ID2221 Wikipedia Project

Group DOST

Omid Hazara (hazara@kth.se)
Djiar Salim (djiar@kth.se)

October 2020

1 Wikipedia

Wikipedia is an essential tool in our society, it might not be the most trustworthy source of information but nonetheless it is an encyclopedia of most of our knowledge. Most articles on Wikipedia have references and in particular, books and their International Standard Book Numbers (ISBN) . Our proposal was to create a relationship graph between articles and books, with articles on Wikipedia as data and Hadoop MapReduce as the tool of computation. The goal was to explore or discover the most important books in our society, and to identify different domains of knowledge. These goals were reached with satisfying results.

2 Dataset

For the purpose of this project we use Wikimedia dump service (Link), where we can download large XML dump files of articles. The files can easily be downloaded and unzipped. We chose to preprocess the XML files before putting them on HDFS. By default, mappers in Hadoop MapReduce treat each line in the input files as a record. This was problematic for us because each page/article in our XML files is structured on multiple lines and we want to keep track of the name of the article containing each reference. After some tries and tedious research, we found out that there are two options to solve our problem. The first option is to preprocess each file before adding it to HDFS in such a way that each article is on a separate line. The second option is to override the default RecordReader and modify the predicate for splitting files so that each mapper process an entire file. The side-effect of the latter option is that our XML dump files are skewed in size. Some contain 8,000 articles while other files contain 500,000+ articles. This was the main reason we chose to preprocess the files first, before putting them on HDFS for MapReduce.

3 Method

After putting the preprocessed dump files to HDFS, we can do a straight-forward MapReduce. The code for our MapReduce is well-commented and can be found in Wiki.java.

3.1 Map

Each record is one line, each line is one page/article. The Map function takes the value, which is the page in Text format, and then finds the title of that page, and for each referenced book (ISBN) it writes to the reducers with a key-value pair of (ISBN, page title). This means that for each record, the mapper (possibly) writes multiple times to the reducers.

3.2 Reduce

From the perspective of the reducer, each key-value pair is (ISBN, [page titles where the ISBN was found]). The desired output file will contain two nodes on each line which represents an edge between these two. For each record reducers receive, multiple key-value pairs are outputted as the final result. This output pair is in the form of (ISBN, page title). If the received value, the set of page titles of an ISBN, is smaller than 3 then we chose to not output that ISBN because we are mostly interested in ISBN's that are found in multiple articles.

4 Result

For this project, we chose to visualize the results using the Graphstream library. In this phase of the project we have our output file that consists of multiple lines of an ISBN and an article name separated by a tab. Then we simply read the output file line by line, for each line we then create a node for the ISBN and one for the article, these nodes are then connected with an edge. Afterwards, the graph is styled using CSS to present an user-friendly visualization. Book titles are fetched through Google Books API for each ISBN that have more than 4 edges (which means that the ISBN was found in more than 4 articles). We exported our GraphStream application as a JAR file called Wiki-Graph.jar so that the user does not have to compile the code and import external libraries. Figure 1 shows the visualization of a third of the entire Wikipedia dump (80GB) as of the first of October 2020.

5 How to run the code

In the project's main directory there is Wiki.java which is the main file of the project that contains the MapReduce. ProcessWikiDumps.java processes all unzipped dump files it can find under the directory wiki_dump. The directory

