Wine Reviews - PRI - FEUP

Group 25

NUNO SANTOS, RODRIGO ABRANTES, and NUNO MARQUES

1 INTRODUCTION

This 1º semester of 21/22 school year, for the class of PRI, we're tasked to develop an information search system, being able to query and retrieve information from a dataset previously collected and prepared, besides evaluating said retrieval. From all the themes we thought made sense to build an information system based on, we chose wine, containing many details from country of origin to its rating, and also reviews for every wine.

2 DADASET DEFINITION, PREPARATION AND PIPELINE PREPARATION

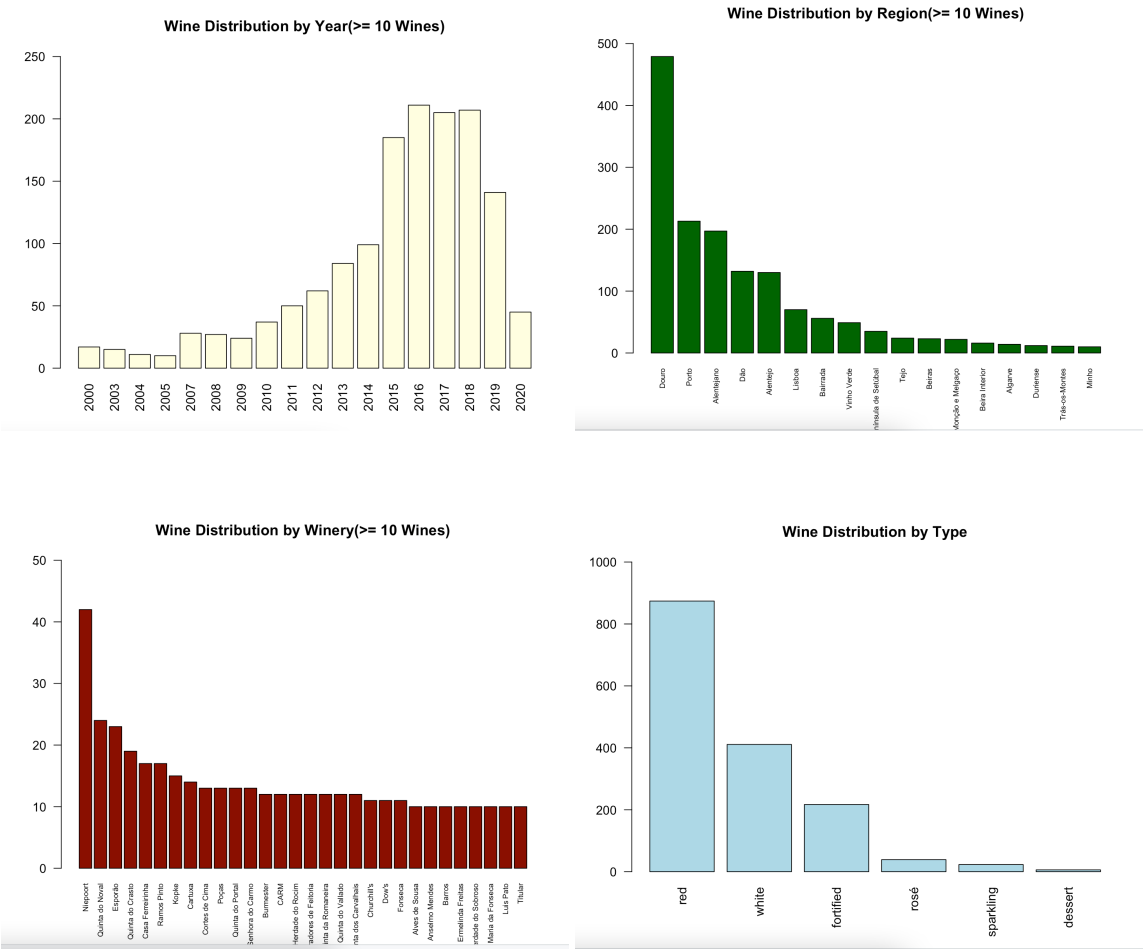
2.1 Final Dataset

Our group chose this theme not only for the motive and logic of it's existence but also for the available data we could find and how easily and efficient we could build a final dataset based on it, if the types of information would be interesting and complete for a later search. With that said we found several portfolios of wine available to compose the final data but the problem was every repository had different attributes so combining all sets would not be possible because on a search the result would have different types of information based on what set was the wine described. To overcome that we decided to use scrapping on Vivino, a website with many wine entries and many information about each wine including several reviews. From ruffly 10 thousand entries we interpreted, we eliminated duplicates, selected the ones that had several English written reviews, eliminated those with missing or wrongly parsed attributes and got to a final Dataset consisting just short of 2000 wines and 15 reviews per wine.

2.2 Data analysis

After the definition of the final dataset we needed to analyse the distribution on the data we had collected, to make sure we had wines present from several regions aside from several types, different ratings,etc.

Authors' address: Nuno Santos, up201405774@up.pt; Rodrigo Abrantes, up201506561@up.pt; Nuno Marques, up201708997@up.pt.



2.3 Present attributes in dataset

As for the actual type of data available in our dataset, it is devided in two files. The first, represented in figure 1, we have knowledge of its id(the one it received when scrapped from vivino), the winery that produced it, the year it was produced, its name, the type(red,white,rosé,etc),the rating,price, country of origin and then more specific, the region and a link to an image of said wine. The second file(Fig 2) contains all the reviews for all the wines, with their rating, note and the user that made it.

