Università degli Studi di Verona

Authorship Attribution

Big Data project report

Davide Bianchi VR424505
Matteo Danzi VR424987

A.A. 2018/2019

Contents

1 Introduction

The project aim was to design a tool which could establish the authorship of a manuscript by using specific criteria described later. The used architecture is based on Hadoop, a distributed filesystem simulator, running in a docker container.

2 Background and System Description

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is an executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.

Docker containers images become containers when they run on Docker Engine. They isolate software from its environment and ensure that it works uniformly despite differences for instance between development and staging.

The Cloudera Docker image used in this project contains an Hadoop Distributed File System (HDFS) partition. Cloudera provides a scalable, flexible, integrated platform that makes it easy to manage rapidly increasing volumes and varieties of data in an enterprise. It is distributed by Apache Hadoop. The Hadoop version used in this project is `Hadoop 2.6.0-cdh5.7.0`.

For the installation of Docker, the Cloudera Docker image and the execution of jobs using jar file it has been used the instructions of the course.

3 Project Workflow

The first step the project is to analyze an amount of manuscripts, extracting relevant information from them in order to create a "dictionary" of known authors. Starting from these data, the program should take unknown manuscripts as input and establish a possible author, comparing the extracted data with the "dictionary" previously created.