University of Pisa

SCUOLA DI INGEGNERIA

Corso di Laurea in Artificial Intelligence and Data Engineering

Documentation of the "STAD" Application

Candidati

Alice Nannini

Marco Parola

Relatori

Prof. Francesco Marcelloni

Prof. Pietro Ducange

Anno Accademico 2019–2020

Contents

1 Introduction

The goal of this application is to prevent and detect situations potentially dangerous, caused by huge amounts of rain, scraping twitter and analyzing each tweet, in order to discover some tweets containing information related to these critical situations.
We analyze the data and develop the application using Python and Sklearn library.

2 The Data

2.1 Retrieve the data

The data, on which this application works, are tweets. In order to collect enought tweets, we scraped twitter, using **twint**.
Twint is an advanced opensource Twitter scraping tool, written in Python, thanks to which it's very easy to collect data, according to some criteria, and store them in csv files. For more information about *twintproject* visit the Github repository: `https://github.com/twintproject`.

```
twint -s <WORDS> -o tweetPioggia.csv --csv

<WORDS> : 'pioggia', 'piove', 'allerta', 'meteo', 'alluvione', 'maltempo'
```

Examples:

- "Ora piove a dirotto per la gioia di yuki che non può andare al parco"
- "Ma dai, ma piove sul bagnato! Povera Antonella!!!!! #GFVIP"

Moreover we added to the dataset some posts randomly downloaded, not related to any weather phenomenons (without specifying any keywords).

2.2 Prepare the dataset

After collectioning the tweets (894), we assigned each of them to a class, in order to prepare the dataset, thanks to which we can build some classifiers.
We decided to map tweets in 3 classes:

- 0 -> the tweet is not related to a weather condition (368)
- 1 -> the tweet is about rain or some weather condition not dangerous (224)
- 2 -> the tweet is about some dangerous situation caused by the rain (302)

3 Preprocessing

In this phase, we delete some tweets, in order to manage only the italian tweets.
Moreover we clean the text of each tweet removing eventual URLs.