SEN163A Fundamentals of Data Analytics

August 10, 2020

Introduction

Tabularazor Inc. is a large national Dutch news-paper that provides news on nearly everything, exclusively written in the fancy language Latin. With the use of metadata, corresponding to the articles, one is able to investigate several publishing patterns of the articles. Those patterns reveal personal information about the employees and the company Tabularazor Inc. itself. These patterns will be retrieved and are thoroughly examined in this paper in order to make statements.

1 Description of the data set

1.1 Data Gathering

The actual data set used is composed by scraping the archives stated on the website of Tabularazor Inc.: `https://news.tabularazor.org/`. To scrape this data an algorithm is written in Python, which makes use of Chrome web driver together with the packages selenium and BeautifulSoup. These packages enable to request websites by giving an url. All web links that refer to pages of Tabularazor INC. are structured and therefore it is possible to iterate the whole website efficiently. While iterating through every web page of Tabularazor, every web page that shows an article will be scraped. On these web pages, the name of the author that published the article as well as the exact date and time the article was posted are scraped. In total the scraping took approximately 6 hours. In theory, this could be way less when multiprocessing is used.

1.2 Description of the data

The data set consists of information retrieved from the archives of the newspaper. All data gathered consists of articles published from 2012 up to 2020. The scraped variables are more extensively explained below:

- **First and last name**: The first and last name of the employee who have worked on the article.
- **Date**: The exact date on which the article is published.
- **Time**: The exact time on which the article is published.

The first findings show that in total 328360 articles are scraped and that the articles are published by 50 distinct authors. Even though, initially the time of publishing is collected, this variable is deleted because it was not necessary to investigate the time of publishing in order to answer the research questions.

2 Employees

To determine the period in which an employee is working at Tabularazor Inc., the first and last date of publishing for each author is retrieved. These ultimate publishing dates help to construct a set for each employee that contains the actual dates this particular employee has published articles. In addition, another set for each particular employee is created, that contains all possible dates this employee could have worked within the working period of the employee. Hence, for each employee two different sets with dates created. To obtain the absences of a particular employee, the difference in these sets is taken.

Table 1: co-absence days

Pair	Absence
(Vonk Billips, Julieta Knapp)	122
(Augusta Beltrami, Grover Gibbons)	88

The resulting set entails all dates a particular employee is absent during his active working period, excluding weekends. Furthermore, the absence patterns of every employee are plotted in figure 1, using the latter constructed set.

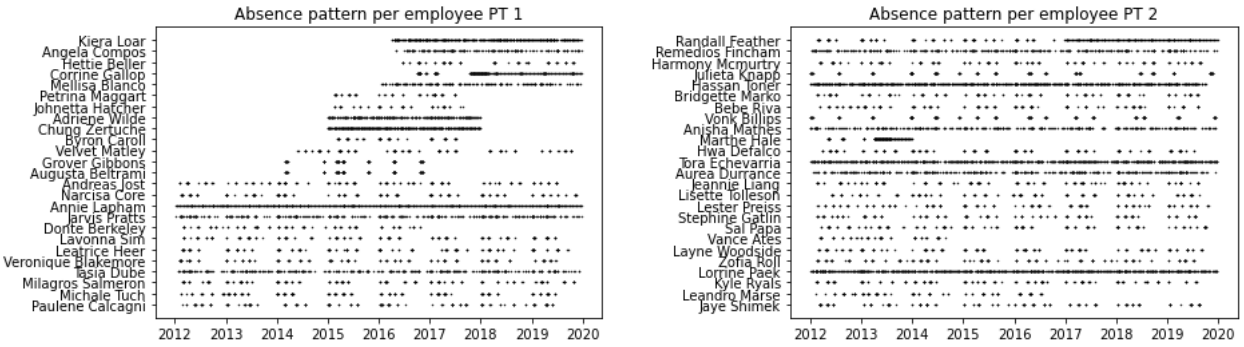

Figure 1: Absence patterns employees Tabularrazor Inc.

