Text File Format Identification

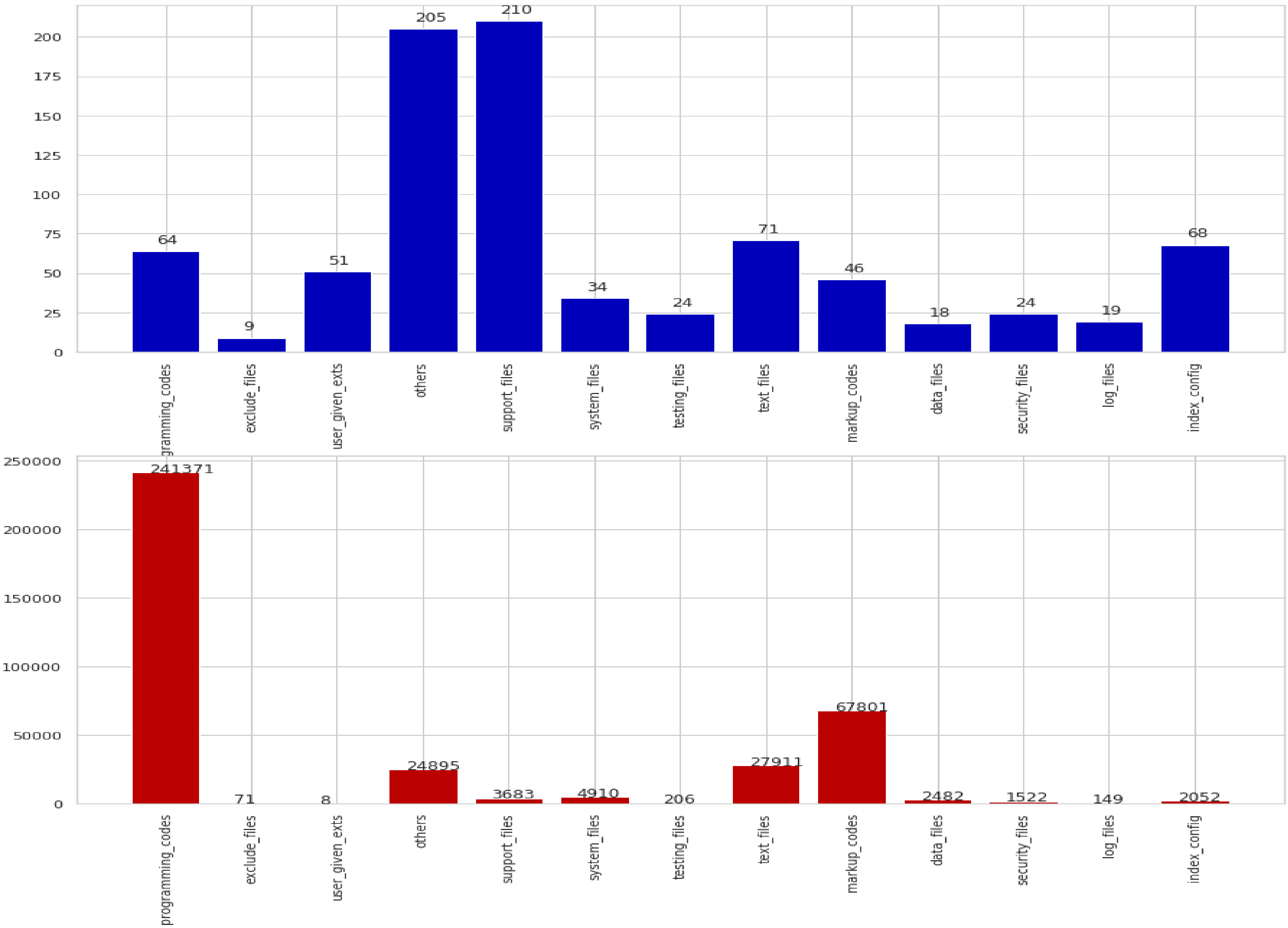
Santhilata K.V.

