CPSC 5330 -- Spring 2023
Lab 6: Spark RDDs

Implementing TF-IDF Calculations using Spark RDDs

Your solution will be contained in a notebook `tfidf.ipynb`, starting with the notebook supplied in the repository.
Please pay attention to the following details.

1. You must implement TF-IDF using the formula below and use the line-processing (termification) code supplied
2. Your RDD that does the final TF-IDF calculation must generate tuples of this form
   `((term, docid) tfidf-value)`
3. Your notebook will also report on TFIDF values for some selected test cases. Instructions are in the supplied notebook.

To Hand In

A Zip file containing (only) these files

- The file `tfidf.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?

TF-IDF Formula

$$\text{TF-IDF}(\text{doc\_id}, \text{term}) = 1000000 * \frac{\frac{(\#\ of\ times\ \text{term}\ appears\ in\ \text{doc\_id})}{(total\ \#\ of\ terms\ in\ \text{doc\_id})}}{(\#\ \text{of documents}\ term\ \text{appears in})}$$