SYNTAX ASSIGNMENT
NLP Technology 2021

Deadline: May 14.
For this assignment you can obtain 34 points, divided as follows:

Part A (7)	Q1: 3 Q2: 1 Q3: 1 Q4: 1 Q5: 1
Part B (10)	Q6: 5 Q7: 5
Part C (12)	Q8: 4 Q9: 4 Q10: 4
Part D (5)	Q11: 5

Part A: Understanding the representation

The table below contains a sentence represented in conll format with the syntactic dependency information provided in Columns 5 and 6. Column 5 contains information about the head of each token, while Column 6 contains information about the dependency label, so information about the syntactic relation established between the token and its head.

Important: to answer the questions below you only need to use the information provided in the table, you don't have to run any parser or provide alternative syntactic analyses for the sentence.

Token #	Word	Lemma	POS	Dependency Head	Dependency Label
1	Meanwhile	meanwhile	RB	10	ADV
2	,	,	,	10	P
3	September	september	NNP	5	NMOD
4	housing	housing	NN	5	NMOD
5	starts	start	NNS	10	SBJ
6	,	,	,	5	P
7	due	due	JJ	5	APPO
8	Wednesday	wednesday	NNP	7	AMOD
9	,	,	,	5	P
10	are	be	VBP	0	ROOT
11	thought	think	VBN	10	VC
12	to	to	TO	11	OPRD
13	have	have	VB	12	IM
14	inched	inch	VBN	13	VC
15	upward	upward	RB	14	ADV
16	.	.	.	10	P

Q1-DRAW-GRAPH: Draw the dependency graph for this sentence. You can draw it manually and provide a picture of the tree. You can also draw it using some application, but it will take you much more time. It is not necessary.

You should follow the notational variant shown in Figure 1.

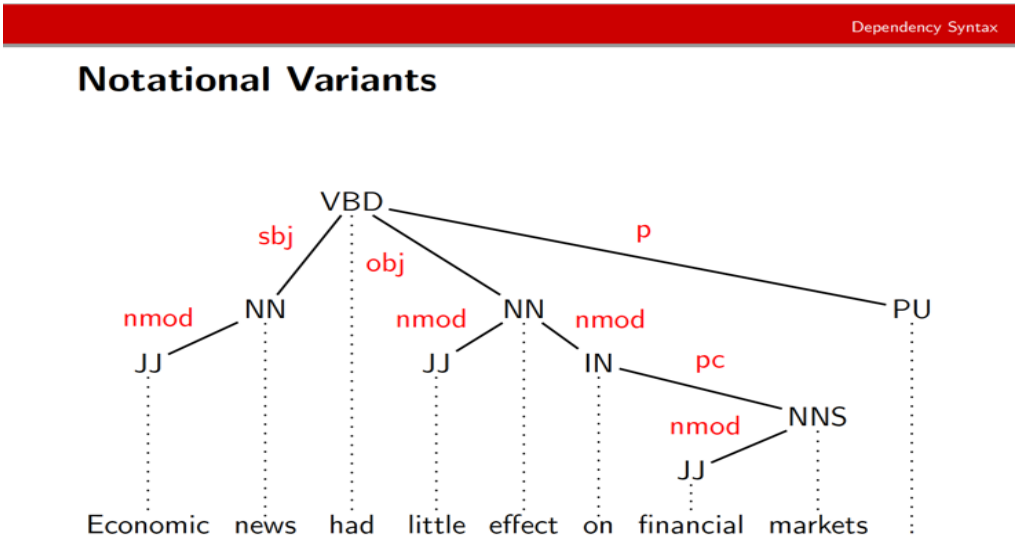

Figure 1. Notational variant to draw a dependency graph.

Q2-HEAD: According to the conll representation in Figure 1, which token is the head of the noun phrase “September housing starts”?

Q3-DEPENDENTS: List the Token # and Words that are the immediate dependents of node 10 (“are”).

Q4-SDP: Provide the shortest syntactic path that goes from token 15 “upward” to token 10 “are” and the length of the path. (See https://towardsdatascience.com/how-to-find-shortest-dependency-path-with-spacy-and-stanfordnlp-539d45d28239).

Example: in the tree represented in Figure 1 above, the shortest dependency path between "economic" and "little" has length 4 and is the following:

economic:JJ↑news:NN↑had:VBD↓effect:NN↓little:JJ

The arrows indicate whether the path goes from dependent to head (↑) or from head to dependent.

Q5-SDP: Provide the shortest dependency path that goes from token 1 (“meanwhile”) to token 13 (“have”) and the length.

Part B: Using a parser

For this exercise you are going to parse text with the dependency parser from the spaCy library. Information about the library is to be found here:

https://spacy.io/

You will use the dependency parser component of the pipeline:

https://spacy.io/api/dependencyparser

Parse the text provided in the file conllst.2017.trial.simple.conll which is provided on Canvas with the assignment.

Q6-PARSER-OUTPUT: Provide the output of the parser.

Q7-PARSER-DESCRIPTION: Describe the parser in max. 10 lines. Provide the references and resources that you used to obtain the information.

1. What is the input that the parser accepts?
2. What is the output that the parser provides?
3. Is it a data-driven parser or a grammar-based parser?
4. Is it a graph-based parser, a transition-based parser or something else?
5. Does the parser make predictions from right to left or from left to right?

