

COSC 589 - Web Search and Sense-Making

Assignment 3

Due Wednesday 2/21/18, 11:59pm

Task: Clean Wikipedia

Introduction:

In this assignment, we will perform initial cleaning of the Wikipedia data.

Requirements:

>100GB free disk space in your machine.

Instructions:

1. 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.

It is not required, but 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
  }
}
```

2. Please see sample input and output files on Piazza
3. Print the total number of pages in English Wikipedia to the screen

COSC 589 - Web Search and Sense-Making

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 output (by using 'head -n20' to show the first 20 lines)

What NOT to Submit:

- Your input or output files

Where to submit:

- Canvas

When:

- Due on 2/21/18, 11:59pm.