



POLITECNICO DI TORINO

Master degree course in Computer Engineering

Master Degree Thesis

Time prediction of software development via machine learning

Artificial Intelligence applied to Software Engineering

Supervisors

prof. Maurizio Morisio

Candidates

Jacopo NASI [255320]

Internship Tutor

dott. Davide Piagneri

ANNO ACCADEMICO 2019-2020

This work is subject to the Creative Commons Licence

Summary

La pressione barometrica di Giove viene misurata mediante un metodo originale messo a punto dai candidati, che si basa sul rilevamento telescopico della pressione.

Acknowledgements

Un ringraziamento speciale ai cavalieri di Smirnuff, luce della mia battaglia.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 7 |
| 1.1 | General Problem | 7 |
| 1.2 | Tools used | 7 |
| 2 | Datasets | 9 |
| 2.1 | SEOSS33 | 9 |
| 2.2 | San Francisco Bike Sharing | 10 |
| 2.3 | San Francisco Fire Department | 11 |
| 3 | Machine Learning models | 13 |
| 3.1 | Random Forest | 13 |
| 3.2 | Neural Networks | 13 |
| 4 | Forecasting | 15 |
| 4.1 | Goal introduction | 15 |
| 4.2 | Feature extraction | 15 |
| 4.3 | Models detail | 15 |
| 4.4 | Forecasting horizons | 15 |
| 4.5 | Feature application | 15 |
| 4.6 | Application over different projects | 15 |
| 5 | Model abstraction | 17 |
| 5.1 | CommonDB | 17 |
| 5.2 | SFBS and literature comparisons | 17 |
| 5.3 | SFFD | 17 |
| 6 | Conclusion | 19 |
| | Bibliography | 21 |

Chapter 1

Introduction

1.1 General Problem

Forecasting is one of the most critic part of a company, it could drive to easily success as well as drive to failure. A software project is not different from a manufacturing product, its development, infact, require analysis of different kind, from resources needed to costs and time required.

The software development experience shows that the process of analysis is really difficult, due to the nature of the problem, coding is a mind product and the time required to produce it can varying in accord to a lot of different factors.

1.2 Tools used

This work is mainly conducted using software tools, here a list of the tools used:

Python The main programming language of the thesis project. Used for data management, feature extraction, machine learning models and for interfaction with other softwares. The specific version used is the v3.7.0

Pandas Open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

NumPy Scientific computing with Python.

Matplotlib 2D Plotting library for Python.

Seaborn Another plotting library for Python.

Tensorflow Platform for machine learning.

Keras High level API for neural networks.

SciKit-Learn Tools and libraries for machine learning.

GitLab Sourcing platform based on Git. Used for the code of the project, available here: <https://gitlab.com/EiS-Projects/analytics/temp/thesisProjectJN>.

GitHub Sourcing platform based on Git. Used for the thesis and calendar sourcing:

- Thesis: <https://github.com/Jacopx/Thesis>
- Calendar: <https://github.com/Jacopx/ThesisCalendar>

JetBrains IDEs Student-free IDE for different language development, product used:

- PyCharm: <https://www.jetbrains.com/pycharm/>
- DataGrip: <https://www.jetbrains.com/datagrip/>

Chapter 2

Datasets

The following section illustrate the structure of the all the principal datasets used during this thesis project.

2.1 SEOSS33

The SEOSS33[1] is a [dataset](#) collecting bug, issue, reports, commit and lot of other information of 33 open source project, following their progress via sourcing platform. At today there are no other public research conducted over this datasets, this works seems to be first.

Is fundamental to understand the structure of this dataset, the majority of the forecasting operation tests are conducted using the data stored by this research. Each project is stored in a SQLITE db file, a SQL offline database, the structure is based on the entity of the *issue*, identified by an *issue_id*, the other tables are used to link additional information, like the number of commit, the version referred, comments and others features. The figure 2.1 show the database schema.

The dataset is composed of 33 different software projects with some common characteristics (reference in [1] chapter 2.1), almost single programming language (Java), usage of versioning software, tracebility of issue and other similiar information. Among these products we have selected five of them, because of size, as shown in table 2.1:

Table 2.1. Project data distribution

| Project | Month | Issue |
|-----------|-------|-------|
| Hadoop | 150 | 39086 |
| Hbase | 131 | 19247 |
| Maven | 183 | 18025 |
| Cassandra | 106 | 13965 |
| Hive | 113 | 18025 |

