COT-6930 Natural Language Processing, Spring 2019
Assignment 1
Christian Garbin

Given a news article, after installing opennlp, the following steps should be completed using opennlp:

1. *Detect sentences in the given news article (2 points)*
2. *Tokenize each sentence into words (2 points)*
3. *Perform Part-of-Speech (POS) on each sentence (3 points)*
4. *Find name entities including person's name entities and locations (3 points)*

How this report is organized:

- The first section, "Results" describe the results from the experiments. There is one sub-section for each requested task. Results are abbreviated to make the report more legible. The full results are in the accompanying file, `assignment1-full-results-cgarbin.txt`.
- The appendix has the code and screenshots.
- The source code is available as accompanying files. The appendix section describes the organization of the code.

Results

Part 1 - Sentence Detection

1. Detect sentences in the given news article (2 points)

The code detected 16 sentences, listed below, in the order they appear in the document. Sentences are abbreviated to make the results more legible. Full results are available in the attached file `assignment1-full-results-cgarbin.txt`.

```
[00] Anglo-French Channel Tunnel operator Eurotunnel Monday ... one billion pounds ($1.56 billion) of its debt.
[01] The long-awaited restructuring brings to an end months ... nearly nine billion pounds ($14.1 billion).
[02] The deal, announced simultaneously in Paris and London,... only 54.5 percent of the company.
[03] "The restructuring plan provides Eurotunnel with the ... " Eurotunnel co-chairman Alastair Morton said.
[04] The firm was now making a profit before interest, he added.
[05] Although shareholders will see their interests diluted,...  negotiated during the tunnel's construction.
[06] Eurotunnel, which has taken around half the cross-Channel ... dividend within the next 10 years.
[07] French co-chairman Patrick Ponsolle said shareholders ... the benefits of the company's success.
[08] He called the debt restructuring plan "an acceptable compromise" for holders of Eurotunnel shares.
[09] The company said there was still considerable work to be ... for approval, probably early in 1997.
[10] Monday's announcement followed two weeks of highly secretive ... Eurotunnel and its six leading banks.
[11] This was extended to the 24 "instructing banks" at a meeting late last week in London.
[12] Eurotunnel said the debt-for-equity swap would be at 130 pence, or 10.40 francs, per share.
[13] That is considerably below the level of around 160 pence ... billion) by 1.0 billion ($1.56 billion).
[14] The company said a further 3.7 billion pounds ($5.8 billion)...  be able to participate in this issue.
[15] If they choose not to take up free warrants entitling them... the company by the end of December 2003.
[16] Eurotunnel's shares, which were suspended last week at 113.5... trading on Tuesday, the company said.
```

Parte 2 - Tokenization

2. Tokenize each sentence into words (2 points)

OpenNLP has three tokenizers:

- `SimpleTokenizer`: based on character classes - *"sequences of the same character class are tokens".*
- `WhitespaceTokenizer`: *"non whitespace sequences are identified as tokens"*
- `TokenizerME`: a probabilistic tokenizer - *"detects token boundaries based on probability model"*

The table below shows the difference between the tokenizers. The table has the results for each tokenizer on the first sentence of the document. Full results are available in the attached file `assignment1-full-results-cgarbin.txt`.

It is clear that `TokenizerME` does a better job of identifying tokens as humans would do. For example, it performs better when extracting numbers from the surrounding text, as shown in the highlighted tokens in the table. For this reason, `TokenizerME` was used in the subsequent tasks (part-of-speech and named entities).

`TokenizerME`	`WhitespaceTokenizer`	`SimpleTokenizer`
[00] Anglo-French	[00] Anglo-French	[00] Anglo
[01] Channel	[01] Channel	[01] -
[02] Tunnel	[02] Tunnel	[02] French
[03] operator	[03] operator	[03] Channel
[04] Eurotunnel	[04] Eurotunnel	[04] Tunnel
[05] Monday	[05] Monday	[05] operator
[06] announced	[06] announced	[06] Eurotunnel
[07] a	[07] a	[07] Monday
[08] deal	[08] deal	[08] announced
[09] giving	[09] giving	[09] a
[10] its	[10] its	[10] deal
[11] creditor	[11] creditor	[11] giving
[12] banks	[12] banks	[12] its
[13] 45.5	[13] 45.5	[13] creditor
[14] percent	[14] percent	[14] banks
[15] of	[15] of	[15] 45
[16] the	[16] the	[16] .
[17] company	[17] company	[17] 5
[18] in	[18] in	[18] percent
[19] return	[19] return	[19] of
[20] for	[20] for	[20] the
[21] wiping	[21] wiping	[21] company
[22] out	[22] out	[22] in
[23] one	[23] one	[23] return
[24] billion	[24] billion	[24] for
[25] pounds	[25] pounds	[25] wiping
[26] (	[26] ($1.56	[26] out
[27] $	[27] billion)	[27] one
[28] 1.56	[28] of	[28] billion
[29] billion	[29] its	[29] pounds
[30])	[30] debt.	[30] (
[31] of		[31] $
[32] its		[32] 1
[33] debt		[33] .
[34] .		[34] 56
		[35] billion
		[36])
		[37] of
		[38] its
		[39] debt
		[40] .

