

COMP 6751 Natural Language Analysis

Project 1 Report 2 (Demo)

Student: Yixuan Li 40079830

Table of Contents

I. Input and Outputs.....	2
1. Test case 1: “reuters/training/267”	2
Test case 1 result explanation:.....	3
2. Test case 2: “reuters/training/279”	5
Test case 2 result explanation:.....	6
II. Interesting case: “reuters/training/6”	8

Expectations of originality:

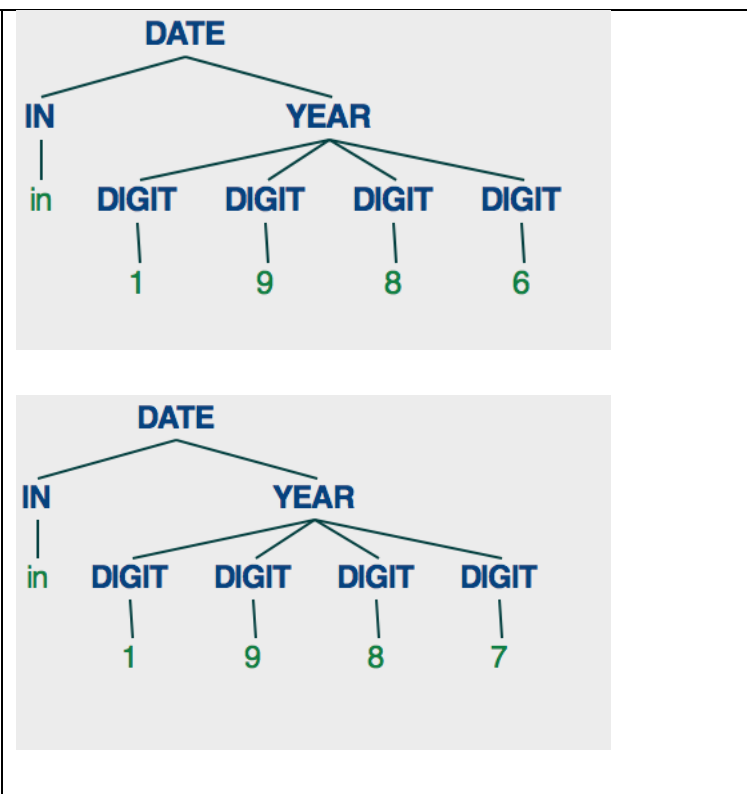
I, student 40079830, certify that this submission is my original work and meets the Faculty’s Expectations of Originality.

Date: October 4, 2020

I. Input and Outputs

1. Test case 1: "reuters/training/267"

	Input file and content	Output on console and parse tree diagram	
reuters fileid	'training/267'	<p>Tokenization results: ['INDONESIA', 'UNLIKELY', 'TO', 'IMPORT', 'PHILIPPINES', 'COPRA', ['Indonesia', 'is', 'unlikely', 'to', 'import', 'copra', 'from', 'the', 'Philippines', 'in', '1987', 'after', 'importing', '30,000', 'tonnes', 'in', '1986', ',', 'the', 'U.S.', 'Embassy', '"', 's', 'annual', 'agriculture', 'report', 'said', '.', 'The', 'report', 'said', 'the', '31', 'pct', 'devaluation', 'of', 'the', 'Indonesian', 'rupiah', ',', 'an', 'increase', 'in', 'import', 'duties', 'on', 'copra', 'and', 'increases', 'in', 'the', 'price', 'of', 'Philippines', 'copra', 'have', 'reduced', 'the', 'margin', 'between', 'prices', 'in', 'the', 'two', 'countries', ',', 'Indonesia', '"', 's', 'copra', 'production', 'is', 'forecast', 'at', '1.32', 'mln', 'tonnes', 'in', 'calendar', '1987', ',', 'up', 'from', '1.30', 'mln', 'tonnes', 'in', '1986', '.']</p> <p>-----</p> <p>Sentences splitting results: ["Indonesia is unlikely to import copra from the Philippines in 1987 after importing 30,000 tonnes in 1986, the U.S. Embassy's annual agriculture report said.", 'The report said the 31 pct devaluation of the Indonesian rupiah, an increase in import duties on copra and increases in the price of Philippines copra have reduced the margin between prices in the two countries.', 'Indonesia's copra production is forecast at 1.32 mln tonnes in calendar 1987, up from 1.30 mln tonnes in 1986.']</p> <p>-----</p> <p>Part-of-speech tagging results: [[('Indonesia', 'NNP'), ('is', 'VBZ'), ('unlikely', 'JJ'), ('to', 'TO'), ('import', 'VB'), ('copra', 'NN'), ('from', 'IN'), ('the', 'DT'), ('Philippines', 'NNPS'), ('in', 'IN'), ('1987', 'CD'), ('after', 'IN'), ('importing', 'VBG'), ('30,000', 'CD'), ('tonnes', 'NNS'), ('in', 'IN'), ('1986', 'CD'), (',', ','), ('the', 'DT'), ('U.S.', 'NNP'), ('Embassy', 'NNP'), ('"', 'POS'), ('s', 'POS'), ('annual', 'JJ'), ('agriculture', 'NN'), ('report', 'NN'), ('said', 'VBD'), (',', ','), (',', ',')], [('The', 'DT'), ('report', 'NN'), ('said', 'VBD'), ('the', 'DT'), ('31', 'CD'), ('pct', 'JJ'), ('devaluation', 'NN'), ('of', 'IN'), ('the', 'DT'), ('Indonesian', 'NNP'), ('rupiah', 'NN'), (',', ','), ('an', 'DT'), ('increase', 'NN'), ('in', 'IN'), ('import', 'JJ'), ('duties', 'NNS'), ('on', 'IN'), ('copra', 'NN'), ('and', 'CC'), ('increases', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('price', 'NN'), ('of', 'IN'), ('Philippines', 'NNPS'), ('copra', 'NNS'), ('have', 'VBP'), ('reduced', 'VBN'), ('the', 'DT'), ('margin', 'NN'), ('between', 'IN'), ('prices', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('two', 'CD'), ('countries', 'NNS'), (',', ',')], [('Indonesia', 'NNP'), ('"', 'POS'), ('s', 'POS'), ('copra', 'NN'), ('production', 'NN'), ('is', 'VBZ'), ('forecast', 'VBN'), ('at', 'IN'), ('1.32', 'CD'), ('mln', 'NN'), ('tonnes', 'NNS'), ('in', 'IN'), ('calendar', 'NN'), ('1987', 'CD'), (',', ','), ('up', 'RB'), ('from', 'IN'), ('1.30', 'CD'), ('mln', 'NN'), ('tonnes', 'NNS'), ('in', 'IN'), ('1986', 'CD'), (',', ',')]]</p> <p>-----</p> <p>Measured entity detection: ['30,000 tonnes', 'two countries', '1.32 mln tonnes', '1.30 mln tonnes']</p> <p>-----</p> <p>Date recognition: {'in 1987', 'in 1986'}</p> <p>-----</p> <p>Date parsing: (DATE (IN in) (YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 6))) (DATE (IN in) (YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 7)))</p>	Console Output

