Homework 3, Part 2: Programming in Spark

[50 points]

Prof. Yanlei Diao

Question 3 [50 points] Spark Exercise using the DBLP Dataset

This is a programming assignment based on a Spark cluster.

Each student will be provided with one Azure subscription, allowing the creation of a HDInsight Spark cluster composed of multiple cores.

To connect (via ssh) to the cluster, follow the tutorial in the lab session.

A. DBLP Data Analysis using Spark

We will first use our familiar dataset on DBLP publications. The schema is the following:

```
authors (id: INTEGER, name: VARCHAR(200))

venue (id: INTEGER, name: VARCHAR(200) NOT NULL, year: INTEGER NOT NULL, school:
VARCHAR (200), volume: VARCHAR(50), number: VARCHAR(50), type: INTEGER NOT NULL)

papers (id: INTEGER, name: VARCHAR NOT NULL, venue: INTEGER REFERENCES VENUE(id),
pages: VARCHAR(50), url: VARCHAR);

paperauths (paperid: INTEGER, authid: INTEGER)
```

The dataset is available in Dropbox. To download the dblp files from Dropbox:

```
wget "https://www.dropbox.com/s/rry1is62g1ays1l/dblp_tsv.tar.gz?dl=0" -O
dblp_tsv.tar.gz


tar -xzvf dblp_tsv.tar.gz
```

You will need to import the DBLP files into your Hadoop File System (HDFS). For more instructions, see the Spark Tutorial.

Please write Spark programs for the following tasks:

A1) Find the names of the top-k authors who have published the most in the DBLP dataset. For this task, *k* is an argument to your spark program.

A2) Find the set of authors who frequently write papers together. For this analytic task, you are expected to find the set of authors who have written at least X papers together, where X = 0.0001 * total_num_papers.

Your program will involve RDD/DataFrame operations as well as the FP-growth package in MLlib for frequent pattern mining (unless you want to implement your own). FP-growth is an improvement of the Apriori algorithm, which we learned in the lecture on "Scalable Machine Learning".

Please output the frequent co-author lists.

A3) Find the top-5 words whose frequency (in papers titles) varies the most between year 2000 and 2015 for SIGMOD conferences.

For each word, illustrate the frequency per year.

A4) Find the 20 clusters of topics from the titles of the papers published at SIGMOD conferences.

For feature engineering, you can use the Word2Vec package from the ML library:
https://spark.apache.org/docs/latest/mllib-feature-extraction.html
https://spark.apache.org/docs/latest/ml-features.html

For clustering, you can use KMeans from the ML library:
https://spark.apache.org/docs/latest/mllib-clustering.html#k-means
https://spark.apache.org/docs/latest/ml-clustering.html#k-means

For each cluster, print the most representative words in the cluster.

Submission

Moodle. Please include your solutions to all the above exercises, including either the notebook file (when a notebook such as Jupyter Notebook is used) or the commands you have executed in the Spark shell and output for the Spark exercise in a pdf document and submit through Moodle.