


# Text File Format Identification

Santhilata K.V.

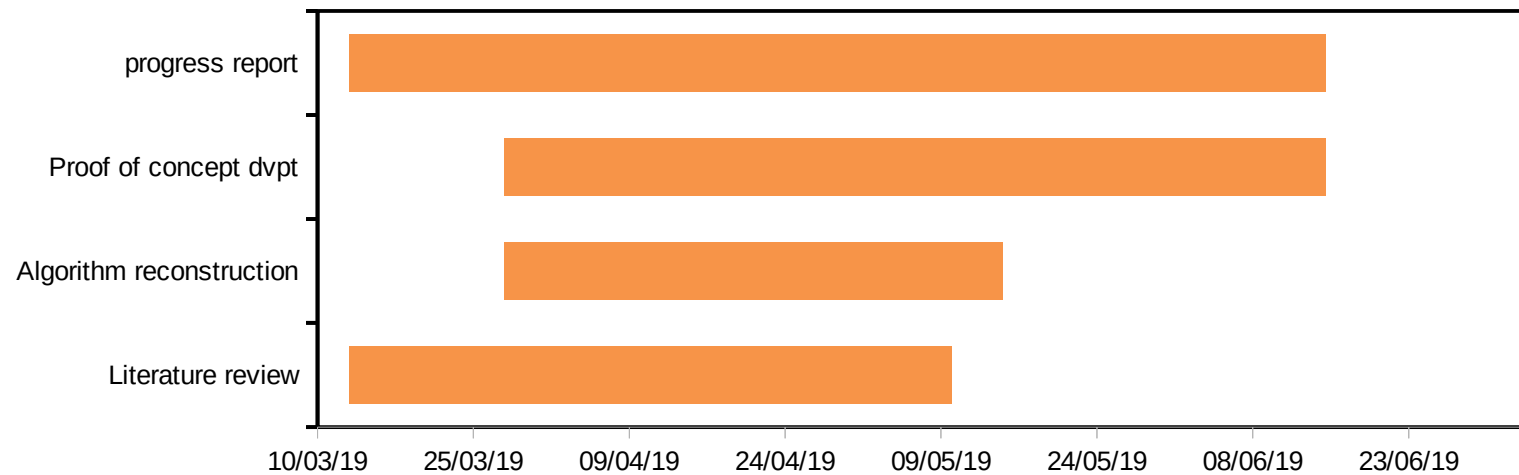
07/05/2019



