GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

File Changes Classification
for analysis

Ankita Bajpai
Atefeh Khajeh
Hamed Jalali
Mohsen Haghaieghshenasfard

Situation

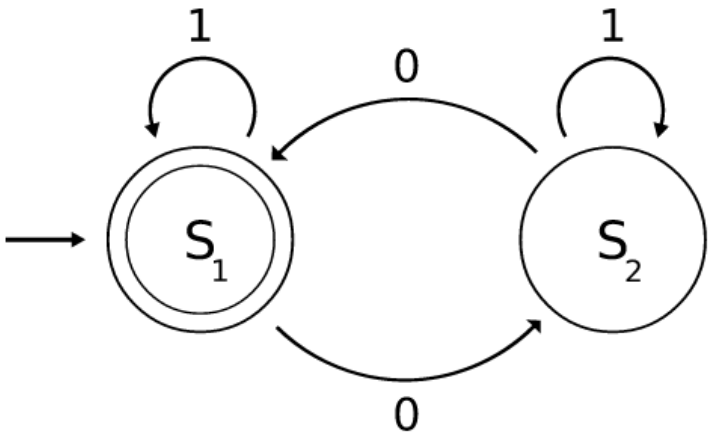
- Designing a model for a software project depends on several criteria and conditions.
- There is a direct relationship between the model of the project and the level of complexity regarding to diverse parts and phases of it.
- Although each model is designed accurately, it is needed to be altered during the implementation phase.

Project Goal

- Control of metrics and relationships between files and merge or split them to make the model more robust

More Description

- Sampling of data based on the hunk collection
- Combine its columns and define new features
- Designing and implementing the proposed model both based on machine learning algorithms



Ref. https://en.wikipedia.org/wiki/Automata_theory

Raw Data

- Hunk Collection

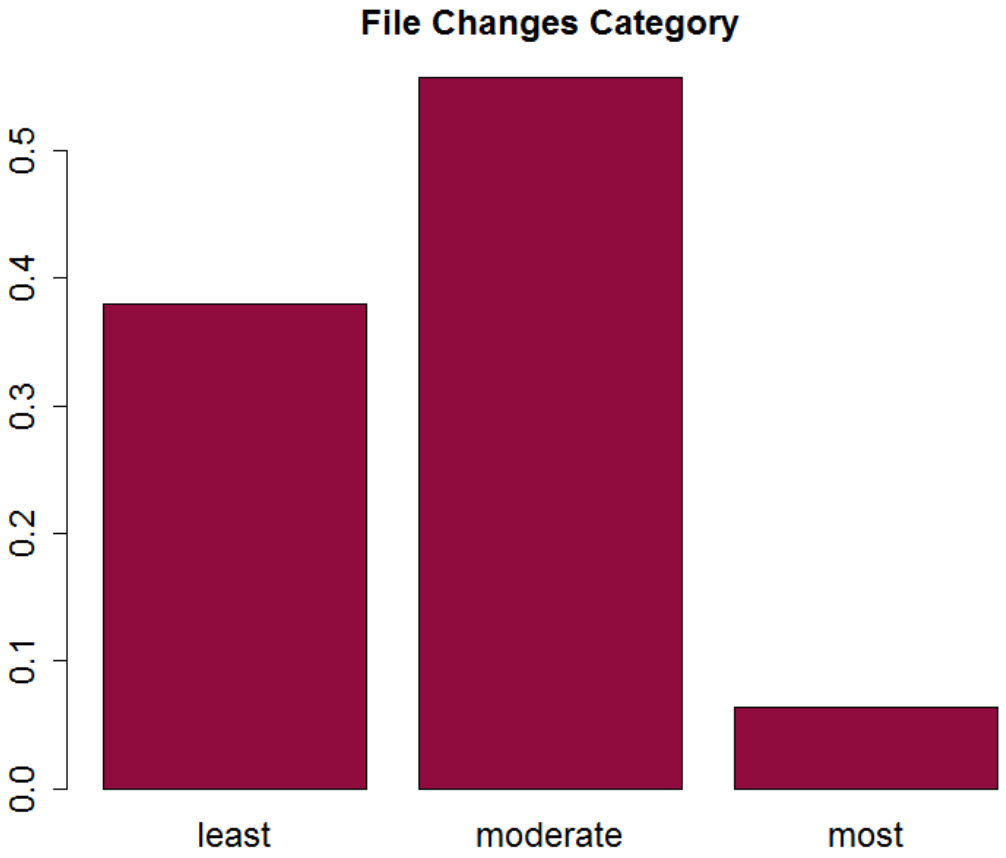
file_action_id	new_start	new_lines	old_start	old_lines	content
5853eb373ee1b95d6d8826f3	7	4	7	2	+ ZOOKEEPER-1433. improve
5853eb373ee1b95d678826f3	18	2	17	0	+ ZOOKEEPER-1339. C client
5853eb373ee1b95d6d8826f6	439	1	439	1	- if (pyw->permanent == 0 &