file content	<p>INDONESIA UNLIKELY TO IMPORT PHILIPPINES COPRA</p> <p>Indonesia is unlikely to import copra from the Philippines in 1987 after importing 30,000 tonnes in 1986, the U.S. Embassy's annual agriculture report said.</p> <p>The report said the 31 pct devaluation of the Indonesian rupiah, an increase in import duties on copra and increases in the price of Philippines copra have reduced the margin between prices in the two countries.</p> <p>Indonesia's copra production is forecast at 1.32 mln tonnes in calendar 1987, up from 1.30 mln tonnes in 1986.</p>		Parse Tree Diagram
--------------	--	---	--------------------

Test case 1 result explanation:

1) Sentence splitting

There are 3 sentences in total in “training/267” raw text in body content, and the result of sentence splitting matches the 3 sentences, showing below (copied from the table above) :

Sentences splitting results:

["Indonesia is unlikely to import copra from the Philippines in 1987 after importing 30,000 tonnes in 1986, the U.S. Embassy's annual agriculture report said."],

'The report said the 31 pct devaluation of the Indonesian rupiah, an increase in import duties on copra and increases in the price of Philippines copra have reduced the margin between prices in the two countries.',

"Indonesia's copra production is forecast at 1.32 mln tonnes in calendar 1987, up from 1.30 mln tonnes in 1986."]

2) Tokenization

The result of tokenization should be a list of separate words (tokens) and maintaining the numbers, abbreviations, percentages etc. not to be split. The results is shown below (copied from the table above) :

Tokenization results:

['INDONESIA', 'UNLIKELY', 'TO', 'IMPORT', 'PHILIPPINES', 'COPRA']

['Indonesia', 'is', 'unlikely', 'to', 'import', 'copra', 'from', 'the', 'Philippines', 'in', '1987', 'after', 'importing', '30,000', 'tonnes', 'in', '1986', ',', 'the', 'U.S.', 'Embassy', '"', 's', 'annual', 'agriculture', 'report', 'said', '.', 'The', 'report', 'said', 'the', '31', 'pct', 'devaluation', 'of', 'the', 'Indonesian', 'rupiah', ',', 'an', 'increase', 'in', 'import', 'duties', 'on', 'copra', 'and', 'increases', 'in', 'the', 'price', 'of', 'Philippines', 'copra', 'have', 'reduced', 'the', 'margin', 'between', 'prices', 'in', 'the', 'two', 'countries', '.', 'Indonesia', '"', 's', 'copra', 'production', 'is', 'forecast', 'at', '1.32', 'mln', 'tonnes', 'in', 'calendar', '1987', ',', 'up', 'from', '1.30', 'mln', 'tonnes', 'in', '1986', '.']

