**Project Book**

**Contents:**

* **Abstract**
* **Project Proposal**
  + Add to part of previous work
  + Update the project aim in the end
  + Add how the project is going to be (write about the phases: aim and tasks in each one)
* **Useful tools used**
  + Pdfid
  + Peepdf
  + Jast
  + Analyze pdf
* First phase: preparations
  + Extracting telemetry
    - Code & explanations
  + Extracting text from picture
    - Code & explanations
  + Extracting text from pdf file (using PDFMiner)
    - Code & explanations
  + Extracting URLs using pyPDF
    - Code & explanations
  + Extracting URLs from JS in the files tags (using peePDF)
    - Code & explanations
  + Extracting preview of a PDF file (thumbnail) using PIL + pdf2image
    - Code & explanations
* Second phase: Image based classification machine
  + Vector & Features:
    - Pixel vector
    - Histogram vector
    - Blur (Laplacian method)
  + ML Algorithms:
    - KNN
    - KNN with clustering
      * Elbow method (StandartScaler, MinMaxScaler
  + Results with the different approaches
  + Code & explanations
* Third phase: Text based classification machine
  + Reading text or preview:
    - By checking num of characters
  + Creating the vector
    - Doc2vec
    - Word2vec
      * GoogleNews-vectors-negative300-SLIM.bin.gz
      * GoogleNews-vectors-negative300.bin.gz
    - TF-IDF
  + ML Algorithms:
    - Logistic regression
      * DBOW
      * Distributed Memory with Averaging
      * Combining DBOW & DMA
    - Random Forest
    - K-Nearest-Neighbors
    - SVM
    - NB
    - LR
    - MLP
  + Results with the different approaches
  + Code & explanations
* Fourth phase: JS, URL, objects and streams based classification machine
  + Creating the vector – features:
    - PDF Tags:
      * Each feature:
        + How was it extracted
        + Why is it important
        + Reading materials about it (if there are)
    - JS:
      * Each feature:
        + How was it extracted
        + Why is it important
        + Reading materials about it (if there are)
    - URLs:
      * Each feature:
        + How it was extracted
        + Why is it important
        + Reading material about it (if there are)
    - Objects & Streams:
      * Each feature:
        + How it was extracted
        + Why is it important
        + Reading material about it (if there are)
  + Improving the vector
    - Random choice method
    - Summing method
    - Combining features
  + ML algorithms:
    - SVM
    - LR
    - MLP
    - KNN
    - RF
  + Results with the different approaches
  + Code & explanations
* Fifth phase: Creating the ensemble classifier
  + Combining machines
  + Classification method
  + Results
* Future Work
  + Phase 1 Improvements
    - Extraction of text to work for more languages
  + Phase 2 Improvements
    - Additional methods for picture classification
      * Near similar image matching
  + Phase 3 Improvements
    - Additional methods
      * Word2vec with cosin similarity
  + Phase 4 Improvements
    - Feature selection improvement (whatever we did not get to do in phase 4)
    - Additional methods
  + Phase 5 Improvements
    - Additional methods