OpenNLP for Text Preprocessing

February 16, 2024

1 Assignment 1 - OpenNLP

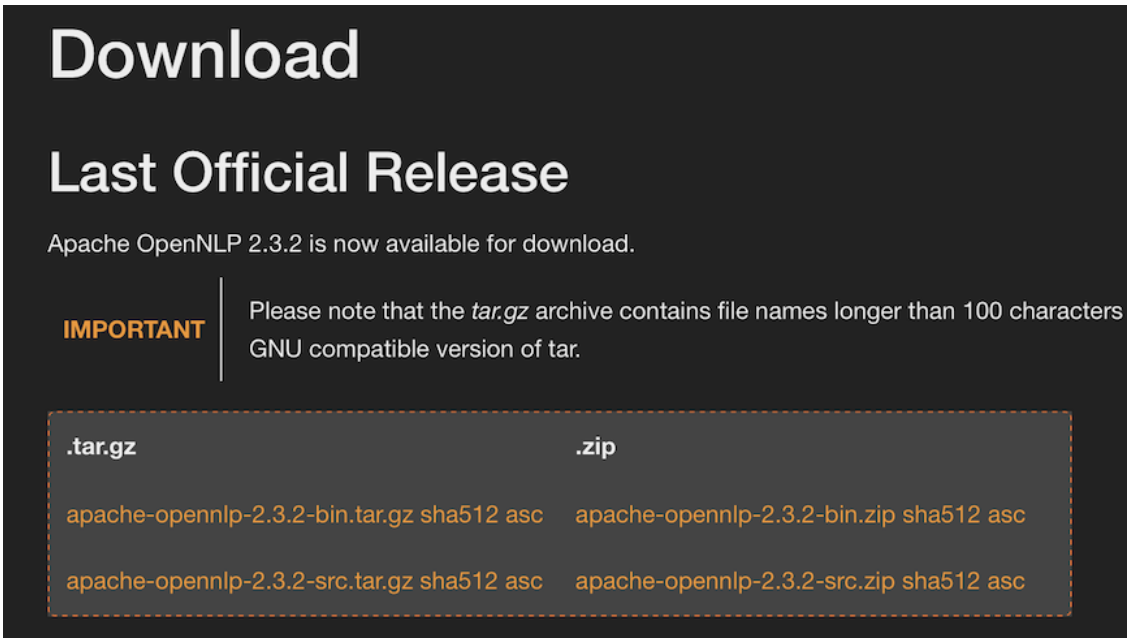
This assignment uses the Apache OpenNLP library to process natural language text, including tasks such as detecting sentences, tokenizing words, performing Part-of-Speech (POS) tagging, and identifying named entities from a provided news article.

Matthew Acs

1.1 I | Configure Apache OpenNLP, Java, and Eclipse

Step one involves configuring Apache OpenNLP, Java, and Eclipse on my Mac. The screenshots below walk through the steps that I took to setup the environment on my machine.

I downloaded Apache OpenNLP from the Apache archive, which I accessed from the Apache OpenNLP main site.



I downloaded the bin and src files of the latest version of OpenNLP.

Index of /dist/opennlp/opennlp-2.3.2

Name	Last modified	Size	Description
Parent Directory		-	
apache-opennlp-2.3.2-bin.tar.gz	2024-01-31 14:31	13M	
apache-opennlp-2.3.2-bin.tar.gz.asc	2024-01-31 14:31	833	
apache-opennlp-2.3.2-bin.tar.gz.sha512	2024-01-31 14:31	162	
apache-opennlp-2.3.2-bin.zip	2024-01-31 14:31	16M	
apache-opennlp-2.3.2-bin.zip.asc	2024-01-31 14:31	833	
apache-opennlp-2.3.2-bin.zip.sha512	2024-01-31 14:31	159	
apache-opennlp-2.3.2-src.tar.gz	2024-01-31 14:31	2.4M	
apache-opennlp-2.3.2-src.tar.gz.asc	2024-01-31 14:31	833	
apache-opennlp-2.3.2-src.tar.gz.sha512	2024-01-31 14:31	162	
apache-opennlp-2.3.2-src.zip	2024-01-31 14:31	3.7M	
apache-opennlp-2.3.2-src.zip.asc	2024-01-31 14:31	833	
apache-opennlp-2.3.2-src.zip.sha512	2024-01-31 14:31	159	

I also don't normally use Java, so I needed to get Java on my Mac M1 via the Java website.

Java Download Developer Resources Help

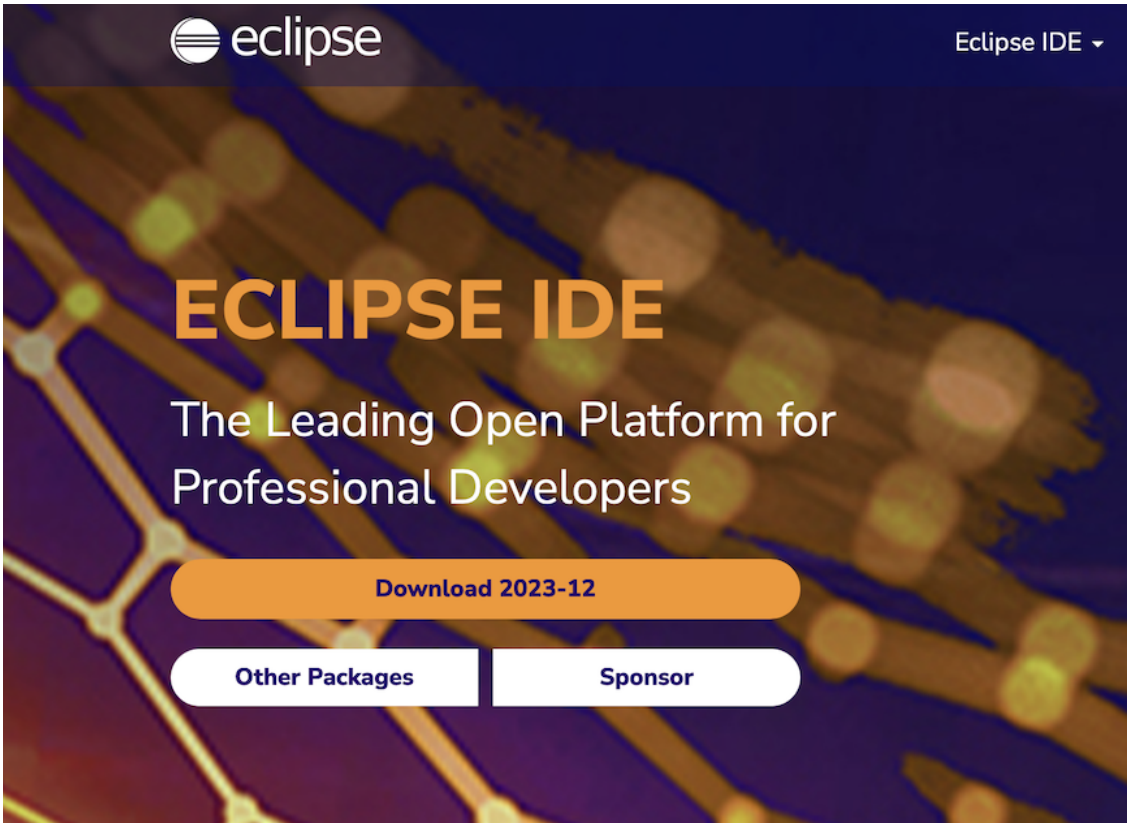
Manual update required for some Java 8 users on macOS

