# Pipeline overview for patient segment cortellis text.

Code location:

SageMake instance

**Data:** Sampled from AACT dataset. Another Dataset: 4 indications – merged data with AACT database on ID.

* **Input file:** raw\_aac.csv
* **Type:** csv
* **Source:** AACT

**Files:**

Pipeline uses pretrained models and regex / text matching rules.

**Rule Files:**

* **models/editable/fp\_dict.csv** – Rules for string matching
* **models/editable/replace\_dict.csv** – Rules for standardizing and merging entity labels.

**Code Files:**

* **ctdi\_treatment/[env\_setup\_start.py](http://localhost:8800/edit/hasham/pocs/bms/bms_hasham_repo/bms/ctdi_treatment/env_setup_start.py" \t "_blank)** –Setup environment.
* **ctdi\_treatment/treatment\_pipeline.py** – main pipeline for processing data.
* **ctdi\_treatment/resolver\_pipeline.py** – standalone pipeline for resolution.

To run pipeline in the spark env (with installed spark-nlp and spark-nlp-jsl):

* cd Main\_Directory/notebooks
* Define csv file path in the file on line # 14
* python end\_code.py

**Notebooks**:

* **notebooks/New\_Code.ipynb** – This notebook notebook runs the treatment pipeline on an input file and generates a json file having output format specified by BMS.
* **notebooks/Running\_Creating\_Data\_for\_Eval.ipynb** – This notebook runs the treatment pipeline on an input file and creates a json that can directly be uploaded to annotation lab for evalauation.

**Output file sample**:

* **notebooks/output.json**

**NLP Pipeline:**

NLP Pipeline includes the following tasks on excerpts of natural language text:

* Clinical Word Embeddings
* Named Entity Recognition models for:
  + Ner\_Posology
    - Drug
    - Strength
    - Duration
    - Dosage
    - Form
    - Route
    - Frequency
  + ner\_clinical
    - Treatment
  + ner\_clinical\_large
    - Treatment
  + ner\_jsl
    - Treatment
* Regex Matcher for:
  + CycleLength
* RxNorm Entity Resolution for Drug names and dispositions.