OPEN DATA SCIENCE CONFERENCE

Boston | May 1 - 4 2018

#ODSC

@ODSC

#ODSC

Industrial-Strength Natural Language Processing

IN PYTHON

spaCy 2.0: Key Features

SpaCy Tutorial

- Installation
- Purpose & Main Features
- Multi-Language Models
- SpaCy Pipeline & Architecture
- Use cases & sample applications:
  - Part-of-Speech Tagging
  - Named Entity Recognition
  - Sentence Boundary Detection

SpaCy 2.0: From Linear to Neural Models

- Trained models for tagger, parser and entity recognizer.
- 10x smaller, 20% more accurate, run less resource intensive
- Built-in support for
  - English, German, Spanish, Portuguese, Italian, French, Dutch
- Large models include word vectors

SpaCy: Features

Feature	Description
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech Tagging	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Label syntactic dependencies between tokens, like subject <=> object.
Lemmatization	Assigning the base forms of words like "was" => "be", "rats" => "rat".
Sentence Boundary Detection	Finding and segmenting individual sentences.
Named Entity Recognition	Labelling named "real-world" objects, like persons, companies or locations.
Similarity	Evaluate similarity of words, text spans and documents.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Find sequences of tokens based on their texts and linguistic annotations.
Training	Updating and improving a statistical model's predictions.
Serialization	Saving objects to files or byte strings.

The material for today’s workshop

- SpaCy Tutorial
  - https://github.com/stefan-jansen/spaCy-tutorial
- Datasets (included in repo):
  - BBC
  - TED English-Spanish