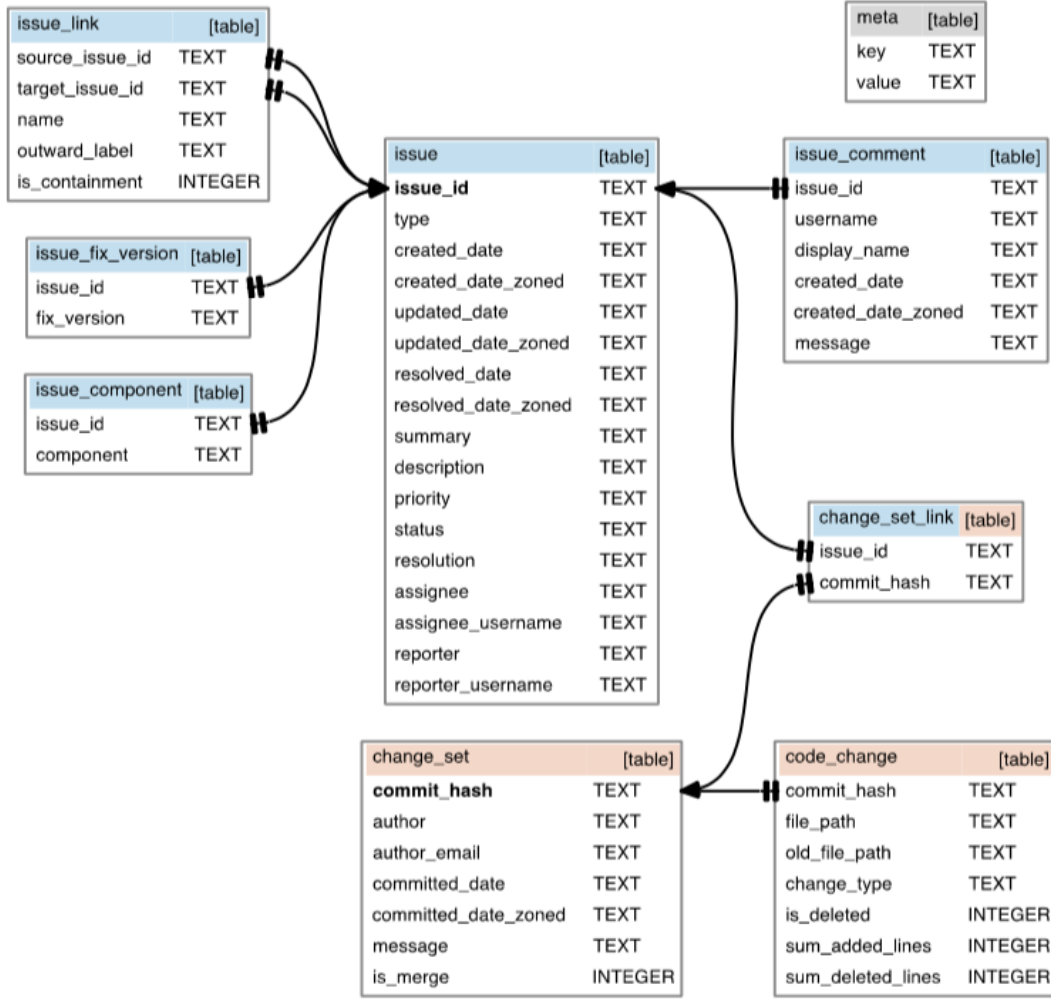


Figure 2.1. SEOSS33 data model

2.2 San Francisco Bike Sharing

The big US city of San Francisco provide a lot of public available dataset related to the service proposed, on of them is about the BS (*Bike Sharing*) system of the Bay Area. Is huge [dataset](#) holding information about 5 years, for 70 stations, with more than 5000 bikes. The structure is simple and is divided among three tables:

- **status**: A log storing the information of number of bikes and docks available at each minute of the day for each station
- **trips**: Each ride performed with the source and destination station, the starting and arrival time, the duration, the number of the bike used and the type

of user subscription.

- **stations:** Store generical informations about the station, name, address, latitude, longitude and number of docks.

2.3 San Francisco Fire Department

Chapter 3

Machine Learning models

3.1 Random Forest

3.2 Neural Networks

Chapter 4

Forecasting

4.1 Goal introduction

4.2 Feature extraction

4.3 Models detail

4.4 Forecasting horizons

4.5 Feature application

4.6 Application over different projects

Chapter 5

Model abstraction

5.1 CommonDB

5.2 SFBS and literature comparisons

5.3 SFFD

Chapter 6

Conclusion

Bibliography

- [1] M. Rath, P. Mäder, “The SEOSS 33 Dataset — Requirements, Bug Reports, Code History, and Trace Links for Entire Projects” in *Data in Brief*, v. 25, p. 104005, 05 2019. [Online]: <https://doi.org/10.7910/DVN/PDDZ4Q>