Fig. 1. File 1

	vivino_id	rating	note	user
1	2484848	4.0	A very nice Alvarinho 😊 An interesting fruit (and ber…	wine & mind
2	2484848	5.0	Had this wine in Lagos at the hotel we stayed in ... rea…	Caroline Hulsman
3	2484848	4.0	Great Alvarinho, loved it. Intense yellowish colour, no…	Martim Amaral Neto
4	2484848	4.0	Very very calm. Excellent balance, light and calm. Frui…	Aleksey Lisenkov
5	2484848	3.0	Candied lemon, hay, beeswax, chalky. Doesn't fulfil it…	Peter Arijs
6	2484848	3.5	Delicate aromas, very floral, light white melon, crisp a…	Rosanna Bucknill
7	2484848	3.5	✅ Still delivering after 5 years, this Alvarinho made b…	Marco Carmini
8	2484848	4.0	Citrusy and slightly sweet, smooth. A wonderful wine	Lauren Adie
9	2484848	3.5	Interesting nose of orange peel.	Mielies
10	2484848	4.5	A wonderful surprise. Aged and supreme.	Sergio Raposo Frade
11	2484848	4.0	Like a LDH white. Didn't know you could age a Alvahri…	Winston Chen
12	2484848	3.5	Excellent wine. Very good value for money.	Luis Ricardo Silva Viegas
13	2658364	4.0	This wine from alentejo has a soft elegant nose with a…	EdTheWineAdvocate
14	2658364	3.5	VERONA BELO HORIZONTE MG 3.3 Good QPR. Ruby r…	MARCELO BRANDÃO
15	2658364	3.5	Fine red at the price. Good first taste...but doesn't hol…	Jens Blomgren-Hansen

Fig. 2. File 2

2.4 Pipeline

There are many programming languages and ways to acquire data and process it but for this project, based on the combined experience of our group on previous projects, we chose Python for data acquisition and processing, and R to data cleaning and statistics/graphics creation, understanding and study. Our pipeline then consists of using Python to amass information through scrapping from Vivino, processing and storage in the appropriate files for later use. Several packages were needed to make this work, such as "Pandas". With the access to those packages we were able to clean the resulting dataset, removing duplicates, missing values,outlies,etc. Although the part of R is not included in the pipeline, we would then use R to build the statistics and grphs to better understang and comprehend the data we had for the rest of the project.

```
all: setup collect-data display-analysis

setup:
        pip install matplotlib
        pip install requests
        pip install pandas
        pip install progress

collect-data:
        python vivino_info_scraper.py
        python vivino_reviews_scraper.py

display-analysis:
        python vivino_analysis.py
```

