1. Harsha Thulasi
2. I spent about 6 hours watching the demo videos and working on the solution.I did do the extra credit. A third of the time was trying out different combinations for extra credit. Pulling the data and putting this document together took me 1 hour.
3. I tried to do the extra credit trying out different regex and other combinations to identify the anomalies but figured I just only stick to few changes. I'll explain more below.
4. Here's the output for the three part files from s3 bucket. Not sure why they are split (if possible can you explain in class on tuesday)?

- part-r-00000

```
austen-emma   158128 104592 8903 6.498991332934848 friendshipindignationabominable
austen-persuasion   83259 54722 5555 6.531485439955146 eightandtwentieth
bryant-stories      45971 29866 3640 5.529152437131326
imitaterightinthemiddleofeveryone
chesterton-brown    71624 44397 7801 6.266096154552466 laughtercatastrophism
shakespeare-macbeth 17649 9010 3304 5.529227919898136 ayredrawnedagger
whitman-leaves      121375 65972 14081 6.1615616482861935 faithfulesthardiestlast
```

- part-r-00001

```
blake-poems   6817 3519 1331 5.206791995148575 chimneysweepers
carroll-alice 26382 17923 2380 5.816881428064783
importantunimportantunimportantimportant
edgeworth-parents   166000 107206 9095 6.034391264414737 sigheddrankhesitateddrank
milton-paradise     79645 40799 8958 5.869793543736807 transgressiondeath
```

- part-r-00002

```
austen-sense  118620 78146 6882 6.745268567475416 letterwritingdelicatetender
austin-persuasion   83259 54722 5555 6.531485439955146 eightandtwentieth
bible-kjv     790029 494645 12338 5.5733181214960865 mahershalalhashbaz
burgess-busterbrown 15864 10394 1234 5.419012797074954 selfrespecting
chesterton-ball     81505 50847 8154 6.292191271446279
ftpftpibiblioorgpubdocsbooksgutenbergetext
chesterton-thursday 57911 36118 6108 6.187124305969807 conspiratorbecause
melville-moby_dick  211802 123344 19288 6.328042686924868 matchestindergunpowderwhat
shakespeare-caesar  20387 10923 2832 5.435862214708369 tempestdroppingfire
shakespeare-hamlet  29578 15998 4516 5.502356406480118
pastoricallcomicallhistoricallpastorall
```

Extra credit:
Looking through some of the books by manually searching the substring from the output I noticed that they contained a "-" symbol inside it (ex: "eight-and-twentieth" in austin-persuasion.txt). I am not sure if a string contains a "-" its a single word or multiple (like "self-respecting"); google says they are single so it makes sense to keep them together. I added a regex check (2nd line that's italicized and bold) to replace certain character combinations in lines before tokenizing

```
    outToken.set(fileName);
    value.set(value.toString().replaceAll("[-]{2,}", " "));
    StringTokenizer itr = new StringTokenizer(value.toString());
```

```
    while (itr.hasMoreTokens()) {
```

Also updated the regex to include the “-” (highlighted in yellow below) as not replace with “”:

```
      nextToken = nextToken.replaceAll("[^-a-zA-Z]", "").toLowerCase();
```

In addition I added a check to remove any of bibliography check (in “chesterton-ball”) that showed up in string combos:

```
    while (itr.hasMoreTokens()) {
      String nextToken = itr.nextToken();
      if (nextToken.contains("ftp://") || nextToken.contains("http://"))
{
        continue;
      }
```

Not sure if that made a difference other than in one book but needed to go through all the words more extensively.

Here are the different regex combos I tried:

- [-]{2,} →two or more dashes

```
austen-emma   160418 107425 6948 6.404713830128507 conscience-stricken
austen-persuasion   83308 54779 5514 6.5378386904553265 eight-and-twentieth
bryant-stories      46049 29950 3586 5.532020622398907 delicious-smelling
chesterton-brown    71796 44600 7670 6.268973378438005 lilies-of-the-valley
shakespeare-macbeth 17650 9010 3307 5.540162037037037 ayre-drawne-dagger
whitman-leaves      122252 66760 13373 6.140290492323218 inter-transportation

blake-poems   6817 3519 1331 5.212856276531231 chimney-sweepers
carroll-alice 26546 18111 2274 5.779016004742146 bread-and-butter
edgeworth-parents   166911 108230 8355 6.00570883250115 powdering-slippers
milton-paradise     79716 40860 8912 5.871216800494132 greatly-instructed

austen-sense  119537 79330 6072 6.705797497948119 self-mortification
austin-persuasion   83308 54779 5514 6.5378386904553265 eight-and-twentieth
bible-kjv     790029 494645 12343 5.573389215394199 mahershalalhashbaz
burgess-busterbrown 15889 10426 1212 5.420831045213252 self-respecting
chesterton-ball     81723 51102 7969 6.281800071846119 crowderbblankslatenet
chesterton-thursday 58001 36211 6044 6.186232216613125 fellow-conspirators
melville-moby_dick  213386 124932 17975 6.297013136771655 standers-of-mast-heads
shakespeare-caesar  20389 10921 2842 5.442543303760034 tempest-dropping-fire
shakespeare-hamlet  29578 15998 4523 5.510972017673049
pastoricall-comicall-historicall-pastorall
```

- - -> all the single dashes. But I don’t know this is correct as compounded words might be legit.

```
austen-emma   160992 107693 6754 6.340606765605359 disinterestedness
austen-persuasion   83615 54970 5386 6.4765229533950075 acknowledgements
bryant-stories      46319 30001 3517 5.432773624218655 monocotyledonous
chesterton-brown    72370 44731 7419 6.129310032924491 respectablebesides
shakespeare-macbeth 17747 9033 3275 5.471654808354372 voluptuousnesse
whitman-leaves      124011 67205 12334 5.941291412878921 circumnavigation
```

```
blake-poems   6837 3524 1324 5.178991850286749 lamentations
carroll-alice 26687 18157 2241 5.68042203985932 affectionately
edgeworth-parents   167700 108499 8067 5.92616678772318 contradistinction
milton-paradise     80050 40966 8710 5.818519087094463 hiskithetmroboscis

austen-sense  119904 79516 5940 6.6531643062295736 disqualifications
austin-persuasion   83615 54970 5386 6.4765229533950075 acknowledgements
bible-kjv     790050 494655 12335 5.573012407115896 mahershalalhashbaz
burgess-busterbrown 15963 10446 1210 5.346383904295813 uncomfortable
chesterton-ball     82104 51180 7803 6.198971672487389 crowderbblankslatenet
chesterton-thursday 58295 36268 5923 6.096835701638898 incomprehensible
melville-moby_dick  215941 125424 16711 6.106775522829966 uninterpenetratingly
shakespeare-caesar  20455 10943 2816 5.402859545836837 enfranchisement
shakespeare-hamlet  29688 16030 4496 5.464709327866452 transformation
```