04/06/2019

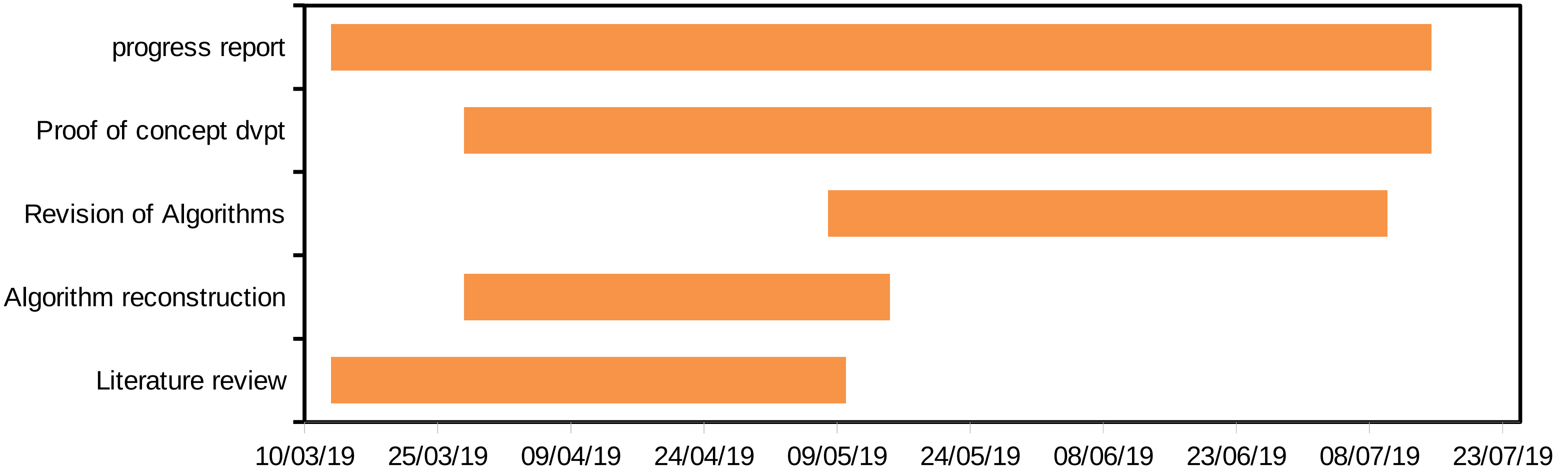
Agenda

- Research question
  - How to correctly identify the file format of a text file from its contents ?
- Time line
- Training dataset - text Files
- Work done /To do