As we can see, the words like “30,000”, “U.S.”, “1.30” are not split. (number normalization)

3) POS tagging

The results of part-of-speech tagging is a list of lists, where each list is the POS tags for each sentence. The result is shown below (copied from the table above) :

Part-of-speech tagging results:

```
[[('Indonesia', 'NNP'), ('is', 'VBZ'), ('unlikely', 'JJ'), ('to', 'TO'), ('import', 'VB'), ('copra', 'NN'), ('from', 'IN'), ('the', 'DT'), ('Philippines', 'NNPS'), ('in', 'IN'), ('1987', 'CD'), ('after', 'IN'), ('importing', 'VBG'), ('30,000', 'CD'), ('tonnes', 'NNS'), ('in', 'IN'), ('1986', 'CD'), (',', ','), ('the', 'DT'), ('U.S.', 'NNP'), ('Embassy', 'NNP'), ('"', 'POS'), ('annual', 'JJ'), ('agriculture', 'NN'), ('report', 'NN'), ('said', 'VBD'), ('.', '.'), (['The', 'DT'], ('report', 'NN'), ('said', 'VBD'), ('the', 'DT'), ('31', 'CD'), ('pct', 'JJ'), ('devaluation', 'NN'), ('of', 'IN'), ('the', 'DT'), ('Indonesian', 'NNP'), ('rupiah', 'NN'), (',', ','), ('an', 'DT'), ('increase', 'NN'), ('in', 'IN'), ('import', 'JJ'), ('duties', 'NNS'), ('on', 'IN'), ('copra', 'NN'), ('and', 'CC'), ('increases', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('price', 'NN'), ('of', 'IN'), ('Philippines', 'NNPS'), ('copra', 'NNS'), ('have', 'VBP'), ('reduced', 'VBN'), ('the', 'DT'), ('margin', 'NN'), ('between', 'IN'), ('prices', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('two', 'CD'), ('countries', 'NNS'), ('.', '.'), (['Indonesia', 'NNP'], ('"', 'POS'), ('copra', 'NN'), ('production', 'NN'), ('is', 'VBZ'), ('forecast', 'VBN'), ('at', 'IN'), ('1.32', 'CD'), ('mln', 'NN'), ('tonnes', 'NNS'), ('in', 'IN'), ('calendar', 'NN'), ('1987', 'CD'), (',', ','), ('up', 'RB'), ('from', 'IN'), ('1.30', 'CD'), ('mln', 'NN'), ('tonnes', 'NNS'), ('in', 'IN'), ('1986', 'CD'), ('.', '.')]]
```

4) number normalization

(see the result in 2nd step tokenization)

5) measured entity detection

The result of measured entity is shown below (copied from the table):

Measured entity detection:

['30,000 tonnes', 'two countries', '1.32 mln tonnes', '1.30 mln tonnes']

6) date recognition

In the original raw text, there are two valid dates which are “in 1986” and “in 1987”.

The result of date recognition shown below matches the expectation.

Date recognition:

{'in 1987', 'in 1986'}

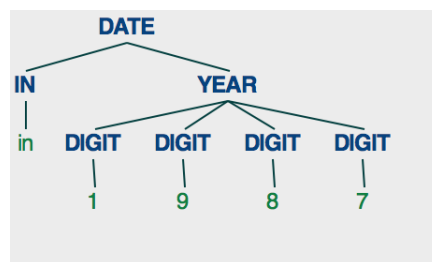
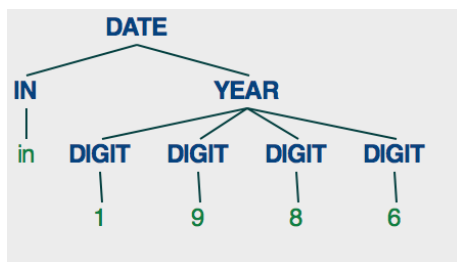
7) date parsing

Date parser parses the dates outputted from date recognition, and prints out the parse tree.

Date parsing:

(DATE (IN in) (YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 6)))

(DATE (IN in) (YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 7)))