Part 3 - Part-of-Speech (POS)

Part-of-speech in OpenNLP requires tokenization first. Based on the results of the previous section, `TokenizerME` was used as the tokenizer.

The results for the first three sentences in the document are shown in the table below. Full results are available in the attached file `assignment1-full-results-cgarbin.txt`.

Sentence 1

[00]	Anglo-French	JJ
[01]	Channel	NNP
[02]	Tunnel	NNP
[03]	operator	NN
[04]	Eurotunnel	NNP
[05]	Monday	NNP
[06]	announced	VBD
[07]	a	DT
[08]	deal	NN
[09]	giving	VBG
[10]	its	PRP$
[11]	creditor	NN
[12]	banks	NNS
[13]	45.5	CD
[14]	percent	NN
[15]	of	IN
[16]	the	DT
[17]	company	NN
[18]	in	IN
[19]	return	NN
[20]	for	IN
[21]	wiping	VBG
[22]	out	RP
[23]	one	CD
[24]	billion	CD
[25]	pounds	NNS
[26]	(	-LRB-
[27]	$	$
[28]	1.56	CD
[29]	billion	CD
[30]	)	-RRB-
[31]	of	IN
[32]	its	PRP$
[33]	debt	NN
[34]	.	.

Sentence 2

[00]	The	DT
[01]	long-awaited	JJ
[02]	restructuring	NN
[03]	brings	VBZ
[04]	to	TO
[05]	an	DT
[06]	end	NN
[07]	months	NNS
[08]	of	IN
[09]	wrangling	VBG
[10]	between	IN
[11]	Eurotunnel	NNP
[12]	and	CC
[13]	the	DT
[14]	225	CD
[15]	banks	NNS
[16]	to	TO
[17]	which	WDT
[18]	it	PRP
[19]	owes	VBZ
[20]	nearly	RB
[21]	nine	CD
[22]	billion	CD
[23]	pounds	NNS
[24]	(	-LRB-
[25]	$	$
[26]	14.1	CD
[27]	billion	CD
[28]	)	-RRB-
[29]	.	.

Sentence 3

[00]	The	DT
[01]	deal	NN
[02]	,	,
[03]	announced	VBD
[04]	simultaneously	RB
[05]	in	IN
[06]	Paris	NNP
[07]	and	CC
[08]	London	NNP
[09]	,	,
[10]	brings	VBZ
[11]	the	DT
[12]	company	NN
[13]	back	RB
[14]	from	IN
[15]	the	DT
[16]	brink	NN
[17]	of	IN
[18]	insolvency	NN
[19]	but	CC
[20]	leaves	VBZ
[21]	shareholders	NNS
[22]	owning	VBG
[23]	only	RB
[24]	54.5	CD
[25]	percent	NN
[26]	of	IN
[27]	the	DT
[28]	company	NN
[29]	.	.

Part 4 - Named Entities

Named entities in OpenNLP require tokenization first. Based on the results of the previous section, `TokenizerME` was used as the tokenizer.

The results for the first four sentences in the document are shown below, with the identified spans highlighted. Full results are available in the attached file `assignment1-full-results-cgarbin.txt`.

> Anglo-French Channel Tunnel operator Eurotunnel Monday announced a deal giving its creditor banks **<PERCENTAGE> 45.5 percent </PERCENTAGE>** of the company in return for wiping out one billion pounds (**<MONEY> $ 1.56 billion </MONEY>**) 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 (**<MONEY> $ 14.1 billion </MONEY>**) .
>
> The deal , announced simultaneously in **<LOCATION> Paris </LOCATION>** and **<LOCATION> London </LOCATION>** , brings the company back from the brink of insolvency but leaves shareholders owning only **<PERCENTAGE> 54.5 percent </PERCENTAGE>** 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 **<PERSON> Alastair Morton </PERSON>** said .

References

The following references were consulted to write the code and to produce this report:

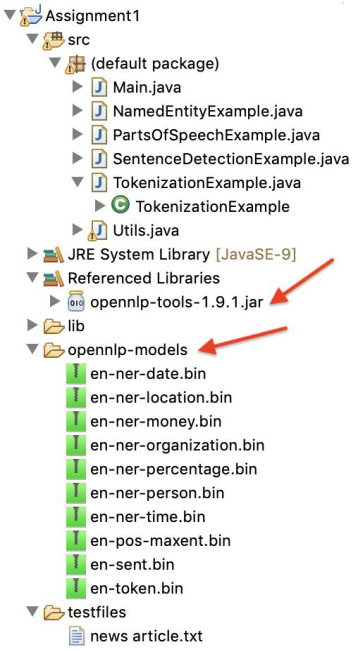
- https://opennlp.apache.org/docs/1.5.3/manual/opennlp.html
- https://opennlp.apache.org/models.html
- https://www.tutorialkart.com/opennlp/how-to-setup-opennlp-java-project/
- http://www.sfs.uni-tuebingen.de/~keberle/NLPTools/presentations/OpenNLP/OpenNLP%20BorenshteinDaeubler.pdf

