Machine learning entity recognition with Spacy

1 Description

A named entity is a "real-world object" that's assigned a name; for example, a person, a country, a product or a book title. Name entity recognition (NER) is the problem of automatically recognizing these entities from the text. This is a common problem in natural language processing (NLP). It can be posed as a particular classification problem in which the entities that we would like to predict for the words are the classes of our problem. As such, it can be solved with different classification approaches. For this project the student will use the Python library `Spacy` (`https://spacy.io/`) which contains several implementation of NLP algorithms and ML techniques.

2 Objectives

The goal of the project is to create a NER algorithm in Spacy able to recognize a set of "Machine learning" entities from the analysis of ML text, for example, from the analysis of a ML paper. Examples of entities could be:

- Supervised ML algorithm (decision tree, random forest, SVM, etc.)
- Unsupervised ML algorithm (clustering, knn, event detection)
- ML software (keras, sklearn, etc.)

The student should: 1) Define a set of ML entities comprising at least 8 entities. 2) Create one or more NER models using Spacy. 3) Design a validation method to evaluate the accuracy of the method or methods for labeling the document.

As in other projects, a report should describe the characteristics of the design, implementation, and results. A Jupyter notebook should include calls to the implemented function that illustrate the way it works.

3 Suggestions

- Review the methods used by Spacy to recognize and create name entities `https://spacy.io/usage/linguistic-features#section-named-entities`
- Create the rules to detect the entities using Spacy.
- You could use the name entity visualizer displaCy to visualize the learned entities `https://explosion.ai/demos/displacy-ent`