Names: Bob Skowron, Jason Walker
Keys: rskowron, jwalker
SVN: jwalker: https://svn.seas.wustl.edu/repositories/jwalker/cse427s_fl17/

1. a. Test input: No now is definitely not the best time
Mapper output:
(N, 2)
(n, 3)
(i, 2)
(d, 10)
(n, 3)
(t, 3)
(b, 3)
(t, 4)
Reducer input:
(N, 2)
(b, 3)
(d, 10)
(i, 2)
(n, [3, 3])
(t, [3, 4])
Reducer output:
(N, 2.0)
(b, 4.0)
(d, 10.0)
(i, 2.0)
(n, 3.0)
(t, 3.5)

b. There are no differences in execution. The outputs return the same results. The only nuance is that instead of running the job as part of the *main* method, the job is run by ToolRunner via the *run* method.

c. Yes, the ToolRunner parses the parameter *-D caseSensitive=false* and sets it in the configuration where the mapper uses it to determine how to set the case sensitivity of the words. In this case, it adjusts them all to lower case and the output reflects the face that N and n are both considered the same key.

d. A: 3.891394576646375
W: 4.464014043300176
a: 3.0776554817818575
t: 3.733261651336357
z: 4.672727272727273

e. Command: *hadoop jar awl.jar stubs.AvgWordLength -D caseSensitive=false shakespeare awl_caseinsensitive*
In the run method we specifically check if there are only 2 arguments passed. Since the toolrunner assumes that its parameters are passed first (via the keys) then it passes <input> <output> <params> through to the run method where it will throw an error after counting 3.
a: 3.275899648342265
w: 4.373096283946263
z: 5.053333333333334

2. a. See SVN

   b. See SVN

   c.

   - Positive Words: 405
   - Negative Words: 805
   - Neutral Words: 5215
     Sentiment Score = $\frac{405-805}{405+805} \approx -0.33$
     Positivity Score = $\frac{405}{405+805} \approx .33$
     Based on these statistics, I would say that Shakespeare's poems exhibit an overall negative sentiment.

   d. First, the above statistics do not take into contextual polarity. Namely, we do not consider modifiers to any of the words. These could negate or intensify the primary positive or negative word identified. Secondly, the statistics do not take into account the frequency of the positive or negative words. A simple way to correct for the latter shortcoming would be to simply use the sum of the wordcounts for positive and negative words instead of merely the count of positive and negative words. The former shortcoming is much more difficult. A first pass could be to simply check the word that appears prior to the word of interest and determine if it is negating, intensifying or minimizing the target word. This still will not fully capture the sentiment however, as total context is necessary. For instance, a but clause could negate a positive and negative sentiment (e.g. I like this place but hate the crowds). There are numerous more complicated examples, but handling all of them requires work in mapping words to parts of speech, identifying the sentence structure and then analyzing.