NER & Information Extraction

03 September 2022

NEUROTECH

Hello!

I am Anthony

- → Software engineer
- → Zindi Ambassador

1.
Topic Elaboration

Let’s start jump right in

Information extraction

extracting structured information from unstructured text, including entities and relations between them, sometimes also connecting to an existing knowledge base.

NER

NER ie Named Entity Recognition or Entity Detection.

The subtask of information extraction, extracting proper names and identifying their classes, such as people, locations, organizations, etc.

2.
Named Entity Recognition

Let’s start jump right in

NER Processes

- **Identifying named entities:**

Tanzania remembered father of the nation *Julius Nyerere* on October 14, 22 years since he passed away.

- **Classifying entities:**
  - Location (Tanzania)
  - Person (Julius Nyerere)
  - Date (October 14)

NER
How does it works?

- ➔ NER works by **identifying notable objects** in a structured or unstructured text.
- ➔ The process is Useful for analyzing a **wide variety** of texts.
- ➔ NER equal to **identify** & **categorizing** informations

NER
How does it works?

- **NER involve two models:** ontology-based models and Deep Learning-based models.
- **Ontology** uses a knowledge based recognition process that relies on lists of datasets. This method work well in medical field.
- **Deep Learning** uses trained NNs consisting of thousands, millions, or even billions of parameters to understand the semantic and syntactic relationship between words and phrases in the input text.

NER Techniques

➔ Rule-based

- Labor intensive
- Need linguistic knowledge of the language
- Gazetteers can be useful when combined with other techniques

➔ Statistical ML

- Different ML classifiers can be used: HMM, Decision Trees, SVM, CRF
- Classifier where context is taken into account (one of the input features is the previous element's label), hence best suited for such a task

➔ Clustering

- Different clustering methods used here like KNN, etc

NER Techniques

➔ Word embedding

- Word embeddings: vector representations of words
- Generating embeddings is easy using a corpus using existing Python libraries (word2vec)
- Lots of NER systems use word embeddings as input features
- Reference can be found here

➔ Deep-learning

- Advantages of deep learning: minimal feature engineering
- Disadvantages: require lots of labeled data
- Many use word embeddings

NER Frameworks

Open source frameworks for **NER**:

- ➔ SpaCy
- ➔ Natural Language ToolKit (NLTK)
- ➔ Stanford Named Entity Recognizer(SNER)

4.
NER real world applications

Let’s start jump right in

Applications

- Classifying content for news providers
- Automatically Summarizing resumes
- Optimizing Search Engine Algorithms
- Simplifying Customer Support

5.
SpaCy

Let’s start jump right in

SpaCy

SpaCy is a **free** and **open source** library for advanced **Natural Language Processing**.

Designed specifically for production use and to support building application that process and **understand large volume of Text**.

With spaCy you can build **information extraction** of natural language understanding or text processing.

https://spacy.io/

SpaCy Features

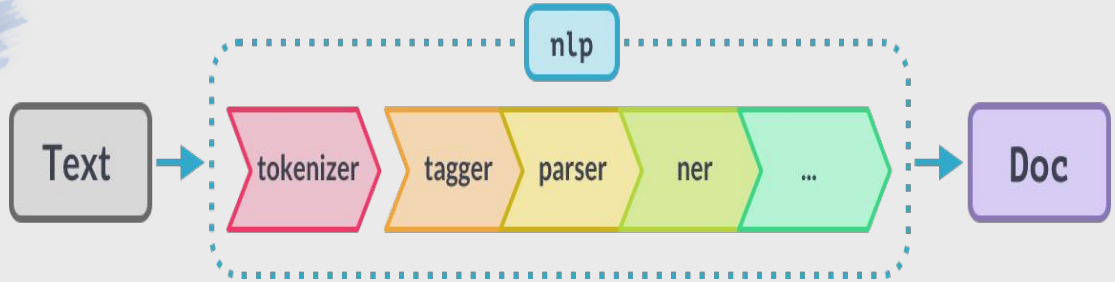
NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) **Tagging**	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of "was" is "be", and the lemma of "rats" is "rat".
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named "real-world" objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a knowledge base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model's predictions.
Serialization	Saving objects to files or byte strings.

Spacy features

SpaCy Processing Pipelines

Invoke **nlp** - text as parameter, processing steps start from **tokenizing** the **text**, **tagger** ....

Final is **document object**(Doc)



6.
Coding Time

Let’s start jump right in

Resources:

- About Named Entity Recognition
- Get started with spaCy
- Understanding SpaCy Processing Pipeline
- Training custom Named Entity Recognition
- Applications of Named Entity Recognition

THANKS!

Any questions?

You can find me
@LoytTony
anthony@neurotech.africa