uc3m | Universidad Carlos III de Madrid

HULAT Human Language & Accessibility Technologies

Cutting-edge Deep Learning for NLP learners

Basic NLP tasks

Isabel Segura-Bedmar

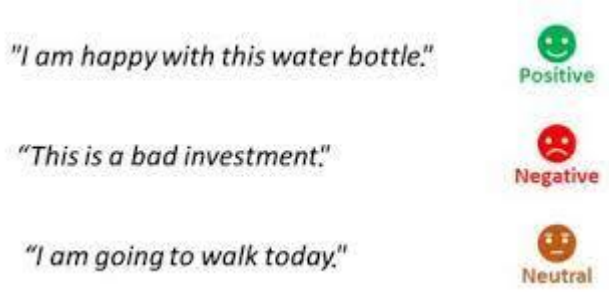
6 June 2022
Máster Universitario en Inteligencia Artificial, UPM

Outline

- Introduction
- Basic Tasks
  - Tokenization
  - Word Normalization
  - StopWords
  - PoS tagging and Parsing
  - NLP libraries
-

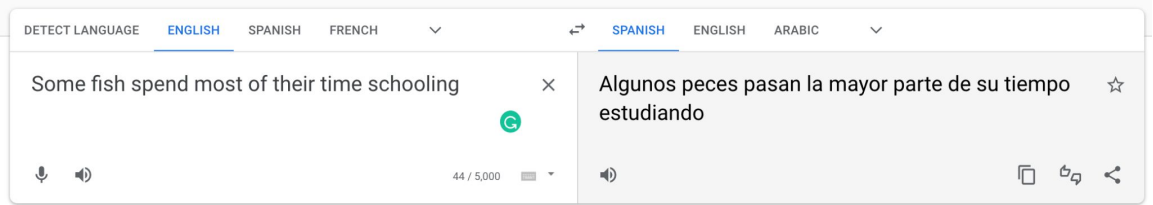
NLP Applications: different outputs

Text classification

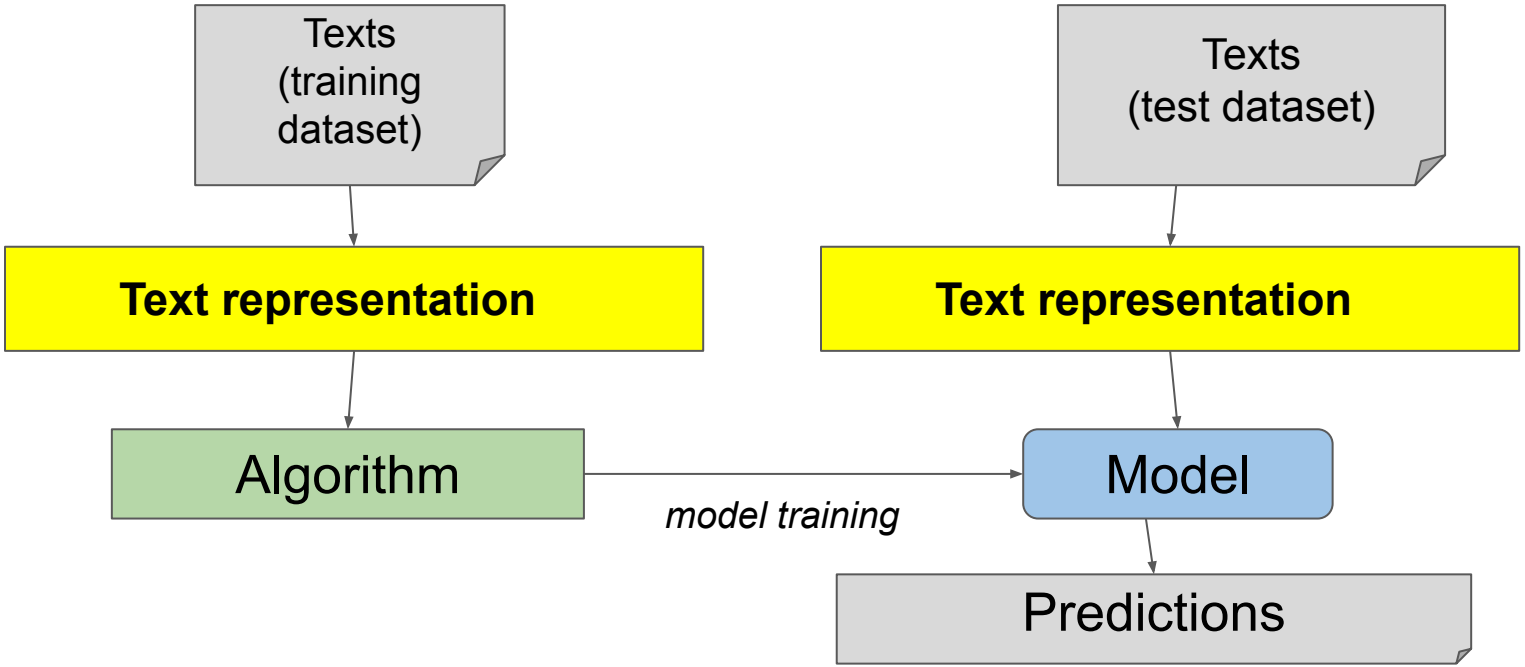


Information Extraction

Machine translation

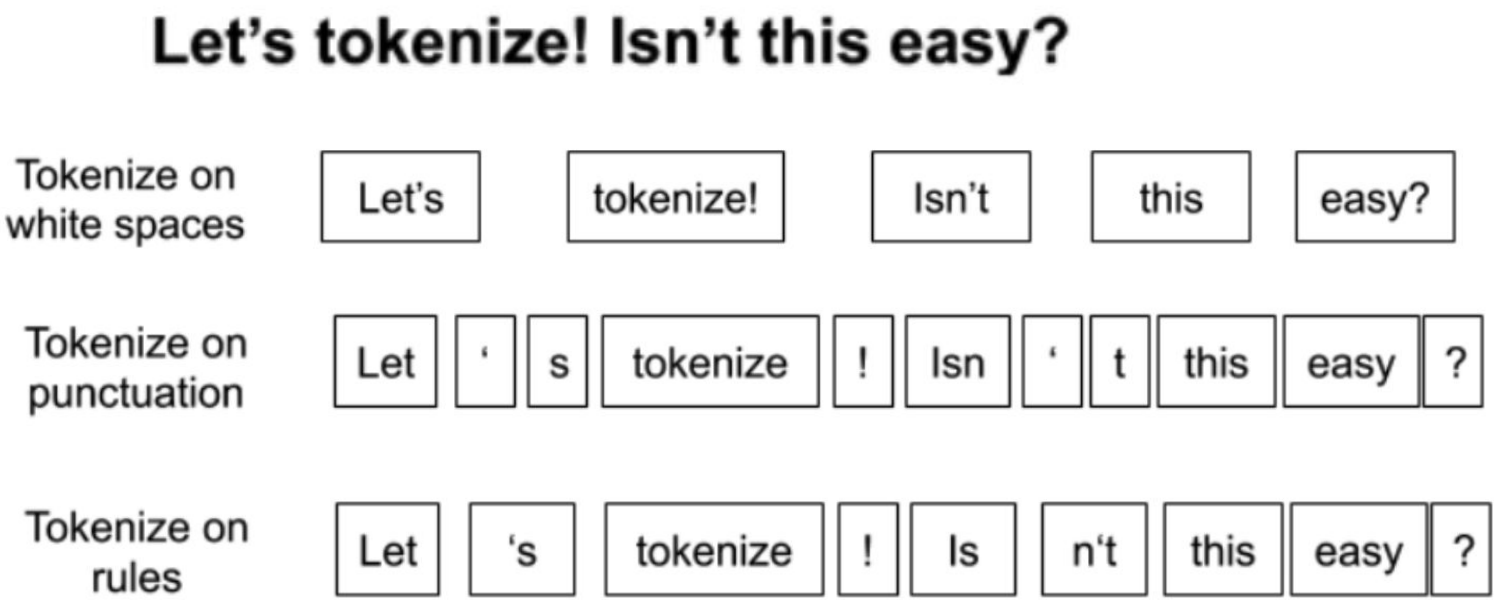


Machine Learning Pipeline



Tokenization

- Split the input text into tokens (words)



Tokenizers

- Please, don’t implement your own tokenizer.
- Many NLP libraries already provide great tokenizers:
  - NLTK (https://www.nltk.org)
  - Spacy (https://spacy.io/api/tokenizer)
  - Hugginface (https://github.com/huggingface/tokenizers)

Word Normalization techniques

1. **Reduce lexical variability** (decrease the size of the vocabulary)
   - *C.C.O.O -> CCOO*
   - *niño, niña, niños, niñas, Niños, Niñas… -> niño*
   - *jugaré, jugarás, jugaremos, … -> jugar*
2. Help text classification, machine translation, etc.
3. most used tasks:
   - remove case sensitive
   - removing non-alphanumeric characters (accents)
   - **lemmatization** and **stemming**