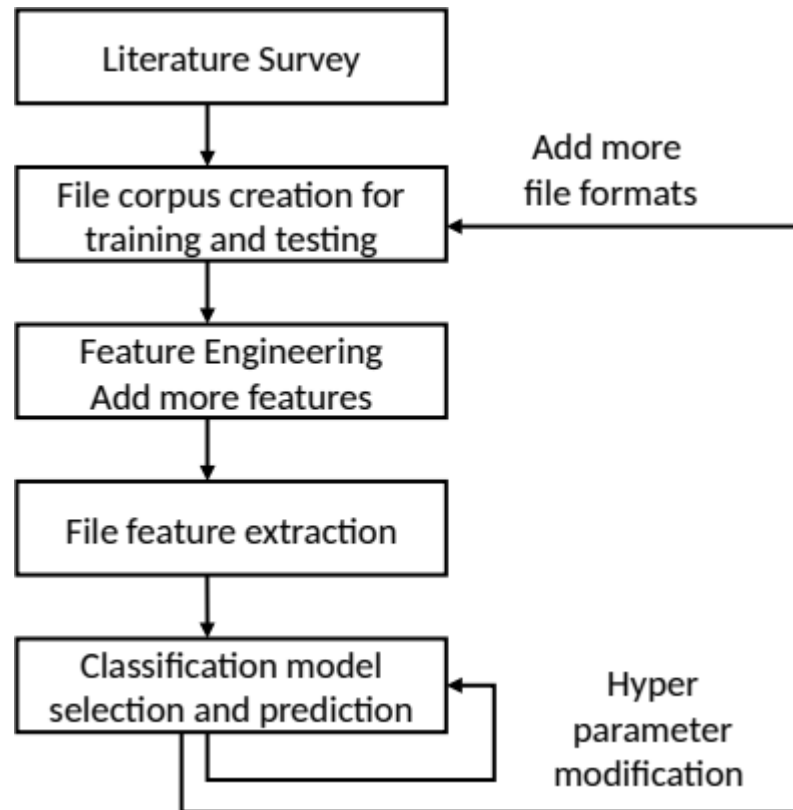
# Agenda

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