- newHunk Dataframe

(file_action_id)	sumNewLines	sumOldLines	totalChanges	frequency
5853eb373ee1b95d618826f3	6	0	6	1
5853eb373ee1b95d618826f6	1	1	2	1
5853eb373ee1b95d628826f3	2	0	2	1

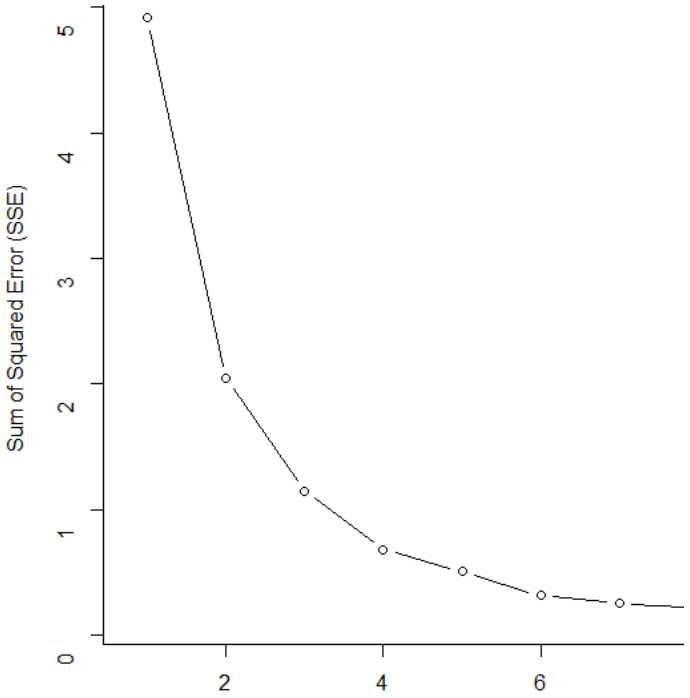
Key Points

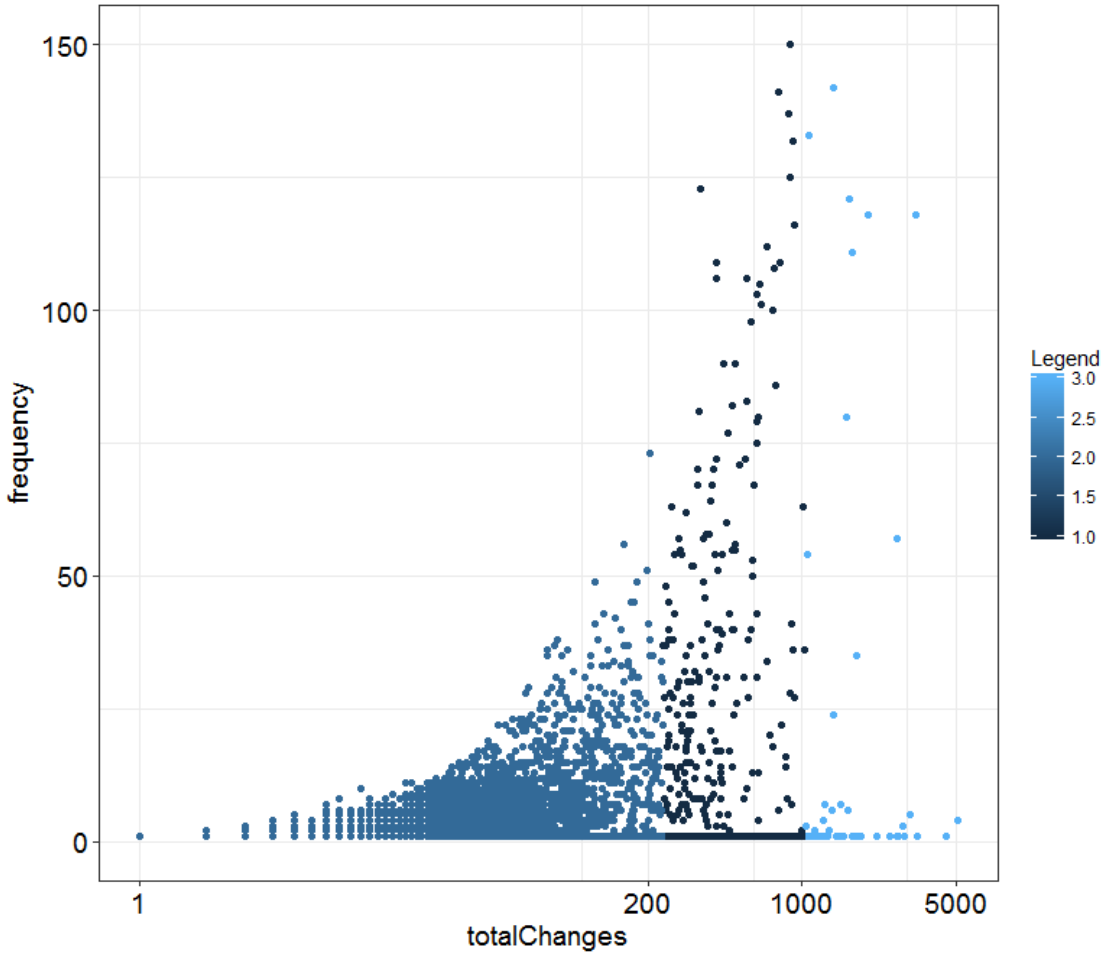
- Classifying of file changes based on some analytic statistics i.e. Min, 1stQu., Median, Mean, 3rd Qu., Max



Key Points (cont.)

- k-means algorithm





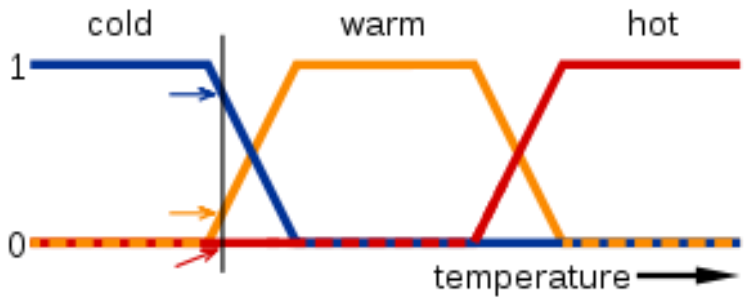
Recommendation

- Designing a boosting model based on several lazy models constructing by different features regarding their correlations

- Run the predictive model hourly and daily to be proactive on the future complexity of project

Further Works

- Find the relationships between different files and identify which files are highly correlated using “Graph Mining”
- Designing a Fuzzy model for controlling the complexity of the project based on “Computing with Words” as weights of the model are determined by words and Fuzzy Memberships



- Online Reverse Modeling in Software Engineering

Ref. https://en.wikipedia.org/wiki/Fuzzy_logic

QUESTION?