ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Big Data

Project Report

Weather Dashboard

Authors:
Aubry Cholleton
Jonathan Duss
Anders Hennum
Alexis Kessel
Quentin Mazars-Simon
Cédric Rolland
Orianne Rollier
David Sandoz
Amato Van Geyt

Professor:
Christoph Koch
Supervisor:
Amir Shaikhha

May 20, 2014

Contents

1 Data Retrieval

1.1 Looking for data

Initially, the project idea was to get Alps weather information from social networks using natural language processing. Hence we first looked how to mine Twitter, Facebook, Instagram or Camptocamp. We also searched for other datasources such as the Institute for Snow and Avalanche Research (SLF), MeteoSuisse or ski resorts. After discussion and clarification with our TA, the scope and goal of our project changed in order to get closer to a *Big Data* project. Hence we needed to process a big dataset and the NCDC dataset from NOAA seems to be the one.

1.2 The NCDC dataset from NOAA

The *National Oceanic and Atmospheric Administration* (NOAA) is the US agency responsible for the weather surveillance and forecast. They appear to have a *National Climatic Data Center* (NCDC), "responsible for preserving, monitoring, assessing, and providing public access to the Nation's treasure of climate and historical weather data and information".

1.3 Downloading the data

2 Nearest Neighbors

2.1 Why?

The main purpose of implementing nearest neighbor, is to find similar weather periods in the history and possibly find patterns in the climate. This is useful in terms of understanding the climate and to forecast specific events.

As an example, here in Switzerland, it could be useful in terms of avalanche danger. If most of the nearest neighbors of the current winter so far, led to many and massive avalanches, it is reason to believe that also the current winter can become a such a winter. By knowing this, it would be possible to take action before the events occur. The same method can be used to forecast events like floods or drought.

It can also be used to find patterns in the climate such as cycles and periodic structures.

TODO: Rewrite to better language

2.2 Implementation

The k-nearest neighbor is a fairly simple algorithm. You choose a metric, calculate the distance from the reference node to all the other's, and pick the k nearest neighbors. This has been done more or less straight forward.

In this case a node is a period of time. We fixed the period to one month, such that you can pick one month for a specific year, and then find the most similar months for the other years. Often climate data is summarized by month, so this makes it easy to find data to compare with.

Next question is to determine how many intervals the period of time should be divided into. Should we compare averages for an hour, a day, a week, or just one value for the whole month? For a small region (a few stations) and a short period of time (a day), then comparing hour by hour could make sense and give good results. But for a large region (thousands of stations) and for a longer period (a month), this will mostly just return noise. Here, the method is implemented with a period option so one can choose how many periods a month should be divided into. The final results is run with this set to 1, meaning that averages for the whole month is used.

TODO: add figure shows how this was implemented with map reduce on Hadoop. Write some more and clean language.

2.3 Results

TODO: Present some results with some figures to show that it works.