PEC 1: Final work definition and planning

Author: Teguayco Gutiérrez González
Studies: Master's Degree in Data Science

1. Title proposal

The chosen title for the final work will be "*A Product Data Matching System for an e-Commerce Aggregator*".

2. Keywords

Product Matching, e-Commerce, Machine Learning, Natural Language Processing.

3. Abstract

In this work, we propose the construction of a system using Machine Learning and Natural Language Processing (NLP) techniques to automatically detect matches between products whose data come from different sources.

4. Proposal description and relevance

Nowadays we can find on the Web a bunch of different e-Commerce aggregators, which are websites that collect data of products from different online shops such as Amazon, MediaMarkt or ToysRUs with the aim of showing their customers the best options to buy a certain product. Google Shopping or Idealo are two famous examples of e-Commerce aggregators.

However, the problem of identifying the same products collected across many sources, known as "Product Matching", must be addressed in order to offer an optimal service. In most cases this is not a trivial task, as data is presented in different formats depending on the source they come from.

5. Personal motivation

There are two main reasons involved in the choice of this proposal:

- Once built, if the system gets good results matching products, it can be used to improve the quality of an e-Commerce aggregator which has been developed in the company the student works for (from now on, the ***client company***). This is a task which is currently being performed manually by a group of matching operators.

- The problem to be solved induces to implement different tools and techniques which have been seen throughout these studies (Machine Learning, NLP, web scraping, …) and during the previous studies of the student as well (Computer Science Degree).

6. Objectives

The main objective of the work is to develop one or more than one Machine Learning classifier that can be able to detect matches, with considerable confidence levels, between products coming from different web shops and products known for the client company contained in its master-data storage. Thus, the matching will be mainly performed based on the product titles.

These products to be matched will be initially electronical products such as computers or mobile phones, which are the predominant types of products the client company operates with.

Furthermore, the following can be taken as secondary objectives:

- Develop an app to input data for a certain product to output the best match for it. This app can serve as a proof of concept, so that it can be extended in the future according to the specific needs of the company.

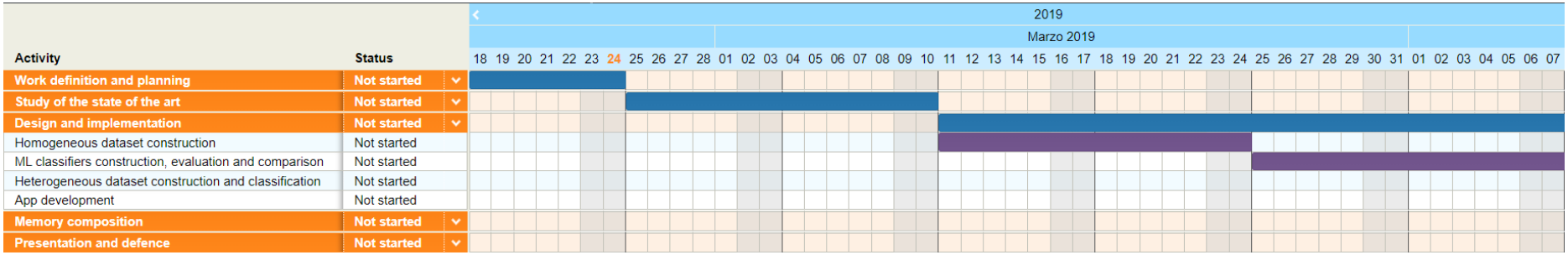
- Extend the ML classifiers to work with more heterogeneous products, e.g. toys, CDs, televisions, etc.

