Natural Language Lab
Tagus and Alameda
Part 2 – Lab 13

Goals of this lab:

- Practice symbolic representation of language
- Play with a corpus with the BIO notation
- Make some experiments regarding a sequence labeling task with a "classic" feature-based approach

1. (from Eisenstein book) Considering the G2 grammar (next figure), find the meaning representation of the phrase "likes a dog" and sentence "Alex likes every dog".

S	$\to$ NP VP	NP.sem@VP.sem
VP	$\to$ V_t NP	V_t.sem@NP.sem
VP	$\to$ V_i	V_i.sem
NP	$\to$ Det Nn	Det.sem@Nn.sem
NP	$\to$ Nnp	$\lambda P.P(\text{Nnp.sem})$
Det	$\to$ *a*	$\lambda P.\lambda Q.\exists x P(x) \wedge Q(x)$
Det	$\to$ *every*	$\lambda P.\lambda Q.\forall x(P(x) \Rightarrow Q(x))$
V_t	$\to$ *likes*	$\lambda P.\lambda x.P(\lambda y.\text{LIKES}(x,y))$
V_i	$\to$ *sleeps*	$\lambda x.\text{SLEEPS}(x)$
Nn	$\to$ *dog*	DOG
Nnp	$\to$ *Alex*	ALEX
Nnp	$\to$ *Brit*	BRIT

2. There is a new criminal in town. This time he/she put a bomb in the city center. This bomb will release a toxic gas that will transform everybody into mushrooms. Apparently, the criminal loves semantic parsing, as a creepy message (see below) was sent to Inspector Morcela. Knowing that you are an expert in Natural Language Processing, a stressed Inspector Morcela calls you. Can you help him find the code?

Dear Morcela,
I'm giving you a change to save the city. The meaning of the sequence "4 3 3 4 2" is the code you need, but you will never be able to solve this. Muahahahahahahahahahahah!

*A →B C D {(B.sem + C.sem – D.sem)*10}(A is the initial symbol)*
B →E F {E.sem F.sem}*
C →F E {F.sem + E.sem}
D →1 {1} | 2 {2}   // the semantics of 1 is 1 and the semantics of 2 is 2
E →4 {4} | 6 {6} | 8 {8}
F →3 {3} | 5 {5} | 7 {7} | 9 {9}

Meaning of the sequence "4 3 3 4 2"	

3. You have just solved the previous challenge and your heart rate is almost normal, when inspector Morcela calls you again: another bomb in the city center and another terrifying message (see below). This time, instead of a code you have to find out which are the wires you should cut. The bomb has a red, a blue, a violet, a green, another blue, a pink and a yellow wire. Can you save the city again?

Dear Morcela,
In the improbable case you have found the previous code, and you are not a mushroom right now, here goes a more complicated challenge, with lambda calculus. Once again, I'm giving you the chance to save the city. The meaning of the sequence "every red cut" tells you what you need to do. Which are the wires you should cut?

S → NP VP {NP.sem@VP.sem}
NP → DET N {DET.sem@N.sem}
DET → every {λP. λQ. ∀x (P(x) => Q(x))}
N → red {BLUE}
VP → Vi {Vi.sem}
Vi → cut {λx. CUT(x)}

========================== Application ============================

4. Consider the following corpora:

- train-original.txt is the original dataset, with BIO notation
- train.txt is the dataset with "intentions"
- train.csv is the slots part of the corpora

Let us spend some time understanding these corpora. Notice that train-original.txt was split so that a classification algorithm could be used to classify the intents (train.txt) and a sequence prediction algorithm could help with the slots (train.csv).

4.1 How many different categories ("intentions") are there in the corpus?

4.2 Notice that in order to identify the "intention" of a given sentence you could (probably) use the same code you used in your project. Try it if you feel like it.

4.3. Try to run the code given to you.

You will need to have pandas, numpy, sklearn and sklearn_crfsuite. Read the README.txt to see how to create a virtual environment and install the needed modules. Then, do:

```
> python3 BIO-CRF.py (it will perform sequence labelling)
```

4.4 Play with the number of instances (vary N in df = df[:N]).

4.5. Try to understand the results.