# Spark NLP

### **Overview**

The idea is to use Spark NLP ( <https://github.com/JohnSnowLabs/spark-nlp> ) to extract relationships between entities and their Part of Speech tags.

1. Follow the instructions to install Spark NLP, either using Python 3 (Jupyter Notebook) or Scala 2.11.
2. Read the given dataset using Spark
3. Create a Spark ML Pipeline using the following annotators (Use English pretrained models):
   1. DocumentAssembler
   2. Tokenizer
   3. WordEmbeddingsModel (Word Embeddings, Glove)
   4. PerceptronModel (Part of Speech)
   5. NerCrfModel (Named Entity Recognition)
4. Print the transformed DataFrame showing only the POS column and the NER column. **BONUS:** Show only the ***result*** attribute of these Annotations
5. Collect the result attribute of NER and POS, find a way to explain any relationship (if exists) between found entities and their part of speech attributes.

Note: An Annotation column is an Array of Annotation objects. Annotation objects have the following scheme:

**Annotation(annotatorType, begin, end, result, metadata, embeddings, sentenceEmbeddings)**

More documentation here:<https://nlp.johnsnowlabs.com/docs/en/annotators>and examples can be found here: <https://github.com/JohnSnowLabs/spark-nlp-workshop>

**Basic Imports:**

*from sparknlp.base import \**

*from sparknlp.annotator import \**

*from sparknlp.embeddings import \**

Datasourse*:*

data/spark\_nlp\_dataset.parquet