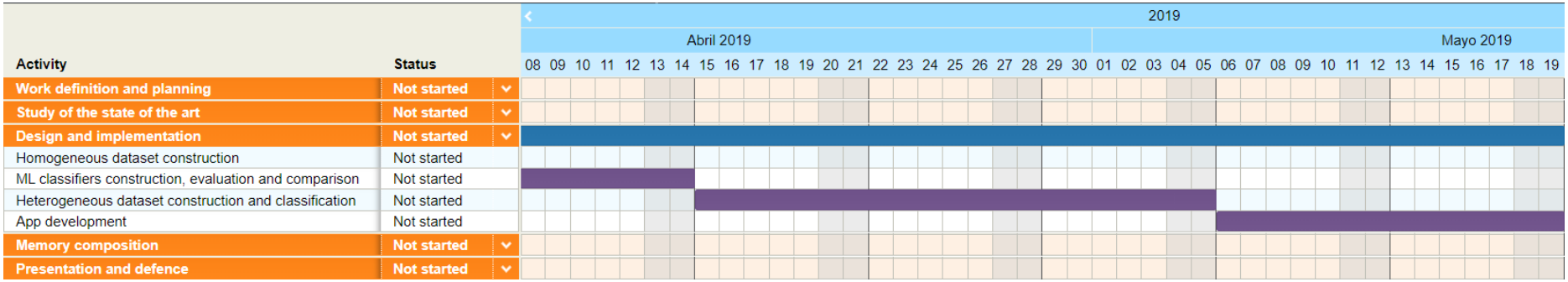
7. Methodology

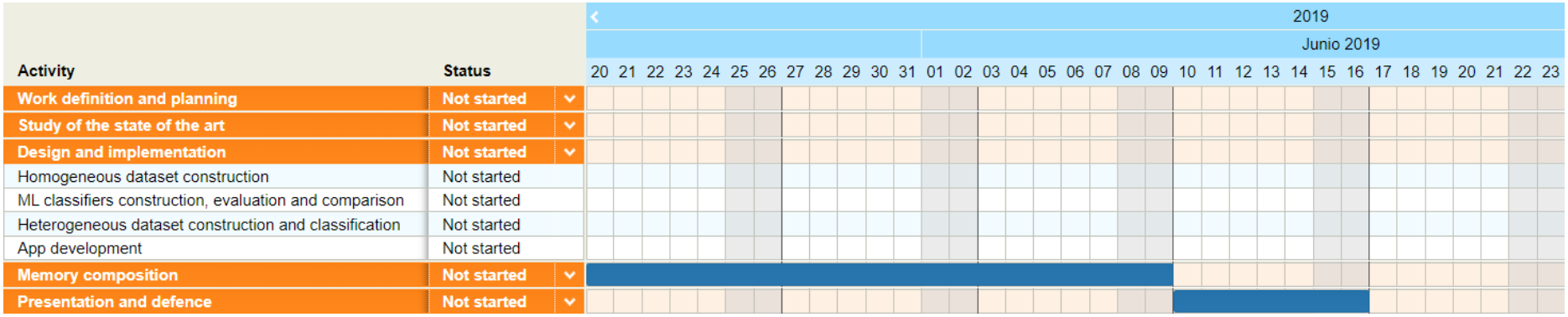
Given the large variety of products the client company handles and the amount of Machine Learning classifiers that can be applied, a progressive approach will be used throughout the development of this work. Thus, if time allows, more heterogeneous datasets will be generated and used to train different ML models each time.

Furthermore, as there are a lot of works addressing the “Product Matching” problem, the already-used approaches and techniques will be thoroughly analysed to then apply the subset of them which best fit for the specific product data the client company owns.

8. Planning (Gantt Diagram)







9. Bibliography

Some of the literature consulted so far include the following works and papers:

[1] "A Machine Learning Approach for Product Matching and Categorization". Petar Ristoski, Petar Petrovski, Peter Mika, Heiko Paulheim. 2017. http://www.semantic-web-journal.net/content/machine-learning-approach-product-matching-and-categorization-0.

[2] "Attribute Extraction from Product Titles in eCommerce". Ajinkya More. 2016. https://arxiv.org/abs/1608.04670.

[3] “Matching unstructured product offers to structured product specifications”. Anitha Kannan, Inmar E. Givoni, Rakesh Agrawal, Ariel Fuxman. 2011. https://www.researchgate.net/publication/221654670_Matching_unstructured_product_offers_to_structured_product_specifications.