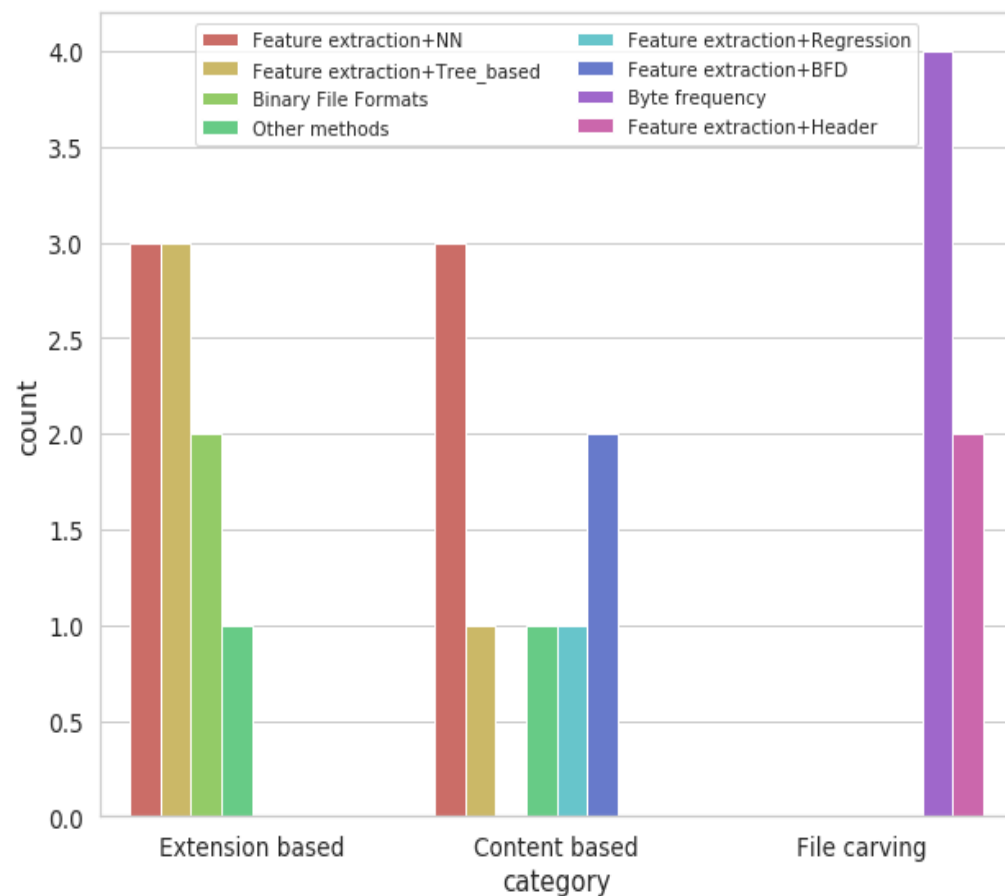
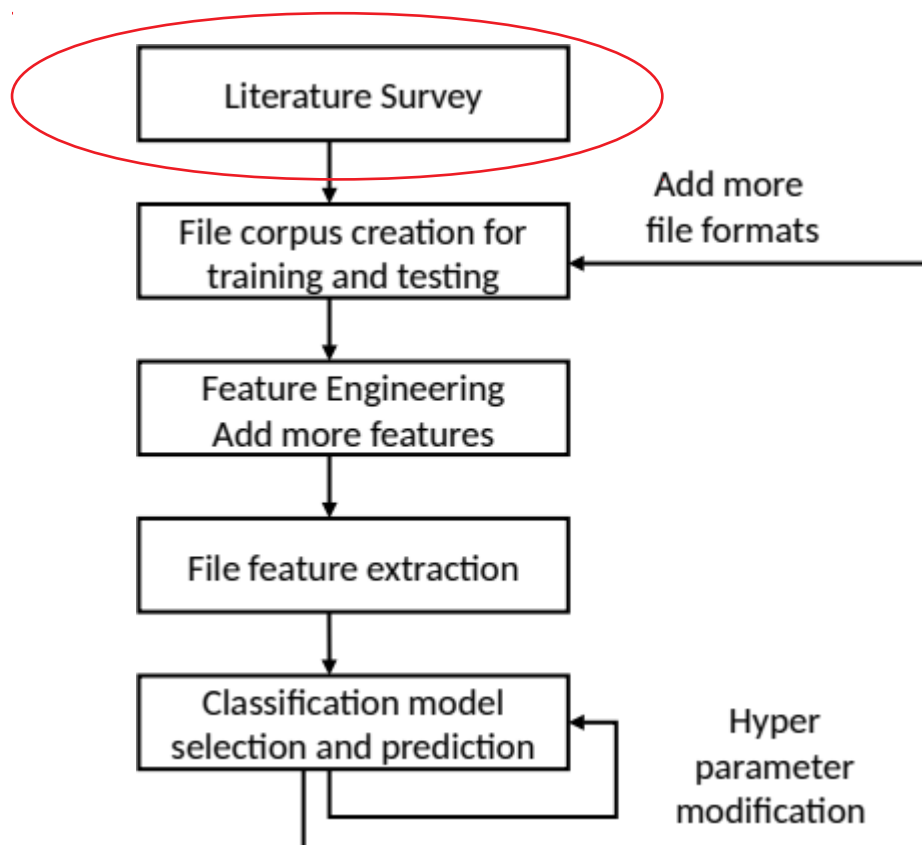
# Time line



