MASSIVE DATA PROCESSING

MapReduce programing

Introduction

Objectives.

This practice aims to design and implement a small MapReduce application to calculate the trending topics on twitter and feelings associated with such hashtags. For its realization many of the concepts presented in this course will apply.

Activity Presentation.

You have to submit the files with the source code made and a brief explanation of the design and implementation done. Also you have to explain how to run each of the steps. All coding is done in Java using the MapReduce framework.

Activity Statement

The mapreduce appointment will work with twitter data. You can download your tweets or use the tweets data available in the hadoop cluster (/shared/nando/data/tweets/ directory in HDFS file system) in text and json format. Also you have a subset file with only the English and Spanish tweets.

1. Trending topics (2.5%)

The #symbol, called a hashtag, is used to mark keywords or topics in a tweet. It was created as a way to categorize messages. By counting the frequency with which a hashtag is mentioned over a given time period, we can determine which topics are trending in the social network.

To find the trending topics, you have to count the occurrence of the hashtags in a similar way that wordcount application. In this exercise you have to count the occurrences of the tweet hashtags.

You can use the following regular expression "(?:\\s|\\A|^)[##]+([A-Za-z0-9-_]+)" to access the hashtag in the text version.

As in the wordcount application, the user will pass two parameters to the job, specifying the directory for the tweets files to be processed and the directory to output the result (trending topics list).

Tips:

- You can use a hadoop-provider mapper for filter the hashtags.

2. Text clean-up using chain mapper (2.5%)

In the exercise we are going focus in the normalization and cleanup of the input data. Three common components of this normalization step are:

- Change the tweets letter case to either lower-case
- Removal of tweets with lacking some of the processed fields (hashtag, text)

- Filter the fields not used in the processing (we use only the following fields: hashtag, text and language).
- Filter the tweets with a different language ("lang":"es" or "lang":"en")

Tips:

- Use the chain mapper to perform the tweets clean-up.
- You can use the hadoop-predefined FieldSelectionMapper mappers to filter the fields.

3. The Top-N pattern (2.5%)

To obtain the tweets trending topics we have to select the Top-N hashtags. Use the Top N patter to obtain the Top N trending topics (N will be passed by arguments by the user).

In the Top N pattern, we keep data sorted in a local data structure. Each mapper calculates a list of the top N records in the split and sends its list to the reducer. A single reducer task finds the top N global records.

Tips:

- Each Mapper has to emit its Top-N hashtags. However it cannot know which hashtags are the top N until it has processed all its records (tweets). There in the map you only have to build sorted list of the hashtags without emitting nothing and once the mapper has processed all its input (cleanup method), emits the values of the top N hashtags (Top N elements in the sorted list).
- To get a better approximation to the Top-N hastags, it is recommendable that the mapper calculates the Top-2N hastags and the reducer the Top-N.
- You can use a TreeMap to sort the top N hashtags.
- In order the reduce can process together all the TopN partial list generated by each mapper you have to generate the same key for all partial result or generate a NullWritable key.
- Only one reducer is mandatory to obtain a global Top N order.

4. Join all previous mapreduce Jobs in a single application (2.5%)

Finally, you have to join all the previous MR jobs in a single application that receives three parameters (input tweets file, output dir, N) and generates as result the Top N trending topics.

You have to decide in with order you have to execute the different jobs and how you have to ingrate it.

Tips:

- Use the chain mapper to perform some pipeline integrations of some mappers / reducers
- Use the job dependencies to define the job order.

Optional Features

5. Sentiment of hashtags (2.5%)

The process of identifying subjective information in a data source is commonly referred to as sentiment analysis. In the previous exercise, we detect trending topics in a social network; we'll now analyze the text shared around those topics to determine whether they express a mostly positive or negative sentiment.

We are going to use the bag-of-words method in order to calculate the topic sentiment. Although simplistic in nature, it can be used as a baseline to mine opinion in text. The bag-of-words method works as follows: for each tweet and for each hashtag, we will count the number of times a positive or a negative word appears and normalize this count by the text length.

In the Mapper step, we emit for each hashtag in the tweet the overall sentiment of the tweet (simply the positive word count minus the negative word count) and the length of the tweet. We will use these values in the reducer to calculate an overall sentiment ratio weighted by the length of the tweets to estimate the sentiment expressed by a tweet on a hashtag.

The user will pass two additional parameters with the positive and the negative words files.

In the "data.zip" folder you will find the lexicon files (with the positive and negative words) for English and Spanish. You have more languages lexicons in the following URL: https://sites.google.com/site/datascienceslab/projects/multilingualsentiment

Tips:

- Use the Distributed Cached files to pass the positive/negative words files to mappers.
- Use the setup method to load the positive/negative words files

Extra Features (optional):

- Use Combiners and Partitioners to improve the application performance (+0,5%)
- Define your custom Writable type for the tweet record. (+0,5%)

Use the compressed Sequence file format for the partial results (+0,5%)

Performance Analysis

Analyze the final performance of your application, and the speed-up when increasing both the input dataset size and the number of nodes (map & reduce tasks).

Delivery Date

Delivered through SAKAI at May 24th, video-conference delivery at May 25th in the same schedule (18h-21h) as the classes. I will assign you a time slot for the presentation of the activity.