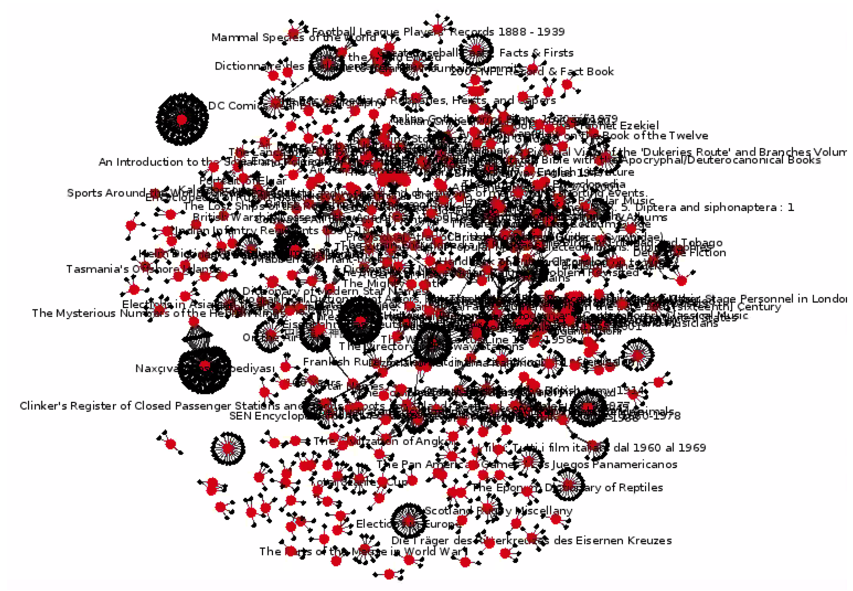

Mammal Species of the World
Football League Players' Records 1888 - 1939
Facts & Firsts
Prophet Ezekiel
Book of the Twelve
Apocryphal/Deuterocanonical Books
Diptera and siphonaptera : 1
Tasmania's Offshore Islands
Clinker's Register of Closed Passenger Stations and
SEN Encyclop
i film italiani dal 1960 al 1969
Los Juegos Panamericanos
The Eponym Dictionary of Reptiles
Scotland Rugby Miscellany
Die Träger des Ritterkreuzes des Eisernen Kreuzes

Figure 1: Graphical representation of about 7 million records (30 GB)

Wiki-Graph_source contains the source files for Wiki-Graph.jar which is our application that is used to visualize a file called "graph".

Preprocessing: Download and unzip the files from the Wikimedia and put them in the directory of wiki_dumps. From the terminal, we can compile and run ProcessWikiDumps.java. This will process the dumps and delete the old dump files we do not need anymore.

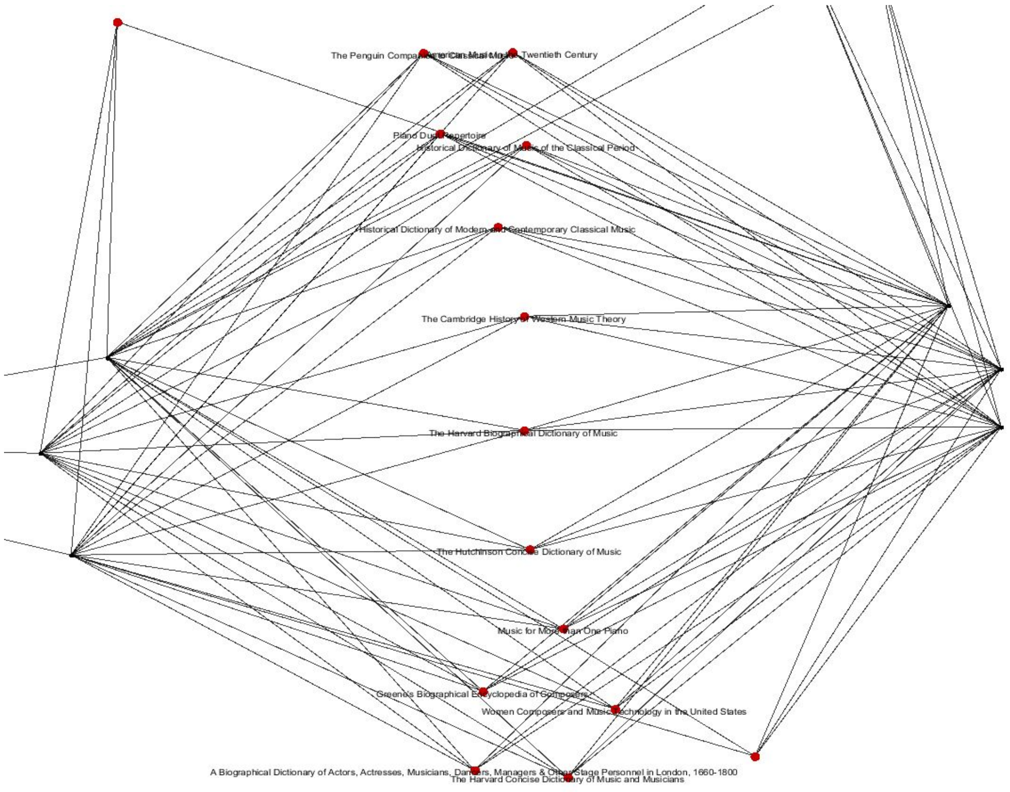
The instructions for using MapReduce in our project is similar to the lab instructions of lab1. To run this application you simply need to use the following commands in your terminal:

```
#start HDFS namenode and datanode
$HADOOP_HOME/bin/hdfs --daemon start namenode
$HADOOP_HOME/bin/hdfs --daemon start datanode
#compile and run preprocessing
javac ProcessWikiDumps.java
java ProcessWikiDumps
#put all files in wiki_dump to new folder called input
$HADOOP_HOME/bin/hdfs -put wiki_dumps/ /input
#build MapReduce JAR file and run
export HADOOP_CLASSPATH=$($HADOOP_HOME/bin/hadoop classpath)
mkdir wiki_classes
javac -cp $HADOOP_CLASSPATH -d wiki_classes Wiki.java
```

```
jar -cvf wiki.jar -C wiki_classes/ .
$HADOOP_HOME/bin/hadoop jar wiki.jar id2221.wiki.Wiki /input output
#save a local copy of the output and save it as "graph"
$HADOOP_HOME/bin/hdfs dfs -get -f output/part-r-00000 graph
#visualize the graph by running the Wiki-Graph JAR
chmod +x Wiki-Graph.jar
java -jar Wiki-Graph.jar
```