# Methodology

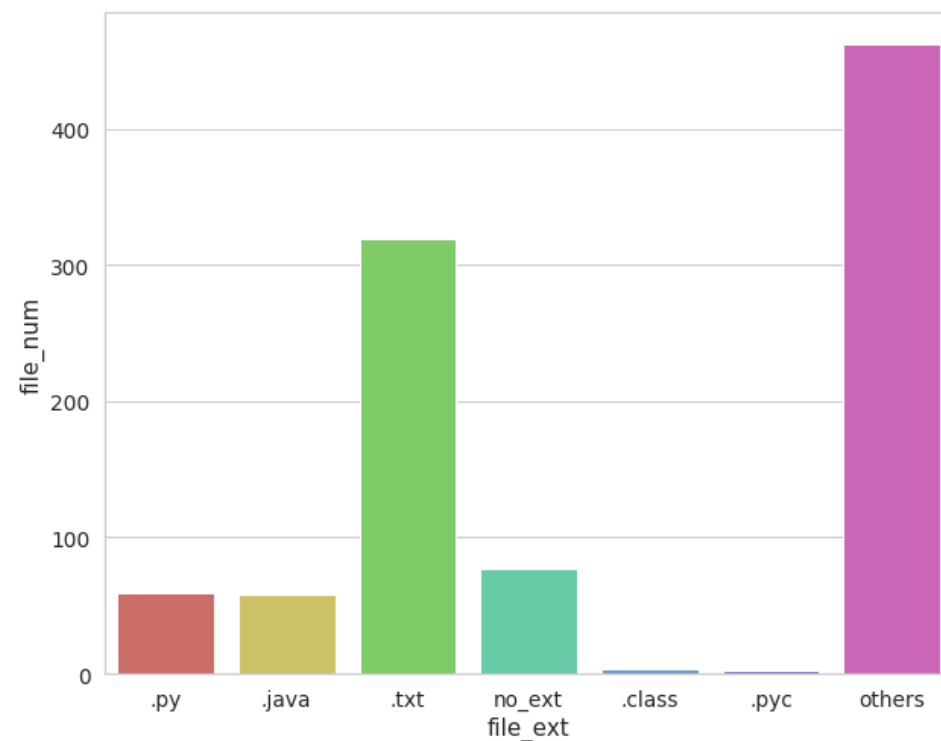
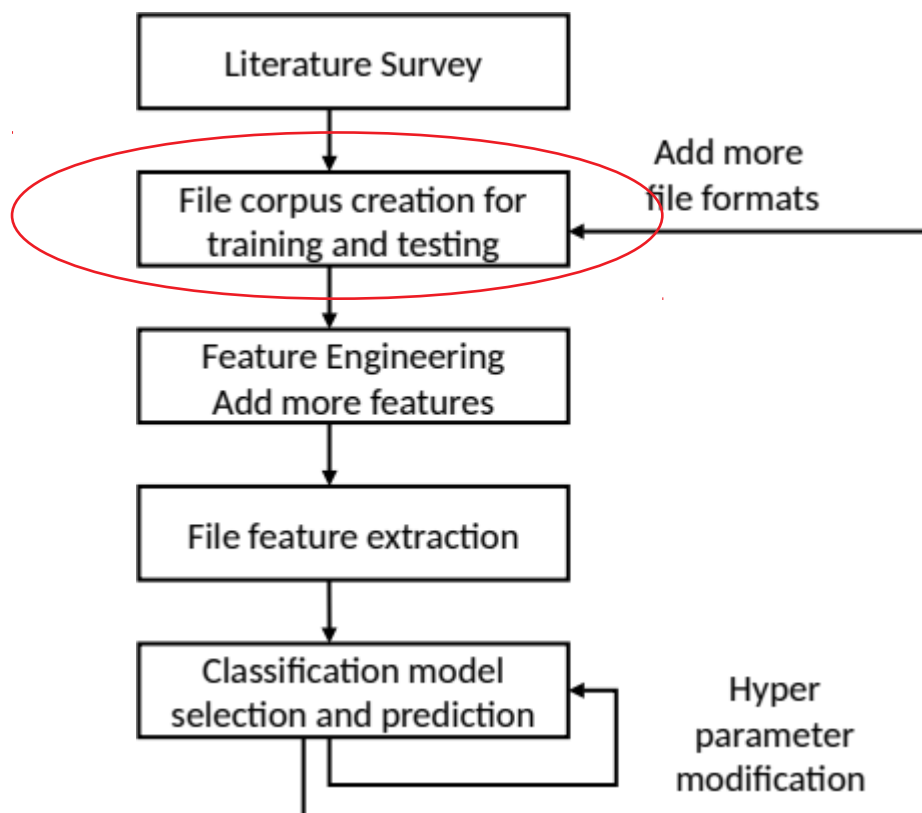


