Big Data Technologies

2021/2022

Spark Project

Rules:

- The code will be presented in a Zeppelin notebook. To start Zeppelin, run the command "zeppelin-daemon.sh start" in a terminal, then open the webpage http://localhost:8090/
- If Zeppelin is not running on your virtual machine (for example, if your laptop has not enough RAM), you can do the project with the spark shell
- You **must not use** explicit loop (for, while, etc.) but you must use “map” and/or “reduce” and/or “join” functions (flatMap, reduceByKey, fullOuterJoin, etc.).
- You must not create temporary files.
- The data should be processed as distributed datasets.

Context:

We will explore the Ling-Spam email dataset. The archive “ling-spam.zip” from the Jalon website contains two directories, /spam/ and /ham/, containing the spam and legitimate emails, respectively. The dataset contains 2412 ham emails and 481 spam emails, all of which were received by a mailing list on linguistics. We want to extract the words that are most informative of whether an email is spam or ham. These words are called the top words.

The first steps in any natural language processing workflow are to remove stop words and lemmatization. Removing stop words involves filtering very common words such as the, this and so on. Lemmatization involves replacing different forms of the same word with a canonical form: both “colors” and “color” would be mapped to “color”, and “organize”, “organizing” and “organizes” would be mapped to “organize”. Removing stop words and lemmatization is very challenging, and beyond the scope of this project. The Ling-Spam e-mail dataset has been already cleaned and lemmatized.

When we are looking for the top words, we will use the presence of a particular word in an email as the feature for our model. We will use a bag-of-words approach: we consider which words appear in an email, but not the word order. Intuitively, some words will be more important than others when deciding whether an email is spam. For instance, an email that contains “language” is likely to be ham, since the mailing list was for linguistics discussions, and “language” is a word unlikely to be used by spammers. Conversely, words which are common to both message types, for instance “hello”, are unlikely to be much use.
One way of quantifying the importance of a word in determining whether a message is spam is the Mutual Information (MI). The mutual information is the gain in information about whether a message is ham or spam if we know that it contains a particular word. Consider a particular word *w*. The email can belong to two classes (*spam* or *ham*) and the word *w* can occur in the email or not. The mutual information *MI* of the word *w* whether that email is spam or ham is then defined by:

$$MI(w) = \sum_{\substack{occurs \in \{true, false\} \\ class \in \{spam, ham\}}} P(occurs, class) \log_2 \left(\frac{P(occurs, class)}{P(occurs)P(class)} \right)$$

The goal of this project is to find the most informative words (the top words) which can be exploited to distinguish spam and ham emails.

Materials:

1. The zip file "*ling-spam.zip*".
2. The notebook template "*BigDataTechnologiesProject2021template.zpln*" file.

Useful commands:

Some Scala commands should be **useful** for the the **project** and the **exam**: *wholeTextFiles, map, flatMapValues, mapValues, filter, case, reduce, reduceByKey, Ordering.by, fullOuterJoin, join, leftOuterJoin, getOrElse, math.log, toDouble, toSet, takeOrdered, foreach, swap.*

These commands are described on http://spark.apache.org/docs/latest/api/scala/index.html

Questions:

You must complete the template file by performing the following tasks. For each task, you must **describe carefully the type** returned by each command (and also the type of each element within the command). You must not create any new files, except the file containing the top words.

1. Download the Ling-Spam email data set *"ling-spam.zip"* from the Moodle website. The data set is described on https://aclweb.org/aclwiki/Spam_filtering_datasets

2. The data set must be uncompressed and stored in the HDFS directory /tmp/ling-spam in the two subdirectories /tmp/ling-spam/ham and /tmp/ling-spam/spam

3. Upload the template notebook in Zeppelin.

4. You must complete the function *probaWordDir*. It must return the couple (*probaWord*, *nbFiles*).
    a. This function reads all the text files within a directory named *filesDir*

    b. The number of files is counted and stored in a variable *nbFiles*

    c. Each text file must be splitted into a set of unique words (if a word occurs several times, it is saved only one time in the set).

    d. Non-informative words must be removed from the set of unique words. The list of non-informative words is ".", ":", ",", " ", "/", "\", "-", "''", "(", ")", "@"

    e. For each word, you must count the number of files in which it occurs. This value must be stored in a RDD, named *wordDirOccurency*, with the map structure: word => number of occurrences of this word.

    f. You must compute the probability of occurrence of a word. The probabilities of all the words must be stored in the RDD, named *probaWord*, with the map structure: word => probability of occurrences of the word. For each word, this probability is computed as the ratio:

       ratio = number of occurrences of the word / nbFiles

    g. Warning, the probaWordDir will be called twice: the first time for the spam emails and the second time for the ham emails.

5. You must complete the function "*computeMutualInformationFactor*". A factor is a term

$$P(occurs, class)\log_2\left(\frac{P(occurs, class)}{P(occurs)P(class)}\right)$$

in the mutual information formula. The variable *probaWC* is a RDD with the map structure: word => probability the word occurs (or not) in an email of a given class. The variable *probaW* has the map structure: word => probability the word occurs (whatever the class). The variable *probaC* is the probability that an email belongs to the given class. When a word does not occur in both classes but only one, its probability $P(occurs, class)$ must take on the default value *probaDefault*. This function returns the factor of each words (so it returns a RDD) given a class value (spam or ham) and an occurrence value (true or false).

6. You must complete the function “main”.
   a. You must compute the couples (*probaWordHam*, *nbFilesHam*) for the directory “ham” and (*probaWordSpam*, *nbFilesSpam*) for the directory “spam”.

   b. You must compute the probability $P(occurs, class)$ for each word. There are two values of *class* (“ham” and “spam”) and two values of *occurs* (“true” or “false”). Hence, you must obtain 4 RDDs, one RDD for each case: (true,ham), (true, spam), (false, ham) and (false, spam). Each RDD has the map structure: word => probability the word occurs (or not) in an email of a given class.

   c. You must compute the mutual information of each word as a RDD with the map structure: word => MI(word). This computation must exploit the function *computeMutualInformationFactor*. Warning! If a word occurs in only one class, its joint probability with the other class must take on the default value = 10^{-5} and not the zero value. The function *computeMutualInformationFactor* will be called 4 times for each possible value of *(occurs, class)*: (true,ham), (true, spam), (false, ham) and (false, spam).

   d. The main function must finally print on screen the 20 top words (maximizing the mutual information value) which can be used to distinguish a spam from an ham email by using the mutual information.

   e. Save all the words and their mutual information values, sorted and the corresp top words must be also stored on HDFS in the file “/tmp/topWords.txt”.