CPSC 5330 -- Spring 2023
Lab 7: Spark DataFrames

Implementing TF-IDF Calculations using Spark DataFrames

This week you will implement your IR system using Spark DataFrames.
You will implement both the indexing phase and the query phase.

Important note – you need to use EMR Version 6.5.0 when solving this lab. You select the EMR version when you create your cluster. Using the default version for EMR will break.

Indexing

Write a function `indexDocuments(path)` where `path` will point to your (S3) location of the document corpus.

This function returns a `DataFrame` with this schema:

```
  StructType(List(StructField(term,StringType,true),
                  StructField(docid,StringType,true),
                  StructField(tfidf,FloatType,true)))
```

where tfidf will be computed using the same formula as appeared in the Lab 6 document.

You must use DataFrames entirely to build this frame -- you may use the `wholeTextFiles` method to produce an RDD with the filenames and documents, but you must immediately make it into a DataFrame, and all subsequent operations must be on DataFrames

---

Relevance

Write a function `relevance(query, index, n=5)` where query is a string of query words, and `index` is a DataFrame produced by your `indexDocuments` function. The last parameter controls the number of results returned.

The return value for this function is a list of length n or fewer, containing the ids of the documents with The list is a list of tuples of the form `(docid, relevanceValue)`. Relevance value is computed as defined in Lab 5.

You will use these functions to implement searches on several sample queries.

For details, hints, and additional instructions, please refer to the template notebook supplied in the course repository.

To Hand In

A Zip file containing (only) these files

- The file `lab7_tfidf_data_frame.ipynb` containing your solution
- A retrospective report in a file `retrospective.pdf` – a reflection on the assignment, with the following components
  - Your name
  - How much time you spent on the assignment
  - If parts of the assignments are not fully working, which parts and what the problem(s) are
  - Were there aspects of the assignment that were particularly challenging? Particularly confusing?
  - What were the main learning take-aways from this lab – that is, did it introduce particular concepts or techniques that might help you as an analyst or engineer in the future?