Funded by the European Union

ASSISTANT

Natural Language Processing

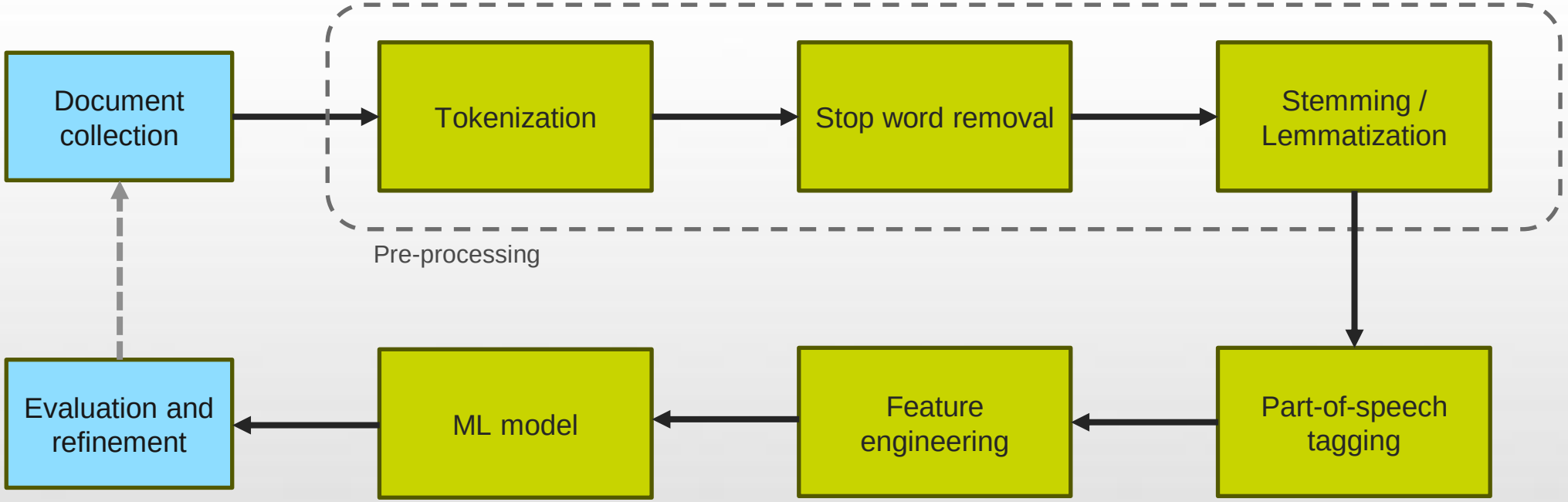
Vitor Rocio

(vitor.rocio@uab.pt)

What is NLP?

- A set of techniques to imbue computers with the ability to understand and generate text or speech in human languages, based on 2 kinds of approaches:
  - Symbolic: dictionaries and rules built by humans
  - Statistical: using corpus-based data, including machine learning
- NLP tasks:
  - Part of speech tagging
  - Sense disambiguation
  - Named entity recognition
  - Anaphora resolution
  - Sentiment analysis
  - Language generation

NLP Pipeline



Pre-processing

- Tokenization: separation into words
- Stop word removal
  - Stop words: **the**, **and**, **of**, **to**, etc.
- Word normalization:
  - Stemming
  - Lemmatization

Stop word removal

- Frequent functional words like prepositions and conjunctions do not have semantic value *per se*;
- In algorithms that consider co-ocurrence of words as meaningful, stop words can be discarded, since they usually co-occur with any word;
- After tokenization, stop words can be removed by referring to a stop word list (language-dependent).

Stemming and lemmatization

- Word variations like **walk**, **walked**, **walks**, or **country**, **countries** can be considered the same for modelling purposes, as we are more interested in their relation with other words;
- Stemming: stripping words of variant affixes
  - **walked -> walk**; **walks -> walk**
- Lemmatization: reducing a word to its normal form:
  - **countries -> country**

Language Modelling

- Part-of-speech tagging
- Feature extraction
- Supervised vs. Unsupervised training

Funded by
the European Union

Thank you for your attention