Get Java for desktop applications

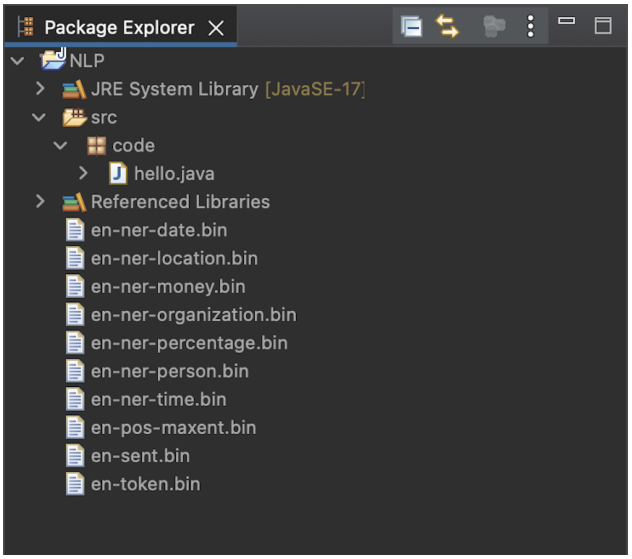
Download Java

What is Java? | Uninstall help

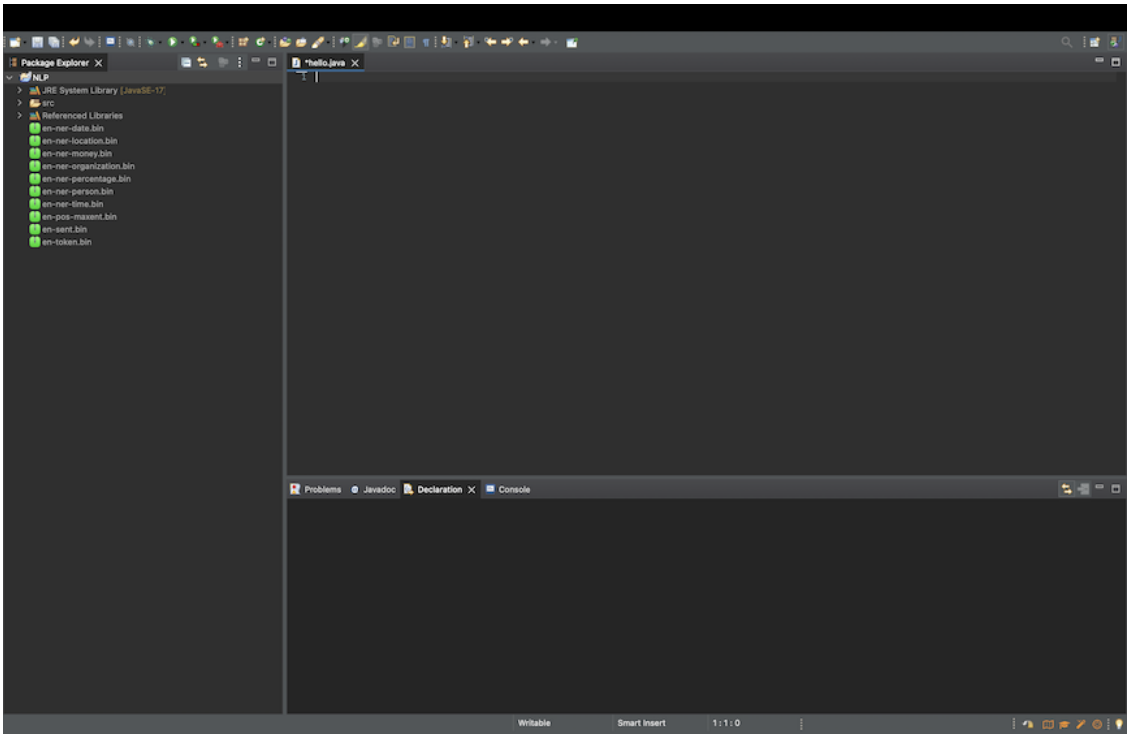
I also don't have a Java IDE, so I downloaded Eclipse for Java.



I set up the Eclipse directory as shown in the screenshot. The src folder contains my source code and the referenced libraries folder contains the OpenNLP library files. I included English language models for named entity recognition, POS tagging, sentence identification, and tokenization in the root of the project.



The screenshot below shows the IDE with Java and OpenNLP installed. No code has been written yet and the console has not output.



The code below shows the library imports and basic code setup for the assignment.

[ ]:
```
package code;

import opennlp.tools.sentdetect.SentenceDetectorME;
import opennlp.tools.sentdetect.SentenceModel;
import opennlp.tools.tokenize.TokenizerME;
import opennlp.tools.tokenize.TokenizerModel;
import opennlp.tools.util.Span;
import opennlp.tools.postag.POSModel;
import opennlp.tools.postag.POSTaggerME;
import opennlp.tools.namefind.NameFinderME;
import opennlp.tools.namefind.TokenNameFinderModel;
import java.io.FileInputStream;
import java.io.IOException;

public class hello {

    public static void main(String[] args) {

        // Parts 2-5 go here

    }

}
```

The following steps will explore each assignment task in closer detail by looking at code fragments along with screenshots of the output that pertains to that task. These fragments are not standalone and need to be combined with more code (i.e loops, print statetments, etc) to compile, which are included in the final code given at the end.

1.2 II | Detect sentences in the given news article

The first task is to detect sentences in the given news article. This involves loading the news article and using the sentence detection model.

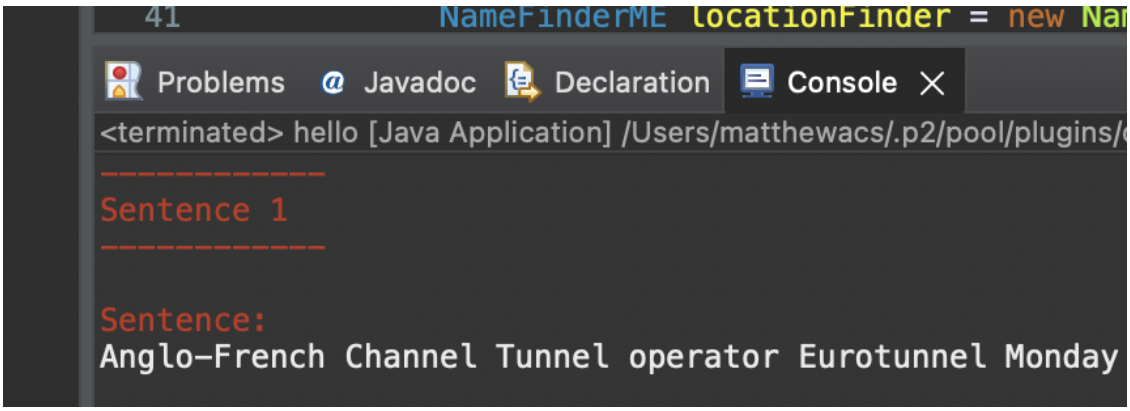
To load the news article, I assigned the entire text to a string called News.

```
[ ]: String News = "Anglo-French Channel Tunnel operator Eurotunnel Monday announced␣
 ↪a deal giving its creditor banks 45.5 percent of the company in return for␣
 ↪wiping out one billion pounds ($1.56 billion) of its debt. The long-awaited␣
 ↪restructuring brings to an end months of wrangling between Eurotunnel and the␣
 ↪225 banks to which it owes nearly nine billion pounds ($14.1 billion). The␣
 ↪deal, announced simultaneously in Paris and London, brings the company back␣
 ↪from the brink of insolvency but leaves shareholders owning only 54.5 percent␣
 ↪of the company. \"The restructuring plan provides Eurotunnel with the␣
 ↪medium-term financial stability to allow it to consolidate its substantial␣
 ↪commercial achievements to date and to develop its operations,\" Eurotunnel␣
 ↪co- chairman Alastair Morton said. The firm was now making a profit before␣
 ↪interest, he added. Although shareholders will see their interests diluted,␣
 ↪they were offered the prospect of a brighter future after months of␣
 ↪uncertainty while Eurotunnel wrestled to reduce crippling interest payments␣
 ↪negotiated during the tunnel's construction. Eurotunnel, which has taken␣
 ↪around half the cross-Channel market from the European ferry companies, said a␣
 ↪strong operating performance could allow it to pay its first dividend within␣
 ↪the next 10 years. French co-chairman Patrick Ponsolle said shareholders would␣
 ↪have to be patient before they could reap the benefits of the company's␣
 ↪success. He called the debt restructuring plan \"an acceptable compromise\"␣
 ↪for holders of Eurotunnel shares. The company said there was still␣
 ↪considerable work to be done to finalise and agree on the details of the plan␣
 ↪before it can be submitted to shareholders and the full 225 bank syndicate for␣
 ↪approval, probably early in 1997. Monday\'s announcement followed two weeks of␣
 ↪highly secretive negotiations between Eurotunnel and its six leading banks.␣
 ↪This was extended to the 24 \"instructing banks\" at a meeting late last week␣
 ↪in London. Eurotunnel said the debt-for-equity swap would be at 130 pence, or␣
 ↪10.40 francs, per share. That is considerably below the level of around 160␣
 ↪pence widely reported before announcement of the deal, and will reduce␣
 ↪outstanding debt of 8.7 billion pounds ($13.6 billion) by 1.0 billion ($1.56␣
 ↪billion). The company said a further 3.7 billion pounds ($5.8 billion) of debt␣
 ↪would be converted into new financial instruments and existing shareholders␣
 ↪would be able to participate in this issue. If they choose not to take up free␣
 ↪warrants entitling them to subscribe to this, Eurotunnel said shareholders\'␣
 ↪interests may be reduced further to just over 39 percent of the company by the␣
 ↪end of December 2003. Eurotunnel's shares, which were suspended last week at␣
 ↪113.5 pence ahead of Monday\'s announcement, should resume trading on Tuesday,␣
 ↪the company said.";
```

