**INTRODUCTION TO DATA SCIENCE**

**Applications reviews analysing**

Vardit Arkash

Gal Turgeman – 308222595  
[gal.turgeman@mail.huji.ac.il](mailto:gal.turgeman@mail.huji.ac.il)

Yarden Yagil - 311549083  
[yarden.yagil@mail.huji.ac.il](mailto:yarden.yagil@mail.huji.ac.il)

**Introduction:**

Our goal in this project is to find the main problems in an application, in order to enable the creators to improve their application in an easier way.  
We were focused on games applications from Google Play website. In order to implement it we have collected reviews from the applications, and by using NLP tools we tried to deduce what are the problems in the application, if there are any.

We defined 6 main problems to look for in the applications:

* Ads: there are to many advertisements in the application.
* Money: users think the app doesn’t worth the money it cost.
* Flow: the app is slow or getting stuck
* Interest: the app is boring
* Bug: there are bugs in the application
* No problem: the application doesn’t contain any of the above.

We analysed the reviews according to this 6 problems.

**The Data**

For the project we needed to get a data of reviews of different game apps. For this purpose, the first part of the project is collecting the apps reviews by crawling the Google Play website (https://play.google.com/store/apps). In this part we go through the various categories of games on the website and collected the users reviews of each application. The data we created consists of X applications and about 40 reviews for each one. The data size is X GB.

First the cawler goes through the html pages of the various categories of games (i.e. action, board, cards, adventure, sport) and collects the urls of the applications. Afterwards it parses the html pages of the applications and collects the user's reviews of each application. Finally, all the reviews are saved in a dictionary which is the database that will serve us in the project. The dictionary is structured so that the key is the application name and the value is a list of all the reviews of that application. The data collection is done once and the dictionary is saved in a file called "reviews.txt" which is located in the "files" folder.

**The solution description:**

In order to evaluate the negative feedback for the apps, we use 2-parts algorithm and some manual process that was done. The description will refer to each of the parts.

**Level 1: Preprocessing the data.**

In the first part, we focus on the extracting the data and extracting some useful information from it. The data extraction was done by crawling the website pages that contains the applications' reviews. Later we use the data we extracted in order to create two files with the information we need for the next level:

1. reviews.txt - contains all the reviews. Ordered by a json format, each app has its own reviews applied to it.
2. info.txt – contains two lists. The first one has apps' names and the second has the popular subjects that appears in all the apps' reviews.

In the code -

We have a preprocessing module that is responsible for creating the mentioned files. It is done by running two other modules:

1. crawler.py – scrapping the web in order to extract the data.
2. topic\_extractor.py – given the data, it finds the popular tokens in it.

*All the code files for this level can be found in the 'preprocessing\_step' folder and the files that were created can be found in the 'files' folder.*

**Level 2: Manual Process.**

The work in this level was done manually. We received the popular tokens that appears in all the apps' reviews, after we removed stop words and the punctuations from the list. And later we kept only words that are nouns or adjectives. We chose to locate more tokens than needed so we could sort out what tokens would be useful for us.

We started with this list and tried to figure out what words would be useful for extracting info about the application's quality. We had word such as: "game", "play" and "coins" that couldn't help us figure out nothing about what the player liked or didn't liked in the game. In contrary, we had words like "time", "ads" and "money" that were informative.

We chose six main topics and for each topic we added some negative words that were related to it and saved them in a new file:

* info\_manual.txt – contains two lists. The first one has apps' names (untouched by us) and the other was built as a dictionary, each topic name was the key and its value was a list that contains the added word relating to the subject.

*The new file can be found in the 'files' folder.*

**Level 3: Analyzing the data.**

This part includes some parts inside of it. The user can choose one application from a given list. For the chosen application, we go over all its reviews and extract all the informative trigrams from it. For each trigram, we decide what its sentiment and if it's negative – we save it in a list.

In the next step, we go over the list of the negative trigrams and check for each topic (from the list we created in the previous level) how many trigrams are referring to this topic and creating a graph based on that info.

In the code –

We have a manager module that is responsible for creating the mentioned graph for the chosen application. It is done by running other modules:

1. reviews\_extraction.py – extracts the reviews for the application from the file that was created in the second level.
2. Info\_extraction.py – extracts the informative phrases as trigrams.
3. sentiment\_analyzer.py – it analyzes its phrases and determines its negative trigrams.
4. result\_analyzer.py – plot the analysis results in graph for the given app.

*All the code files for this level can be found in the 'analyzing\_step' folder and the graphs that were created can be found in the 'Graphs' folder.*

Important remarks:

* The popular subjects were taken from all the apps' reviews and not popular subjects for every app, so we have one list that includes the generally popular topics.
* Stop words are

To add

1. explanation - Why some words were informative and other weren't
2. packages we used, built-in algorithms

**Experiments**

* Evaluation Criteria: The success of the project we measured by a survey among X people. Each one read the reviews of some application and decided according to the reviews what the problematic topics (The topics which selected in the first stage of the program). Afterwards we compared the participants' decisions with the graphs obtained by running our program.
* Setup: X people get X reviews of X applications. according to the reviews the participant selects the topics that were problematic in the app from the following topics: Money, UI, Advertising, Interest, Flow, Bug, No problem.
* Results:
* Visualization:
* Impediments: