Deliverable 2 Guide

This project has to be sent by email by a single member of each group. The project needs to be in a zip file cotaining all the code to reproduce the results.

The zip file containing all the work needs to have the following form:

name1_name2_name3.zip

Where name1,name2,... are the names of the members of each group. Please write your full name.

BASIC RULES OF THE DELIVERABLE

- The zip file does not need to inlcude the training data.
- The zip file needs a "maindoc.pdf" file with a description of your work.
- The zip needs a notebook `reproduce_results.ipynb` that loads the work and evaluates it.
    - Accuracy on the train set and test set.
    - Confusion matrix on the train and test set.
    - Number of sentences without any label error in both train and test sets (for each sequence you score a 1 if all the words have the correct label and a zero otherwise).
    - The sentences from TEST_SENTENCES (look below) need to be used to evaluate your work (besides the accuracy, confusion matrix and number of sentences without any label error).

- The data used for this project is `ner_dataset.csv` and you will use it to create a train test split.
-
    - Train set: From "Sentence: 1" to "Sentence: 35970"
    - Test set: From "Sentence: 35971" to "Sentence: 47959"
    - You can choose any validation set using data within your train set
- Each requirement from this document that is not satisfied in your work will imply an reduction on your final mark.

Format of the code

I will put all your projects inside a `deliverable_2` folder which will contain the folder `data` with the csv in the following format: `data/kaggle_ner/ner_dataset.csv`

Example:

```
deliverable_2/name1_name2_name3
deliverable_2/name4_name5
deliverable_2/data/kaggle_ner/ner_dataset.csv
```

Therefore:

- You don't need to include the csv in your project
- Your code has to assume the data is in the parent folder of your `name1_name2_name3` folder. That is in `../data`.

Zip organization

- The zip file should have 2 notebooks: `train_models.ipyng` and `reproduce_results.ipynb` .
    - Notebook `train_models.ipyng` chould contain the code for training (and model selection if you want).
    - Notebook `reproduce_results.ipynb` should load the trained models and evaluate them in train and test. Make sure the notebook that loads from disk and evaluates results gets the same results that you have.
- A simple document `maindoc.pdf` should be included containing basic information about the models tested and results.

About the models you need to test....

- A structured perceptron has to be tested, and features added to boost performance.
- Another model needs be tested (you can choose any technique for this part and provide a simple explanation).

About maindoc.pdf

Your document needs to discuss if any feature generation has been tested and what it does (and a link/path to the source code). In particular you should answer:

- How new features affect performance ?

Note that if you add 3 features you don't need to test all combinations of features it is enough to test "the addition of feature 1" , "the addition of features 1 and 2" etc.. Please explain features well in the document and give a motivation of why you decided to use them.

TEST_SENTENCES

Test your trained models with the following phrases (DO NOT change any misspelled words)

Notebook `reproduce_results.ipynb` should print the output sequences of the `TEST_SEQUENCES` in the following format:

```
w1/t1 w2/t2 w3/t3 w4/t4
```

where `wi` is a word and `ti` the tag associated to word `i`.

```
# TEST_SENTENCES

The programmers from Barcelona might write a sentence without a spell
checker.

The programmers from Barchelona cannot write a sentence without a spell
checker.

Jack London went to Parris.

Jack London went to Paris.

We never though Microsoft would become such a big company.

We never though Microsof would become such a big company.

The president of U.S.A though they could win the war

The president of the United States of America though they could win the war

The king of Saudi Arabia wanted total control.

Robin does not want to go to Saudi Arabia.

Apple is a great company.

I really love apples and oranges.
```

maindoc.pdf

Your document needs to discuss if any feature generation has been tested and what it does (and a link/path to the source code).

In particular you should answer:

- How new features affect performance ?

Note that if you add 3 features you don't need to test all combinations of features it is enough to test "the addition of features 1" , "the addition of features 1 and 2" etc.. Please explain features well in the document and give a motivation of why you decided to use them.

Optional (individual) work

This is a completly optional work that might add your overall mark 1 point out of 10. This means that your final score of NLP can be an 11, or that if you have an average of 9 you can get a 10.

This part has to be done INDIVIDUALLY. That means if `person_x` decides to do it it will be send and evaluated only for `person_x`.

Why on earth would you give us more work?

- Some people migh need a very high score for schollarships, this might help
- Some people might want to learn in detail the structured perceptron model, this might help.

Your optional (individual) task consist on providing a new structured perceptron implementation (or a copy paste of the provided implementation) that speeds training and/or evaluation of the model. In class you have learned a bit of cython that might help you get there.