2.1 Parttimers

Most remarkable in figure 1 are the thick lines. These lines represents a continuous absence of the corresponding employee and could be interpreted as the absence pattern of a part-time worker. These patterns are enlarged in figure 2 ,where the most trivial patterns of part-time workers are plotted.

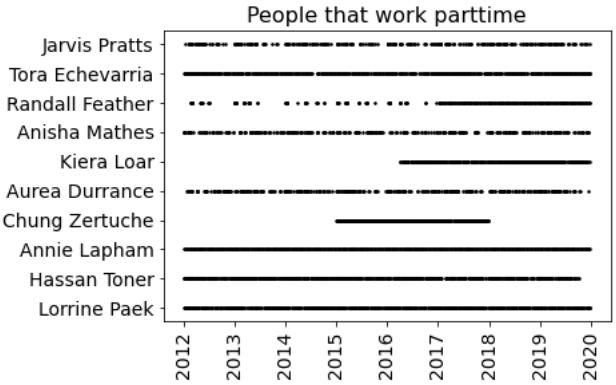

Figure 2: Absence Pattern of parttimers

2.2 Couples among employees

Couples tend to have a similar absence pattern since they like to go on holiday together. To identify whether pairs of employees have similar absence patterns, the sets that contain all absent days for each employee come in handy. For every possible pair of employees, the intersection of their corresponding absence days are used. Subsequently, these co-absence days are counted and stored in a data frame where every row represents an employee pair and its corresponding absence days. Finally, for identifying possible couples, all pairs that contain one or more part-time workers are removed from the data frame. Since they have a high co-absence with most other employees. The couples that stood out are Julieta Knapp Vonk Billips and Augusta Beltrami and Grover Gibbons as visible in table 1. This is also illustrated by the heatmap in figure 6 stated in the appendix.

The absence patterns of these identified couples, visible in figure 3, show that the patterns highly correlate. However, from 2018 the absences of Julieta Knapp and Vonk Billips start to differ slightly. There are myriad reasons for the difference in their absences, but it could indicate possible trouble in their relationship.

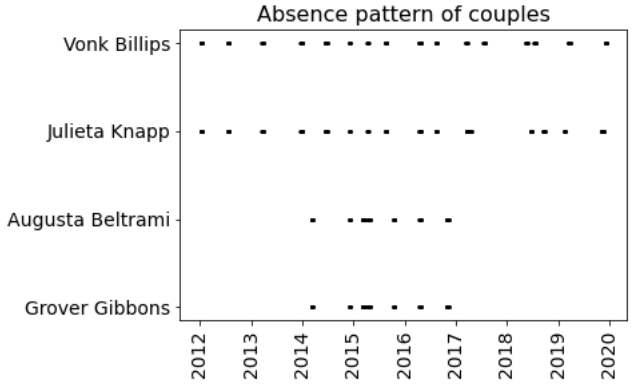

Figure 3: Grover and Augusta and their absence pattern

2.3 Maternity leave

In order to find out if someone among the employees had a child when working at Tabularazor Inc., it is necessary to look for maternity leave among women. The absence patterns of all the employees are used to determine this. A normal maternity leave would consists of a 4 to 5 months break, after which the woman would come back to work for Tabularazor Inc.

The absence patterns are identified by examining particular patterns occurring in figure 1. More specifically, thick lines that end up in thinner lines could identify a period of maternity leave. As a result, two women are found which absence pattern matched this description. The names of those women are Marthe Hale and Corrine Gallop, with their enlarged absence pattern during maternity leave shown in figure 4.

