You are given a dataset *smoker_status.csv* – clinical notes containing records of patients' smoking status. The goal is to **extract and classify each given patient's smoking status** into the four categories: *smoker, non-smoker, former smoker, unknown*.

Task Description

Using the language / framework of your choice, write a rule or script to extract patients' smoking status. Tag the extraction output as ***Smoker***, ***Non Smoker***, ***Former Smoker*** or ***Unknown***.

Build a basic processing pipeline / finite-state transducer that will include a NER annotator and do the following:

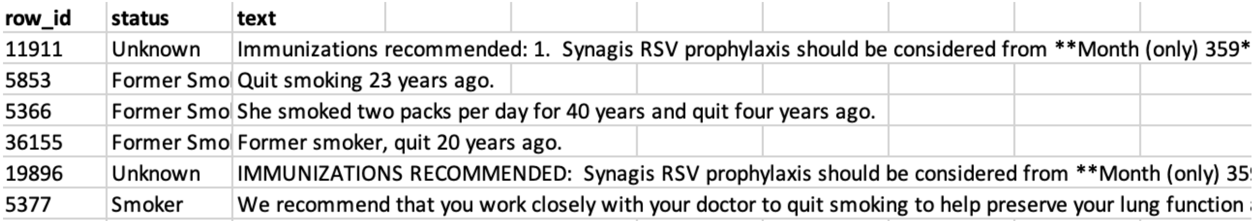
- Read the input file (see below for detailed description), take the **text** column values as input to the annotator.
- Load the rule as a processing resource, run it on the *smoker_status* dataset to extract and tag smoking status for the given number of patients.
- Store your output in a file.
- Calculate the accuracy, precision, recall and F1

Note: Make sure your code can be run on unseen data with the structure as below

Sample Input

The dataset contains 70 sentences that mention patients' smoking status and has the following columns:

- **row_id** ***`Int`***: each patient's unique identifier
- **status** ***`Str`***: ground truth smoking status to be used for evaluation
- **text** ***`Str`***: sample sentences



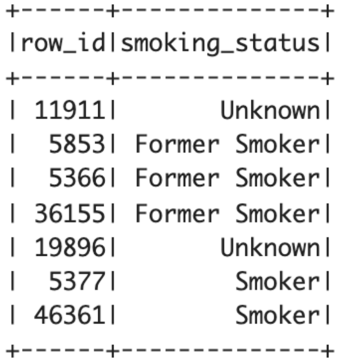
row_id	status	text
11911	Unknown	Immunizations recommended: 1. Synagis RSV prophylaxis should be considered from **Month (only) 359*
5853	Former Smo	Quit smoking 23 years ago.
5366	Former Smo	She smoked two packs per day for 40 years and quit four years ago.
36155	Former Smo	Former smoker, quit 20 years ago.
19896	Unknown	IMMUNIZATIONS RECOMMENDED: Synagis RSV prophylaxis should be considered from **Month (only) 35
5377	Smoker	We recommend that you work closely with your doctor to quit smoking to help preserve your lung function

Feel free to transform or convert the input file into a format you are comfortable working with.

Sample Output

Output your results as key – value pairs:

- **row_id** ***Int***: the given patient’s unique identifier
- **smoking_status** ***Str***: each patient’s corresponding smoking status. **smoking_status** will only contain the following values: ***Smoker***, ***Non Smoker***, ***Former Smoker***, ***Unknown***.



```
+------+-------------+
|row_id|smoking_status|
+------+-------------+
| 11911|      Unknown|
|  5853| Former Smoker|
|  5366| Former Smoker|
| 36155| Former Smoker|
| 19896|      Unknown|
|  5377|       Smoker|
| 46361|       Smoker|
+------+-------------+
```

Note: We use Apache Spark as part of our processing pipeline, hence the dataframe output. Feel free to store your extraction results in a format you are comfortable working with (e.g., txt, csv, JSON, etc.).

Guidelines checklist

- Is the code readable?
- Where code is not enough, are there comments to explain the “why”?
- Are there parts of the code that need to be refactored before submitting the solution?

Deliverables

Create a repository on GitHub or GitLab with your solution, add your name as a collaborator, and send us the link to the repository.