Appendix

Source code overview

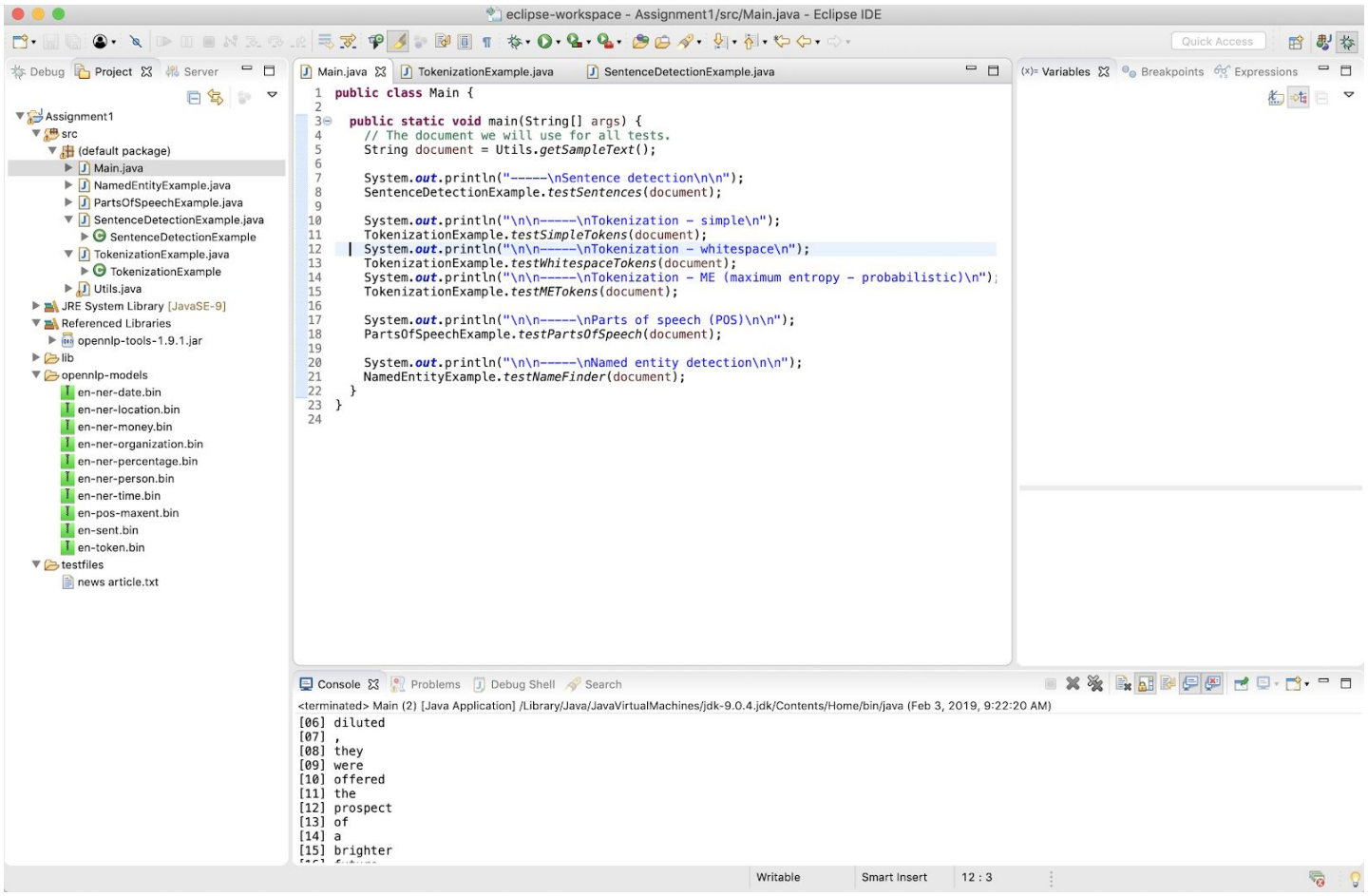
The code is an eclipse project, structured as in the picture below. Each task is in a separate file named accordingly, e.g. NamedEntityExamples.java has the code for the named entity code.