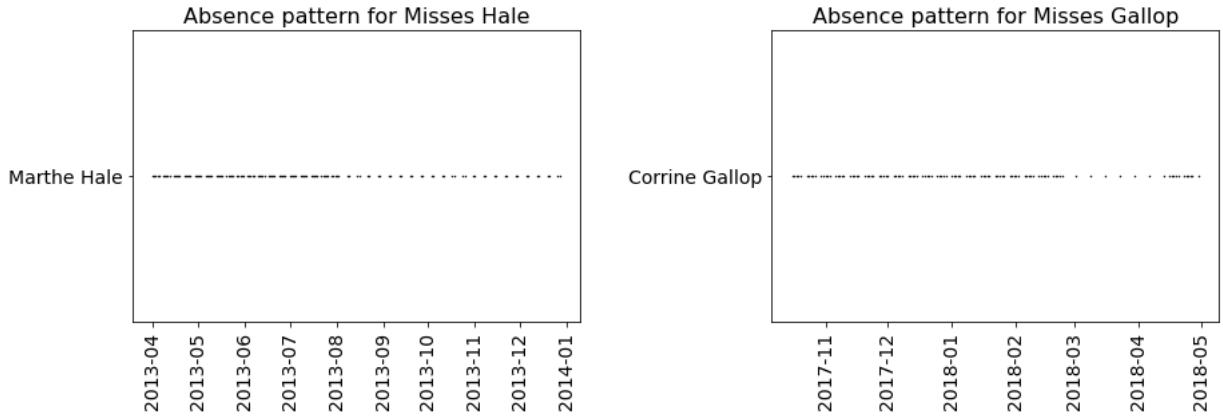

Figure 4: Left: Maternity leave pattern Misses Hale Right: Maternity leave pattern Misses Gallop

Marthe Hale shows a 4 months maternity leave, which started in April 2013 and ended in august. However, when she came back to work, she started working part-time for Tabularazor Inc. The same holds for Corrine Gallop, although her maternity leave started around october 2017 and ended around march 2018.

3 Holidays at Tabularazor Inc.

To investigate the amount of holidays, one can expect when working for Tabularazor Inc. a new data set is constructed which contains the name of the employee, the amount of total absent days for this particular employee, the start- and end date this employee started publishing and the average amount of absent days per year. For calculating the average amount of absent days an average of 7 days per year is subtracted. Since there are approximately 6 to 7 national holidays per year. When sorting the data set, a table like the one shown below follows:

Table 2: Average amount of holidays per year

Name	Absence	Start	End	Avg abs/y
Layne Woodside	172	2012-01-02	2019-12-31	14.500
Narcisa Core	173	2012-01-02	2019-12-31	14.625
Milagros Salmeron	174	2012-01-02	2019-12-31	14.750
Leatrice Heer	174	2012-01-02	2019-12-31	14.750
Michale Tuch	175	2012-01-02	2019-12-31	14.875
Lester Preiss	176	2012-01-02	2019-12-31	15.000
Vance Ates	66	2012-01-02	2014-12-31	15.000
Stephine Gatlin	177	2012-01-02	2019-12-31	15.125
Julieta Knapp	179	2012-01-02	2019-12-31	15.375
Harmony Mcmurtry	180	2012-01-02	2019-12-31	15.500

The most frequent amount of average holidays per year range from 11 up to 16, since 32 out of 50 employees have an average between 11 and 16. Hence, when working for Tabularazor Inc. one can expect at least 11 to 16 holidays per year. To estimate the exact average amount of holidays per year, a subset of 32 employees with the lowest amount of absence days per year are selected. In this manner, part-time workers are left out. The average of the absent days from the resulting people is between 14 and 15. Since the amount of national holidays per year varies, it is more plausible that employees from Tabularazor Inc. have 15 days per year off.

4 Conclusions

Most of Tabularazor inc.'s employees are loyal because most worked from at least 2012 up to 2020. A significant share of the employees works parttime. Since in 2012 two couples worked at Tabularazor Inc.. One of the two couples stopped working and from the other couple it remains unclear whether they still are together. In addition, at least two employees had a child during their working period. Lastly, the estimated holidays per year for employees at Tabularazor Inc. is 15.

5 Discussion

Most findings in this report are based on visual inspection of graphs. Even though this method is relatively reliable, more proof is needed to strengthen the conclusions derived out of the findings. For example, an improved correlation plot that illustrates the correlation degree with respect to co-absences would be suitable to identify couples. Further research should therefore incorporate these kind of plots to provide proper proof.