With GraphStream, it is possible to zoom in and zoom out with Shift+PgUp and Shift+PgDn, movement is done with the arrow keys. See Appendix for some of our exploration results. As of this writing, we have yet not tried the entire Wikipedia dump which is 80GB.

6 Appendix



The Penguin Companion to Classical Music
American Music in the Twentieth Century
Piano Duet Repertoire
Historical Dictionary of Music of the Classical Period
Historical Dictionary of Modern and Contemporary Classical Music
The Cambridge History of Western Music Theory
The Harvard Biographical Dictionary of Music
The Hutchinson Concise Dictionary of Music
Music for More than One Piano
Greene's Biographical Encyclopedia of Composers
Women Composers and Music Technology in the United States
A Biographical Dictionary of Actors, Actresses, Musicians, Dancers, Managers & Other Stage Personnel in London, 1660-1800
The Harvard Concise Dictionary of Music and Musicians

Figure 2: Graphical representation of references in music domain

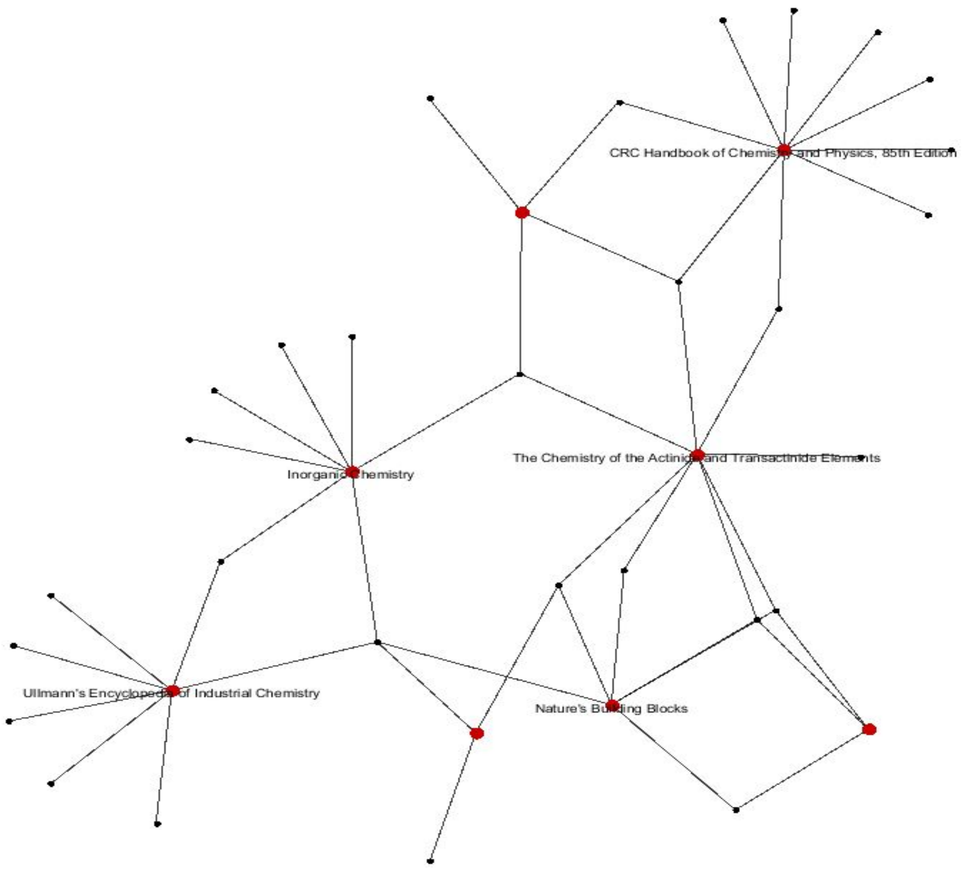

CRC Handbook of Chemistry and Physics, 85th Edition
Inorganic Chemistry
The Chemistry of the Actinide and Transactinide Elements
Ullmann's Encyclopedia of Industrial Chemistry
Nature's Building Blocks

Figure 3: Graphical representation of references in chemistry domain