanford.nlp.tagger.maxent.TaggerConfig.setProperties(TaggerConfig.java:196)
at edu.stanford.nlp.tagger.maxent.TaggerConfig.(TaggerConfig.java:155)
at edu.stanford.nlp.tagger.maxent.TaggerConfig.(TaggerConfig.java:128)
at edu.stanford.nlp.tagger.maxent.MaxentTagger.main(MaxentTagger.java:1837)

I could not understand why im getting this error message. Please help me out?

Thanks

Reply
sanju says:
March 27, 2015 at 4:57 am
Exception in thread "main" java.lang.RuntimeException: 'model' parameter must be specified
at edu.stanford.nlp.tagger.maxent.TaggerConfig.setProperties(TaggerConfig.java:196)
at edu.stanford.nlp.tagger.maxent.TaggerConfig.(TaggerConfig.java:155)
at edu.stanford.nlp.tagger.maxent.TaggerConfig.(TaggerConfig.java:128)
at edu.stanford.nlp.tagger.maxent.MaxentTagger.main(MaxentTagger.java:1837)

I could not understand why im getting this error message. Please help me out?

Thanks

Reply
waqas ahmad says:
June 12, 2015 at 2:19 pm
Hi rajvardhan!!!
you seemed to be the hope for me.
I'm doing android project on Homeopathy. User enters few symptoms form android then it has to match with the data in DB. It contains the column in which the data is given in paragraph form.The data in column contains almost One Lakh entries. I have to match all the all data..and i have to use POS TAGGER using webservices.
Kindly help me in this regard at earliest. As i have to give Final Demo after a Week.

Reply

Blog at WordPress.com.

The Enterprise Theme.