To detect sentences in the article, I initialize a sentence detection model using the code below. I then use the pretrained model to detect sentences from the news article, which I store in an array called sentence.

[ ]:
```
// Load the sentence detection model
SentenceModel sentenceModel = new SentenceModel(new FileInputStream("en-sent.
 ↪bin"));
SentenceDetectorME sentenceDetector = new SentenceDetectorME(sentenceModel);

// Detect sentences
String[] sentences = sentenceDetector.sentDetect(News);
```

The screenshot below shows the first sentence detected as output on the console.



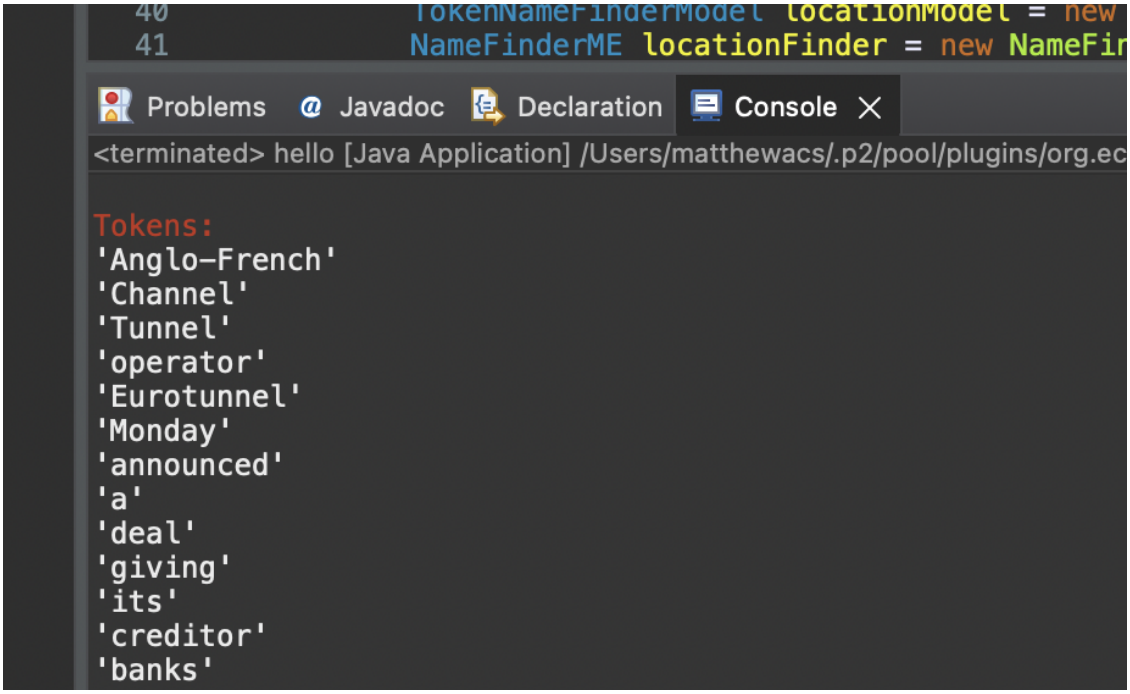
1.3 III | Tokenize each sentence into words

The second task is to tokenize each sentence in the given news article. This involves taking each sentence from the sentence detection and tokenizing it using the tokenizer model.

To tokenize a sentence, I initialize a tokenizer model using the code below. I then use the pretrained model to tokenize each sentence from the sentence detector, which I store in an array called tokens.

[ ]:
```
// Load the tokenizer model
TokenizerModel tokenizerModel = new TokenizerModel(new FileInputStream("en-token.
 ↪bin"));
TokenizerME tokenizer = new TokenizerME(tokenizerModel);

// Tokenize each sentence
String[] tokens = tokenizer.tokenize(sentence);

// Output tokens
System.out.println("\u001B[31mTokens:\u001B[0m");
for (String token : tokens) {
    System.out.println("'" + token + "'");
}
```

The screenshot below shows the tokens of the first sentence as output on the console.



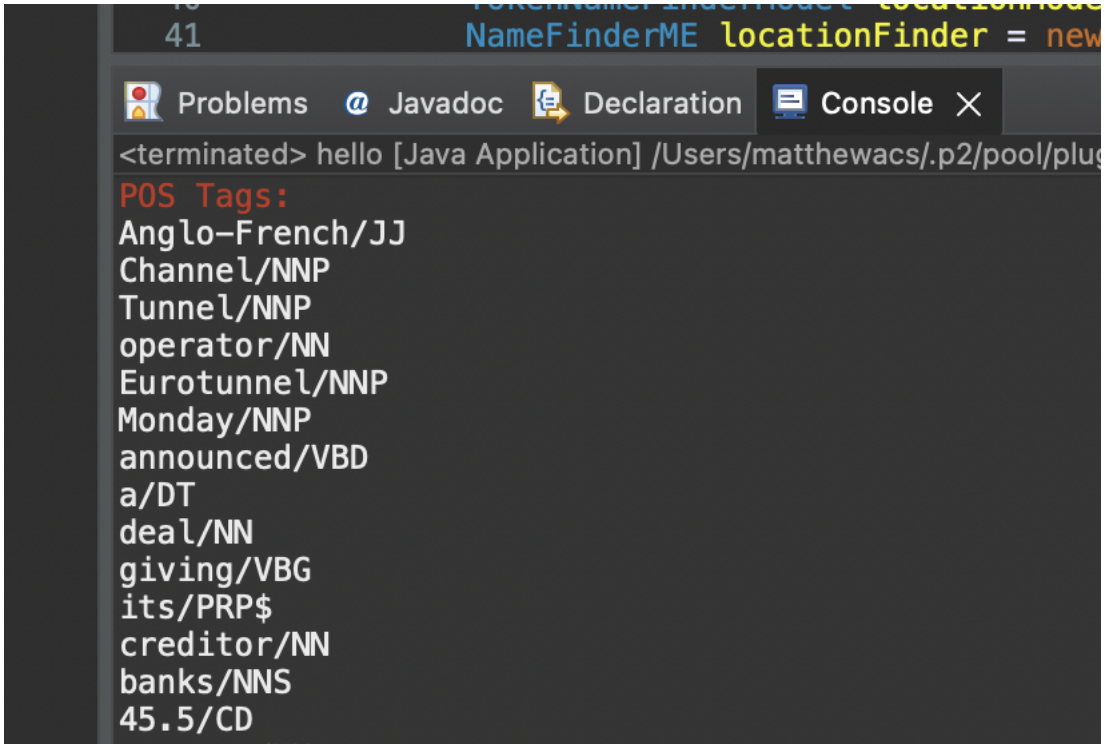
1.4 IV | Perform Part-of-Speech (POS) on each sentence

The third task is to perform POS tagging on each sentence in the given news article. This involves taking each sentence from the sentence detection and tagging it using the POS tagger model.