2. Test case 2: "reuters/training/279"

	Input file and content	Output on console and parse tree diagram	
fileid	'training/279'	<p>Tokenization results:</p> <p>['JAPAN', 'S', 'NTT', 'FORECASTS', 'PROFITS', 'FALL', 'IN', '1987', '88']</p> <p>['<', 'Nippon', 'Telegraph', 'and', 'Telephone', 'Corp', '>', '(', 'NTT', ')', 'expects', 'its', 'profits', 'to', 'fall', 'to', '328', 'billion', 'yen', 'in', 'the', 'year', 'ending', 'March', '31', ',', '1988', 'from', 'a', 'projected', '348', 'billion', 'this', 'year', ',', 'the', 'company', 'said', ',', 'Total', 'sales', 'for', 'the', 'same', 'period', 'are', 'expected', 'to', 'rise', 'to', '5,506', 'billion', 'yen', 'from', 'a', 'projected', '5,328', 'billion', 'this', 'year', ',', 'NTT', 'said', 'in', 'a', 'business', 'operations', 'plan', 'submitted', 'to', 'the', 'Post', 'and', 'Telecommunications', 'Ministry', ',', 'NTT', 'said', 'it', 'plans', 'to', 'make', 'capital', 'investments', 'of', '1,770', 'billion', 'yen', 'in', '1987', '88', ',', 'including', '109', 'billion', 'for', 'research', 'and', 'development', ',', 'as', 'against', 'a', 'total', 'of', '1,600', 'billion', 'this', 'year', ',', 'An', 'NTT', 'spokesman', 'said', 'increased', 'competition', 'from', 'new', 'entrants', 'to', 'the', 'telecommunications', 'field', 'and', 'the', 'effect', 'of', 'a', 'sales', 'tax', 'scheduled', 'to', 'be', 'introduced', 'next', 'January', ',', 'were', 'the', 'major', 'factors', 'behind', 'the', 'projected', 'decrease', 'in', 'profits', ',', 'The', 'Japanese', 'telecommunications', 'industry', 'was', 'deregulated', 'in', '1985', '']</p> <p>-----</p> <p>Sentences splitting results:</p> <p>['<Nippon Telegraph and Telephone Corp> (NTT) expects its profits to fall to 328 billion yen in the year ending March 31, 1988 from a projected 348 billion this year, the company said.', 'Total sales for the same period are expected to rise to 5,506 billion yen from a projected 5,328 billion this year, NTT said in a business operations plan submitted to the Post and Telecommunications Ministry.', 'NTT said it plans to make capital investments of 1,770 billion yen in 1987/88, including 109 billion for research and development, as against a total of 1,600 billion this year.', 'An NTT spokesman said increased competition from new entrants to the telecommunications field and the effect of a sales tax scheduled to be introduced next January, were the major factors behind the projected decrease in profits.', 'The Japanese telecommunications industry was deregulated in 1985.']</p> <p>-----</p> <p>Part-of-speech tagging results:</p> <p>[['<', 'JJ'], ('Nippon', 'NNP'), ('Telegraph', 'NNP'), ('and', 'CC'), ('Telephone', 'NNP'), ('Corp', 'NNP'), ('>', 'NNP'), ('(', 'P'), ('NTT', 'NNP'), (')', 'P'), ('expects', 'VBZ'), ('its', 'PRP\$'), ('profits', 'NNS'), ('to', 'TO'), ('fall', 'VB'), ('to', 'TO'), ('328', 'CD'), ('billion', 'CD'), ('yen', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('year', 'NN'), ('ending', 'VBG'), ('March', 'NNP'), ('31', 'CD'), (',', ','), ('1988', 'CD'), ('from', 'IN'), ('a', 'DT'), ('projected', 'VBN'), ('348', 'CD'), ('billion', 'CD'), ('this', 'DT'), ('year', 'NN'), (',', ','), ('the', 'DT'), ('company', 'NN'), ('said', 'VBD'), (',', ','), [(['Total', 'JJ'], ('sales', 'NNS'), ('for', 'IN'), ('the', 'DT'), ('same', 'JJ'), ('period', 'NN'), ('are', 'VBP'), ('expected', 'VBN'), ('to', 'TO'), ('rise', 'VB'), ('to', 'TO'), ('5,506', 'CD'), ('billion', 'CD'), ('yen', 'NNS'), ('from', 'IN'), ('a', 'DT'), ('projected', 'VBN'), ('5,328', 'CD'), ('billion', 'CD'), ('this', 'DT'), ('year', 'NN'), (',', ','), ('NTT', 'NNP'), ('said', 'VBD'), ('in', 'IN'), ('a', 'DT'), ('business', 'NN'), ('operations', 'NNS'), ('plan', 'NN'), ('submitted', 'VBN'), ('to', 'TO'), ('the', 'DT'), ('Post', 'NNP'), ('and', 'CC'), ('Telecommunications', 'NNP'), ('Ministry', 'NNP'), (',', ','), [(['NTT', 'NNP'), ('said', 'VBD'), ('it', 'PRP'), ('plans', 'VBZ'), ('to', 'TO'), ('make', 'VB'), ('capital', 'NN'), ('investments', 'NNS'), ('of', 'IN'), ('1,770', 'CD'), ('billion', 'CD'), ('yen', 'NNS'), ('in', 'IN'), ('1987', 'CD'), ('88', 'CD'), (',', ','), ('including', 'VBG'), ('109', 'CD'), ('billion', 'CD'), ('for', 'IN'), ('research', 'NN'), ('and', 'CC'), ('development', 'NN'), (',', ','), ('as', 'IN'), ('against', 'IN'), ('a', 'DT'), ('total', 'NN'), ('of', 'IN'), ('1,600', 'CD'), ('billion', 'CD'), ('this', 'DT'), ('year', 'NN'), (',', ','), [(['An', 'DT'), ('NTT', 'NNP'), ('spokesman', 'NN'), ('said', 'VBD'), ('increased', 'VBN'), ('competition', 'NN'), ('from', 'IN'), ('new', 'JJ'), ('entrants', 'NNS'), ('to', 'TO'), ('the', 'DT'), ('telecommunications', 'NNS'), ('field', 'NN'), ('and', 'CC'), ('the', 'DT'), ('effect', 'NN'), ('of', 'IN'), ('a', 'DT'), ('sales', 'NNS'), ('tax', 'NN'), ('scheduled', 'VBN'), ('to', 'TO'), ('be', 'VB'), ('introduced', 'VBN'), ('next', 'JJ'), ('January', 'NNP'), (',', ','), ('were', 'VBD'), ('the', 'DT'), ('major', 'JJ'), ('factors', 'NNS'), ('behind', 'IN'), ('the', 'DT'), ('projected', 'JJ'), ('decrease', 'NN'), ('in', 'IN'), ('profits', 'NNS'), (',', ','), [(['The', 'DT'), ('Japanese', 'JJ'), ('telecommunications', 'NNS'), ('industry', 'NN'), ('was', 'VBD'), ('deregulated', 'VBN'), ('in', 'IN'), ('1985', 'CD'), (',', ','),]]]]</p> <p>-----</p> <p>Measured entity detection:</p> <p>['328 billion yen', '348 billion', '5,506 billion yen', '5,328 billion', '1,770 billion yen', '109 billion', '1,600 billion']</p> <p>-----</p> <p>Date recognition:</p> <p>{ 'in 1987', 'in 1985', 'March 31, 1988' }</p> <p>-----</p> <p>Date parsing:</p> <p>(DATE (IN in) (YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 7)))</p> <p>(DATE (IN in) (YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 5)))</p> <p>(DATE</p> <p>(MONTH_STR March)</p> <p>(DAY (DIGIT 3) (DIGIT 1))</p> <p>(SEP .)</p> <p>(YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 8)))</p>	Console Output