# Methodology



Total # Papers = 23 (published after 2008)

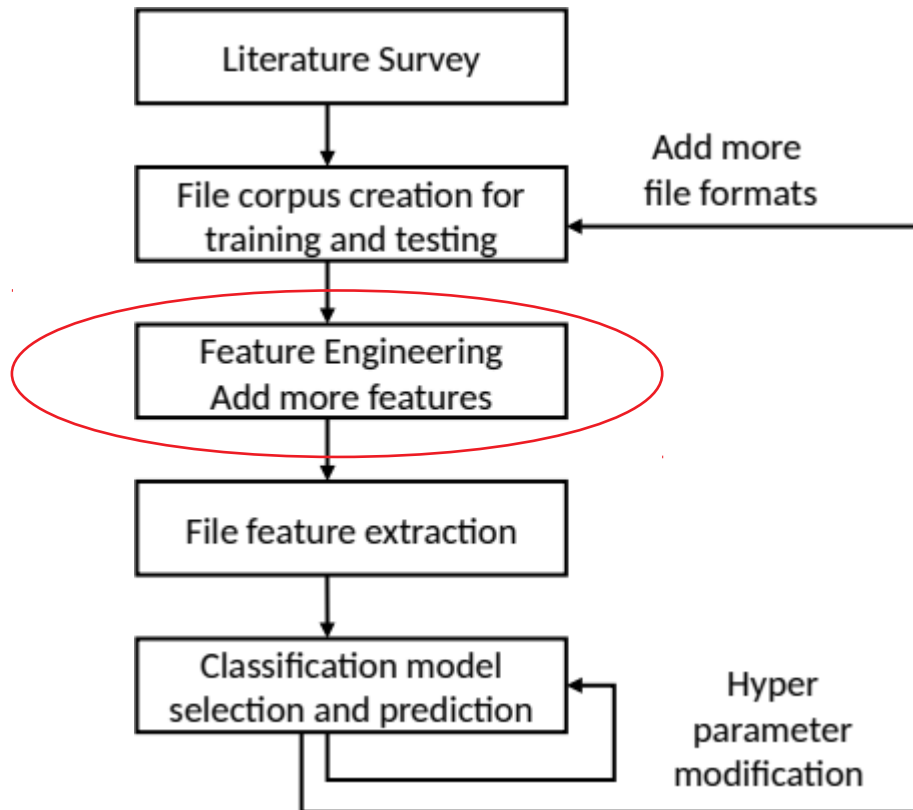
# Methodology



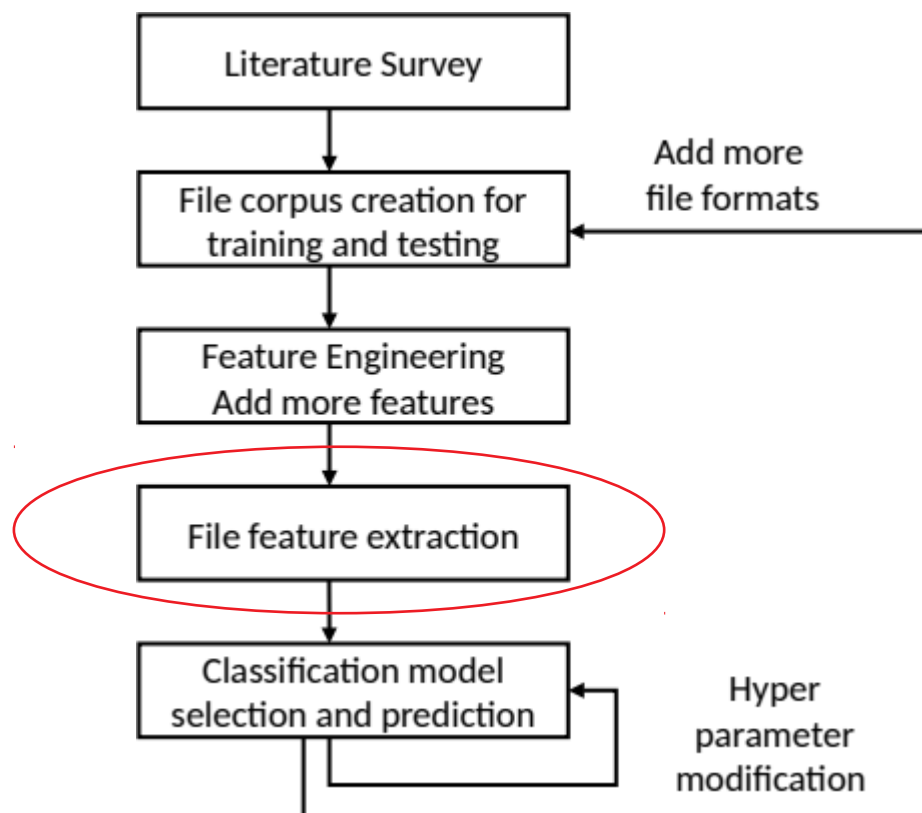
Corpus size = 980

# Methodology

- # features so far = 16
- key\_words
- file\_extension
- Header\_trailer info
- file\_structure
  - Indentation
  - SOL & EOL markers
  - Naming (case sensitive)
- Commenting style