To POS tag a sentence, I initialize a POS tagger model using the code below. I then use the pretrained model to POS tag each sentence from the sentence detector, which I store in an array called tags.

[ ]:
```
// Load the POS tagging model
POSModel posModel = new POSModel(new FileInputStream("en-pos-maxent.bin"));
POSTaggerME posTagger = new POSTaggerME(posModel);

// Perform POS tagging
String[] tags = posTagger.tag(tokens);

// Output POS tags
System.out.println("\n\u001B[31mPOS Tags:\u001B[0m");
for (int j = 0; j < tokens.length; j++) {
    System.out.println(tokens[j] + "/" + tags[j]);
}
```

The screenshot below shows the POS tags of the first sentence as output on the console.



1.5 V | Find name entities including person's name entities and locations

The final task is to perform named entity recognition on each sentence in the given news article. This involves taking each sentence from the sentence detection and recognizing various named entities using the finder models.

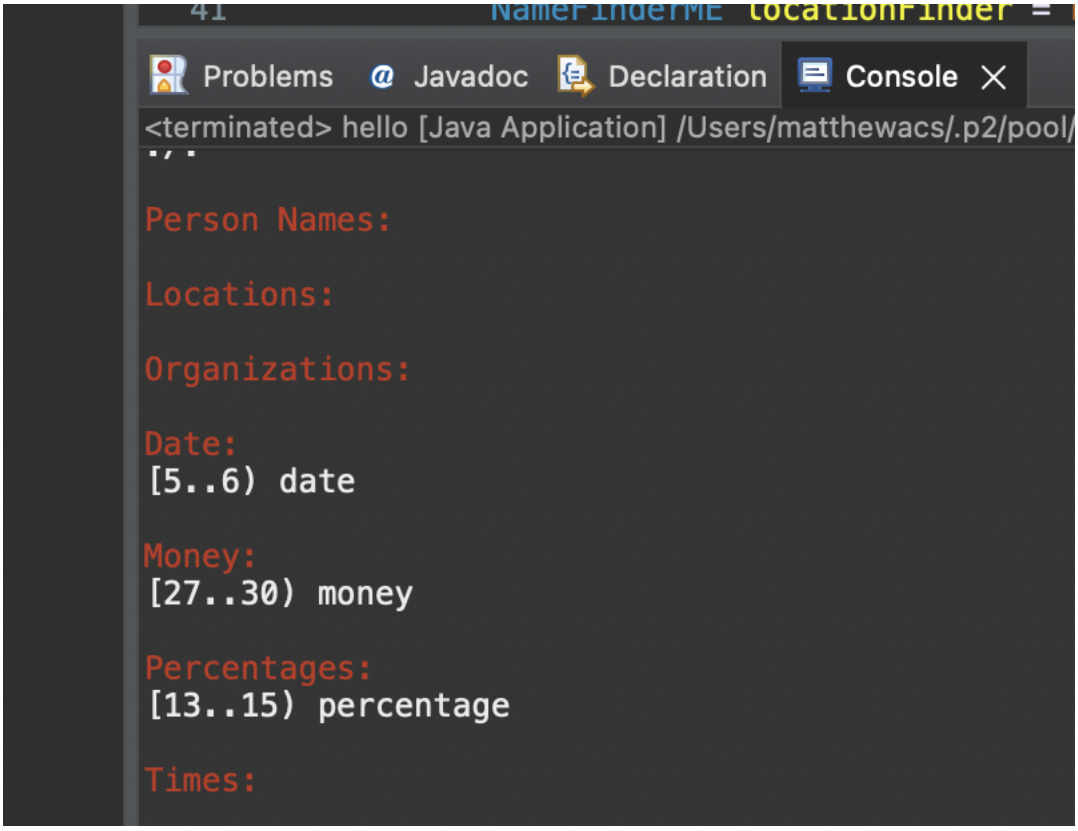
To perform named entity recognition on each sentence, I initialize the finder models using the code below. I then use the pretrained models to perform named entity recognition on each sentence from the sentence detector, which I store in an array called names, locations, etc. I used the name, location, organization, date, money, percentage, and time finder models.

[ ]:
```
// Load the name finder model for person names
TokenNameFinderModel personModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-person.bin"));
NameFinderME personFinder = new NameFinderME(personModel);

// Load the name finder model for locations
TokenNameFinderModel locationModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-location.bin"));
NameFinderME locationFinder = new NameFinderME(locationModel);

// Load the name finder model for organizations
TokenNameFinderModel organizationModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-organization.bin"));
```

```java
NameFinderME organizationFinder = new NameFinderME(organizationModel);

// Load the name finder model for dates
TokenNameFinderModel dateModel = new TokenNameFinderModel(new␣
 ↪FileInputStream("en-ner-date.bin"));
NameFinderME dateFinder = new NameFinderME(dateModel);

// Load the name finder model for money
TokenNameFinderModel moneyModel = new TokenNameFinderModel(new␣
 ↪FileInputStream("en-ner-money.bin"));
NameFinderME moneyFinder = new NameFinderME(moneyModel);

// Load the name finder model for percentages
TokenNameFinderModel percentageModel = new TokenNameFinderModel(new␣
 ↪FileInputStream("en-ner-percentage.bin"));
NameFinderME percentageFinder = new NameFinderME(percentageModel);

// Load the name finder model for time
TokenNameFinderModel timeModel = new TokenNameFinderModel(new␣
 ↪FileInputStream("en-ner-time.bin"));
NameFinderME timeFinder = new NameFinderME(timeModel);

// Perform named entity recognition for persons
Span[] personNames = personFinder.find(tokens);

// Output person names
System.out.println("\n\u001B[31mPerson Names:\u001B[0m");
for (Span name : personNames) {
    System.out.println(name);
}

// Perform named entity recognition for locations
Span[] locations = locationFinder.find(tokens);

// Output locations
System.out.println("\n\u001B[31mLocations:\u001B[0m");
for (Span location : locations) {
    System.out.println(location);
}

// Perform named entity recognition for organizations
Span[] organizations = organizationFinder.find(tokens);

// Output organizations
System.out.println("\n\u001B[31mOrganizations:\u001B[0m");
for (Span organization : organizations) {
    System.out.println(organization);
```

```java
}

// Perform named entity recognition for dates
Span[] dates = dateFinder.find(tokens);

// Output dates
System.out.println("\n\u001B[31mDate:\u001B[0m");
for (Span date : dates) {
    System.out.println(date);
}

// Perform named entity recognition for money
Span[] money = moneyFinder.find(tokens);

// Output money
System.out.println("\n\u001B[31mMoney:\u001B[0m");
for (Span dollar : money) {
    System.out.println(dollar);
}

// Perform named entity recognition for percentages
Span[] percentages = percentageFinder.find(tokens);

// Output percentages
System.out.println("\n\u001B[31mPercentages:\u001B[0m");
for (Span percent : percentages) {
    System.out.println(percent);
}

// Perform named entity recognition for time
Span[] times = timeFinder.find(tokens);

// Output time
System.out.println("\n\u001B[31mTimes:\u001B[0m");
for (Span time : times) {
    System.out.println(time);
}
```

The screenshot below shows the named entities of the first sentence as output on the console.



1.6 VI | Full Source Code

This section contains the full executable source code.

