Named Entity Recognition

IS407 Intro to Data Science

What we will discuss today

- Section-1: Introduction
- Section-2: Use cases and Results
- Section-3: Implementation in Python
- Section-4: What lies ahead?
- Section-5: Further References

Introduction

What is NER?

Named entity recognition (NER) is an NLP based technique to identify and assign labels named entities in text

Labels like a person, location, organisation, product etc.

What is a Named entity?

A real world object denoted by proper noun

Named Entities

Examples:

- Champaign - <city>
- UIUC - <university>
- Sharvi Tomar - <person>
- Grainger Engineering Library - <location>

Use Cases

- Document Categorisation
- Tracking mentions of particular entities in documents
- For question answering, answers are usually named entities
- Extracting wanted information using named entities
- Support Chatbots
- NER for Medical Purposes

Results

Below is an screenshot of how a NER algorithm can extract particular entities from a given text:

When [Sebastian Thrun PERSON] started working on self - driving cars at [Google ORG] in [2007 DATE], few people outside of the company took him seriously . " I can tell you very senior CEOs of major [American NORP] car companies would shake my hand and turn away because I was n't worth talking to , " said [Thrun PERSON], in an interview with [Recode ORG]

Intuition: Example Sentence

'European authorities fined Google a record $5.1 billion on Wednesday for abusing its powers'

Step-1: Tokenize

```
['European',
 'authorities',
 'fined',
 'Google',
 'a',
 'record',
 '$',
 '5.1',
 'billion',
 'on',
 'Wednesday',
 'for',
 'abusing',
 'its',
 'powers']
```

Step-2: Part-of-Speech(POS) Tagging

```
[('European', 'JJ'),
 ('authorities', 'NNS'),
 ('fined', 'VBD'),
 ('Google', 'NNP'),
 ('a', 'DT'),
 ('record', 'NN'),
 ('$', '$'),
 ('5.1', 'CD'),
 ('billion', 'CD'),
 ('on', 'IN'),
 ('Wednesday', 'NNP'),
 ('for', 'IN'),
 ('abusing', 'VBG'),
 ('its', 'PRP$'),
 ('powers', 'NNS')]
```

Step-3: Chunking Noun Phrases

- Identify named entities using a regular expression(regex)
- Regex rules indicate how sentences should be chunked
- Our chunk pattern consists of one rule, that a noun phrase, NP, should be formed whenever the chunker finds an optional determiner, DT, followed by any number of adjectives, JJ, and then a noun, NN.

```
pattern = 'NP: {<DT>?<JJ>*<NN>}'
```

Step-4: Add category labels to named entities

```
(S
  European/JJ
  authorities/NNS
  fined/VBD
  Google/NNP
  (NP a/DT record/NN)
  $/$
  5.1/CD
  billion/CD
  on/IN
  Wednesday/NNP
  for/IN
  abusing/VBG
  its/PRP$
  (NP power/NN)
```

SpaCy's Pipeline
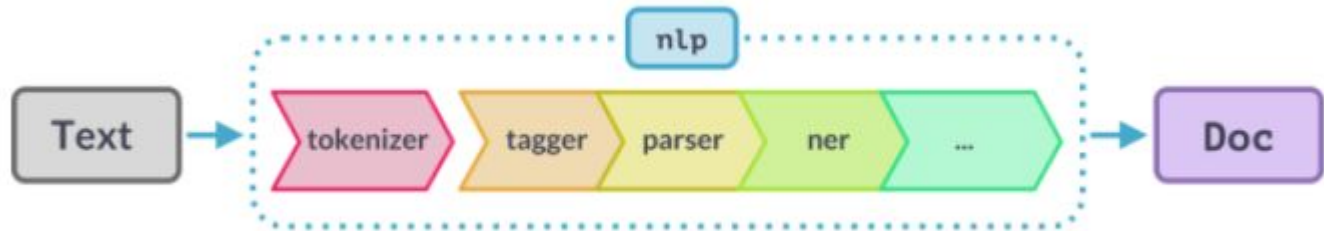
Section-2 Approaches

A. Traditional approaches :

- Rule-based solutions: have an obvious flaw, i.e. they require a well-defined and exhaustive lexicon set
- Feature-based supervised machine learning approaches: requires carefully hand-crafted features like lookup list, case/morphology/POS tag etc to make it work.

B. Deep Learning-based models

- Training Neural Network models for NER
- Rather than training models from scratch, the new paradigm in NLP is to select an off-the-shelf model that has been trained on the task of "language modelling" (predicting which words belong in a sentence), then "fine-tuning" the model with data from your specific task.

Section-3: Implementation

Let's jump to experiment notebook

Section-4: What's next for you?

- Advanced courses at UIUC
  - IS 567 - Text Mining
  - CS 447 - Natural Language Processing
  - CS 410 - Text Information Systems
  - CS 510 - Advanced Information Retrieval

Section-5: References for further study

R book with chapters and examples with NER (need university library to access)

https://learning.oreilly.com/library/view/mastering-text-mining/9781783551811/ch07.html#ch07lvl1sec41

https://learning.oreilly.com/library/view/mastering-text-mining/9781783551811/ch07s03.html#ch07lvl3sec59

Implementation in R

https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/#:~:text=Spacy%20is%20an%20open%2Dsource,very%20easily%20for%20NER%20tasks.

https://towardsdatascience.com/named-entity-recognition-ner-using-spacy-nlp-part-4-28da2ece57c6

https://towardsdatascience.com/custom-named-entity-recognition-using-spacy-7140ebbb3718

https://medium.com/in-pursuit-of-artificial-intelligence/named-entity-recognition-using-spacy-ner-da6eebd3d08

https://nanonets.com/blog/named-entity-recognition-with-nltk-and-spacy/

https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da

Questions/comments?