6 Appendix: Heatmap

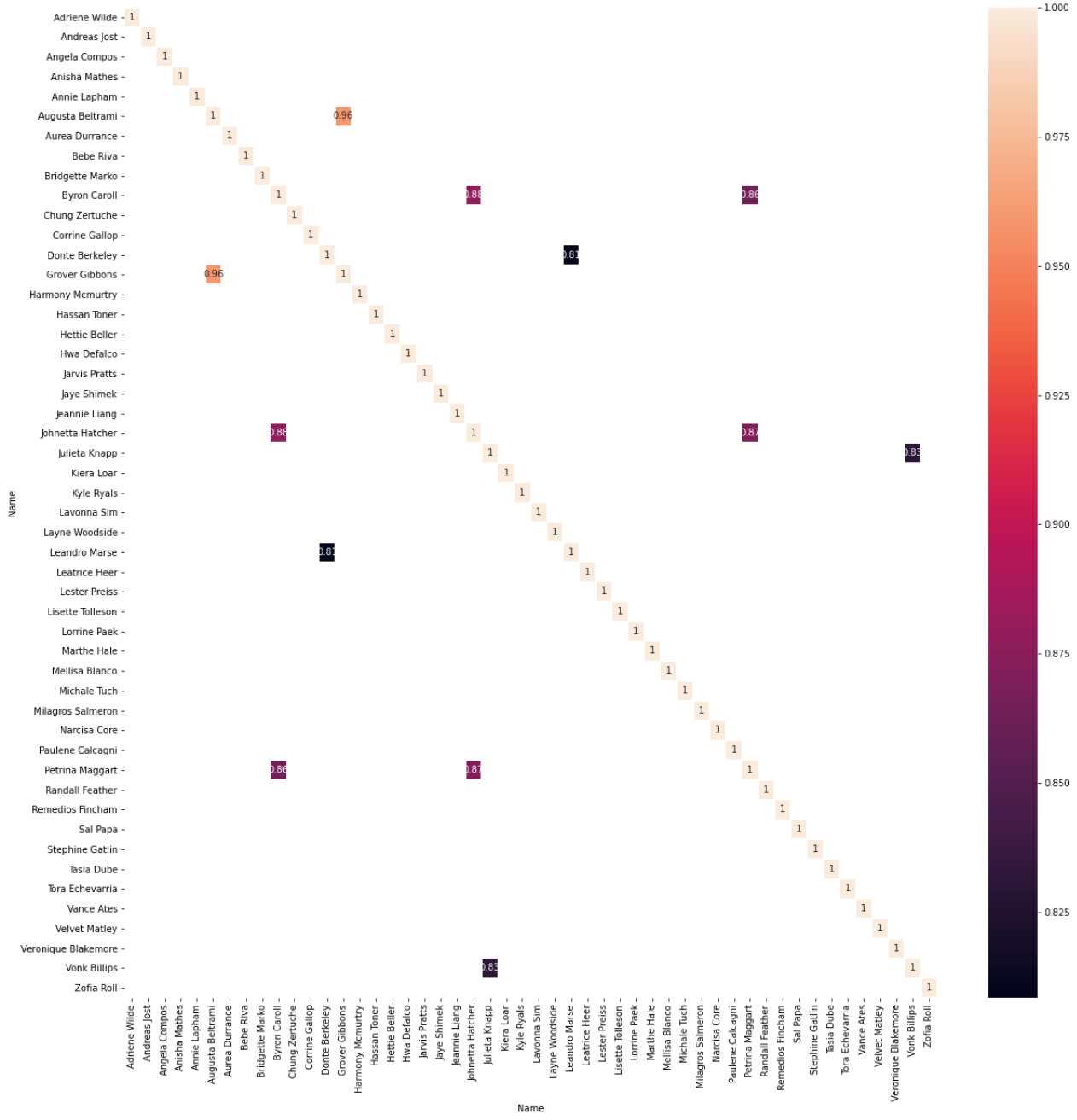

Figure 5: Heatmap of couples that correlate