Lemmatization

- Decrease the vocabulary size and improve information retrieval

  smaller, smallest -> small
- Obtain the lemma of a word.
  - houses -> house, women -> woman
  - youngest -> young
  - wearing -> wear, are -> be

- Online lemmatizers:
  - Spanish lemmatizer
  - https://cst.dk/tools/index.php#output

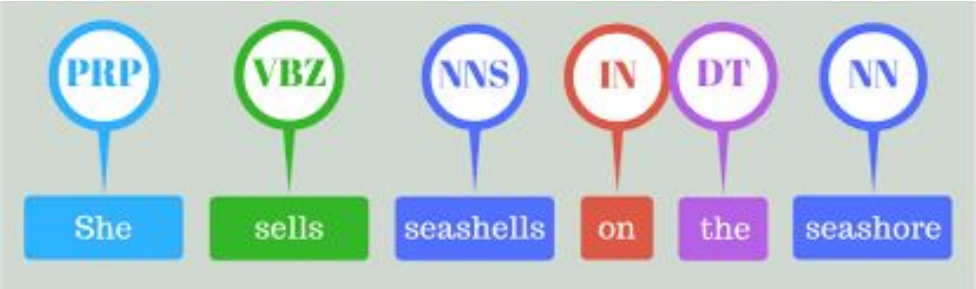
Stemming

- Porter Algorithm, Stemmer online
  - *heart**s** -> heart*
  - *sing**ing** -> sing*
  - *sing**er** -> sing*
  - *sang -> sang*
  - *learn**ed** -> learn*
  - *learnt -> learnt*
  - *wrote -> wrote*
  - *hous**es** -> house, hou**se** -> hous*
  - *poni**es** -> poni*
  - *are -> are*
  - *women -> women*

StopWords

- Most common words in any language (articles, prepositions, pronouns, conjunctions, etc)
- Does not add much information to the text
- In many NLP tasks (text classification), stopwords are removed
- link

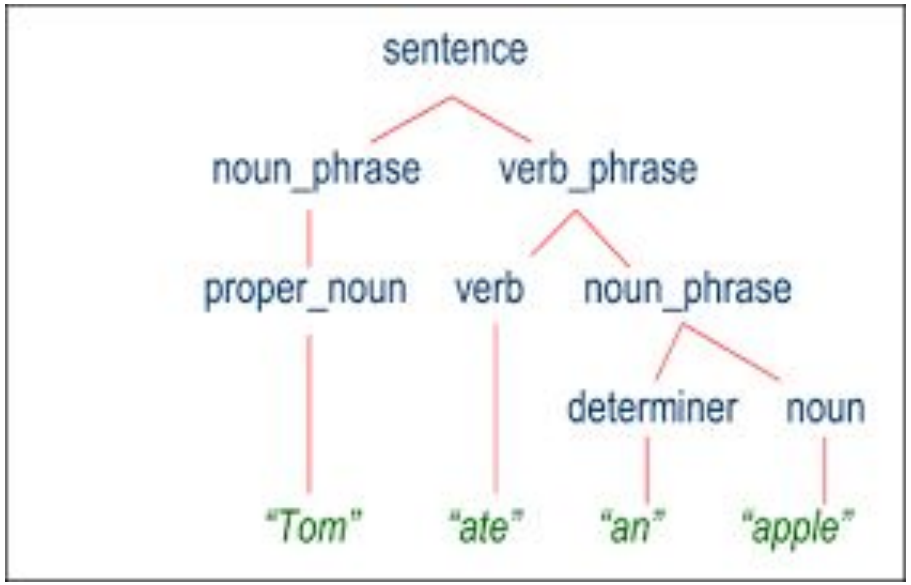
Part-of-speech (PoS) tagging



https://github.com/dhirajhr/POS-Tagging

PoS tagging helps IE (NER, RE), Question Answering

Parsing



https://forum.huawei.com/

Parsing helps IE (NER, RE), Question Answering, Machine Translation, Text summarization, etc.

Hands-on NLP Libraries

- NLTK
- Spacy
- Normalization

Exercises

- https://github.com/isegura/MUIA2022
- Please, download the files:
  - SpaCy.ipynb
  - 2-ExerciseSpaCy.ipynb
- Save them into google colab
- Practice!!!

Thank you
Question time!!!

isegura@inf.uc3m.es
https://hulat.inf.uc3m.es/nosotros/miembros/isegura
https://github.com/isegura