


# Text File Format Identification

Santhilata K.V.

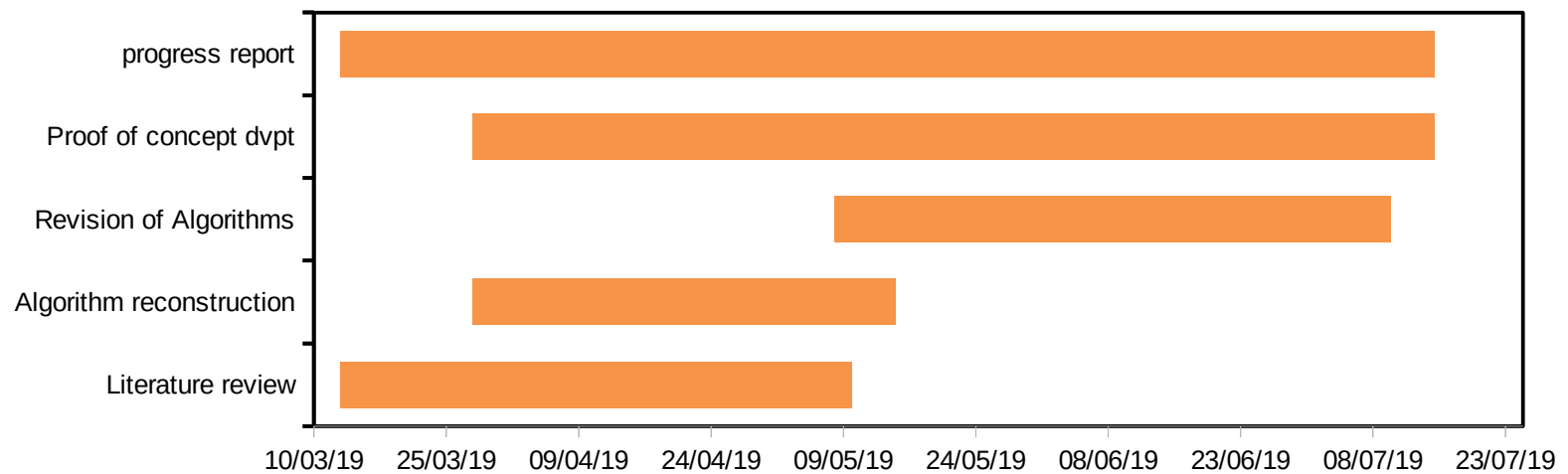
06/08/2019



# Agenda

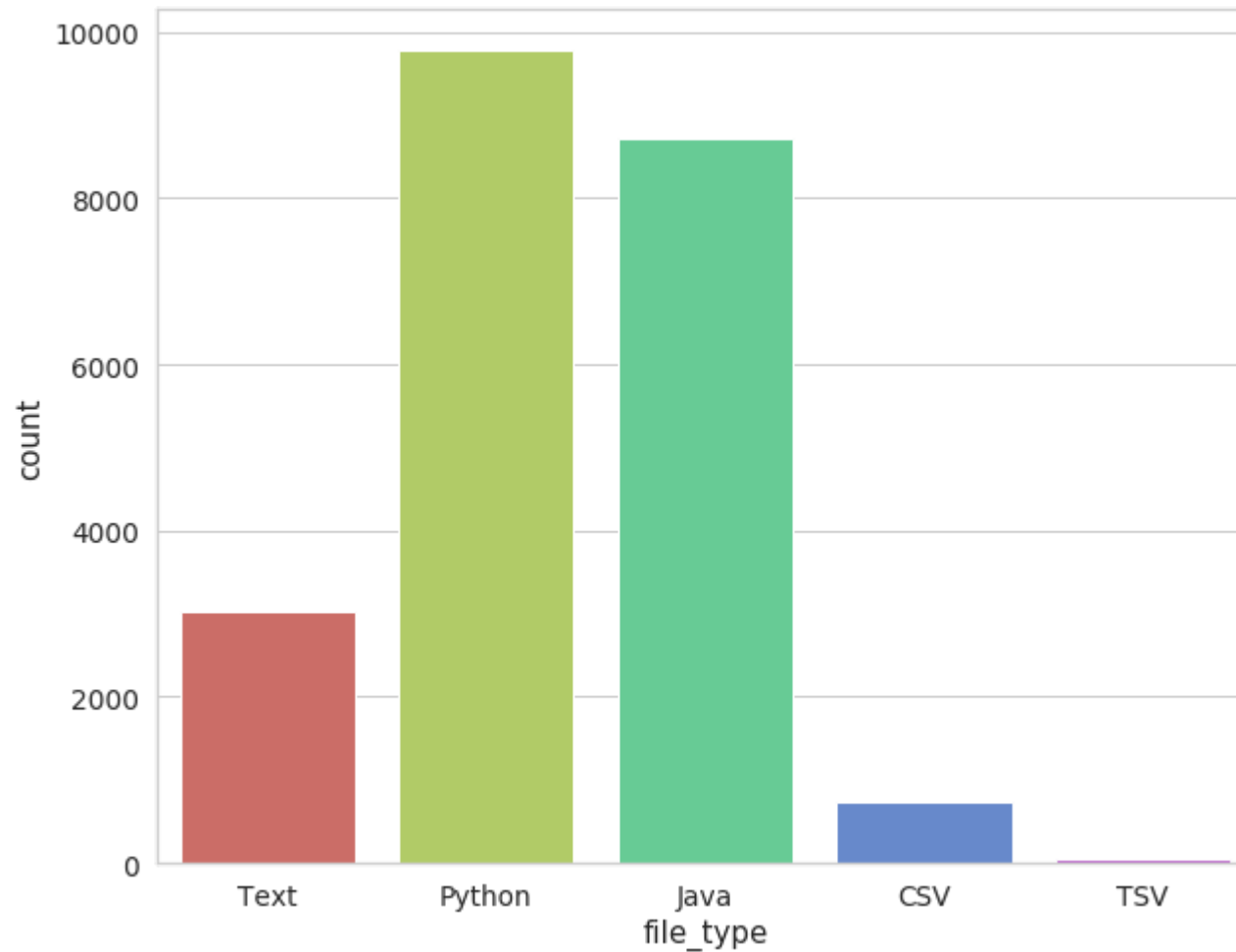
- Research question
  - \_ How to correctly identify the file format of a text file from its contents ?
- Time line
- Discussion on results
- Conclusion and Future directions

# Time line



# Discussion

File corpus – Total # of files = 22292



# Comparison of 3 classification algorithms

Decision Tree Classification  
(Train to test → 80:20)

	CSV	Java	Python	TSV	Text
CSV	139	0	0	0	19
Java	0	1734	20	0	0
Python	0	2	1923	0	0
TSV	1	0	0	3	7
Text	10	0	0	4	597

k-NN Classification  
(Train to test → 80:20)

	CSV	Java	Python	TSV	Text
CSV	117	5	4	2	30
Java	4	1677	61	0	12
Python	6	54	1852	0	13
TSV	4	0	1	2	4
Text	19	41	24	1	526