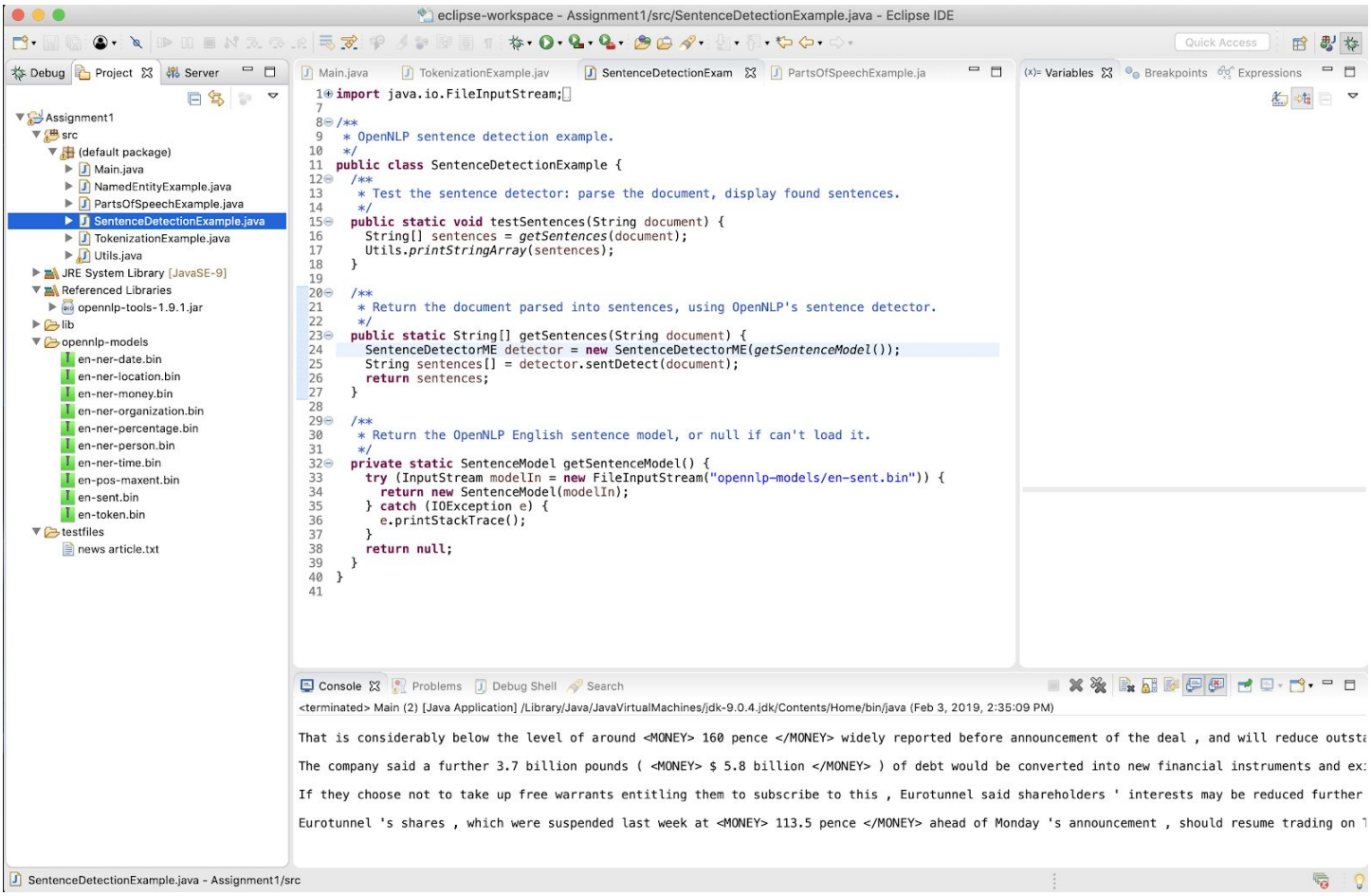
The Java code used for this report is submitted in a separate file. It does not contain OpenNLP components. The arrows indicate OpenNLP libraries and models that have to be downloaded and added to the project.