[ ]:
```
package code;

import opennlp.tools.sentdetect.SentenceDetectorME;
import opennlp.tools.sentdetect.SentenceModel;
import opennlp.tools.tokenize.TokenizerME;
import opennlp.tools.tokenize.TokenizerModel;
import opennlp.tools.util.Span;
import opennlp.tools.postag.POSModel;
import opennlp.tools.postag.POSTaggerME;
import opennlp.tools.namefind.NameFinderME;
import opennlp.tools.namefind.TokenNameFinderModel;
import java.io.FileInputStream;
import java.io.IOException;

public class hello {

    public static void main(String[] args) {
```

```
        String News = "Anglo-French Channel Tunnel operator Eurotunnel Monday␣
↪announced a deal giving its creditor banks 45.5 percent of the company in␣
↪return for wiping out one billion pounds ($1.56 billion) of its debt. The␣
↪long-awaited restructuring brings to an end months of wrangling between␣
↪Eurotunnel and the 225 banks to which it owes nearly nine billion pounds ($14.
↪1 billion). The deal, announced simultaneously in Paris and London, brings the␣
↪company back from the brink of insolvency but leaves shareholders owning only␣
↪54.5 percent of the company. \"The restructuring plan provides Eurotunnel with␣
↪the medium-term financial stability to allow it to consolidate its substantial␣
↪commercial achievements to date and to develop its operations,\" Eurotunnel␣
↪co- chairman Alastair Morton said. The firm was now making a profit before␣
↪interest, he added. Although shareholders will see their interests diluted,␣
↪they were offered the prospect of a brighter future after months of␣
↪uncertainty while Eurotunnel wrestled to reduce crippling interest payments␣
↪negotiated during the tunnel's construction. Eurotunnel, which has taken␣
↪around half the cross-Channel market from the European ferry companies, said a␣
↪strong operating performance could allow it to pay its first dividend within␣
↪the next 10 years. French co-chairman Patrick Ponsolle said shareholders would␣
↪have to be patient before they could reap the benefits of the company's␣
↪success. He called the debt restructuring plan \"an acceptable compromise\"␣
↪for holders of Eurotunnel shares. The company said there was still␣
↪considerable work to be done to finalise and agree on the details of the plan␣
↪before it can be submitted to shareholders and the full 225 bank syndicate for␣
↪approval, probably early in 1997. Monday\'s announcement followed two weeks of␣
↪highly secretive negotiations between Eurotunnel and its six leading banks.␣
↪This was extended to the 24 \"instructing banks\" at a meeting late last week␣
↪in London. Eurotunnel said the debt-for-equity swap would be at 130 pence, or␣
↪10.40 francs, per share. That is considerably below the level of around 160␣
↪pence widely reported before announcement of the deal, and will reduce␣
↪outstanding debt of 8.7 billion pounds ($13.6 billion) by 1.0 billion ($1.56␣
↪billion). The company said a further 3.7 billion pounds ($5.8 billion) of debt␣
↪would be converted into new financial instruments and existing shareholders␣
↪would be able to participate in this issue. If they choose not to take up free␣
↪warrants entitling them to subscribe to this, Eurotunnel said shareholders\'␣
↪interests may be reduced further to just over 39 percent of the company by the␣
↪end of December 2003. Eurotunnel's shares, which were suspended last week at␣
↪113.5 pence ahead of Monday\'s announcement, should resume trading on Tuesday,␣
↪the company said.";

        try {

            // Load the sentence detection model
            SentenceModel sentenceModel = new SentenceModel(new␣
↪FileInputStream("en-sent.bin"));
            SentenceDetectorME sentenceDetector = new␣
↪SentenceDetectorME(sentenceModel);
```

```
            // Load the tokenizer model
            TokenizerModel tokenizerModel = new TokenizerModel(new␣
↪FileInputStream("en-token.bin"));
            TokenizerME tokenizer = new TokenizerME(tokenizerModel);

            // Load the POS tagging model
            POSModel posModel = new POSModel(new FileInputStream("en-pos-maxent.
↪bin"));
            POSTaggerME posTagger = new POSTaggerME(posModel);

            // Load the name finder model for person names
            TokenNameFinderModel personModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-person.bin"));
            NameFinderME personFinder = new NameFinderME(personModel);

            // Load the name finder model for locations
            TokenNameFinderModel locationModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-location.bin"));
            NameFinderME locationFinder = new NameFinderME(locationModel);

            // Load the name finder model for organizations
            TokenNameFinderModel organizationModel = new␣
↪TokenNameFinderModel(new FileInputStream("en-ner-organization.bin"));
            NameFinderME organizationFinder = new␣
↪NameFinderME(organizationModel);

            // Load the name finder model for dates
            TokenNameFinderModel dateModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-date.bin"));
            NameFinderME dateFinder = new NameFinderME(dateModel);

            // Load the name finder model for money
            TokenNameFinderModel moneyModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-money.bin"));
            NameFinderME moneyFinder = new NameFinderME(moneyModel);

            // Load the name finder model for percentages
            TokenNameFinderModel percentageModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-percentage.bin"));
            NameFinderME percentageFinder = new NameFinderME(percentageModel);

            // Load the name finder model for time
            TokenNameFinderModel timeModel = new TokenNameFinderModel(new␣
↪FileInputStream("en-ner-time.bin"));
            NameFinderME timeFinder = new NameFinderME(timeModel);
```

```java
        // Detect sentences
        String[] sentences = sentenceDetector.sentDetect(News);

        int i = 1;

        // Output detected sentences
        for (String sentence : sentences) {

                System.out.println("\u001B[31m------------\u001B[0m");
                System.out.println("\u001B[31mSentence " + i + "\u001B[0m");
                System.out.println("\u001B[31m------------\u001B[0m");
                System.out.println("");
            System.out.println("\u001B[31mSentence:\u001B[0m");
            System.out.println(sentence);
            System.out.println("");

            // Tokenize each sentence
            String[] tokens = tokenizer.tokenize(sentence);

            // Output tokens
            System.out.println("\u001B[31mTokens:\u001B[0m");
            for (String token : tokens) {
                System.out.println("'" + token + "'");
            }

            // Perform POS tagging
            String[] tags = posTagger.tag(tokens);

            // Output POS tags
            System.out.println("\n\u001B[31mPOS Tags:\u001B[0m");
            for (int j = 0; j < tokens.length; j++) {
                System.out.println(tokens[j] + "/" + tags[j]);
            }

            // Perform named entity recognition for persons
            Span[] personNames = personFinder.find(tokens);

            // Output person names
            System.out.println("\n\u001B[31mPerson Names:\u001B[0m");
            for (Span name : personNames) {
                System.out.println(name);
            }

            // Perform named entity recognition for locations
            Span[] locations = locationFinder.find(tokens);
```

