**E Health Methods and Application: Project Report Part 1 GROUP 04**

**DATABASE**

STEP 1: Download Google-Playstore.csv dataset from Kaggle

We decided to work on the Google-Playstore database provided by Kaggle, because it was the more complete file describing the Google applications. Moreover, this file contains a column named “App Id” that will be useful when we will work with google-play-scraper, as described in the following steps.

STEP 2: Filtering the applications by relevant educational features

We have implemented a function that filters the applications by certain categories ("Education", "Educational", "Family", "Learn", "4 year old kids" and "4 year olds"), and minimum requirements on rating (4.0 or above), rating counts (>100000). In addition, in order to exclude the apps whose information are not translated in English, we provide a function capable of determine the set of characters that compose the app name and decide if it’s an app written in English or not.

We chose those parameters based on our assumptions of what a “real” serious game should be, and taking into account the feedback provided by the users of the app too, as a reliable source of “good quality”. For instance, we looked at all the categories defined in the dataset, and decided to select only those above-mentioned, as they were the ones related semantically to education and children. We chose a minimum rating of 4.0 because we noticed that many applications had a very high rating. Regarding the number of ratings, we decided to filter them as they play an important role in determining the reliability of the rating. Indeed, if a game has a rating of 5/5 but only has a few ratings, the rating may not be very representative of the quality of the application.

STEP 3: Selection of app classified as “games”

Then, we have implemented a filter to identify the educational games. Indeed, each application has a specific category and sub-category: the latter starts with *GAME\_CATEGORY* in case of a game, thus we checked the subcategory of each educational app with google-play-scraper (using the ID of the apps as a reference) to remove all those that were not games.

Furthermore, since the Google-Playstore.csv was not up-to-date, some apps presented in it were not found with google-play-scraper (they were probably deleted from the Google Playstore): consequently, we identified and removed them from our dataset.

STEP 4: Defining a first function to enrich database with description and reviews

Our dataset now present educational games and specific features. Nevertheless, some important characteristics are missing. Thus, we have implemented a first function that is able to enrich our database with the description and the number of reviews of each educational game, using google-play-scraper. We believe that the description can be very useful to understand the purpose of a game and to judge how serious it is. Moreover, the description will be used in the next steps to extract other important parameters from the games. Regarding the number of reviews, we decided to display them as they can help us, later in the progress of the project, to judge the users’ critics regarding the games.

STEP 5: Using NLP approach to identify learning category and age range

Then, we wanted to classify the different games according to their learning categories and age ranges. Since these parameters could not be found in the initial dataset nor with google-play-scraper, we decided to use the method of Natural Language Processing. For each parameter, we first defined specific categories: ["science", "counting", "language", "creativity", "shape", "food", "music" and "sport"] for the learning categories and [“babies”, “children”, “adolescents”, adults”] for the age ranges. These categories have been selected so as to separate the games precisely without being too widely. Then, for each of these categories, we wrote a list of associated keywords by looking at synonyms or semantically-related words. From those lists of keywords, we wrote a function that counts the maximum number of keywords found in the applications’ descriptions for each category, in order to assign a specific category to each one of them.

STEP 6: Create a new dataset to train and test the model we got

We chose 120 applications from the Google Playstore randomly, including 40 serious games, 40 “fake” serious games (defined as those games that seem “serious”, but that are actually not and that we can consider as misleading serious games), and finally, 40 random applications.

This dataset (entirely built by us) has been compared to the one found by the algorithm in order to measure three parameters:

* *Accuracy*: the ability to recognize both the serious games and the misleading ones;
* *Sensitivity*: the ability to recognize the serious games;
* *Specificity*: the ability to recognize the misleading serious games.

STEP 7: Create the final database