HELLENIC REPUBLIC
National and Kapodistrian University of Athens

DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS - FACULTY OF SCIENCE

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

M111: Big Data Management
Spring '19

Project Progress Report
Title: Restaurants classification using Apache Spark

Christodoulou Kyriakos - cs2180019
Demetris Flouris - cs2180023

List of Contents

Project Description and Purpose

In order to make world a better place for food lovers we decided to make an application that will take as input reviews of restaurants and will decide if each review talks about a good restaurant or a bad restaurant.

In more detail, we will create a Python program that will use pySpark, Apache's Spark API for Python, so that the application can deal with much larger datasets compared to ours. First of all, we will load and preprocess the training dataset. Next we will use 10-fold Cross Validation on the train set to find the classifier with the best accuracy.

The purpose of our project is to use the best classifier to estimate if the reviews of the test set are about good or bad restaurants. Data will be found locally on the hard disk and results will also be saved locally on a csv file on hard disk again.

Dataset Description

Train Set

Our train set consists of a tab separated value file which contains two columns and 82,065 records. The first column is the numerical rating the user has given to the restaurant and it ranges between 1-5. We will consider ratings 1,2,3 as bad and ratings 4 and 5 as

good. The second column is the textual review given which may describe the experience of the user at the restaurant. We will use this dataset in order to determine which classifier has the best accuracy in classification with 10-fold cross validation.

Example

Score	Text
4	Thank you thank you thank you !! I want to thank the people that made this place...
5	A Humane Society store at the Biltmore? Interesting. I had seen an adorable ....
1	Don't buy Nike sneakers if you want to return or exchange them.. I've bought ...
3	I have to say I love most things about Sprouts and especially this particular store ...
5	The tire pressure light came on a day or so ago, so I had my husband fill the tire ...