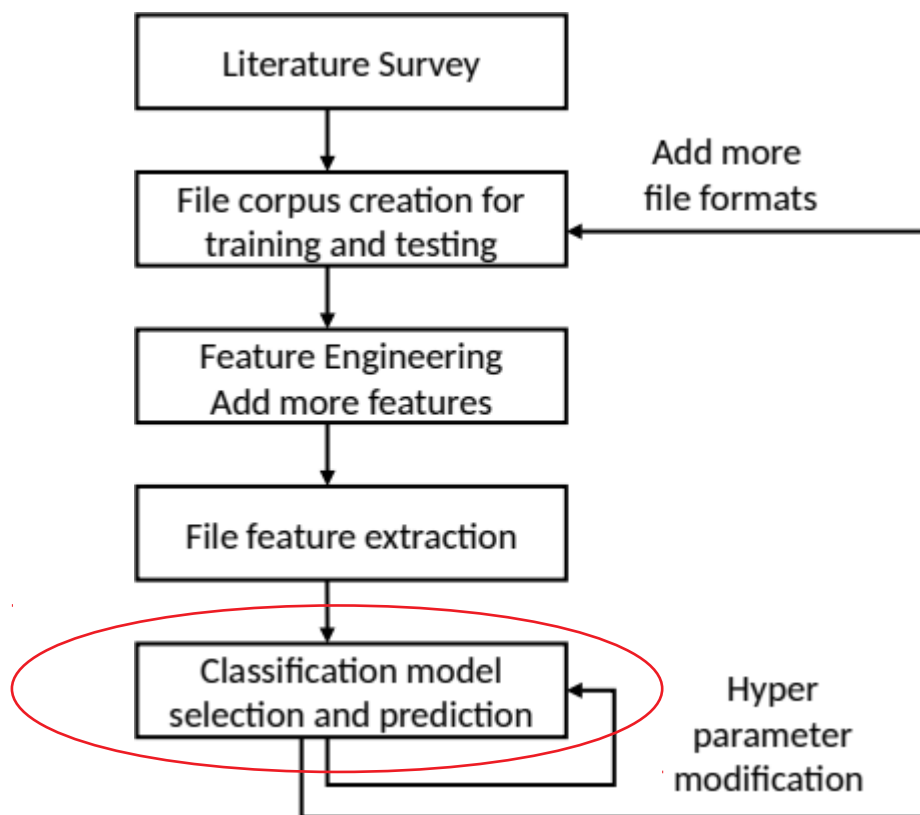
# Methodology



- Laborious process
- Creates sparse matrix



# Methodology

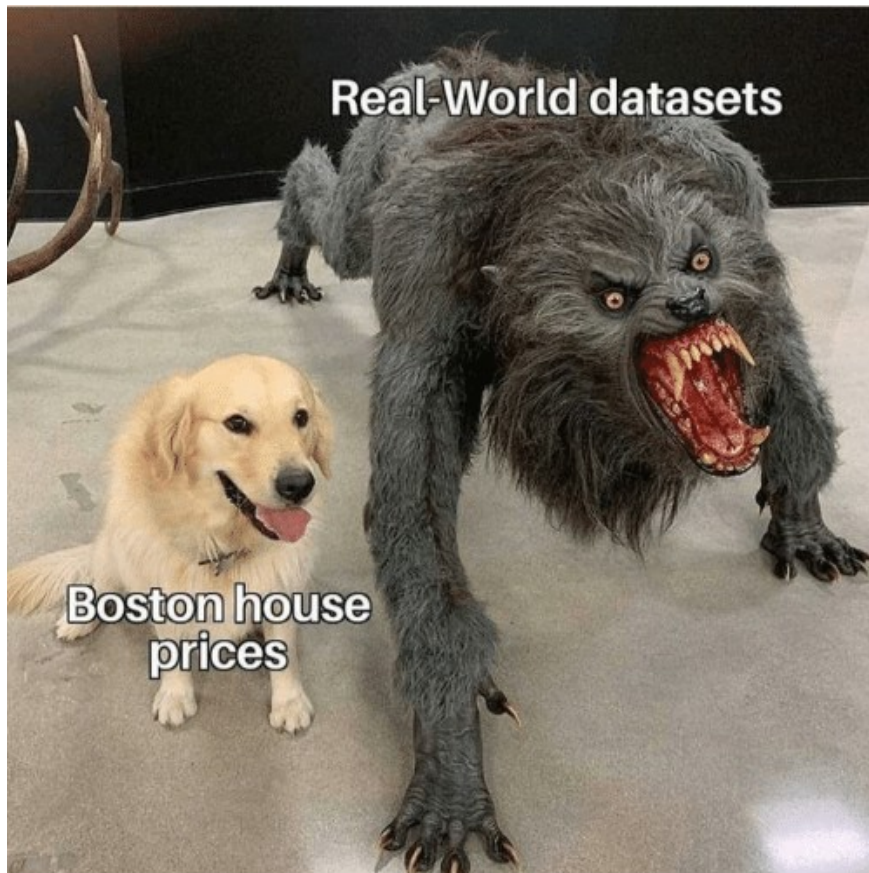


Classification model	Accuracy
Decision tree	92.15
K-Nearest neighbors	93.632

## Reasons for high accuracy:

- features are hand crafted
- corpus size is small
- few no.of classifications

# Conclusion



I think I will stick to the house prices datasets

<https://me.me/i/real-world-datasets-boston-house-oricess-0-gi-i-think-i-81679db8c7ae447baa14cc471b36548d>

- More data to be added
  - Data cleaning is the major issue
  - Need to know specific text files encountered at TNA
- More classes (file\_types)
  - Iteration of feature engineering step
  - Finding feature importance (to do)