


Text File Format Identification

Santhilata K.V.

02/07/2019

A decorative blue wavy line that spans the width of the slide, starting from the left edge and ending at the right edge, positioned below the date.

What did I do all these days?

- CAS Workshop catch up
- Updates on TFFI
- Targets for the next meeting

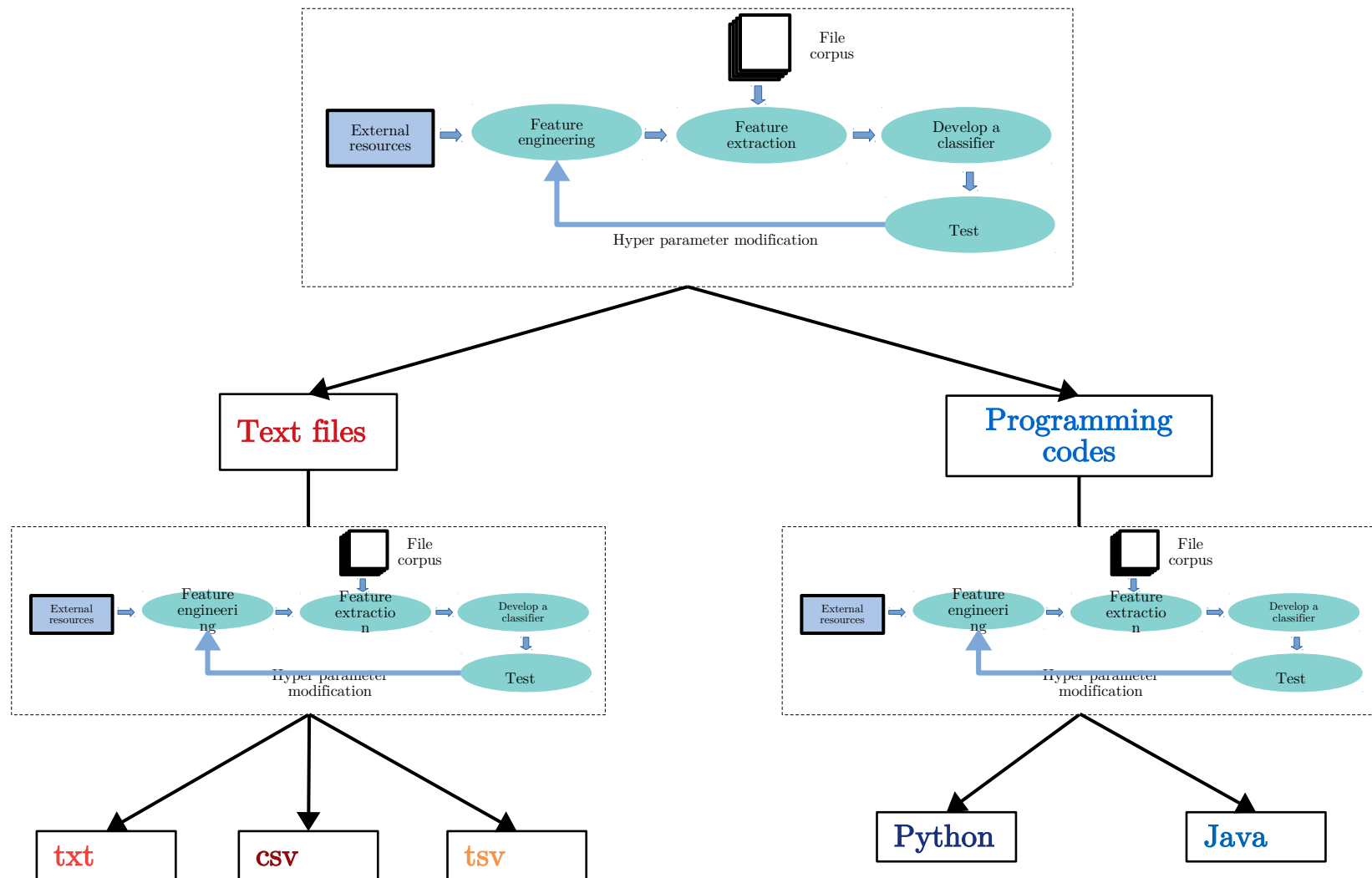
CAS Workshop – Topic Modeling

- Data manipulation
- Revision of code
- Report making

Updates on TFFI

- Status revision
 - Revised targets
 - DT classification algorithm for programming codes is working fine. But needs to be re-addressed
 - Tackling text file classification

Hierarchical process



Text file classification

- Predict normal text files from ['.csv', '.tsv']
 - Create corpus for the identification of specific terms for dominant text files (on going)
 - Idea is to come up with our own word embedding
- Feature extraction for .csv & .tsv (on going)
- Experimenting with different architectures for text classification.
 - Train word embeddings with on the corpus
 - Working with Recurrent Neural Networks

Until next meeting

- Complete the dominant feature identification for .csv & .tsv
- Come up with solid architecture to combine the classification task of structured/ unstructured text files
 - (*use of LSTM /BiLSTM*)