



School of Computer Science

COMP47470

---

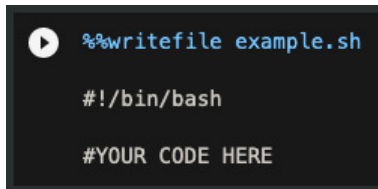
**Coding Project**  
**Bash, Data Management, Hadoop, Graph**  
**Processing, PySpark**

---

Teaching Assistant:	Ruth Holmes
Coordinator:	Anthony Ventresque
Date:	Tuesday 15 <sup>th</sup> November, 2022
Total Number of Pages:	6

## General Instructions

- We ask you to hand in a compressed .zip file containing a notebook for **each** question completed **and** a short report in pdf format.
- The report should be no longer than 10 pages and include:
  - A short introduction
  - A section for each completed exercise explaining your solutions and any written task from that exercise (do not include code in your report).
  - A short conclusion
- While you are encouraged to collaborate with your peers on this project, all written work must be your own. In particular we expect you to be able to explain every aspect of your solutions if asked. All reports will be checked for plagiarism.
- When writing scripts, use the "writefile" method in code cells as show below:



```
%%writefile example.sh

#!/bin/bash

#YOUR CODE HERE
```

- Complete 4 of the 5 exercises. If you complete all 5, you will be graded on your best 4. The breakdown of marks for the project will be as follows:
  - Exercise 1: 15%
  - Exercise 2: 15%
  - Exercise 3: 15%
  - Exercise 4: 15%
  - Exercise 5: 15%
  - Report: 40%
- Submissions should be made via Brightspace before: **11th December at 23:59**

## 1 Bash

Download the file `crimedata-australia.csv` from [here](#). Answer the following questions using Bash or Bash scripts. For each question, include your answer in a code cell and an explanation of how you solved it a proceeding text/markdown cell.

1. How many lines of content (no header) are there in the file? (`tail`, `wc`)
2. How many columns are there in the file? (`sed`, `wc`, ...)
3. For a given city (given as a column number, e.g., 10=Sydney), what is the type of crime that occurred the most in the crime list (`cat`, `cut`, `sort`, `head`)?
4. For a given city (given as a column number, e.g., 10=Sydney), how many crimes were committed? i.e., for a given column, what is the sum of all rows?

**Hint:** In Bash, command substitution (syntax: `$(<command>)`) allows us to execute a command and replace the command with its output.

5. For a given city (given as a column number, e.g., 10=Sydney), how many crimes were committed on average across all types of crimes.

**Hint:** `Q4` gives the sum, and `Q1` the number of lines. You can use `tr` to trim whitespace from outputs and get numbers.

6. Which city has the lowest average crime? You should loop through cities, computing their average crime, and keeping the one with the smallest value. Your output should be this city's name.

## 2 Data Management

Download the Players.csv and Teams.csv datasets from [here](#) and [here](#) respectively. Then complete the following tasks using **either** MySQL or MongoDB in non-interactive mode.

1. Describe the two datasets in a short paragraph (to be included in report).
2. Create a database/collection and tables/documents to represent the data, populating them using a Bash script.

**Hint:** Some player names include an apostrophe (special character), which will cause the insert query to fail for these rows. It might make things easier if you first remove/replace the apostrophes in the csv file using a Bash command.

3. What player, whose team's name contains "ia", played less than 200 minutes and made more than 100 passes? Return the player's surname.
4. What players made more than 20 shots? Return all players' information in descending order of shots made.
5. What goalkeepers belong to a team that played more than four games? List the surname of the goalkeepers, their team's name, and the number of minutes the goalkeeper played.
6. How many players who play on a team with ranking <10 played more than 350 minutes? Return one number labelled 'superstar' (column name).
7. What is the average number of passes made by forwards? By midfielders? Write one query that gives both values with the corresponding position.
8. Which team has the highest ratio of goalsFor to goalsAgainst? What is the value of the ratio for that team?
9. Find all teams whose defenders averaged more than 150 passes. Return the team and average number of passes by defenders, in descending order of average passes.

### 3 Hadoop

You may complete the following tasks using Java **or** Python scripts. The output for each task in this section should be just one line. Upload the following three books to your HDFS:

- <http://www.gutenberg.org/cache/epub/1524/pg1524.txt>
- <http://www.gutenberg.org/cache/epub/1112/pg1112.txt>
- <http://www.gutenberg.org/cache/epub/2267/pg2267.txt>

**Note:** If you are having trouble getting your solutions to run, it is acceptable to reduce the number of input texts.

1. How many words in the corpus begin with the letter Y/y?
2. What is the total number of unique words in the corpus?
3. Which word occurs most frequently in the corpus? How many times does it occur?
4. For that most commonly occurring word, what word most frequently follows it in a line?

## 4 Graph Processing

The dataset in this exercise is a reduced version of U.S. patent citation data, which is maintained by the National Bureau of Economic Research. A line contains information in the format:

$$< \textit{from\_vertex} > \quad < \textit{to\_vertex} >$$

which means that patent *from\_vertex* has a citation to patent *to\_vertex*.

We say that there is a one-hop citation from A to B if A cites B directly. We say that there is a two-hop citation from vertex A to vertex B if there is a patent C such that A cites C and C cites B. For the purpose of this exercise, we define the *significance* of a patent A to equal the number of distinct patents B such that there is either a one-hop citation or a two-hop citation from B to A.

Download the dataset [here](#) before uploading to your HDFS. Then write a MapReduce job in Hadoop that computes the significance of patents. Output the top patent (with largest significance) and its significance.

**Hints:** As we'll be looking for patent citations at a distance of 2, this problem will require 2 MapReduce jobs (2 map functions and 2 reduce functions):

- The first map function will read the graph file and output 2 key, value pairs for each line in the original input file:
  - 'toNode', 'fromNode'
  - 'fromNode', '-toNode'. The "negative" node meaning that the paper is cited.
- The first reduce function will collect positives and negatives in the list of values for a particular patent (key) and emits this as output.
- The second map function will emit the absolute value of each negative node as key, with each of the positive nodes + the key as value.
- The second reduce function will sum the results.
- The main class will need to chain multiple jobs.

## 5 PySpark

In this exercise you will explore a dataset extracted from Twitter, representing tweets of Wordle game results. Download the dataset [here](#). Each line, delimited by tabs, contains a daily Wordle puzzle's id, a tweet's id, the date of the tweet, the username of the tweet's author, and the text of the tweet. You may need to use operations on DataFrames and/or operations on RDDs. If you find either to be suitable, use just one.

1. Launch Spark and create your RDD/DataFrame from the input file.

**Hint:** This file contains multiline (i.e. records including a newline character) records and a header which you will need to consider.

2. Which Wordle puzzle in the dataset was the most tweeted about?
3. How many times do the words "play", "the" and "wordle" occur in the tweet\_text column? Return the results in descending order of count.

**Note:** Your answer should be case-insensitive e.g. "wordle", "Wordle" and "WORDLE" could all be counted as "wordle").

4. On what day of the week did people tweet the most games?

**Hint:** Use the date\_format method and Datetime Patterns.