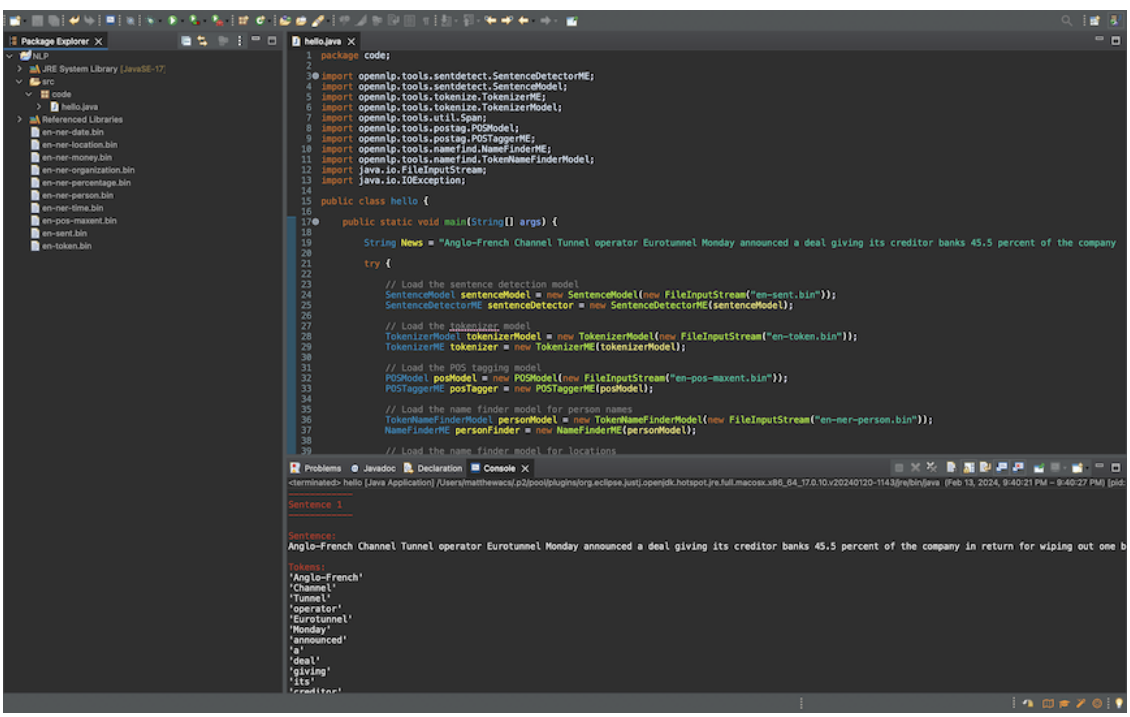
```java
        // Output locations
        System.out.println("\n\u001B[31mLocations:\u001B[0m");
        for (Span location : locations) {
            System.out.println(location);
        }

        // Perform named entity recognition for organizations
        Span[] organizations = organizationFinder.find(tokens);

        // Output organizations
        System.out.println("\n\u001B[31mOrganizations:\u001B[0m");
        for (Span organization : organizations) {
            System.out.println(organization);
        }

        // Perform named entity recognition for dates
        Span[] dates = dateFinder.find(tokens);

        // Output dates
        System.out.println("\n\u001B[31mDate:\u001B[0m");
        for (Span date : dates) {
            System.out.println(date);
        }

        // Perform named entity recognition for money
        Span[] money = moneyFinder.find(tokens);

        // Output money
        System.out.println("\n\u001B[31mMoney:\u001B[0m");
        for (Span dollar : money) {
            System.out.println(dollar);
        }

        // Perform named entity recognition for percentages
        Span[] percentages = percentageFinder.find(tokens);

        // Output percentages
        System.out.println("\n\u001B[31mPercentages:\u001B[0m");
        for (Span percent : percentages) {
            System.out.println(percent);
        }

        // Perform named entity recognition for time
        Span[] times = timeFinder.find(tokens);

        // Output time
        System.out.println("\n\u001B[31mTimes:\u001B[0m");
```

```
                for (Span time : times) {
                    System.out.println(time);
                }

                System.out.println();

                i++;
            }
        } catch (IOException e) {
            e.printStackTrace();
        }

    }

}
```

1.7 VII | Code Output

This section contains the output of the code.

The screenshot below shows the Eclipse IDE with the executed source code and output.



The full output is shown below.

Sentence 1

Sentence: Anglo-French Channel Tunnel operator Eurotunnel Monday announced a deal giving its creditor banks 45.5 percent of the company in return for wiping out one billion pounds ($1.56 billion) of its debt.

Tokens: ‘Anglo-French’ ‘Channel’ ‘Tunnel’ ‘operator’ ‘Eurotunnel’ ‘Monday’ ‘announced’ ‘a’ ‘deal’ ‘giving’ ‘its’ ‘creditor’ ‘banks’ ‘45.5’ ‘percent’ ‘of’ ‘the’ ‘company’ ‘in’ ‘return’ ‘for’ ‘wiping’ ‘out’ ‘one’ ‘billion’ ‘pounds’ ‘(’ ‘$’ ‘1.56’ ‘billion’ ‘)’ ‘of’ ‘its’ ‘debt’ ‘.’

POS Tags: Anglo-French/JJ Channel/NNP Tunnel/NNP operator/NN Eurotunnel/NNP Monday/NNP announced/VBD a/DT deal/NN giving/VBG its/PRP$ creditor/NN banks/NNS 45.5/CD percent/NN of/IN the/DT company/NN in/IN return/NN for/IN wiping/VBG out/RP one/CD billion/CD pounds/NNS (/-LRB- / 1.56/CD billion/CD)/-RRB- of/IN its/PRP$ debt/NN ./.

Person Names:

Locations:

Organizations:

Date: [5..6) date

Money: [27..30) money

Percentages: [13..15) percentage

Times:

Sentence 2

Sentence: The long-awaited restructuring brings to an end months of wrangling between Eurotunnel and the 225 banks to which it owes nearly nine billion pounds ($14.1 billion).

Tokens: ‘The’ ‘long-awaited’ ‘restructuring’ ‘brings’ ‘to’ ‘an’ ‘end’ ‘months’ ‘of’ ‘wrangling’ ‘between’ ‘Eurotunnel’ ‘and’ ‘the’ ‘225’ ‘banks’ ‘to’ ‘which’ ‘it’ ‘owes’ ‘nearly’ ‘nine’ ‘billion’ ‘pounds’ ‘(’ ‘$’ ‘14.1’ ‘billion’ ‘)’ ‘.’

POS Tags: The/DT long-awaited/JJ restructuring/NN brings/VBZ to/TO an/DT end/NN months/NNS of/IN wrangling/VBG between/IN Eurotunnel/NNP and/CC the/DT 225/CD banks/NNS to/TO which/WDT it/PRP owes/VBZ nearly/RB nine/CD billion/CD pounds/NNS (/-LRB- / 14.1/CD billion/CD)/-RRB- ./.

Person Names:

Locations:

Organizations:

Date:

Money: [25..28) money

Percentages:

Times:

Sentence 3

Sentence: The deal, announced simultaneously in Paris and London, brings the company back from the brink of insolvency but leaves shareholders owning only 54.5 percent of the company.

Tokens: ‘The’ ‘deal’ ‘,’ ‘announced’ ‘simultaneously’ ‘in’ ‘Paris’ ‘and’ ‘London’ ‘,’ ‘brings’ ‘the’ ‘company’ ‘back’ ‘from’ ‘the’ ‘brink’ ‘of’ ‘insolvency’ ‘but’ ‘leaves’ ‘shareholders’ ‘owning’ ‘only’ ‘54.5’ ‘percent’ ‘of’ ‘the’ ‘company’ ‘.’

POS Tags: The/DT deal/NN ,/, announced/VBD simultaneously/RB in/IN Paris/NNP and/CC London/NNP ,/, brings/VBZ the/DT company/NN back/RB from/IN the/DT brink/NN of/IN insolvency/NN but/CC leaves/VBZ shareholders/NNS owning/VBG only/RB 54.5/CD percent/NN of/IN the/DT company/NN ./.

Person Names:

Locations: [6..7) location [8..9) location

Organizations:

Date:

Money:

Percentages: [24..26) percentage

Times:

Sentence 4

Sentence: “The restructuring plan provides Eurotunnel with the medium-term financial stability to allow it to consolidate its substantial commercial achievements to date and to develop its operations,” Eurotunnel co- chairman Alastair Morton said.

Tokens: “‘’ ‘The’ ‘restructuring’ ‘plan’ ‘provides’ ‘Eurotunnel’ ‘with’ ‘the’ ‘medium-term’ ‘financial’ ‘stability’ ‘to’ ‘allow’ ‘it’ ‘to’ ‘consolidate’ ‘its’ ‘substantial’ ‘commercial’ ‘achievements’ ‘to’ ‘date’ ‘and’ ‘to’ ‘develop’ ‘its’ ‘operations’ ‘,’ “‘’ ‘Eurotunnel’ ‘co-’ ‘chairman’ ‘Alastair’ ‘Morton’ ‘said’ ‘.’