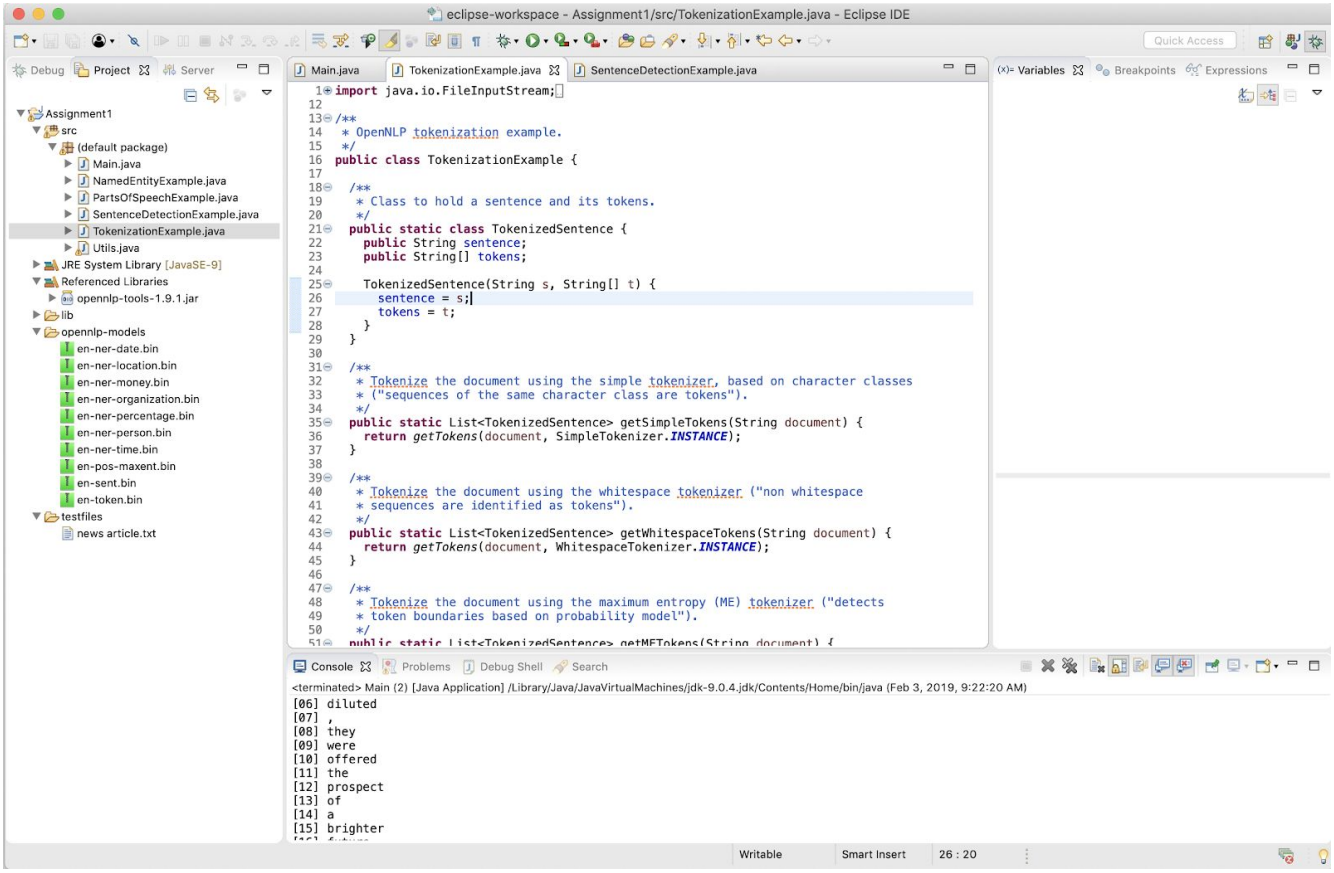


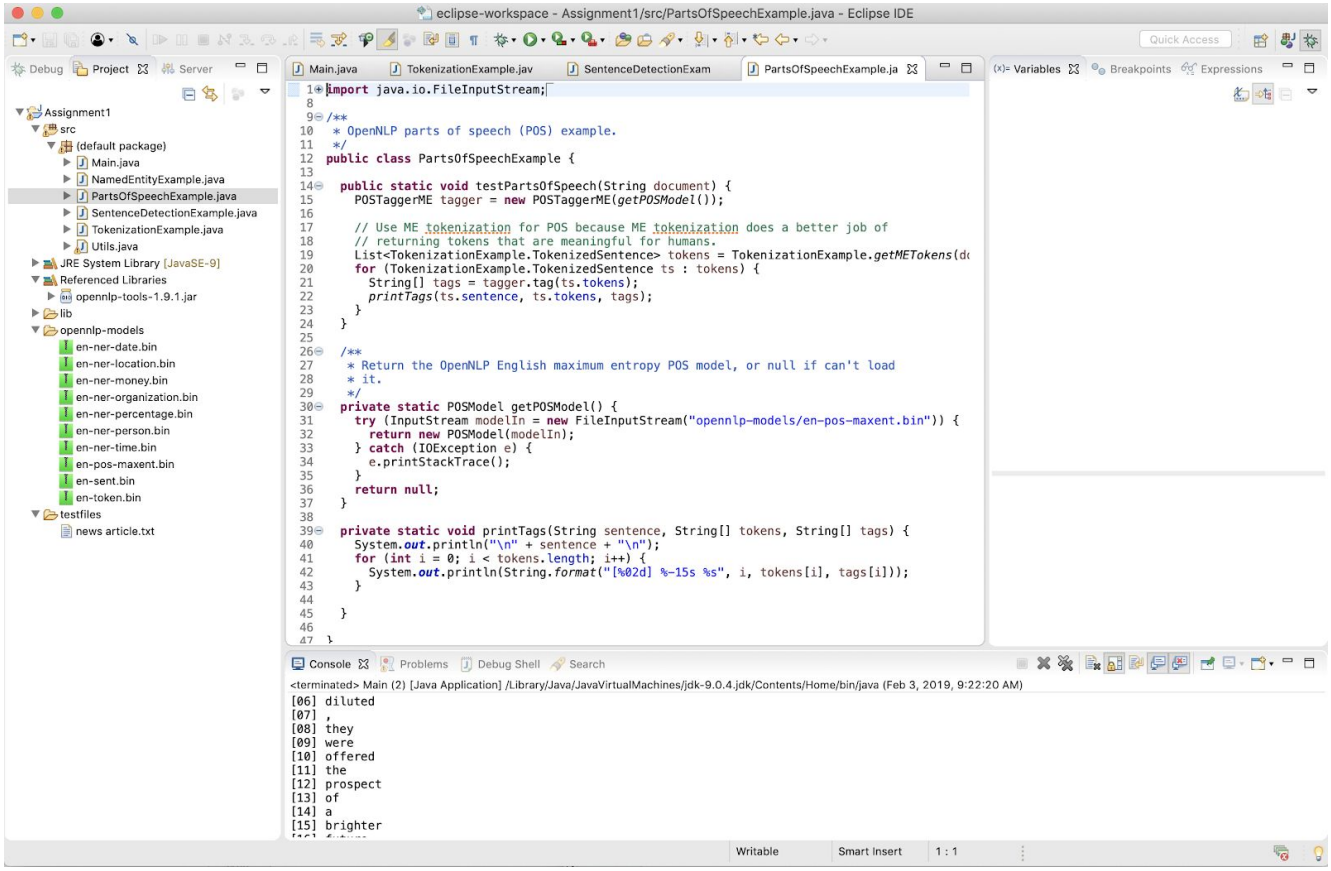
Source code for the tasks is listed below.

