MIS381N (Fall 2018) Homework5

Due on 10/29 before 12 pm

Problem Statement: Compute statistics for the words that appear in one or more messages contained in the input. Each line in the input file represents one message.

Input files: dataSet2a.txt,dataSet2b.txt

Output file content: Message count, total count, sum of squares, min occurrences, max occurrences, mean and variance for the words that appear in the document.

Same statistics for message length.

Output format: Text file with fields: word, \t, message count, total count, sum of squares, min, max, mean, variance.

For your code development use the file dataSet2Small.txt, then apply your solution to the two large files to generate the final output. When defining the input file(s) for the step that runs your map-reduce job (the run that will read both files), enter both filenames joined by a comma:

s3://bucket/data/dataSet2a.txt,s3://bucket/data/dataSet2b.txt

Required elements:

Implement a map program in Python that collects the numbers required to compute the requested statistics.

Implement a reduce program in Python that computes (via sum, min, max or other "math") the requested statistics.

Artifacts to submit:

Assignment2Code.zip - all files (Python) in a flat directory for easy inspection for grading

Assignment2Output.txt - output generated by your map-reduce program

Your Python code should include comments that describe your approach, so that it can be read and easily understood for grading.