IAB BigData HW

데드라인: 9/21 자정

**[과제 목표]**

SparkRDD와 SparkSQL을 사용하여 다양한 데이터 분석을 연습한다.

**[과제 제출 방법]**

etl에서 Spark.ipynb 파일을 다운받아 구현을 완료한 ipynb 파일을 **IAB2022F\_학번\_이름.ipynb** 로 변경하신 후 eTL에 제출하시면 됩니다.

* 이름 변경이 제대로 되지 않은 경우 감점이 있을 수 있습니다.

**[숙제 환경 셋팅]**

제공된 ipynb notebook 을 colab으로 실행

**Problem 1. SparkSQL**

WikipediaTable에서 project='en' 에 해당하는 row들 중 size 항목이 1000 이상 10000 미만인 값들을 title 기준으로 내림차순 정렬한 표

⇒ 결과: project, title, size column을 갖는 table

**Problem 2. SparkSQL**

WikipediaTable과 다음의 OwnerTable을 JOIN한 후 각 owner 당 count의 평균값을 구한 표

⇒ 결과: owner, avg\_count column을 갖는 table

OwnerTable

| title | owner |
| --- | --- |
| Woodkid | Lila |
| Sia | Jane |
| Ryuichi\_Sakamoto | Sam |

**Problem 3. SparkRDD**

wikipedia\_dataset\_sample을 사용하여, project별로 count 필드의 median 값을 구하고, notebook상에 출력하는 RDD 프로그램을 작성하여라.

주의:

* count값들 중 중복된 값이 있는 경우, 단일한 값인 것으로 가정

ex) project=en인 데이터가 모두 (‘en’, 1L), (‘en’, 1L), (‘en’, 2L), (‘en’, 3L) 라고 할 때, (‘en’, 1L), (‘en’, 2L), (‘en’, 3L)로 가정하고 계산할 것

* median을 구해주는 Spark 내장 함수 또는 기타 라이브러리 함수를 사용하면 0점

ex) DataFrame.approxQuantile, NumPy.median, …

* **collect()해서 가져온 값에서 median을 구하면 0점**

Hint:

* 스파크의 groupByKey(), mapValues(..), distinct() 를 사용하셔서 median을 구하시면 됩니다.
* python의 sort()함수를 사용하시는 것은 가능합니다.
* 짝수 개의 element인 경우에는, median 두 개의 값의 평균을 구해주세요.   
  ex) [1,2,3,4] => (2+3)/2

05/25 전병곤 교수님 office hour (~18:00)

질문 있으신 분들은 채팅 혹은 마이크 켜고 말씀해주세요