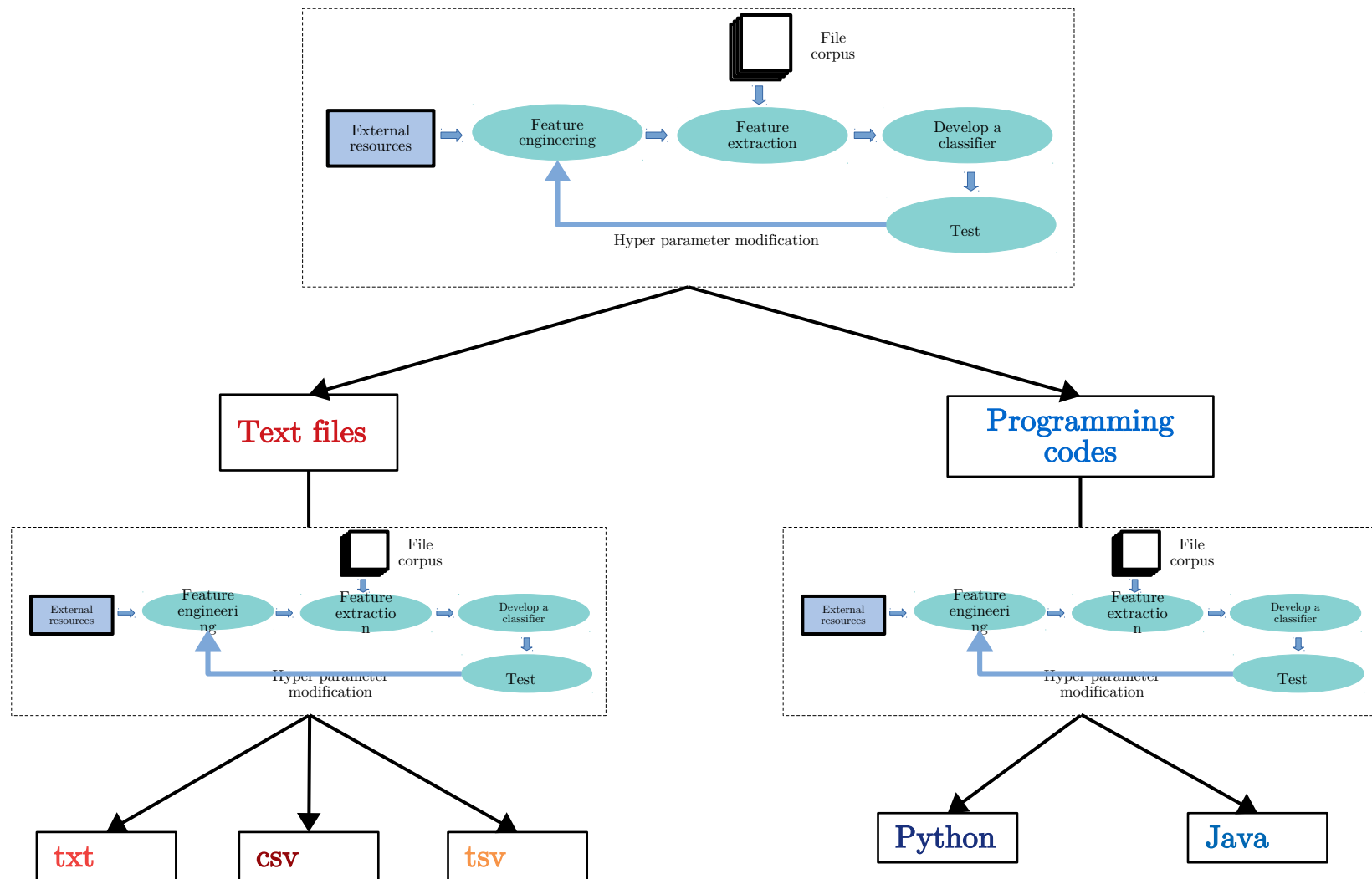
NN – 3 layer (12x12x5) , activation - relu  
train set samples = 18000, test set samples= 4292 (19% approx)

Accuracy = 97.46%

# Conclusion

- A methodology is established to find the file types from their contents.
- Current classification models work well for .py, .java and .txt extensions.
- High accuracy for Python and Java files.
- Models are not trained for tsv files due to its low representation in the corpus.

# TFFI as a Hierarchical classification Problem



# Future Direction

- Complete the dominant feature identification for .csv & .tsv
- .csv files are misidentified as txt due to excessive comments/heading data. This issue should be addressed.
- Multiple tables in a single csv file issue is ignored at the moment.
- Revised feature engineering is to be done to extract features to include new file types.
-