File Type Distribution



Time line

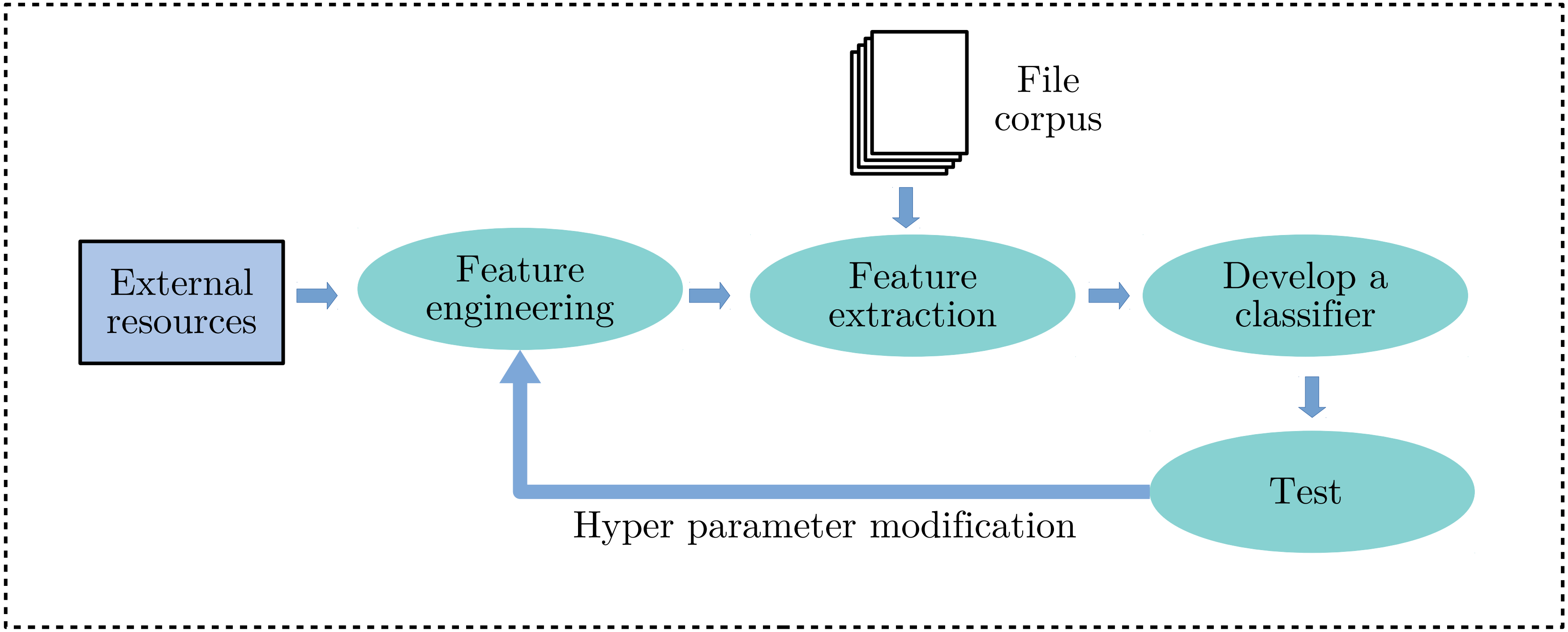

progress report
Proof of concept dvpt
Revision of Algorithms
Algorithm reconstruction
Literature review
10/03/19
25/03/19
09/04/19
24/04/19
09/05/19
24/05/19
08/06/19
23/06/19
08/07/19
23/07/19

