PROJECT

In this problem you will be working with data from a collection of Wikipedia edit logs. The file that you will be working with is enwiki-20080103.main.bz2 .
This file is bzip2 compressed and is about 8.5GB. The file decompresses to a little over 300GB. You will want a part of the file to use while developing your program. Using bzcat you can decompress the file a little at a time. Once you have decompressed some of it you can recompress it using the bz2 command.

The output you create in this part will be used in the second part of the project. You should familiarize yourself with this file before planning out your code.

The data may be found at http://snap.stanford.edu/data/wiki-meta.html. Make sure you get the correct data set as there are more than one at this link. Please use Python streaming Hadoop mapreduce (AWS recommended), Spark/Scala, pyspark or simply python to run on your local machine.

WARNING : Although this data contains only text data there is offensive material contained in it. You are likely to find this once you begin to extractthe link data. Given the size of the dataset I am not completely aware of everything which one might find. If you believe that this is likely to be an issue for you please let me know.

1. Execute a job whose output is a file containing lines that consist of tab separated fields where the fields give the following: Article Name, Number of Edits of the Article, Number of Major Edits of the Article, Number of Outlinks, Number of Inlinks, Number of Distinct Editors, and the time of the earliest edit.
Thus there are seven tab separated fields.

Here by Article Name we mean anything that is an article that was edited or something that was linked to by an actual article. So Article Name can include things like image files. Do not include External links. It is thus the case that some of the fields may not exist. You can use 0 for all missing fields except the earliest edit where you should use something indicating absence, if no edit times can be associated with it.

2. Determine the directed graph which associates articles to the objects towhich they link, excluding External Links. Each line of output will be an article followed by an object to which it links where the two fields are tab separated.

3. In this part of the project you will work with the data you generated from the first part. For these problems you are free to use any method you like. In fact you are encouraged to choose whatever tool you feel best fits the problem. In this problem you will think of, and execute, a series of queries on your table data from problem 1 of the first part. A page refers to something which was the subject of at least one edit. First perform the following:
(a) Determine the page which has been edited the most number of times.
(b) Determine the page which has the largest number of distinct editors.
(c) Determine the page which has the earliest edit time.
(d) Determine the object which has the largest number of inlinks.
(e) Determine the page which has the largest number of outlinks.
(f) Determine the number of pages which have no outlinks.

Now think of four more queries to perform. In your submission include the results of each of the ten queries you performed. Also include a description of the four additional queries you performed.