Lab 1 – Wikipedia Word Count

CC5212-1

March 11, 2015

- Download `http://aidanhogan.com/teaching/cc5212-1/code/mdp-lab1.zip`.
- Download `http://aidanhogan.com/teaching/cc5212-1/data/lab1/es-abstracts-10k.txt.gz` and put it in a local folder you can access. You can unzip it and open it if curious. It contains 10,000 abstracts from articles on Spanish Wikipedia (one abstract per line).
- Download `http://aidanhogan.com/teaching/cc5212-1/data/lab1/es-abstracts.txt.gz` and put it in the same folder as above. Do NOT unzip it. It contains abstracts from all of Spanish Wikipedia. You can continue to the next step(s) while waiting for it to download.
- Unzip the first file (`mdp-lap1.zip`) and open it as a project in Eclipse.
- Now you want to count the number of occurrences of Spanish words in the abstracts and print the top k most popular words.
- `org.mdp.wc.WordParserIterator` will help you. It parses the file and creates a stream of individual words. You can look over it if curious.
- `org.mdp.cli.Main` is just a command-line interface. You can ignore that for now.
- `org.mdp.cli.RunWordCountLocally` is the class in which you need to work. In it you will find two `TODO` comments.
- In the first `TODO` comment, you will need to store a count of words. The loop it is inside iterates over individual words. Every time you see a word, you need to add one to the number of occurrences.
- In the second `TODO` comment, you need to print the top k most frequently occurring words that you have found (in order of most common first). If k is negative, print all words in order.
- Time to test your code! You can run `org.mdp.cli.RunWordCountLocally` within Eclipse for the small file first ...
  - right click > Run As > Run Configurations, set it as the Main class, name the configuration.
  - Click on arguments and enter the following: `-i [dir]/es-abstracts-10k.txt.gz -igz -k 100`, replacing `[dir]` with the right directory.
- Did it work? If so, try running it again for the big file.
- Did the big file work? If so, you're done. ☺
- Submit the `RunWordCountLocally` class to ucursos before the Monday lecture.