Big Data: Data Sets

Distributing input data for the “Big Data” part of the course is challenging. In general, I'll distribute a few different input sets so you can test your programs at different scales. (Working from small to large is a completely realistic strategy.)

Getting/Finding The Data

I will distribute small, medium, and large data sets. You can find them (with consistent names in each location):

- the `.zip` file you're accustomed to downloading with each exercise (small sets only);
- on CSIL Linux workstations, in `/usr/shared/CMPT/data-science` (small and medium sets);
- downloadable at http://cmpt732.csil.sfu.ca/data-science/ *(http://cmpt732.csil.sfu.ca/data-science/)* (small and medium sets, zipped);
- on the Cluster's HDFS in `/courses/353/` (all data sets).

Using The Data: your computer

You can download the data sets (small in the exercise `.zip` or from the URL above). Unzip to a location, and then give that path on the command line, just like…

Using The Data: CSIL and the Cluster

On the CSIL computers and cluster, you can directly use these paths as input to your jobs. For the ones you download, you'll have to unzip, but can then use as input.

Assuming you have a Spark job that takes input and output directories as arguments, on a CSIL Linux workstation, you can run a command like:

```
spark-submit my-task.py /usr/shared/CMPT/data-science/dataset-2/ output
```

And on the cluster, you can use the HDFS data as input:

```
module load 353
spark-submit my-task.py /courses/353/dataset-3/ output
```

On Compression

It's common to keep big data data files compressed: it turns out that the time to decompress them is often less than the time it would take to read the larger uncompressed files from disk.

I will be distributing the files GZIP compressed. It's much more realistic to use a faster compression method *(https://catchchallenger.first-world.info/wiki/Quick_Benchmark:_Gzip_vs_Bzip2_vs_LZMA_vs_XZ_vs_LZ4_vs_LZO)* like LZ4 which creates larger files but takes a fraction of the processor time to decompress.

Unfortunately, the LZ4 libraries aren't part of anybody's standard installation and are non-trivial to get working. They work fine on the cluster, but would be hard to use elsewhere. Thus, I'm sticking with GZIP: it works.

The Data

In case you're interested, the data sets distributed actually are: [Data sizes after compression.]

- `xyz-*`: randomly-generated x, y, z values so we have something to work with while getting started. `-1` [450kB], `-2` [110MB], `-3`. [1.1GB]
- `weather-*`: subsets of the 2016 GHCN *(https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn)* data.
  - `-1`: a random subset of the data for a single day. [115kB]
  - `-2`: data for a single day. [850kB]
  - `-3`: all 2016 data. [270MB]
- `reddit-*` subsets of the Reddit Comment corpus *(http://files.pushshift.io/reddit/)* .
  - `-0`: a fake two-line data set for a minimal experiment [≈0kB]
  - `-1`: six low-volume subreddits from 2011. [1.3MB]
  - `-2`: eight low-volume subreddits for 2014 [74MB]
  - `-3`: five medium-volume subreddits for 2016 [790MB]
- `pagecounts-*`: subsets of the Wikipedia page view statistics *(https://dumps.wikimedia.org/other/pagecounts-raw/)* .
  - `-0`: a fake data set that covers the filtering cases. [≈0kB]
  - `-1`: Titles that start with "Simon_" for two days. [1MB]
  - `-2`: Titles that start with "Fa" for five days. [24MB]
  - `-3`: Titles that start with "A" for five days. [360MB]
  - `-4`: Full data for five days. [9GB]
- `nasa-logs-*`: a collection of web server logs from NASA *(http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html)* from two months in 1995.
  - `-1`: the first 1000 lines of both files.
  - `-2`: the full data set.
- `wordcount-*`: plain (mostly English) text, so we can count the occurrences of words.
  - `-1`: a subset of the `wordcount-2` files. [490kB]
  - `-2`: the Project Gutenberg files distributed in the NLTK gutenberg corpus *(http://www.nltk.org/book/ch02.html)* . [4MB]
  - `-3`: text extracted from the Standard Ebooks *(https://standardebooks.org/)* books with html2text *(https://pypi.python.org/pypi/html2text)* . [19MB]
  - `-4`: all of the text I could get out of the Project Gutenberg 2010 DVD *(http://www.gutenberg.org/wiki/Gutenberg:The_CD_and_DVD_Project)* . [1.9GB]

Updated Fri Aug. 28 2020, 11:21 by ggbaker.