Revised Objectives

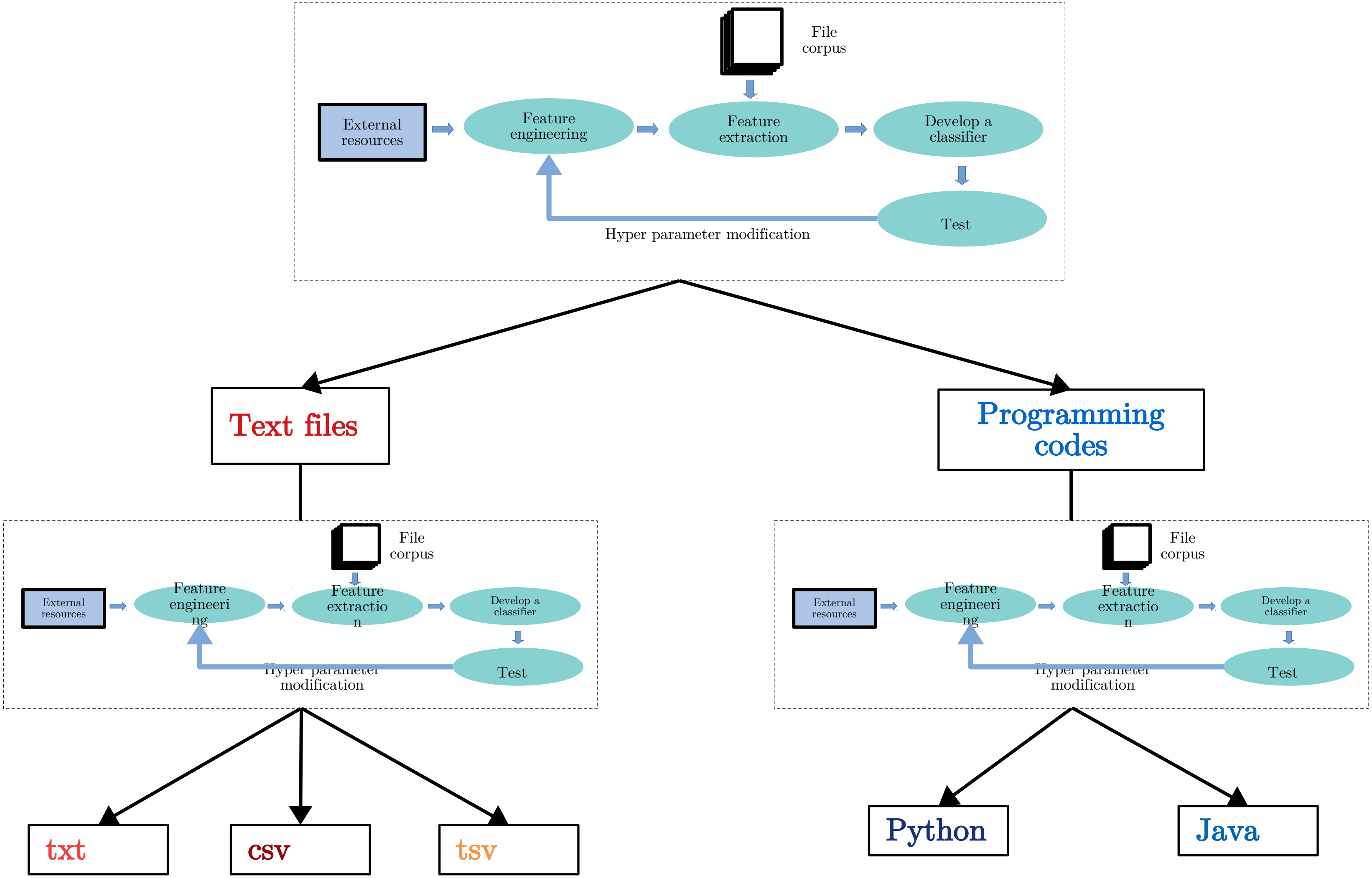
- Identification of file formats

  (I) programming codes (Java and Python) and

  (II) text files (tsv, csv and txt)



Hierarchical process



File corpus

External resources

Feature engineering

Feature extraction

Develop a classifier

Test

Hyper parameter modification

Text files

Programming codes

File corpus

External resources

Feature engineeri ng

Feature extractio n

Develop a classifier

Test

Hyper parameter modification

File corpus

External resources

Feature engineeri ng

Feature extractio n

Develop a classifier

Test

Hyper parameter modification

txt

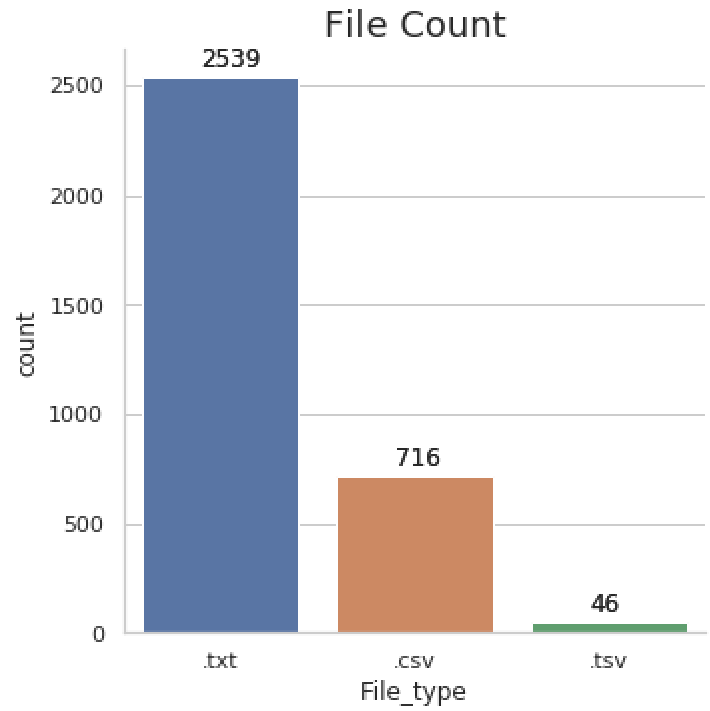
csv

tsv

Python

Java

Training dataset



- Characteristics of csv files:
  - Multiple tables can be present in a single file (if the csv is generated using Excel)
  - Possible set of delimiters:

    [, . \t ; : # | ^]
  - Multiple delimiters are not allowed
  - One record per row

Published resource: **Characteristics of Open Data CSV Files** *by Johann Mitl¨ohner, Sebastian Neumaier, J¨urgen Umbrich, and Axel Polleres*

Work done / to do

- Formulated an algorithm to unearth specific features from csv files
- To design a neural network model that helps to learn the text file characteristics
- To design a hierarchical classification model to divide programming codes from text files
- Complete the report and create publishing material