


Text File Format Identification

Santhilata K.V.

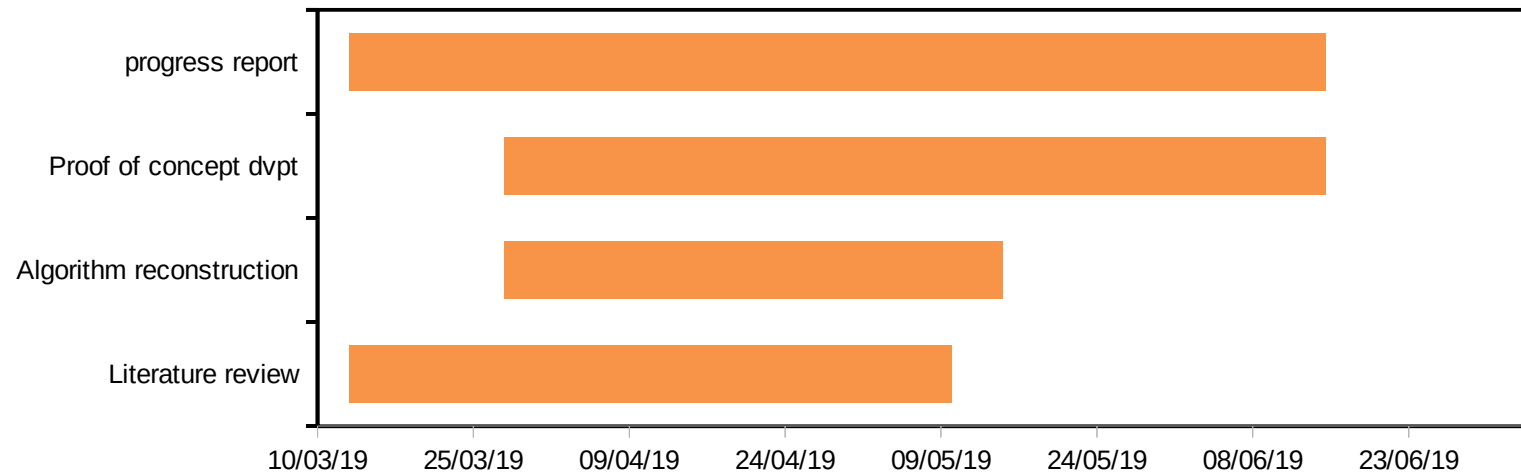
23/05/2019



Agenda

- Research question
 - _ How to correctly identify the file type from the contents of text files?
- Time line discussion
- Training Data
- Conclusion /To do

