James Young

Homework Assignment 2 – INFS 774

Working with Hadoop and MapReduce

**Lab 1**

At the end of the steps 2.1-2.9 the command Hadoop fs -ls issue the following results.

A screenshot of a cell phone

Description automatically generated

Continuing to step 3.3 and 3.4 in Lab 1 the following results were obtained, respectively.

A screenshot of a cell phone

Description automatically generated

3.4 used the “less” command

A screenshot of a cell phone

Description automatically generated

**Lab 2**

Working with MapReduce in lab 2 the following results were obtained.

Step 5:

A close up of a newspaper

Description automatically generated

Step 6: Here we can see it won’t duplicate, which is what we expected.

A close up of a newspaper

Description automatically generated

Step 7:

A screenshot of a cell phone

Description automatically generated

Step 8: Here are the results from the command

Hadoop fs -cat wordcounts/part-r-00000 | less

A screenshot of a cell phone

Description automatically generated

Step 9: Below we repeat the steps of 7 and 8 but for the poems now

Technically this (directly below) is step 6 and not requested directly, but it shows that the program ran before the results were available as expected.

A screenshot of a cell phone

Description automatically generated

A picture containing knife

Description automatically generated

Hadoop fs -cat pwordscounts/part-r-00000 | less

A screenshot of a cell phone

Description automatically generated

**Essay Questions**

Question 1: What is big-data, how is it different from related disciplines, is it over-hyped, what value does it bring to industry and research?

The immediate thought “Big Data” invokes is its implication for data volume, however, volume alone is insufficient to encapsulate the burgeoning “Big Data” environment and demonstrate its’ value. In the book “Understanding Big Data”, the authors, championing Hadoop on behalf of IBM, offer that large volumes of data have been around in enterprise settings for some time now. The authors of this book focus on the 3 V’s to define “Big Data” including Volume, Variety, and Velocity (now 5 V’s including Veracity and Value). The utilization of a new “variety” of data, especially unstructured data such as text and images as opposed to the traditional tabular data, is one differentiator of “Big Data” from just data with large volume. The “velocity” factor alludes to the use of streaming data to make important decisions in real-time, as opposed to cross-sectional or time-stationary data. “Big Data” allows for discovery-based approaches more-so than traditional tabular data. One reason for this suiting of this previously untapped unstructured data to discovery is its sheer size, estimated at ~80% of total data at the writing of the book. Another reason is the information richness that can be extracted. The field may be overhyped on a case by case basis, but generally sounds like it delivers value to industry and research. Some of the cases where it is overhyped may be when trying to solve problems that only require a sampling of the data or smaller scale techniques to solve. In industry, big data may allow the discovery of patterns that influence sales, production, competition, etc. In biology research big data allows for finding new potential disease-causing genes through next generation sequencing (NGS) which can be further tested in the wet-lab. To generalize (hopefully not overfitting), big data seems to lead to actionable and immediate value insights in business that can rapidly translate into money, whereas in the research that I am versed in it is a valuable hypothesis generating process helping the field focus it’s lab work on the most promising genes.

Question 2: What are two motivational uses of big data analytics other than those mentioned in the provided materials?

I will give two examples of motivational uses of “big data” that create value for industry or research. The examples, like my interests, are disparate. The first example, analyzing satellite imagery to forecast store revenue or oil production, is motivational because of its’ creativity but also conceptual simplicity/elegance (although not technically simple). Oil traders (or at least their quants) have used satellite imagery to forecast oil production and demand1. These images are considered unstructured data that represent changing information over time, which improves forecasting. Researchers working with these images can try to quantify anything from the number of cars in a Walmart lot to the amount of diesel derived contaminants in the air to track and project economic activity.

The second example I find motivational is the use of information extraction in scientific literature. It is cursory knowledge to researchers that there is a publishing explosion happening in scientific research. At the same time, some argue that innovation is slowing. How can these two statements be simultaneously true? Perhaps there is some spin on one of those statements or perhaps we as a research community are not fully integrating and synthesizing the firehose of smaller research advances. Using natural language processing (NLP), various startups and open-source technologies aim to distill the increasingly burdensome amount of info into insightful and synthesized outputs2. Hopefully, this will lead to a more efficient use and translation of prior research.

1. <https://www.marketwatch.com/story/how-oil-traders-are-using-satellites-to-keep-an-eye-on-an-increasingly-unpredictable-market-2019-10-04>

2. <https://www.nature.com/articles/d41586-018-06617-5>