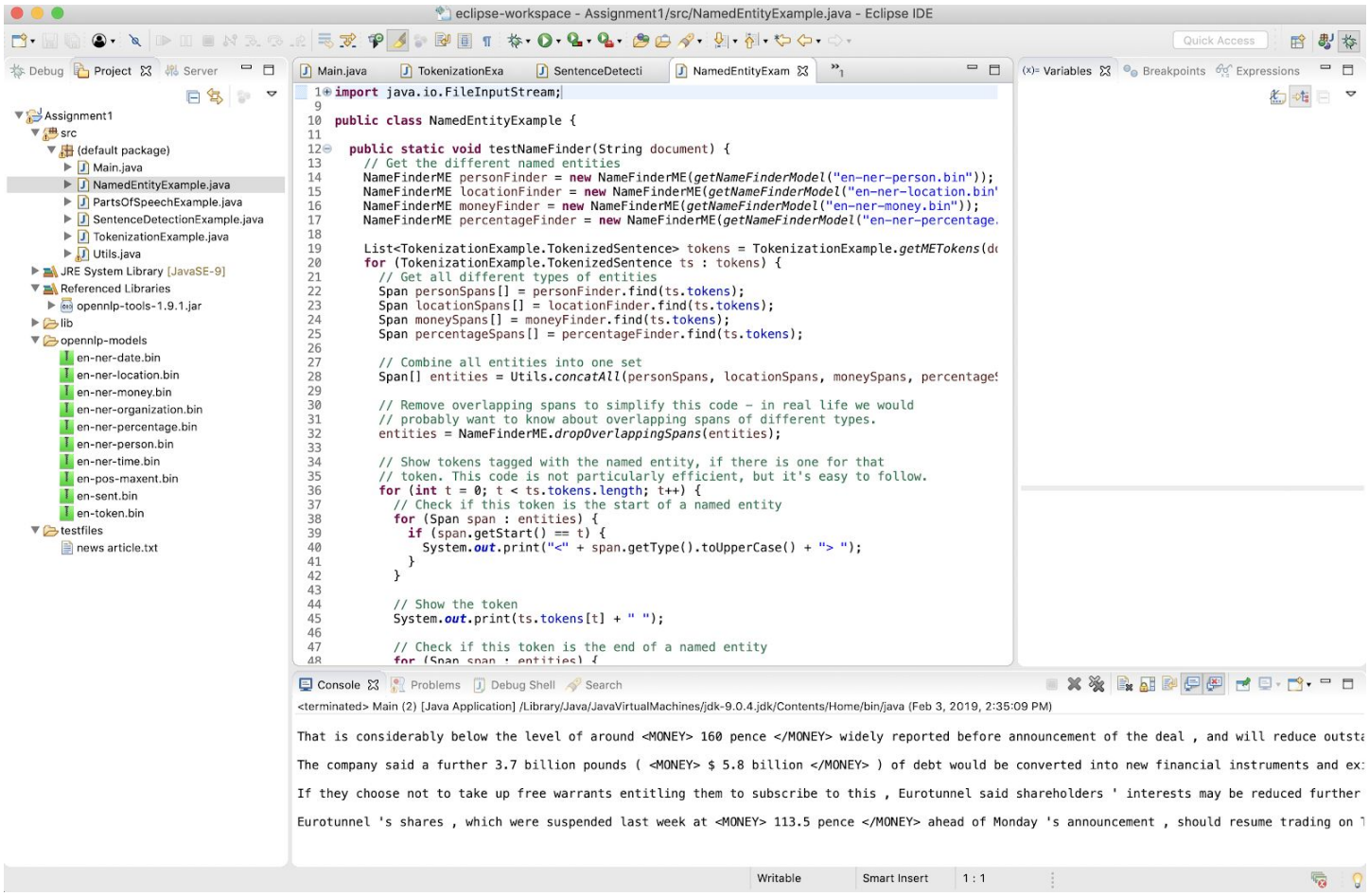
Screenshots




eclipse-workspace - Assignment1/src/SentenceDetectionExample.java - Eclipse IDE
Assignment1
src
(default package)
Main.java
NamedEntityExample.java
PartsOfSpeechExample.java
SentenceDetectionExample.java
TokenizationExample.java
Utils.java
JRE System Library [JavaSE-9]
Referenced Libraries
opennlp-tools-1.9.1.jar
lib
opennlp-models
en-ner-date.bin
en-ner-location.bin
en-ner-money.bin
en-ner-organization.bin
en-ner-percentage.bin
en-ner-person.bin
en-ner-time.bin
en-pos-maxent.bin
en-sent.bin
en-token.bin
testfiles
news article.txt
import java.io.FileInputStream;
/**
* OpenNLP sentence detection example.
*/
public class SentenceDetectionExample {
/**
* Test the sentence detector: parse the document, display found sentences.
*/
public static void testSentences(String document) {
String[] sentences = getSentences(document);
Utils.printStringArray(sentences);
}
/**
* Return the document parsed into sentences, using OpenNLP's sentence detector.
*/
public static String[] getSentences(String document) {
SentenceDetectorME detector = new SentenceDetectorME(getSentenceModel());
String sentences[] = detector.sentDetect(document);
return sentences;
}
/**
* Return the OpenNLP English sentence model, or null if can't load it.
*/
private static SentenceModel getSentenceModel() {
try (InputStream modelIn = new FileInputStream("opennlp-models/en-sent.bin")) {
return new SentenceModel(modelIn);
} catch (IOException e) {
e.printStackTrace();
}
return null;
}
}
Console
Problems
Debug Shell
Search
<terminated> Main (2) [Java Application] /Library/Java/JavaVirtualMachines/jdk-9.0.4.jdk/Contents/Home/bin/java (Feb 3, 2019, 2:35:09 PM)
That is considerably below the level of around <MONEY> 160 pence </MONEY> widely reported before announcement of the deal , and will reduce outsta
The company said a further 3.7 billion pounds (<MONEY> $ 5.8 billion </MONEY>) of debt would be converted into new financial instruments and ex
If they choose not to take up free warrants entitling them to subscribe to this , Eurotunnel said shareholders ' interests may be reduced further
