COSC 589 - Web Search and Sense-Making

Assignment 4
Due Wednesday 3/15/17, 11:59pm
Submit to Blackboard

Task: Download and Clean Wikipedia

Introduction:

In this assignment, we download the latest English Wikipedia dump and perform initial cleaning of the data.

Requirements:

>100GB free disk space in your machine.

Instructions:

1. Following the steps in the lecture notes to download the English Wikipedia Dump at http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles-multistream.xml.bz2

2. Uncompress the Wikipedia dump file, by following the instructions in the lecture notes.

3. Write a PreProc.scala file to preprocess the file. Basically, we will extract all content in <page>…</page> and output each per line into an output file. Please keep the two beginning and closing tags <page> and </page> in your output file.

   You are welcome to use the following code template:

```
import scala.io.Source
import java.io.PrintWriter
import java.io.File
import scala.collection.mutable.StringBuilder

object PreProc {
  def main(args: Array[String]) {
    val inputfile = “your_wikidump_file”
    val outputfile = new PrintWriter(new File(“your_output_file”))
    var a_output_line = new StringBuilder

    // write your code to extract content in every <page> …. </page>
    // write each of that into one line in your output file
    for (inputline <- Source.fromFile(inputfile).getLines) {
        …….
    }
    outputfile.close
}
```

4. Please see sample input and output files on Piazza

5. Print the total number of pages in English Wikipedia to the screen

What to Submit:

- Your code
- Screen capture of the page count results that you print to the screen
- Screen capture of the beginning of your outputs (let us say the first 20 lines)

What NOT to Submit:

- Your input or output files

Where to submit:

- Blackboard

When:

- Due on 3/15/17, 11:59pm.