# Analyze Similarities in Source Code – high-level timeline

|  |  |
| --- | --- |
| Month | Content |
| January | Select Checkmarx project:   * Analyze source code similarities * Automated clustering of vulnerabilities   Have a kick-off meeting @ Checkmarx, with the mentor  Define a high-level plan |
| February | Read relevant algorithms  Learn relevant material in Clustering and distances  Get the code of the first project [Webgoat.Net (open source)] and its csv  “Play” a little with the raw data  **Project presentation in the course** |
| March | Preprocess the data  Define a relevant distance measurement  Run 2-3 different clustering algorithms  Analyze first results |
| April | Use coarse-similarity and fine-similarity for improved performance  Fine-tune similarity values to get better results  Fine-tune the clustering algorithms |
| May | Analyze the clusters to extract outliers  Test outliers to find a fine threshold for outliers  Extract at least one “interesting” result from the outliers |
| June | Analyze a larger codebase in C# (Checkmarx codebase?)  Fine-tune the clustering algorithms  Fine-tune the outliers algorithms  Fine-tune the distance calculations  Extract at least 5 “interesting” results from the outliers |
| July | Stretch goals:   * Analyze some non-C# code [dependent on a csv extractor from Checkmarx] * Find interesting results on open-source code   **Present the results** |