6. Have you found evaluation scores of the parser? For which languages?
7. Can you indicate a positive aspect of the parser?
8. Can you indicate a negative aspect of the parser?

Part C: Error analysis

Once you have run the parser you can perform some error analysis. For that, you have to compare the **output that you obtained from the parser** with the **gold parse information** provided in the file conllst.2017.trial.simple.dep.conll .

Q8-PARSER2GOLD: Provide a table of the differences that you find between the system output and the gold labels (you could use the following columns to structure your answer: "Output / Gold Label / Difference"). You should check the following items:

1. Do all tokens in the sentence have the same token number as in the gold file or has the parser merged some tokens? This might happen if you don't tell the parser to use your tokenized text.
2. Has a ROOT node been found by the parser for each sentence?
3. For every token:
    a. Is the token assigned the same dependency head as in the gold file?
    b. Is the token assigned the same dependency label as in the gold file?
    c. Has the token the same dependents as in the gold file?

Q9-%ERRORS-POS: Calculate the percentage of errors for every **POS** type that occurs in the gold data. For example, if there are 30 tokens with NN in the gold corpus and 5 of them have a wrong dependency label or dependency head assigned by the system, the percentage of errors for NN is 16,6 %.

For this purpose, provide a table with an overview of all types of POS tags in the gold corpus (column 1) and the corresponding error rate (column 2). Structure the table in descending order.

Q10-%ERRORS-LABEL: Calculate the percentage of errors for all types of **dependency label** (SBJ, NMOD, etc.) that occur in the gold data. For example, if there are 100 tokens with SBJ label in the gold file and 10 of them have not been assigned by the system, then the percentage of error is 10 %.

Part D: Evaluating the parser

Given the following sentence with the gold parse information provided:

1	Thursday	thursday	NNP	3	NMOD
2	's	's	POS	1	SUFFIX
3	report	report	NN	10	SBJ
4	on	on	IN	3	NMOD
5	the	the	DT	9	NMOD
6	September	september	NNP	9	NMOD
7	consumer	consumer	NN	9	NMOD
8	price	price	NN	9	NMOD
9	index	index	NN	4	PMOD
10	is	be	VBZ	0	ROOT
11	expected	expect	VBN	10	VC
12	to	to	TO	11	OPRD
13	rise	rise	VB	12	IM
14	,	,	,	13	P
15	although	although	IN	13	ADV
16	not	not	RB	15	ADV
17	as	as	RB	16	AMOD
18	sharply	sharply	RB	16	AMOD
19	as	as	IN	16	AMOD
20	the	the	DT	23	NMOD
21	0.9	0.9	CD	22	NMOD
22	%	%	NN	23	NMOD
23	gain	gain	NN	19	PMOD
24	reported	report	VBN	23	APPO
25	Friday	friday	NNP	24	TMP
26	in	in	IN	24	LOC
27	the	the	DT	30	NMOD
28	producer	producer	NN	30	NMOD
29	price	price	NN	30	NMOD
30	index	index	NN	26	PMOD
31	.	.	.	10	P

And given the following output of a parser:

1	Thursday	thursday	NNP	3	NMOD
2	's	's	POS	**3**	SUFFIX
3	report	report	NN	10	SBJ
4	on	on	IN	3	NMOD
5	the	the	DT	9	NMOD
6	September	september	NNP	9	NMOD
7	consumer	consumer	NN	9	NMOD
8	price	price	NN	7	NN
9	index	index	NN	4	PMOD
10	is	be	VBZ	0	ROOT
11	expected	expect	VBN	10	VC
12	to	to	TO	11	OPRD
13	rise	rise	VB	11	IM
14	,	,	,	13	P
15	although	although	IN	13	ADV
16	not	not	RB	15	ADV
17	as	as	RB	15	PMOD
18	sharply	sharply	RB	16	AMOD
19	as	as	IN	16	NMOD
20	the	the	DT	23	NMOD
21	0.9	0.9	CD	22	NMOD
22	%	%	NN	21	NMOD
23	gain	gain	NN	19	PMOD
24	reported	report	VBN	23	APPO
25	Friday	friday	NNP	24	TMP
26	in	in	IN	25	NMOD
27	the	the	DT	30	NMOD
28	producer	producer	NN	30	NMOD
29	price	price	NN	30	NMOD
30	index	index	NN	26	NMOD
31	.	.	.	10	P

Q11-EVAL: Define and calculate the labeled and unlabeled attachment score, and the label accuracy score (see Martin and Jurafsky 2020, https://web.stanford.edu/~jurafsky/slp3/14, 14.6).

Include not only the end result, but also a formula and a way you arrived at your result so we can more easily understand your approach.

Submit a pdf file on canvas, as well as the file that contains the output of the parser. Please, use the following template to submit the answers:

Template for answers

Group name:
Student name:
Student name:
Student name:

Part A:

Q1: insert picture
Q2: answer
Q3: answer
Q4: answer
Q5: answer

Part B:

Q6: name of uploaded file that contains output of parser.
Q7: answer

Part C:

Q8: answer (list of errors)
Q9: answer
Q10: answer

Part D:

Q11:
Labeled attachment score: definition and calculation.
Unlabeled attachment score: definition and calculation.
Label accuracy score: definition and calculation.