Eurotunnel 's shares , which were suspended last week at <MONEY> 113.5 pence </MONEY> ahead of Monday 's announcement , should resume trading on
SentenceDetectionExample.java - Assignment1/src


eclipse-workspace - Assignment1/src/TokenizationExample.java - Eclipse IDE
Assignment1
src
(default package)
Main.java
NamedEntityExample.java
PartsOfSpeechExample.java
SentenceDetectionExample.java
TokenizationExample.java
Utils.java
JRE System Library [JavaSE-9]
Referenced Libraries
opennlp-tools-1.9.1.jar
lib
opennlp-models
en-ner-date.bin
en-ner-location.bin
en-ner-money.bin
en-ner-organization.bin
en-ner-percentage.bin
en-ner-person.bin
en-ner-time.bin
en-pos-maxent.bin
en-sent.bin
en-token.bin
testfiles
news article.txt
import java.io.FileInputStream;
/**
* OpenNLP tokenization example.
*/
public class TokenizationExample {
/**
* Class to hold a sentence and its tokens.
*/
public static class TokenizedSentence {
public String sentence;
public String[] tokens;
TokenizedSentence(String s, String[] t) {
sentence = s;
tokens = t;
}
}
/**
* Tokenize the document using the simple tokenizer, based on character classes
* ("sequences of the same character class are tokens").
*/
public static List<TokenizedSentence> getSimpleTokens(String document) {
return getTokens(document, SimpleTokenizer.INSTANCE);
}
/**
* Tokenize the document using the whitespace tokenizer ("non whitespace
* sequences are identified as tokens").
*/
public static List<TokenizedSentence> getWhitespaceTokens(String document) {
return getTokens(document, WhitespaceTokenizer.INSTANCE);
}
/**
* Tokenize the document using the maximum entropy (ME) tokenizer ("detects
* token boundaries based on probability model").
*/
Console
Problems
Debug Shell
Search
<terminated> Main (2) [Java Application] /Library/Java/JavaVirtualMachines/jdk-9.0.4.jdk/Contents/Home/bin/java (Feb 3, 2019, 9:22:20 AM)
[06] diluted
[07] ,
[08] they
[09] were
[10] offered
[11] the
[12] prospect
[13] of
[14] a
[15] brighter
Writable
Smart Insert
26 : 20