POS Tags: “/“ The/DT restructuring/NN plan/NN provides/VBZ Eurotunnel/NNP with/IN the/DT medium-term/JJ financial/JJ stability/NN to/TO allow/VB it/PRP to/TO consolidate/VB its/PRP$ substantial/JJ commercial/JJ achievements/NNS to/TO date/NN and/CC to/TO develop/VB its/PRP$ operations/NNS ,/,”/’ ’ Eurotunnel/NNP co-/, chairman/NN Alastair/NNP Morton/NNP said/VBD ./.

Person Names: [32..34) person

Locations:

Organizations:

Date:

Money:

Percentages:

Times:

Sentence 5

Sentence: The firm was now making a profit before interest, he added.

Tokens: ‘The’ ‘firm’ ‘was’ ‘now’ ‘making’ ‘a’ ‘profit’ ‘before’ ‘interest’ ‘,’ ‘he’ ‘added’ ‘.’

POS Tags: The/DT firm/NN was/VBD now/RB making/VBG a/DT profit/NN before/IN interest/NN ,/, he/PRP added/VBD ./.

Person Names:

Locations:

Organizations:

Date:

Money:

Percentages:

Times:

Sentence 6

Sentence: Although shareholders will see their interests diluted, they were offered the prospect of a brighter future after months of uncertainty while Eurotunnel wrestled to reduce crippling interest payments negotiated during the tunnel’s construction.

Tokens: ‘Although’ ‘shareholders’ ‘will’ ‘see’ ‘their’ ‘interests’ ‘diluted’ ‘,’ ‘they’ ‘were’ ‘offered’ ‘the’ ‘prospect’ ‘of’ ‘a’ ‘brighter’ ‘future’ ‘after’ ‘months’ ‘of’ ‘uncertainty’ ‘while’ ‘Eurotunnel’ ‘wrestled’ ‘to’ ‘reduce’ ‘crippling’ ‘interest’ ‘payments’ ‘negotiated’ ‘during’ ‘the’ ‘tunnel’ ‘’s’ ‘construction’ ‘.’

POS Tags: Although/IN shareholders/NNS will/MD see/VB their/PRP$ interests/NNS diluted/VBN ,/, they/PRP were/VBD offered/VBN the/DT prospect/NN of/IN a/DT brighter/JJR future/NN after/IN months/NNS of/IN uncertainty/NN while/IN Eurotunnel/NNP wrestled/VBD to/TO reduce/VB crippling/JJ interest/NN payments/NNS negotiated/VBD during/IN the/DT tunnel/NN ’s/VBZ construction/NN ./.

Person Names:

Locations:

Organizations:

Date:

Money:

Percentages:

Times:

Sentence 7

Sentence: Eurotunnel, which has taken around half the cross-Channel market from the European ferry companies, said a strong operating performance could allow it to pay its first dividend within the next 10 years.

Tokens: ‘Eurotunnel’ ‘,’ ‘which’ ‘has’ ‘taken’ ‘around’ ‘half’ ‘the’ ‘cross-Channel’ ‘market’ ‘from’ ‘the’ ‘European’ ‘ferry’ ‘companies’ ‘,’ ‘said’ ‘a’ ‘strong’ ‘operating’ ‘performance’ ‘could’ ‘allow’ ‘it’ ‘to’ ‘pay’ ‘its’ ‘first’ ‘dividend’ ‘within’ ‘the’ ‘next’ ‘10’ ‘years’ ‘.’

POS Tags: Eurotunnel/NNP ,/, which/WDT has/VBZ taken/VBN around/IN half/PDT the/DT cross-Channel/NN market/NN from/IN the/DT European/JJ ferry/NN companies/NNS ,/, said/VBD a/DT strong/JJ operating/NN performance/NN could/MD allow/VB it/PRP to/TO pay/VB its/PRP$ first/JJ dividend/NN within/IN the/DT next/JJ 10/CD years/NNS ./.

Person Names:

Locations:

Organizations:

Date:

Money:

Percentages:

Times:

Sentence 8

Sentence: French co-chairman Patrick Ponsolle said shareholders would have to be patient before they could reap the benefits of the company’s success.

Tokens: ‘French’ ‘co-chairman’ ‘Patrick’ ‘Ponsolle’ ‘said’ ‘shareholders’ ‘would’ ‘have’ ‘to’ ‘be’ ‘patient’ ‘before’ ‘they’ ‘could’ ‘reap’ ‘the’ ‘benefits’ ‘of’ ‘the’ ‘company’ ‘’s’ ‘success’ ‘.’

POS Tags: French/JJ co-chairman/NN Patrick/NNP Ponsolle/NNP said/VBD shareholders/NNS would/MD have/VB to/TO be/VB patient/JJ before/IN they/PRP could/MD reap/VB the/DT benefits/NNS of/IN the/DT company/NN ’s/POS success/NN ./.

Person Names: [2..4) person

Locations:

Organizations:

Date:

Money:

Percentages:

Times:

Sentence 9

Sentence: He called the debt restructuring plan “an acceptable compromise” for holders of Eurotunnel shares.

Tokens: ‘He’ ‘called’ ‘the’ ‘debt’ ‘restructuring’ ‘plan’ “‘’ ‘an’ ‘acceptable’ ‘compromise’ “‘’ ‘for’ ‘holders’ ‘of’ ‘Eurotunnel’ ‘shares’ ‘.’

POS Tags: He/PRP called/VBD the/DT debt/NN restructuring/NN plan/NN “/“ an/DT acceptable/JJ compromise/NN”/’ ’ for/IN holders/NNS of/IN Eurotunnel/NNP shares/NNS ./.

Person Names:

Locations:

Organizations:

Date:

Money:

Percentages:

Times:

Sentence 10

Sentence: The company said there was still considerable work to be done to finalise and agree on the details of the plan before it can be submitted to shareholders and the full 225 bank syndicate for approval, probably early in 1997.

Tokens: ‘The’ ‘company’ ‘said’ ‘there’ ‘was’ ‘still’ ‘considerable’ ‘work’ ‘to’ ‘be’ ‘done’ ‘to’ ‘finalise’ ‘and’ ‘agree’ ‘on’ ‘the’ ‘details’ ‘of’ ‘the’ ‘plan’ ‘before’ ‘it’ ‘can’ ‘be’ ‘submitted’ ‘to’ ‘shareholders’ ‘and’ ‘the’ ‘full’ ‘225’ ‘bank’ ‘syndicate’ ‘for’ ‘approval’ ‘,’ ‘probably’ ‘early’ ‘in’ ‘1997’ ‘.’

POS Tags: The/DT company/NN said/VBD there/EX was/VBD still/RB considerable/JJ work/NN to/TO be/VB done/VBN to/TO finalise/VB and/CC agree/VB on/IN the/DT details/NNS of/IN the/DT plan/NN before/IN it/PRP can/MD be/VB submitted/VBN to/TO shareholders/NNS and/CC the/DT full/JJ 225/CD bank/NN syndicate/NN for/IN approval/NN ,/, probably/RB early/RB in/IN 1997/CD ./.

Person Names:

Locations:

Organizations:

Date: [40..41) date

Money:

Percentages:

Times:

Sentence 11

Sentence: Monday's announcement followed two weeks of highly secretive negotiations between Eurotunnel and its six leading banks.

Tokens: 'Monday' ''s' 'announcement' 'followed' 'two' 'weeks' 'of' 'highly' 'secretive' 'negotiations' 'between' 'Eurotunnel' 'and' 'its' 'six' 'leading' 'banks' '.'

POS Tags: Monday/NNP 's/POS announcement/NN followed/VBD two/CD weeks/NNS of/IN highly/RB secretive/JJ negotiations/NNS between/IN Eurotunnel/NNP and/CC its/PRP$ six/CD leading/JJ banks/NNS ./.

Person Names:

Locations:

Organizations:

Date: [0..1) date

Money:

Percentages:

Times:

Sentence 12

Sentence: This was extended to the 24 "instructing banks" at a meeting late last week in London.