file content	<p>JAPAN'S NTT FORECASTS PROFITS FALL IN 1987/88</p> <p>&lt;Nippon Telegraph and Telephone Corp></p> <p>(NTT) expects its profits to fall to 328 billion yen in the year ending March 31, 1988 from a projected 348 billion this year, the company said.</p> <p>Total sales for the same period are expected to rise to 5,506 billion yen from a projected 5,328 billion this year, NTT said in a business operations plan submitted to the Post and Telecommunications Ministry.</p> <p>NTT said it plans to make capital investments of 1,770 billion yen in 1987/88, including 109 billion for research and development, as against a total of 1,600 billion this year.</p> <p>An NTT spokesman said increased competition from new entrants to the telecommunications field and the effect of a sales tax scheduled to be introduced next January, were the major factors behind the projected decrease in profits.</p> <p>The Japanese telecommunications industry was deregulated in 1985.</p>	<div data-bbox="667 136 1297 510"> <pre> graph TD DATE[DATE] --> IN[IN] DATE --> YEAR[YEAR] IN --> in[in] YEAR --> DIGIT1[DIGIT] YEAR --> DIGIT2[DIGIT] YEAR --> DIGIT3[DIGIT] YEAR --> DIGIT4[DIGIT] DIGIT1 --> 1[1] DIGIT2 --> 9[9] DIGIT3 --> 8[8] DIGIT4 --> 7[7] </pre> </div> <div data-bbox="667 555 1311 918"> <pre> graph TD DATE[DATE] --> IN[IN] DATE --> YEAR[YEAR] IN --> in[in] YEAR --> DIGIT1[DIGIT] YEAR --> DIGIT2[DIGIT] YEAR --> DIGIT3[DIGIT] YEAR --> DIGIT4[DIGIT] DIGIT1 --> 1[1] DIGIT2 --> 9[9] DIGIT3 --> 8[8] DIGIT4 --> 5[5] </pre> </div> <div data-bbox="667 958 1425 1252"> <pre> graph TD DATE[DATE] --> MONTH_STR[MONTH_STR] DATE --> DAY[DAY] DATE --> SEP[SEP] DATE --> YEAR[YEAR] MONTH_STR --> March[March] DAY --> DIGIT1[DIGIT] DAY --> DIGIT2[DIGIT] DIGIT1 --> 3[3] DIGIT2 --> 1[1] SEP --> comma[,] YEAR --> DIGIT3[DIGIT] YEAR --> DIGIT4[DIGIT] YEAR --> DIGIT5[DIGIT] YEAR --> DIGIT6[DIGIT] DIGIT3 --> 1[1] DIGIT4 --> 9[9] DIGIT5 --> 8[8] DIGIT6 --> 8[8] </pre> </div>	Parse Tree Diagram
--------------	---	--	--------------------

Test case 2 result explanation:

1) Sentence splitting

There are 5 sentences in “training/279” raw text in body content, and the result of sentence splitting matches. The result is shown below (copied from the table):

Sentences splitting results:

['<Nippon Telegraph and Telephone Corp> (NTT) expects its profits to fall to 328 billion yen in the year ending March 31, 1988 from a projected 348 billion this year, the company said.', 'Total sales for the same period are expected to rise to 5,506 billion yen from a projected 5,328 billion this year, NTT said in a business operations plan submitted to the Post and Telecommunications Ministry.', 'NTT said it plans to make capital investments of 1,770 billion yen in 1987/88, including 109 billion for research and development, as against a total of 1,600 billion this year.', 'An NTT spokesman said increased competition from new entrants to the telecommunications field and the effect of a sales tax scheduled to be introduced next January, were the major factors behind the projected decrease in profits.', 'The Japanese telecommunications industry was deregulated in 1985.']