eclipse-workspace - Assignment1/src/PartsOfSpeechExample.java - Eclipse IDE
Assignment1
src
(default package)
Main.java
NamedEntityExample.java
PartsOfSpeechExample.java
SentenceDetectionExample.java
TokenizationExample.java
Utils.java
JRE System Library [JavaSE-9]
Referenced Libraries
opennlp-tools-1.9.1.jar
lib
opennlp-models
en-ner-date.bin
en-ner-location.bin
en-ner-money.bin
en-ner-organization.bin
en-ner-percentage.bin
en-ner-person.bin
en-ner-time.bin
en-pos-maxent.bin
en-sent.bin
en-token.bin
testfiles
news article.txt
import java.io.FileInputStream;
/**
* OpenNLP parts of speech (POS) example.
*/
public class PartsOfSpeechExample {
public static void testPartsOfSpeech(String document) {
POSTaggerME tagger = new POSTaggerME(getPOSModel());
// Use ME tokenization for POS because ME tokenization does a better job of
// returning tokens that are meaningful for humans.
List<TokenizationExample.TokenizedSentence> tokens = TokenizationExample.getMETokens(d
for (TokenizationExample.TokenizedSentence ts : tokens) {
String[] tags = tagger.tag(ts.tokens);
printTags(ts.sentence, ts.tokens, tags);
}
}
/**
* Return the OpenNLP English maximum entropy POS model, or null if can't load
* it.
*/
private static POSModel getPOSModel() {
try (InputStream modelIn = new FileInputStream("opennlp-models/en-pos-maxent.bin")) {
return new POSModel(modelIn);
} catch (IOException e) {
e.printStackTrace();
}
return null;
}
private static void printTags(String sentence, String[] tokens, String[] tags) {
System.out.println("\n" + sentence + "\n");
for (int i = 0; i < tokens.length; i++) {
System.out.println(String.format("[%02d] %-15s %s", i, tokens[i], tags[i]));
}
}
Console
Problems
Debug Shell
Search
<terminated> Main (2) [Java Application] /Library/Java/JavaVirtualMachines/jdk-9.0.4.jdk/Contents/Home/bin/java (Feb 3, 2019, 9:22:20 AM)
[06] diluted
[07] ,
[08] they
[09] were
[10] offered
[11] the
[12] prospect
[13] of
[14] a
[15] brighter
Writable
Smart Insert
1 : 1


eclipse-workspace - Assignment1/src/NamedEntityExample.java - Eclipse IDE
Assignment1
src
(default package)
Main.java
NamedEntityExample.java
PartsOfSpeechExample.java
SentenceDetectionExample.java
TokenizationExample.java
Utils.java
JRE System Library [JavaSE-9]
Referenced Libraries
opennlp-tools-1.9.1.jar
lib
opennlp-models
en-ner-date.bin
en-ner-location.bin
en-ner-money.bin
en-ner-organization.bin
en-ner-percentage.bin
en-ner-person.bin
en-ner-time.bin
en-pos-maxent.bin
en-sent.bin
en-token.bin
testfiles
news article.txt
import java.io.FileInputStream;
public class NamedEntityExample {
public static void testNameFinder(String document) {
// Get the different named entities
NameFinderME personFinder = new NameFinderME(getNameFinderModel("en-ner-person.bin"));
NameFinderME locationFinder = new NameFinderME(getNameFinderModel("en-ner-location.bin'
NameFinderME moneyFinder = new NameFinderME(getNameFinderModel("en-ner-money.bin"));
NameFinderME percentageFinder = new NameFinderME(getNameFinderModel("en-ner-percentage.
List<TokenizationExample.TokenizedSentence> tokens = TokenizationExample.getMETokens(d
for (TokenizationExample.TokenizedSentence ts : tokens) {
// Get all different types of entities
Span personSpans[] = personFinder.find(ts.tokens);
Span locationSpans[] = locationFinder.find(ts.tokens);
Span moneySpans[] = moneyFinder.find(ts.tokens);
Span percentageSpans[] = percentageFinder.find(ts.tokens);
// Combine all entities into one set
Span[] entities = Utils.concatAll(personSpans, locationSpans, moneySpans, percentage
// Remove overlapping spans to simplify this code - in real life we would
// probably want to know about overlapping spans of different types.
entities = NameFinderME.dropOverlappingSpans(entities);
// Show tokens tagged with the named entity, if there is one for that
// token. This code is not particularly efficient, but it's easy to follow.
for (int t = 0; t < ts.tokens.length; t++) {
// Check if this token is the start of a named entity
for (Span span : entities) {
if (span.getStart() == t) {
System.out.print("<" + span.getType().toUpperCase() + "> ");
}
}
// Show the token
System.out.print(ts.tokens[t] + " ");
// Check if this token is the end of a named entity
Console
Problems
Debug Shell
Search
<terminated> Main (2) [Java Application] /Library/Java/JavaVirtualMachines/jdk-9.0.4.jdk/Contents/Home/bin/java (Feb 3, 2019, 2:35:09 PM)
That is considerably below the level of around <MONEY> 160 pence </MONEY> widely reported before announcement of the deal , and will reduce outst
The company said a further 3.7 billion pounds (<MONEY> $ 5.8 billion </MONEY>) of debt would be converted into new financial instruments and ex
If they choose not to take up free warrants entitling them to subscribe to this , Eurotunnel said shareholders ' interests may be reduced further
Eurotunnel 's shares , which were suspended last week at <MONEY> 113.5 pence </MONEY> ahead of Monday 's announcement , should resume trading on
Writable
Smart Insert
1 : 1