Tokens: 'This' 'was' 'extended' 'to' 'the' '24' "'instructing' 'banks' "'' 'at' 'a' 'meeting' 'late' 'last' 'week' 'in' 'London' '.'

POS Tags: This/DT was/VBD extended/VBN to/TO the/DT 24/CD "instructing/NN banks/NNS"/' ' at/IN a/DT meeting/NN late/RB last/JJ week/NN in/IN London/NNP ./.

Person Names:

Locations: [16..17) location

Organizations:

Date: [12..15) date

Money:

Percentages:

Times:

Sentence 13

Sentence: Eurotunnel said the debt-for-equity swap would be at 130 pence, or 10.40 francs, per share.

Tokens: ‘Eurotunnel’ ‘said’ ‘the’ ‘debt-for-equity’ ‘swap’ ‘would’ ‘be’ ‘at’ ‘130’ ‘pence’ ‘,’ ‘or’ ‘10.40’ ‘francs’ ‘,’ ‘per’ ‘share’ ‘.’

POS Tags: Eurotunnel/NNP said/VBD the/DT debt-for-equity/NN swap/NN would/MD be/VB at/IN 130/CD pence/NN ,/, or/CC 10.40/CD francs/NNS ,/, per/IN share/NN ./.

Person Names:

Locations:

Organizations:

Date: [12..13) date

Money: [8..10) money [12..14) money

Percentages:

Times:

Sentence 14

Sentence: That is considerably below the level of around 160 pence widely reported before announcement of the deal, and will reduce outstanding debt of 8.7 billion pounds ($13.6 billion) by 1.0 billion ($1.56 billion).

Tokens: ‘That’ ‘is’ ‘considerably’ ‘below’ ‘the’ ‘level’ ‘of’ ‘around’ ‘160’ ‘pence’ ‘widely’ ‘reported’ ‘before’ ‘announcement’ ‘of’ ‘the’ ‘deal’ ‘,’ ‘and’ ‘will’ ‘reduce’ ‘outstanding’ ‘debt’ ‘of’ ‘8.7’ ‘billion’ ‘pounds’ ‘(’ ‘$’ ’13.6’ ’billion’ ’)’ ’by’ ’1.0’ ’billion’ ’(’ ’$’ ‘1.56’ ‘billion’ ‘)’ ‘.’

POS Tags: That/DT is/VBZ considerably/RB below/IN the/DT level/NN of/IN around/IN 160/CD pence/NN widely/RB reported/VBN before/IN announcement/NN of/IN the/DT deal/NN ,/, and/CC will/MD reduce/VB outstanding/JJ debt/NN of/IN 8.7/CD billion/CD pounds/NNS (/-LRB- / 13.6/CD billion/CD)/-RRB- by/IN 1.0/CD billion/CD (/-LRB- / 1.56/CD billion/CD)/-RRB- ./.

Person Names:

Locations:

Organizations:

Date:

Money: [8..10) money [28..31) money [36..39) money

Percentages:

Times:

Sentence 15

Sentence: The company said a further 3.7 billion pounds ($5.8 billion) of debt would be converted into new financial instruments and existing shareholders would be able to participate in this issue.

Tokens: ‘The’ ‘company’ ‘said’ ‘a’ ‘further’ ‘3.7’ ‘billion’ ‘pounds’ ‘(’ ‘$’ ‘5.8’ ‘billion’ ‘)’ ‘of’ ‘debt’ ‘would’ ‘be’ ‘converted’ ‘into’ ‘new’ ‘financial’ ‘instruments’ ‘and’ ‘existing’ ‘shareholders’ ‘would’ ‘be’ ‘able’ ‘to’ ‘participate’ ‘in’ ‘this’ ‘issue’ ‘.’

POS Tags: The/DT company/NN said/VBD a/DT further/RB 3.7/CD billion/CD pounds/NNS (/-LRB- / 5.8/CD billion/CD)/-RRB- of/IN debt/NN would/MD be/VB converted/VBN into/IN new/JJ financial/JJ instruments/NNS and/CC existing/VBG shareholders/NNS would/MD be/VB able/JJ to/TO participate/VB in/IN this/DT issue/NN ./.

Person Names:

Locations:

Organizations:

Date:

Money: [9..12) money

Percentages:

Times:

Sentence 16

Sentence: If they choose not to take up free warrants entitling them to subscribe to this, Eurotunnel said shareholders’ interests may be reduced further to just over 39 percent of the company by the end of December 2003.

Tokens: ‘If’ ‘they’ ‘choose’ ‘not’ ‘to’ ‘take’ ‘up’ ‘free’ ‘warrants’ ‘entitling’ ‘them’ ‘to’ ‘subscribe’ ‘to’ ‘this’ ‘,’ ‘Eurotunnel’ ‘said’ ‘shareholders’ ’ ’ ‘interests’ ‘may’ ‘be’ ‘reduced’ ‘further’ ‘to’ ‘just’ ‘over’ ‘39’ ‘percent’ ‘of’ ‘the’ ‘company’ ‘by’ ‘the’ ‘end’ ‘of’ ‘December’ ‘2003’ ‘.’

POS Tags: If/IN they/PRP choose/VBP not/RB to/TO take/VB up/RP free/JJ warrants/NNS entitling/VBG them/PRP to/TO subscribe/VB to/TO this/DT ,/, Eurotunnel/NNP said/VBD shareholders/NNS ’/POS interests/NNS may/MD be/VB reduced/VBN further/RB to/TO just/RB over/IN 39/CD percent/NN of/IN the/DT company/NN by/IN the/DT end/NN of/IN December/NNP 2003/CD ./.

Person Names:

Locations:

Organizations:

Date: [35..39) date

Money:

Percentages: [28..30) percentage

Times:

Sentence 17

Sentence: Eurotunnel's shares, which were suspended last week at 113.5 pence ahead of Monday's announcement, should resume trading on Tuesday, the company said.

Tokens: 'Eurotunnel' ''s' 'shares' ',' 'which' 'were' 'suspended' 'last' 'week' 'at' '113.5' 'pence' 'ahead' 'of' 'Monday' ''s' 'announcement' ',' 'should' 'resume' 'trading' 'on' 'Tuesday' ',' 'the' 'company' 'said' '.'

POS Tags: Eurotunnel/NNP 's/POS shares/NNS ,/, which/WDT were/VBD suspended/VBN last/JJ week/NN at/IN 113.5/CD pence/NNS ahead/RB of/IN Monday/NNP 's/POS announcement/NN ,/, should/MD resume/VB trading/VBG on/IN Tuesday/NNP ,/, the/DT company/NN said/VBD ./.

Person Names:

Locations:

Organizations:

Date: [7..9) date [10..11) date [14..15) date [22..23) date

Money: [10..12) money

Percentages:

Times:

1.8 VIII | Discussions and Conclusions

Overall, this code was able to correctly detect each sentence, tokenize it, perform POS tagging, and identify named entities. Extracting this information for the text using OpenNLP can allow for fast identification of important information such as how much money, what time frame, and who was involved. Additionally, text preprocessing using OpenNLP can be a first step is a larger NLP project that involves training a model to classify, generate, or segment a corpus.

One difficulty I faced in completing this assignment was the environment setup and language. I have never used Java before, which posed a syntax barrier for my Python-wired brain. Additionally, getting Eclipse, OpenNLP, and Python working was time consuming and confusing at first. Once the environment was setup, the programming was more straightforward using the resources and documentation listed in the references section.

1.9 IX | References

- https://opennlp.apache.org/
- https://www.java.com/en/
- https://www.eclipse.org/ide/
- https://www.programcreek.com/2012/05/opennlp-tutorial/

- https://www.tutorialspoint.com/opennlp/opennlp_environment.htm

1.10 X | Appendix

Source Code and Output: https://github.com/matthewaaa123/OpenNLP-Assignment/tree/main