Mark as done

Opened: Monday, 23 September 2024, 12:00 AM
Due: Monday, 30 September 2024, 12:00 AM

NOTE: Assignment can be done in a group of 2. You can take help from friends regarding concepts, but you are NOT allowed to see code developed by another group, or show your code to another group. If you get stuck and need help debugging, you can take help from your TAS. Any odd person out can do the assignment individually, but can do a subset of the task OR form a group of 3 (but email me about it)

In this assignment you will write PySpark code to process a set of files (articles) in a dataset in parallel to find co-occurrences of entities in articles. For example, if an article mentions Narendra Modi and Rahul Gandhi, we say that the two entities co-occur. To keep your life simple, we assume that entity names are single words (Modi, Gandhi, etc), and we assume no two entities have the same name. The list of entities to consider is at the end of this page; use it in a constant array in your program.

Input:

- You are provided with a set of news articles as json files in a zip file newsdata.zip on Moodle, that you should save and unzip into a directory.
- Each json file contains metadata about the article and the article content (article_body) itself.

Objective of your code:

In every article some entities may be mentioned. Two different entities mentioned in an article are said to co-occur in the article.

You have the following tasks

1. Create a dataset (RDD) with (entity, article_id) pairs where the article contains the entity. There should not be any duplicate rows. You should tokenize the input, and perform case-insensitive comparison; the wordcount program example will help you with this. Print the dataset.
2. Convert the dataset to Spark DataFrame (with appropriate column names)
3. Use SparkSQL functions on dataframes (not SQL itself) to create a dataframe containing (entity, count) pairs indicating how many articles the entity occurs in. Entities that don't occur in any article should not be output. Print the dataframe.
4. As above, but for all pairs of the given entities, instead of single entities.
5. Find the top-10 entity pairs based on their co-occurrence counts.

NOTE: **Do NOT use pandas DataFrame in this assignment**; even if your code works, if you use pandas dataframe the point of big data processing will be lost, and so will your marks for the assignment!

Submission guidelines: Submit your code file(s) along with a README (in case you want to specify anything about your assignment) as a single tar.gz or zip file. The name of the tar file should be rollnum1_rollnum2.tar.gz or rollnum1_rollnum2.zip.

List of entities:

```
modi
rahul
jaitley
sonia
lalu
nitish
farooq
sushma
tharoor
smriti
mamata
karunanidhi
kejriwal
sidhu
yogi
mayawati
akhilesh
chandrababu
chidambaram
fadnavis
uddhav
pawar
```

Add submission

Submission status

Submission status	No submissions have been made yet
Grading status	Not graded
Time remaining	5 days 3 hours remaining
Last modified	-
Submission comments	▸ Comments (0)

Jump to...

‹ Previous Activity

Next Activity ›

✉ Contact site support

You are logged in as Soumik Dutta (Log out)