Time line



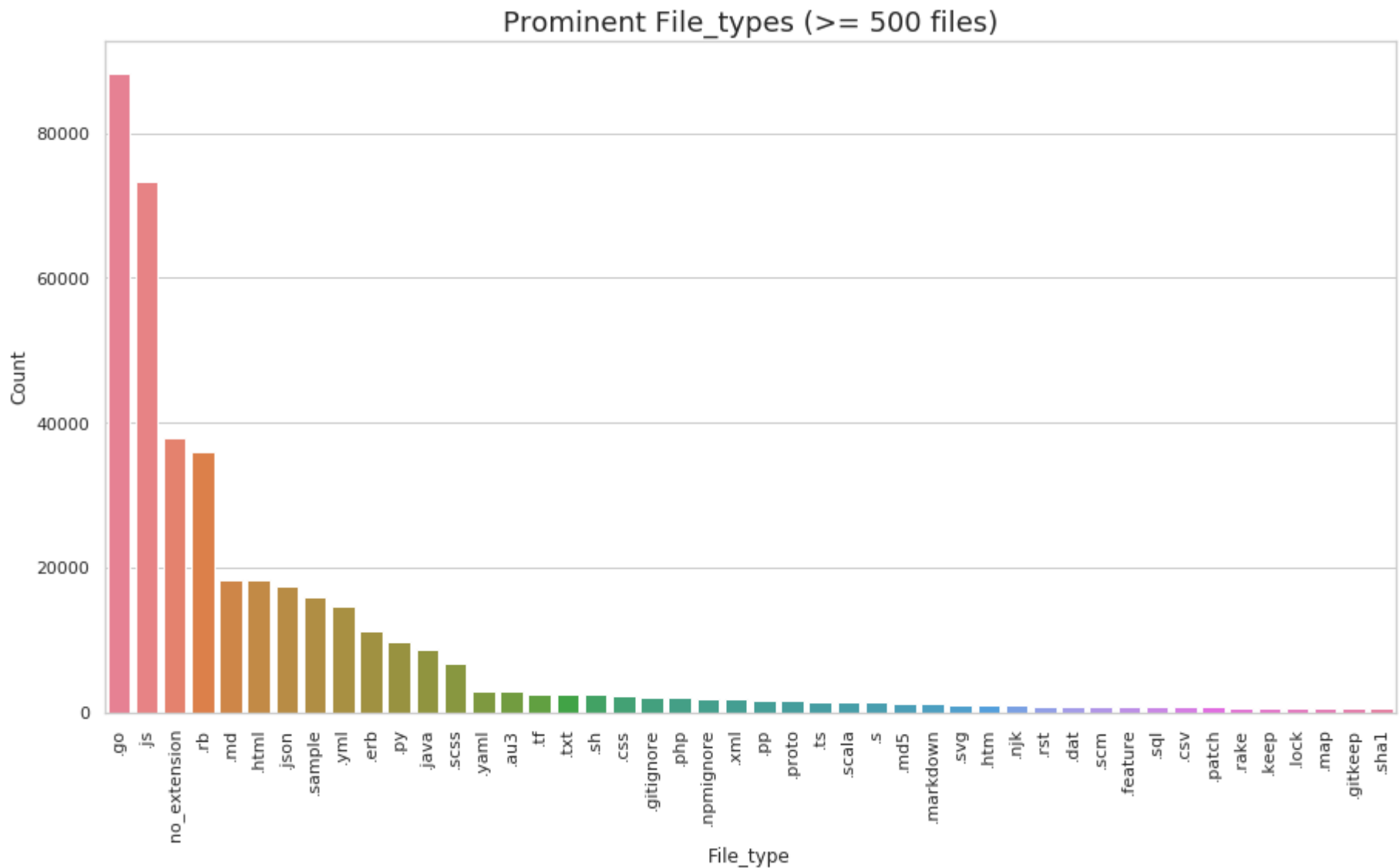
Data sources for Training set

- GITHUB repos of GDS
 - <https://github.com/alphagov>
- GITHUB repos of TNA
 - <https://github.com/nationalarchives>