2) Tokenization

The result of tokenization doesn't split numbers that includes commas or periods such as "5,506", "5,328", and the possessive ending "'s" or "'S". The result is shown below (copied from the table):

Tokenization results:

```
['JAPAN', '"', 'S', 'NTT', 'FORECASTS', 'PROFITS', 'FALL', 'IN', '1987', '88']  
['<', 'Nippon', 'Telegraph', 'and', 'Telephone', 'Corp', '>', '(', 'NTT', ')', 'expects', 'its', 'profits', 'to', 'fall',  
'to', '328', 'billion', 'yen', 'in', 'the', 'year', 'ending', 'March', '31', ',', '1988', 'from', 'a', 'projected', '348',  
'billion', 'this', 'year', ',', 'the', 'company', 'said', '.', 'Total', 'sales', 'for', 'the', 'same', 'period', 'are',  
'expected', 'to', 'rise', 'to', '5,506', 'billion', 'yen', 'from', 'a', 'projected', '5,328', 'billion', 'this', 'year', ',',  
'NTT', 'said', 'in', 'a', 'business', 'operations', 'plan', 'submitted', 'to', 'the', 'Post', 'and',  
'Telecommunications', 'Ministry', '.', 'NTT', 'said', 'it', 'plans', 'to', 'make', 'capital', 'investments', 'of',  
'1,770', 'billion', 'yen', 'in', '1987', '88', ',', 'including', '109', 'billion', 'for', 'research', 'and',  
'development', ',', 'as', 'against', 'a', 'total', 'of', '1,600', 'billion', 'this', 'year', '.', 'An', 'NTT', 'spokesman',  
'said', 'increased', 'competition', 'from', 'new', 'entrants', 'to', 'the', 'telecommunications', 'field', 'and',  
'the', 'effect', 'of', 'a', 'sales', 'tax', 'scheduled', 'to', 'be', 'introduced', 'next', 'January', ',', 'were', 'the',  
'major', 'factors', 'behind', 'the', 'projected', 'decrease', 'in', 'profits', '.', 'The', 'Japanese',  
'telecommunications', 'industry', 'was', 'deregulated', 'in', '1985', '.']
```

3) POS tagging

The part-of-speech result is shown below:

Part-of-speech tagging results:

```
[[['<', 'JJ'), ('Nippon', 'NNP'), ('Telegraph', 'NNP'), ('and', 'CC'), ('Telephone', 'NNP'), ('Corp', 'NNP'), ('>', 'NNP'), ('(', 'NNP'), ('NTT', 'NNP'), (')', 'NNP'), ('expects', 'VBZ'), ('its', 'PRP$'), ('profits', 'NNS'), ('to', 'TO'), ('fall', 'VB'), ('to', 'TO'), ('328', 'CD'), ('billion', 'CD'), ('yen', 'NNS'), ('in', 'IN'), ('the', 'DT'), ('year', 'NN'), ('ending', 'VBG'), ('March', 'NNP'), ('31', 'CD'), (',', 'NNP'), ('1988', 'CD'), ('from', 'IN'), ('a', 'DT'), ('projected', 'VBN'), ('348', 'CD'), ('billion', 'CD'), ('this', 'DT'), ('year', 'NN'), (',', 'NNP'), ('the', 'DT'), ('company', 'NN'), ('said', 'VBD'), (',', 'NNP'), (['Total', 'JJ'), ('sales', 'NNS'), ('for', 'IN'), ('the', 'DT'), ('same', 'JJ'), ('period', 'NN'), ('are', 'VBP'), ('expected', 'VBN'), ('to', 'TO'), ('rise', 'VB'), ('to', 'TO'), ('5,506', 'CD'), ('billion', 'CD'), ('yen', 'NNS'), ('from', 'IN'), ('a', 'DT'), ('projected', 'VBN'), ('5,328', 'CD'), ('billion', 'CD'), ('this', 'DT'), ('year', 'NN'), (',', 'NNP'), ('NTT', 'NNP'), ('said', 'VBD'), ('in', 'IN'), ('a', 'DT'), ('business', 'NN'), ('operations', 'NNS'), ('plan', 'NN'), ('submitted', 'VBN'), ('to', 'TO'), ('the', 'DT'), ('Post', 'NNP'), ('and', 'CC'), ('Telecommunications', 'NNP'), ('Ministry', 'NNP'), (',', 'NNP'), (['NTT', 'NNP'), ('said', 'VBD'), ('it', 'PRP'), ('plans', 'VBZ'), ('to', 'TO'), ('make', 'VB'), ('capital', 'NN'), ('investments', 'NNS'), ('of', 'IN'), ('1,770', 'CD'), ('billion', 'CD'), ('yen', 'NNS'), ('in', 'IN'), ('1987', 'CD'), ('88', 'CD'), (',', 'NNP'), ('including', 'VBG'), ('109', 'CD'), ('billion', 'CD'), ('for', 'IN'), ('research', 'NN'), ('and', 'CC'), ('development', 'NN'), (',', 'NNP'), ('as', 'IN'), ('against', 'IN'), ('a', 'DT'), ('total', 'NN'), ('of', 'IN'), ('1,600', 'CD'), ('billion', 'CD'), ('this', 'DT'), ('year', 'NN'), (',', 'NNP'), (['An', 'DT'), ('NTT', 'NNP'), ('spokesman', 'NN'), ('said', 'VBD'), ('increased', 'VBN'), ('competition', 'NN'), ('from', 'IN'), ('new', 'JJ'), ('entrants', 'NNS'), ('to', 'TO'), ('the', 'DT'), ('telecommunications', 'NNS'), ('field', 'NN'), ('and', 'CC'), ('the', 'DT'), ('effect', 'NN'), ('of', 'IN'), ('a', 'DT'), ('sales', 'NNS'), ('tax', 'NN'), ('scheduled', 'VBN'), ('to', 'TO'), ('be', 'VB'), ('introduced', 'VBN'), ('next', 'JJ'), ('January', 'NNP'), (',', 'NNP'), ('were', 'VBD'), ('the', 'DT'), ('major', 'JJ'), ('factors', 'NNS'), ('behind', 'IN'), ('the', 'DT'), ('projected', 'JJ'), ('decrease', 'NN'), ('in', 'IN'), ('profits', 'NNS'), (',', 'NNP'), (['The', 'DT'), ('Japanese', 'JJ'), ('telecommunications', 'NNS'), ('industry', 'NN'), ('was', 'VBD'), ('deregulated', 'VBN'), ('in', 'IN'), ('1985', 'CD'), (',', 'NNP)]]
```

