Homework 2: Getting Started with Spark and Climate Change

1. Setting up your Spark instance

- Read carefully the instructions uploaded on Canvas, and set up your spark environment in the mode of your preference. I recommend using Google's Colab.
- Go to Chapter 2 of our Spark book, and replicate the *end-to-end example* on page 22. Make sure to use PySpark. **[10 pts]**

2. Climate Change: Project Tasmania

In a remote part of the world, hidden somewhere in Tasmania, researchers of climate change decided to place a steel vault about the size of a school bus. It will operate much like a plane's black box, and its mission will be to inform whoever discovers it about the end of human civilization. In other words, it will serve as a guide for a post apocalyptic society. It will create an archive for the future habitats of earth that could be critical in forming their society by piecing together the missteps of humanity, should humanity be destroyed by climate change. The box will record leaders' actions (or inactions) by scraping the internet for keywords relating to climate change from newspapers, social media and peer-reviewed journals. It will collect daily metrics, including average oceanic and land temperatures, atmospheric carbon dioxide concentration and biodiversity loss. For the purposes of this homework, we will assume that you are the future species of earth that discovered the Tasmanian box, but we will simplify the analysis to just two datasets. More specific:

- On Canvas, you will find two datasets. The first dataset contains temperature data by countries. Date starts from 1750 for average land temperature and goes up to 2015. Answer the following questions:
    a. For which country and during what year, the highest average temperature was observed? **[5 pts]**
    b. Analyze the data by country over the years, and name which are the top 10 countries with the biggest change in average temperature. **[5 pts]**
- The second dataset contains data on CO2 Emissions per capita across countries from 1960 to 2014.
    a. Merge the two datasets by country, and keep the data from 1960 to 2014. **[10 pts]**
    b. What is the correlation between CO2 emissions and temperature change? **[20 pts]**
- Run all the above on Google's Colab. However, I also want you to create a cluster in google cloud and run your code as a spark job (this step will help you in your final project, where you will have to deal with big datasets that cannot just run on Google's Colab). In your report, briefly mention the steps you followed to do this and put a screenshot of your cluster **[20 pts]**
- Reflect on your everyday life activities. What can you personally do to make a positive change for the environment? Write a short paragraph with your thoughts. **[20 pts]**
- **Extra: Once you are done with the homework (and only then!), I suggest you watch the following documentaries on Netflix: 1) "History 101: Episode 1", 2) "History 101: Episode 4" and 3) "Cowspiracy: The Sustainability**

Secret”.

3. Start thinking about your final project

Read the passage found on Canvas from “**Exponential Organizations: why new organizations are ten times better, faster, and cheaper than yours**”.

- Ideate 3 ideas about new sources of information that can underpin new companies, and type a short paragraph describing your ideas (bullet points are accepted). **[10 pts]**