UNIVERSITÀ DI PISA

Advance Data Mining and Machine Learning

Games Genre Prediction

First Semester 2019

Matilde Mazzini — Marsha Gomez

February 5, 2020

Contents

Abstract

Websites like PlayStation Now, Steam and Origin, offer lists of games based on genres, this makes it easier for a user to select the game that interests you based on the genre is motivated towards. Tagging of games is a complex process and usually involves a labor-intensive process where the games are assigned to one or more Genres based on the proposals sent by the users and consumers. If we can systematize this process of game tagging, not only will it be fast, save human effort but it will be more accurate than an untrained human as well.

Chapter 1

Introduction

Public games database such RAWG provides genre information to assist searching. The tagging of games genres is still a manual process which involves the collection of users suggestions and consumers. Games are often registered with inaccurate genres. Automatic genres classification of a game based on its synopsis not only speeds up the classification process by providing a list of suggestion but the result may potentially be more accurate than an untrained human. We will collect data using one of many available apis on internet and compile a data set wich will be primarily based on *GDB (Game Data Base)*. We will rely on text analysis of the Plot/Summary of the movie data collected and train our classifier using text analysis techniques

Chapter 2

Data Collection

The Data Set on this project is a set of text files from GDB.They contain 80,000 games and 65,000 of genre information of games. For this project **65,000** *unique* titles at wich both the description and genre information were available were chosen and randomly split into 80% and 20% sets. The former was used for training while the latter for testing.

Note that a game can be (and often so) associated with **more than one** genres. There are 13 listed genres in the GDB data set and only the **7** (denoted as L) most commong genres were used in this project. The genres names and percentages of games in them are: *action (22%), adventure (15%), puzzle (14%), RPG (10%), simulation (9%), strategy (9%), Shooter (7%), sports (4%), racing (3%) educational (2%), fighting (2%), BoardGames (3%)*

Number	Genre	Count	Porcentage
1	Action	6848	22 %
2	Adventure	4725	15 %
3	Puzzle	4412	14 %
4	RPG	3105	10 %
5	Simulation	2745	9 %
6	Strategy	2660	9 %
7	Shooter	2101	7 %
8	Sports	1122	4 %
9	Racing	1000	3 %
10	Educational	549	2 %
11	Fighting	593	2 %
12	BoardGames	640	2 %
13	Card	262	1 %

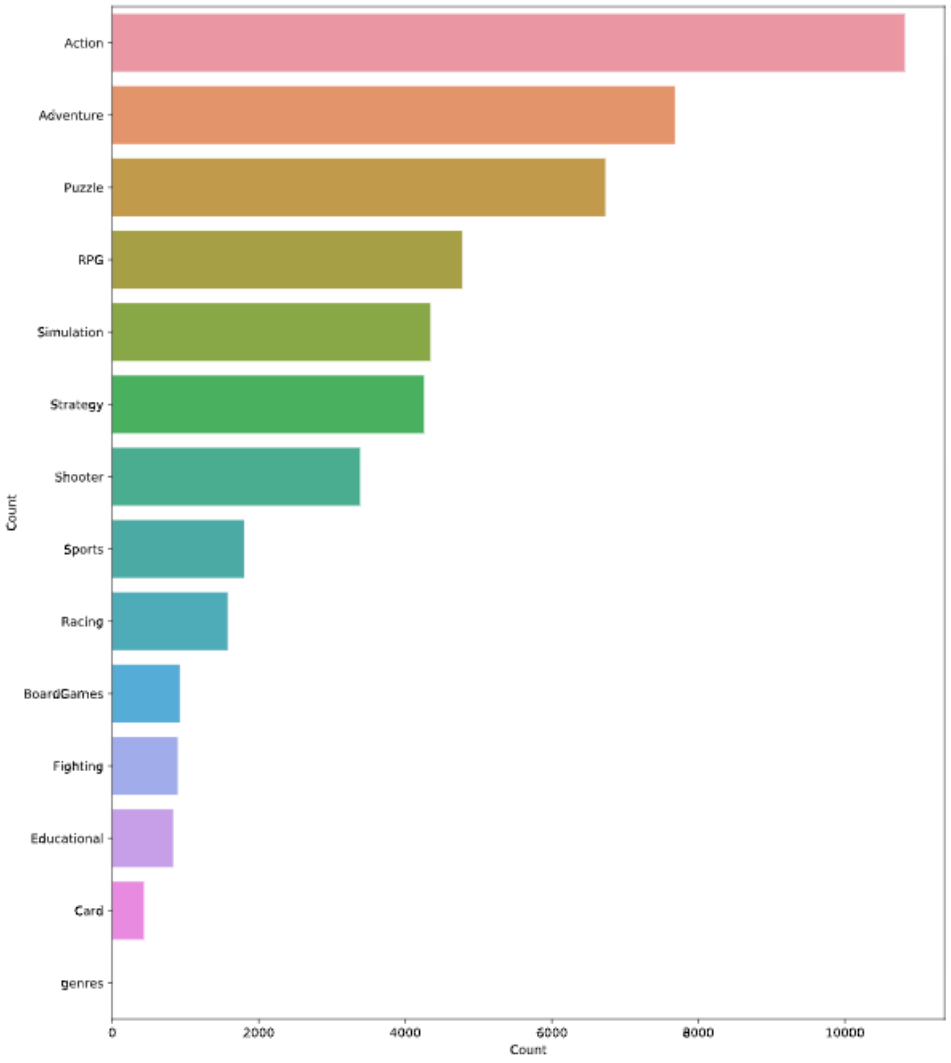
Action
Adventure
Puzzle
RPG
Simulation
Strategy
Shooter
Sports
Racing
BoardGames
Fighting
Educational
Card
genres
Count
0
2000
4000
6000
8000
10000
Count

Figure 2.1: .

Chapter 3

Data Preprocesing

Chapter 4

Data Mining

Chapter 5

Conclusions