4) number normalization

(see the result in 2nd step tokenization)

5) measured entity detection

All the unit of measurement entities in raw text are detected without false positive.

The result is shown below:

Measured entity detection:

```
['328 billion yen', '348 billion', '5,506 billion yen', '5,328 billion', '1,770 billion yen', '109 billion', '1,600 billion']
```

6) date recognition

There are three dates in raw text, and all of them are recognized. However, one was false positive since the date format ("YYYY/YY", e.g. 1987/88) isn't included in the grammar.

The result is shown below:

Date recognition:

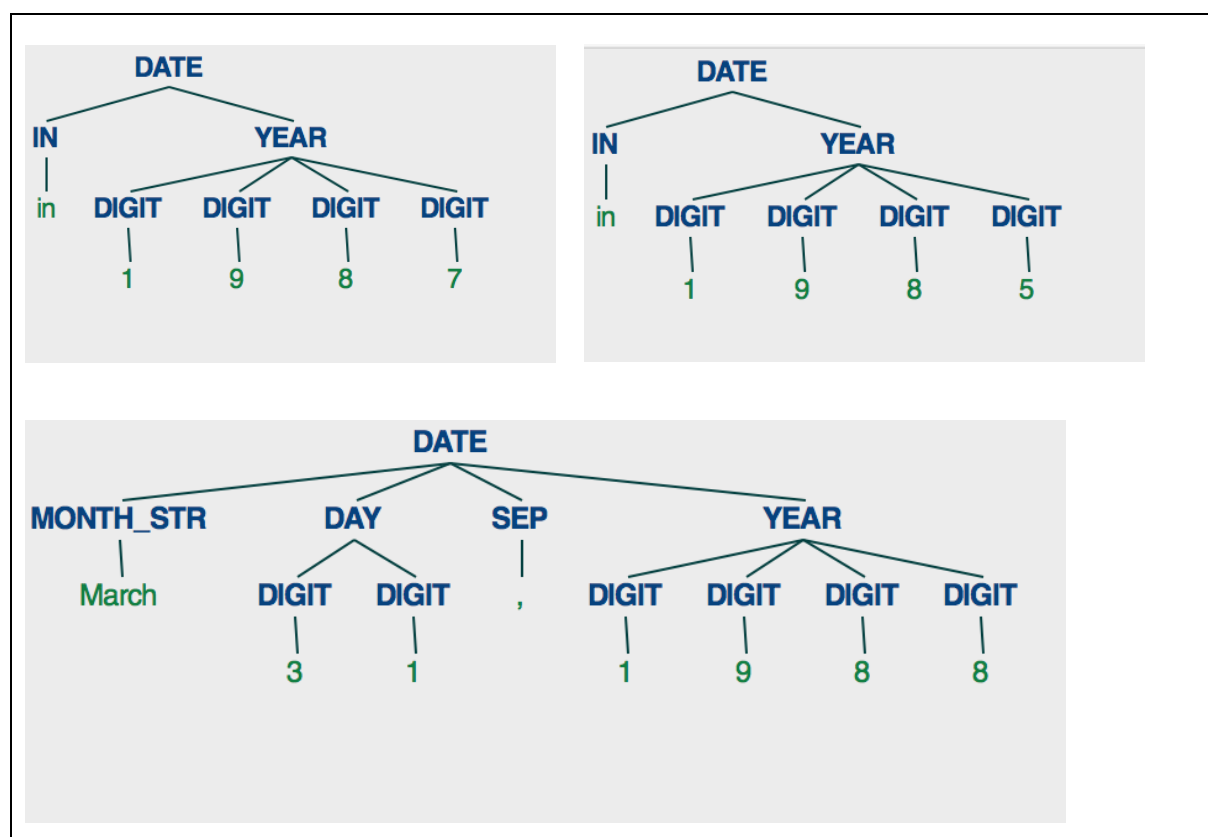
```
{in 1987', 'March 31, 1988', 'in 1985'}
```

The **false positive one** should be 'in 1987/88'

The reason is that in my grammar, the year was defined as 4 digits, "88" in the text also indicates the last two digits of a year, so it was not recognized.