Fig. 3. Project pipeline

2.5 Objectives, search possibilities and quality of information retrieval

Although we are not working in this section as of yet, we theorized some possibilities for the later search on our data. We thought it would better shape the way we map our information and our sources if we knew what would be done with it later. The final data has attributes that must be present as a base search such as origin,rating ,

Winery, year,etc. But we felt that those were very basic searches so we needed to focus more on the reviews part because there the complexity of the search possibilities would be far higher. As seen in the Bigrams and Trigrams graphs(fig 3 and 4) we can formulate the search engine based on the sequence of words present on the reviews, so sequences like "fruity taste", "hint of...",etc, that are present in the query will have good results.

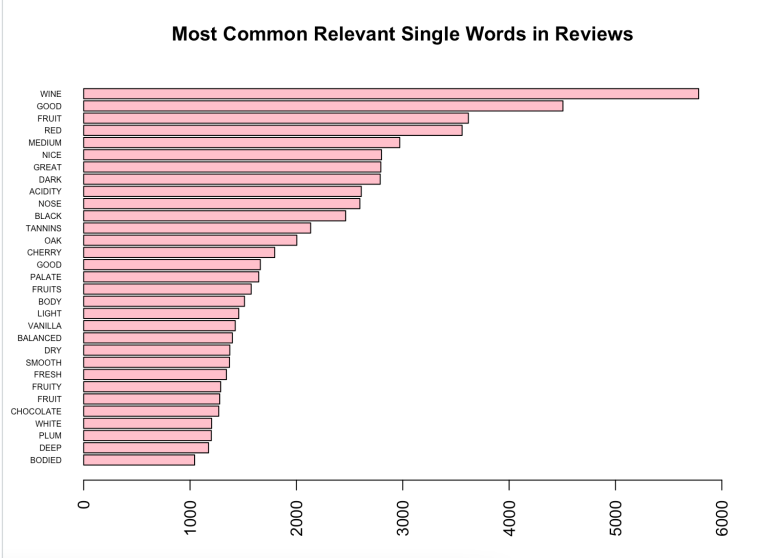


Fig. 4. Reviews statistics

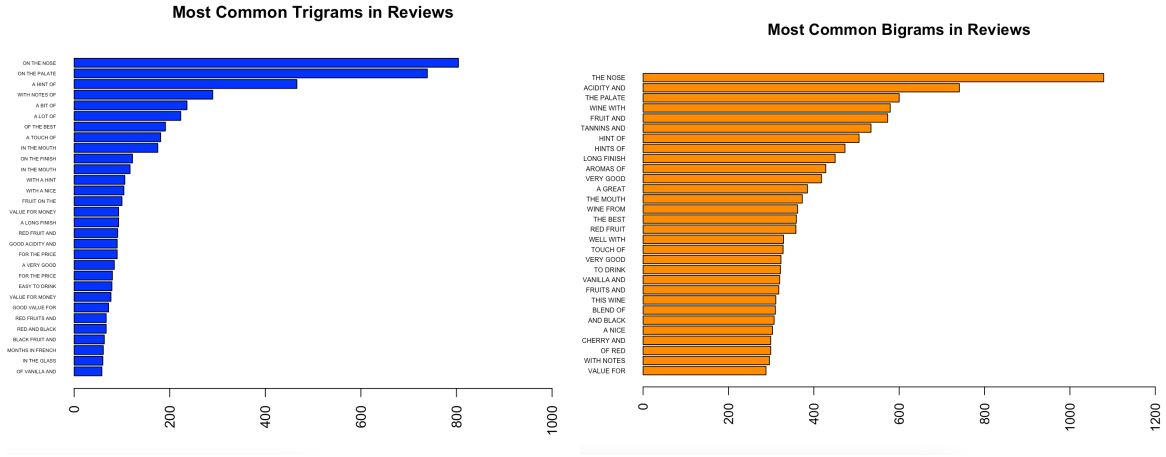


Fig. 5. Reviews statistics