


# Text File Format Identification

Santhilata K.V.

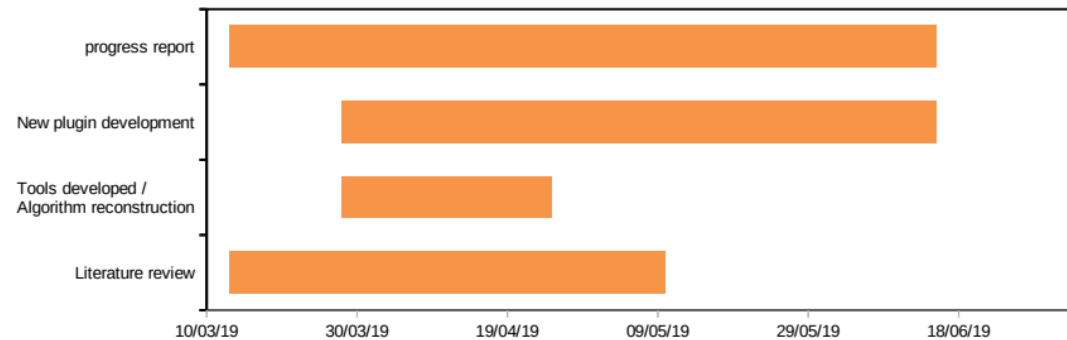
11/04/2019



# Agenda

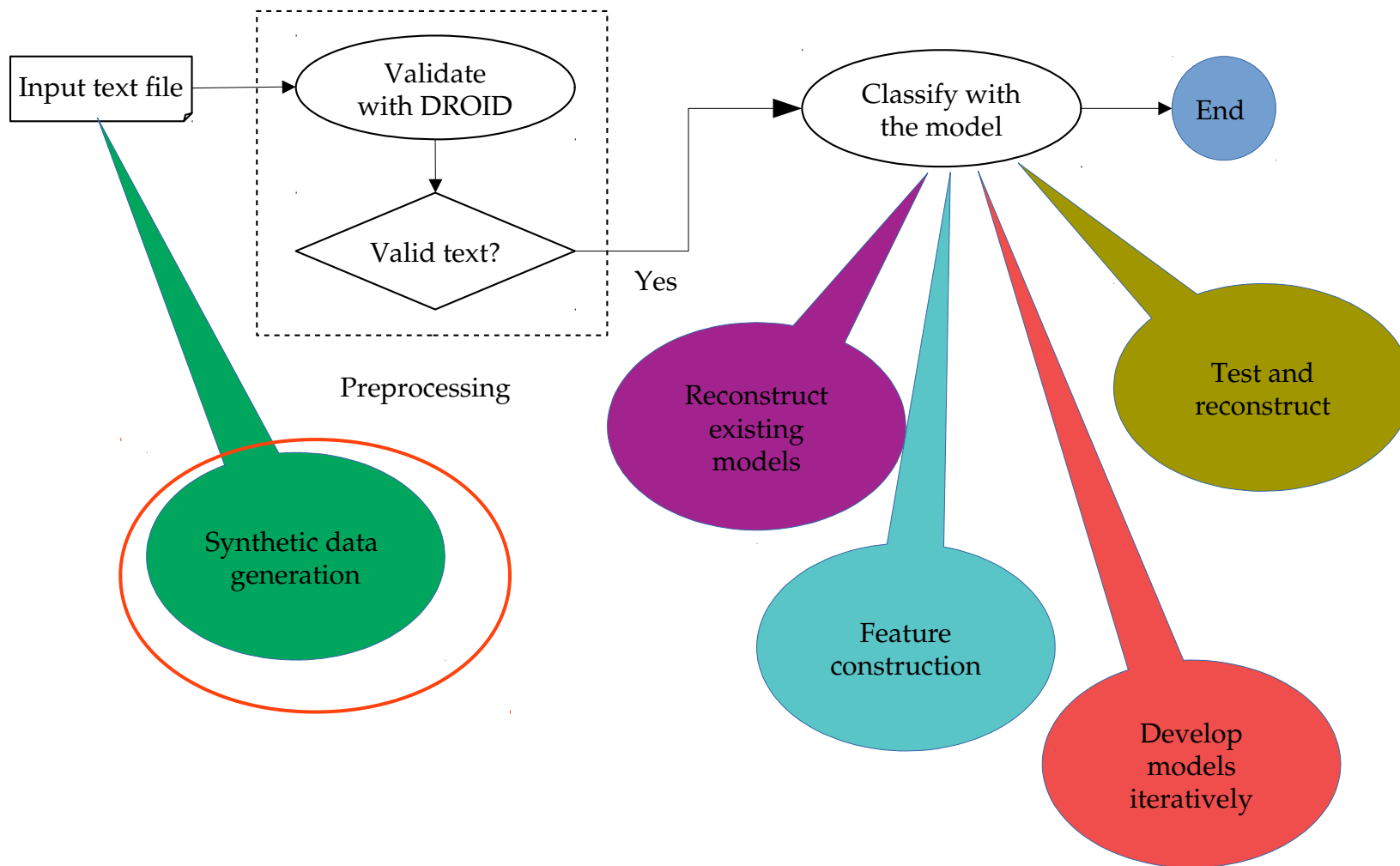
- Research question
  - \_ How to correctly identify the file type from text files?

- Time line



- Work done so far
- Targets for the next meeting

# Workflow (top view)



# Workdone

- Necessary training
  - \_ Enrolled with Coursera in Natural language processing & Text processing
  - \_ Signed for after office hours meetups in London for NLP
- Feature Engineering
  - \_ Added 13 features till now
- Synthetic File corpus creation
  - \_ Created python sample files (9) / Github files(2) with .py extension
  - \_ Stored the same also as different file formats (.txt and no extension)
- Automating the feature extraction for the training files is in progress

# Hickups

- Majority of the papers are published by IEEE
- Number of papers only discuss methods / approach but neither tool nor sample datasets. A challenge to decide whether a particular approach is useful!
- Algorithm reconstruction is going to take some time.
- a