BigML Assignment 3: Streaming Phrase Finding

Runyun Zhang
runyunz@andrew.cmu.edu

1. **The top 20 phrases (sorted by total score) with their phraseness and informativeness scores from the full data set and the apple data set. (2 points) [Full]**

of the 0.0020832626340116797 0.009536130506247345 -0.007452867872235665
in the 0.0016023935628376217 0.005965858046680817 -0.004363464483843196
new york 0.0012755626618492387 0.0017100057514728493 -4.344430896236108E-4
it is 0.0010934197604781156 0.0021801171732069463 -0.0010866974127288307
to be 0.0010018887754239968 0.00220906636833746 -0.0012071775929134632
can be 9.880722143538335E-4 0.0015285884146785506 -5.405162003247171E-4
on the 8.90743818812407E-4 0.002501874484238423 -0.001611130665426016
et al 8.825601618170773E-4 0.0010342538253753737 -1.5169366355829636E-4
united states 8.115412170386163E-4 0.001069125886118662 -2.575846690800456E-4
have been 7.622503395007055E-4 0.001199434259770842 -4.371839202701367E-4
has been 7.079262616844092E-4 0.0010831959214428432 -3.7526965975843404E-4
may be 6.976788534651598E-4 0.0011717739937885292 -4.740951403233694E-4
such as 6.828972959051756E-4 9.907523658276042E-4 -3.0785506992242863E-4
it was 6.575119438777936E-4 0.0012583052824194907 -6.007933385416971E-4
as a 5.872106931827664E-4 0.0013823990466810333 -7.951883534982668E-4
had been 5.577492800972097E-4 8.924299781598284E-4 -3.3468069806261874E-4
should be 5.559485423410916E-4 8.78389990263967E-4 -3.224414479228753E-4

for example 5.400601499809626E-4 7.767459957614331E-4
-2.3668584578047047E-4
as well 5.037790016262232E-4 7.627178133107961E-4
-2.589388116845729E-4
the same 4.889614081592511E-4 0.001007314378377972
-5.183529702187209E-4

[Apple]

the apple 0.494502469911671960.5664374001491665
-0.0719349302374946
an apple 0.361066947841834460.4072569817056597 -0.046190033863825246
apple computer 0.219805381191044060.22591872279502154
-0.006113341603977483
apple pie 0.1712205674445035 0.1845002110858834
-0.013279643641379898
apple juice 0.1548870175813611 0.16303066482486422
-0.008143647243503105
apple tree 0.129131646179121 0.1445218544324387
-0.015390208253317699
and apple 0.116679388973749560.12997681876639972
-0.013297429792650152
of apple 0.112048369031455130.127433679371283 -0.01538531033982788
apple menu 0.107688944789850550.1011622628528523
0.006526681936998256
apple and 0.103792661112283330.11809157560094728
-0.014298914488663958
apple trees 0.102050031686572 0.11374946024350052
-0.011699428556928532
apple macintosh 0.100967132451682780.10114463053376355
-1.7749808208076482E-4
apple cider 0.0800005347881293 0.083121413931737
-0.0031208791436076956
apple ii 0.071314307714958570.08056067552585897-0.009246367810900394
crab apple 0.055738672647432280.058863904081888734
-0.003125231434456455
big apple 0.052764729387795486 0.05708991562100881
-0.004325186233213325
apple orchard 0.042305800648352030.04650300242158007
-0.004197201773228041
apple event 0.039300785792337145 0.02850847214004722
0.010792313652289923

apple of 0.036821977285912987 0.045165373601374556
-0.008343396315461571
with apple 0.0354607461363551 0.039241151143072375
-0.0037804050067172705

2. **What do you notice about the phrases ranked highest in your results for the two data sets? Do they give you any insights into events or trends in the 90s? (2 points)**
   The results from full dataset are basically general phrases with less meaning (of the, in the, etc.) The results from apple dataset are more meaningful and the trend of Apple Inc. product can be observed. Those phrases occurring more often in the 90s tend to have larger negative informativeness scores.

3. **Are there any downfalls you see to using the total phrase score? For example, are there some phrases that are ranked high even though you don't think they should be? Why are they ranked so high? (2 points)**
   The downfalls would be the penalization of *informativeness* has been emphasized to much that the trend of phrase changing cannot be observed. Another downfall would be how to eliminate those common phrases which make less senses.
   It surprises me that 'new york' has a very high score in the full dataset. One assumption would be the dataset chose focuses more on new york area.

4. **How could you improve upon the total score proposed by Tomokiyo and Hurst? (2 points)**
   Give weights to *phraseness* and *informativeness* to adjust their influence according to different use case.

5. **Consider the workflow discussed on 1/27 in class:**
   a) **Answer the questions below (5 points):**
      i. **(a) What are the entries in eventCounts.dat associated with the words "toast", "likes", and "steak"?**
         X=toast^Y=breakfast//3
         X=likes^Y=breakfast//2
         X=likes^Y=diner//2
         X=steak^Y=diner//2
      ii. **What are the entries in words.dat associated with the words "toast", "likes", and "steak"?**
         toast//C[w^Y=breakfast]=3
         likes//C[w^Y=breakfast]=2,C[w^Y=dinner]=2
         steak//C[w^Y=dinner]=2
      iii. **What is the output of requestWordCounts on the test corpus? Please**

write key values pairs as "key//value" so we can see the different parts easily.

Jane//~ctr to id1
ordered//~ctr to id1
eggs//~ctr to id1
and//~ctr to id1
toast//~ctr to id1

iv. **What is the output of answerWordCountRequests on the test corpus?**

id1//~ctr for and is C[w^Y=breakfast]=1, C[w^Y=diner]=1
id1// ~ctr for Jane is C[w^Y=breakfast]=1
id1//~ctr for toast is C[w^Y=breakfast]=3

v. **What is the input to testNBUsingRequests?**

id1// Jane ordered eggs and toast
id1//~ctr for and is C[w^Y=breakfast]=1, C[w^Y=diner]=1
id1// ~ctr for Jane is C[w^Y=breakfast]=1
id1//~ctr for toast is C[w^Y=breakfast]=3

b) Suppose there are K classes, V distinct words in the training corpus, and N tokens in the test corpus. Answer the questions below (7 points):

i. **The number of integers that are stored in eventCounts.dat.**

Around (K*V). [Actual number depend on the appearance of each word appearing in each class]

ii. **The number of key-value pairs that are stored in eventCounts.dat.**

Around (K*V).

iii. **The number of integers that are stored in words.dat.**

Around (K*V).

iv. **The number of key-values pairs that are stored in words.dat.**

Around V.

v. **The number of key-value pairs output by requestWordCounts.**

N.

vi. **The number of key-value pairs read as input by answerWordCountRequests.**

V + N.

vii. **The number of key-value pairs produced as output by answerWordCountRequests.**

N.

6. **Your answers to collaboration policy (on the first page of this handout).**

Did you receive any help whatsoever from anyone in solving this assignment? No.
Did you give any help whatsoever to anyone in solving this assignment? No.