Natural Language Processing

Word categorisation and understanding

TechTalents
sólo es el principio

Steps

1. Cleaning
   Remove unwanted, unneeded, unnecessary data.
2. Tokenization
   Split data pieces that fit the analysis being done.
3. Tagging
   Calculate and add “metadata”.
4. Normalization
   Retrieve the root/source of the word.
5. Contextualization
   Work out the meaning of words depending on their position.
6. Extraction
   Pull new understanding out of the data.

Tagging

Apply a short tag describing the words function and position in a body of text.

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

https://github.com/tt-n-walters/uria-python/blob/master/week7_session3/piece_of_speach_tags.md

Normalization

Stemming

Removing and replacing of suffixes to get the root form of the word, called the **stem**.

Normalization

Stemming

Removing and replacing of suffixes to get the root form of the word, called the **stem**.

run, running -> run
talk, talks -> talk

Normalization

Lemmatization

Performing contextual and “morphological” analysis of the text, to determine closest correct root word.

Normalization

Lemmatization

Performing contextual and “morphological” analysis of the text, to determine closest correct root word.

Lemmatization needs to be given the context of the word, ie. “type”. Verb, Noun, Adjective, or Adverb.