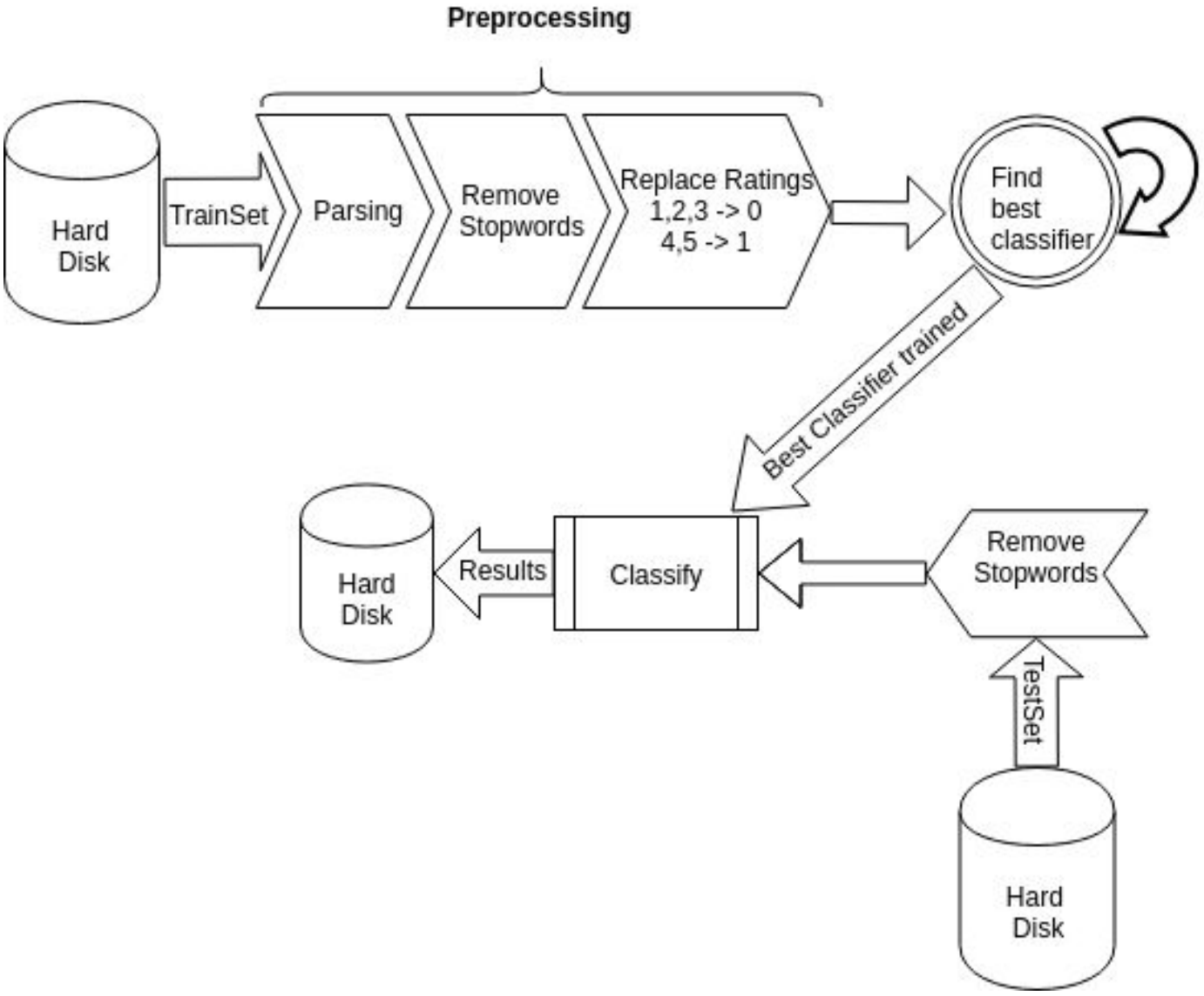
Test Set

Our test set is also a tab separated value which contains only one column, the textual review, and has 34,194 records. We will use this dataset to predict whether the review is bad or good.

Example

Text
My son just loves this place. Weird that he'd ask to come here everytime we go grocery...
We gave it a 9, so we will make that 5-, 4,5 stars. To start with it's just beautiful and ...
After three lunch visits I've come to the conclusion that this restaurant fares just ok in my...
What started out as a simple attempt to find a perfect birthday present turned into a life ...
If they had a Culver's on every street corner, the cardiologists would all have twice as ...

Architecture and Components Description



Components:

1. Preprocessing data

Preprocessing data is consisted of loading data from hard disk and three more sub-components. Parsing will first read data from train set and cache them. Then we will remove stopwords from the textual reviews so that will not affect

classification's results. Then we will change the ratings of 1,2,3 to 0 to represent bad reviews and 4,5 to 1 to represent good reviews.

2. Find best classifier

The data acquired from the previous component will be used in repetition in order to find the best classifier from a set of choices. This will be the result of 10-fold cross validation on the training set and the outcome will be the classifier with the best accuracy.

3. Classification of test set

This component uses the result of the previous component, which is the classifier to be used trained with the train set, and loads the test set from hard disk. Removing stopwords is needed again before classification takes place.

4. Demonstration of results

Results will be saved in a tab separated value file containing only one column with the predicted value for each review, 0 for bad and 1 for good reviews.

Final Report Structure

1. Project Description and Purpose
2. Dataset Description
3. Architecture and Components Description
4. Classifiers comparison and final choice
5. Results demonstration
6. Conclusions

Progress so far

As matters now stand, we have implemented the first sub-components of the preprocessing data component. We read data from the tsv file and remove the stopwords from the training set caching it.

Things to be done

As it matters implementing the application, we still have some work to do in preprocessing data, which is converting numerical ratings 1,2,3 to 0 and 4,5 to 1.
We also have to implement the component which will find the best classifier and train it with the training dataset. Finally we will use this classifier in combination with the removing stopwords sub-component in order to predict test set records and save them.