Summary of Meeting 5 October 2018

# What has been done so far?

Lena has constructed a table of Sources of data from WINE SEARCHER magazine/website

This table has 32 rows. Each row is a source of ratings and columns identify the characteristics of that source.

Each row should be scraped separately. A separate scraping code should be written for each source.

There is a second table of 19 other sources that have also been compiled from various places. Some of them like WINEMAG are already a ready database of 132K ratings. Others are less directly available.

# Hiring

BB will forward the Job Advert to Lena Margaraf

BB will send Delactis’ CV to Stefan

Stefan assigns a work sample to Delactis and if satisfactory then we move on

# Important conceptual and practical issues and problems

* An important and critical problem is to uniquely identify each wine with an ID system. So, is there a standardized unified taxonomy for wines? Importantly, this taxonomy should be one that encompasses ALL wines that we are interested in (French, Italian, CA, …) and not just one naming system for French and another naming system for CA.
* The Basic Structure of our database is going to contain 3 tables (see the photo posted by Stefan in Slack):

1. Critics: each row is a critic. Critic could be of several types: an individual, a group of individuals (Bettane & Desseauve), a society, etc. The columns of the table would identify the specifications of each critic (type, gender, age, years in profession, qualifications etc)
2. Wines: each row is a unique wine. Columns correspond to Year, region, appellation, grape, country, etc. Therefore, Chateaux Margoux 2017 and 2016 would have a separate row each (corresponding to their own unique ID)
3. Ratings: At least 3 columns (Critic ID, Wine ID, Rating). Possible additional column would be Date of the Review.

* When scraping each website, we should also download a full copy (mirror image) of the entire website as backup for our records. This would make it possible to avoid unexpected changes in the data due to changes that may happen in the websites as they revamp themselves across time.
* This raises the problem of being detected by servers as Robots and shut out of the website content. We should assess how serious this problem could be and if necessary, follow guidelines (such as downloading at a rate similar to a human observer or using VPN)
* Many websites we have identified as Source require subscription fees. We may needs a budget of < ~3000 $ for acquiring subscription for around 30 websites at a rate of 100$ each.

08 Feb 2019

What has been done so far:

1. Dilectis has scraped data from the following five sources
   1. Robert Parker
   2. D & D
   3. Campbell
   4. Jancis Robinson
   5. Antonio Belliano (??)
2. This data has 40-50k rows and the following columns:
   * 1. Wine Name
     2. Vintage
     3. Average Price (at the time of publication of rating)
     4. Prodcuer (??)
     5. Region
     6. Appellacion (??)
     7. Country
     8. Blend
     9. Grape
     10. Alcohol
     11. **Rating**
     12. **Source**: the website

There are many repetitions in this data-frame because various wines have been tasted and rated by various sources.

1. Pantelis has collected a number of smaller but more dense databases from various events and websites. This data is available in the shared Excel file Pantelis has sent everyone.

What will be done next

**Extremely Important; Our First Priority**

1. Organize the Documentation in a Python notebook (or R Markdown)
2. History of the scraped data frame, the source URLs, and corresponding wrangling for each source will be documented.

The analysis tables is the next target

1. Documentation for the Analysis tables (Wine, Critic, Rating) will be organized separately as follows:
   1. Wine table: each row will be a unique item from the data-frame described above WITHOUT the rating and the source information. There will be 10 columns corresponding to (i, ii, iii, …, x)
   2. Critic table: each row is a rating source. The columns are:
      1. Name of the Team/Individual whose published data we have scraped
      2. Age
      3. Gender
      4. Nationality
      5. Years of experience
      6. url of the website
      7. rating scale (eg 4 star, 1-20, 1-100, etc)
   3. Rating table: each row will be one published rating. This table will have as many ROWs as our original data-frame described above. The columns are:
      1. Wine index (row in the wine table corresponding to this rating)
      2. Critic (row in the cirtic table corresponding to this rating)
      3. Date of rating (could be inferred from data of publication of data)
      4. Original Rating
      5. Standardized rating
      6. Normalised Rating

Note: once we have the Critic table and have a grasp of the various rating scales employed in the field, then we should converge on a Standardizing Method for bringing all ratings to the same format (for example scale of 1 to 10). Then we should agree on a normalizing score (such as Z-scoring)