7) date parsing

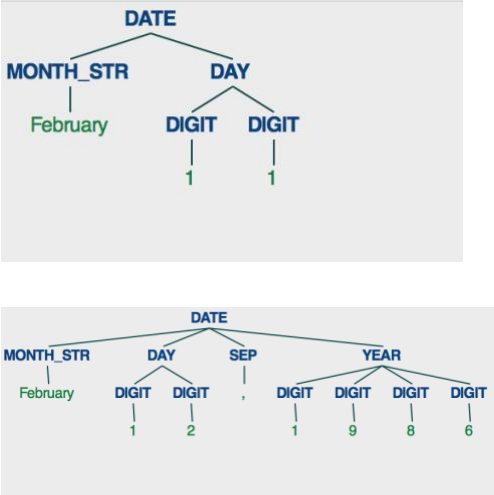
Since there is one false positive during date recognition, hence the corresponding one was also wrongly parsed. The other two dates are parsed correctly.



II. Interesting case: "reuters/training/6"

In grocery figures, it is common to have month followed by numbers indicating the sale performance in that month. Therefore, it is easy to mislabel the chunk as date. In the program, my grammar performs well on this scenarios. I will take "reuters/training/6" as an example, and for saving the space, I will only attach the parse trees of detected dates and omit the output on tokenization, sentence splitting, POS tagging and entity detection.

	Input file and content	Output on console and parse tree diagram	
fileid	'training/6'	<p>Tokenization results:(omitted)..... -----</p> <p>Sentences splitting results:(omitted)..... -----</p> <p>Part-of-speech tagging results:(omitted)..... -----</p> <p>Measured entity detection:(omitted)..... -----</p> <p>Date recognition: {'February 11', 'February 12 , 1986'} -----</p> <p>Date parsing: (DATE (MONTH_STR February) (DAY (DIGIT 1) (DIGIT 1))) (DATE (MONTH_STR February) (DAY (DIGIT 1) (DIGIT 2)) (SEP ,) (YEAR (DIGIT 1) (DIGIT 9) (DIGIT 8) (DIGIT 6)))</p>	Console Output

file content	<p>ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS</p> <p>Argentine grain board figures show crop registrations of grains, oilseeds and their products to February 11, in thousands of tonnes, showing those for future shipments month, 1986/87 total and 1985/86 total to February 12, 1986, in brackets:</p> <p>Bread wheat prev 1,655.8, Feb 872.0, March 164.6, total 2,692.4 (4,161.0).</p> <p>Maize Mar 48.0, total 48.0 (nil).</p> <p>Sorghum nil (nil)</p> <p>Oilseed export registrations were:</p> <p>Sunflowerseed total 15.0 (7.9)</p> <p>Soybean May 20.0, total 20.0 (nil)</p> <p>The board also detailed export registrations for subproducts, as follows,</p> <p>SUBPRODUCTS</p> <p>Wheat prev 39.9, Feb 48.7, March 13.2, Apr 10.0, total 111.8 (82.7) .</p> <p>Linseed prev 34.8, Feb 32.9, Mar 6.8, Apr 6.3, total 80.8 (87.4).</p> <p>Soybean prev 100.9, Feb 45.1, Mar nil, Apr nil, May 20.0, total 166.1 (218.5).</p> <p>Sunflowerseed prev 48.6, Feb 61.5, Mar 25.1, Apr 14.5, total 149.8 (145.3).</p> <p>Vegetable oil registrations were :</p> <p>Sunoil prev 37.4, Feb 107.3, Mar 24.5, Apr 3.2, May nil, Jun 10.0, total 182.4 (117.6).</p> <p>Linoil prev 15.9, Feb 23.6, Mar 20.4, Apr 2.0, total 61.8, (76.1).</p> <p>Soybean oil prev 3.7, Feb 21.1, Mar nil, Apr 2.0, May 9.0, Jun 13.0, Jul 7.0, total 55.8 (33.7). REUTER</p>		Parse Tree Diagram
--------------	--	--	--------------------

In the file content cell above, I highlighted the misleading parts as **yellow** and the correct date parts as **green**. And as we can see in the Parse Tree Diagram cell, the grammar and program only detect the green parts as dates correctly.

Limitations

In the list of recognized dates above, there are two omitted (false negative) which are “1986/87”, “1985/86”. The reason is that it is not easy to differentiate a year or a year’s last two digits and a common integer. Therefore, for the date which only contains a year (or last two digits), my date recognizer grammar only labels it as a date if there is a preposition (e.g. in) in front of it.