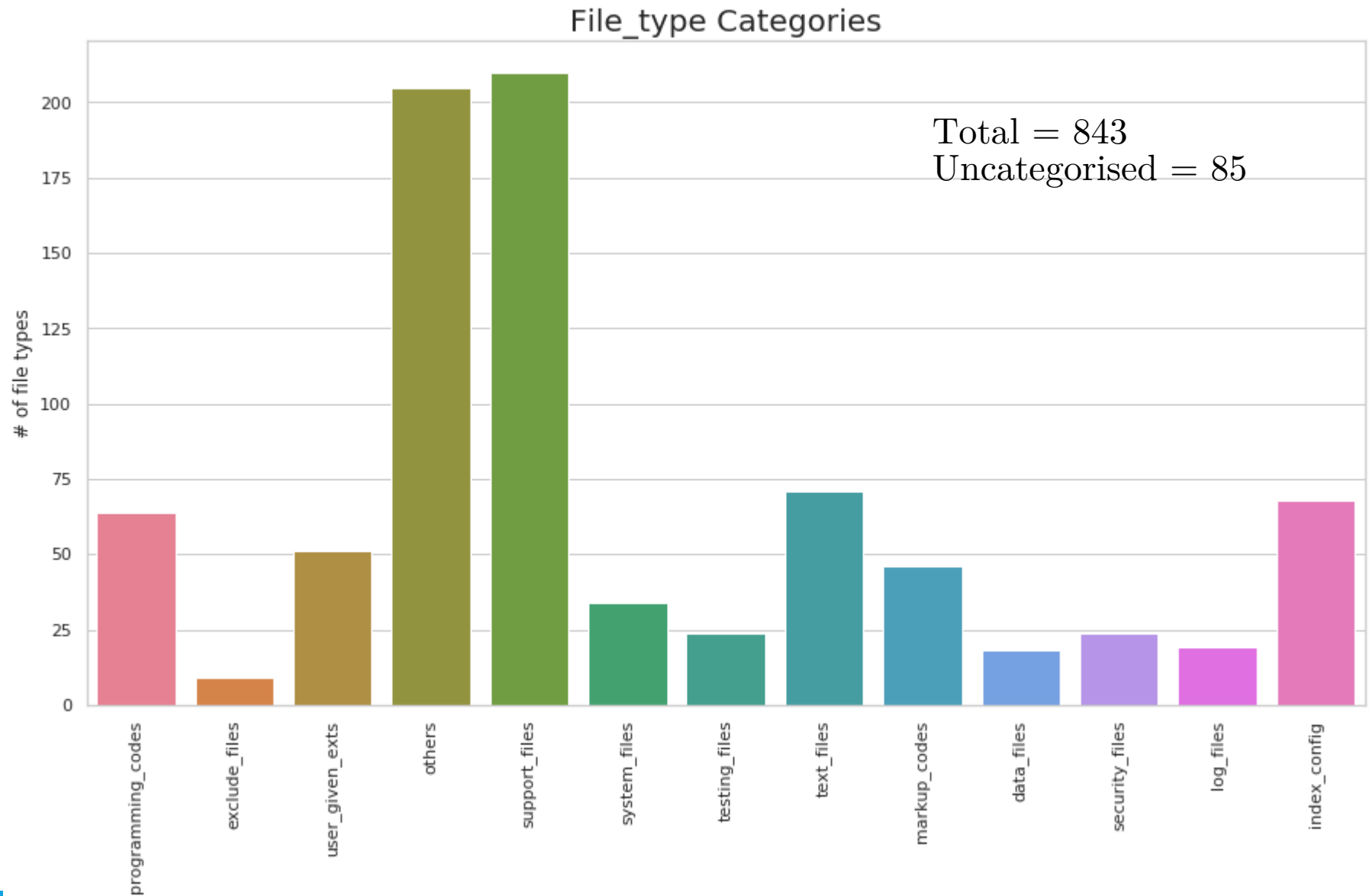
Training Data

Data	Original Stats	Stats after removing non-ascii files
GDS Github repos	1290	
TNA Github repos	167	
Total	1457	
Total Files #	578322	419011
File Types #	1165	928
FileType Count>500	57	46

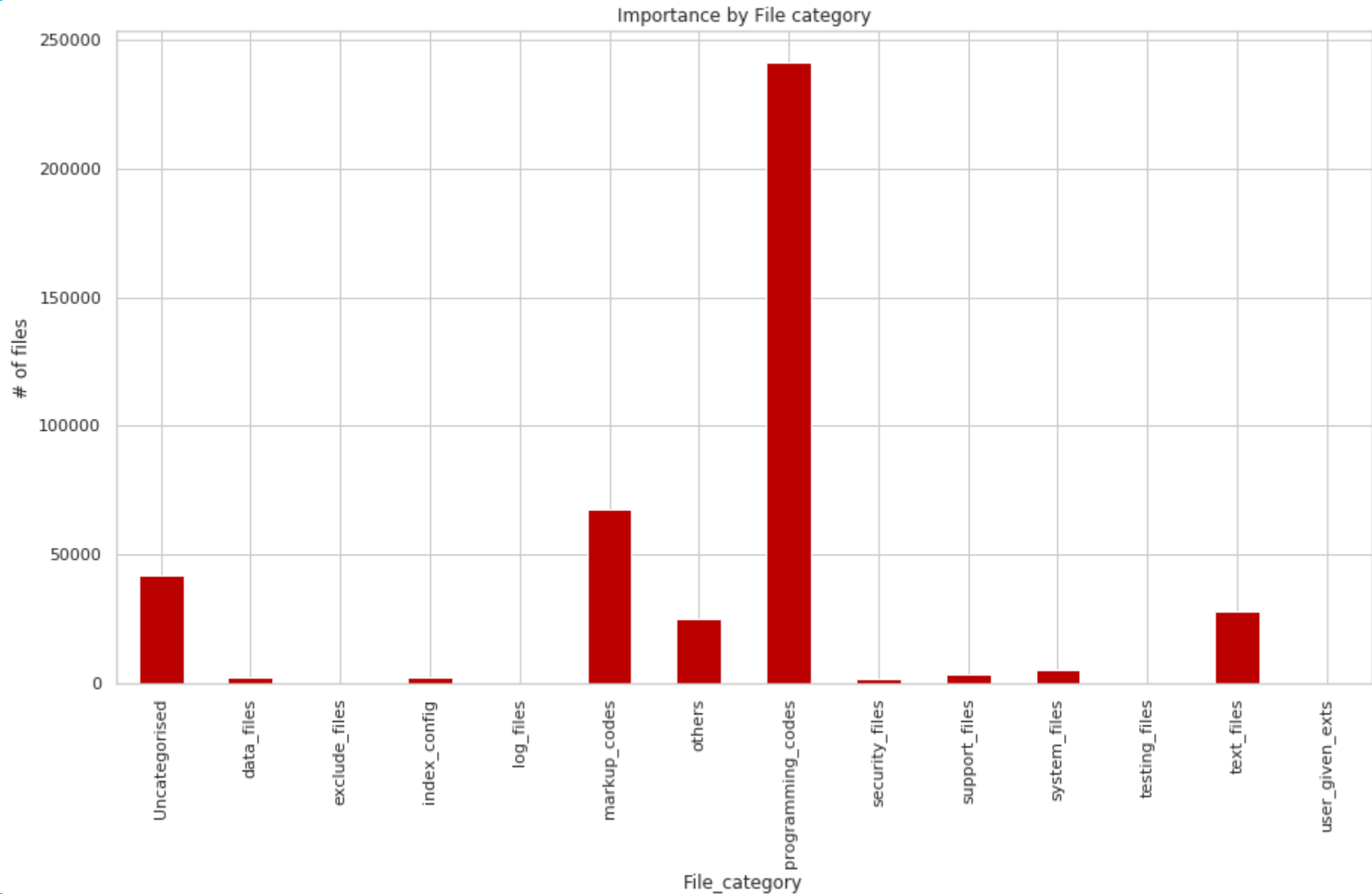
Training data exploration



Training data exploration



Training data exploration



To Do