10-605 HW3 Report

Chen Sun / chens1

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 data set:

of	the	0.015837272	0.017898652	-0.00206138
in	the	0.010797527	0.011300736	-5.03E-04
on	the	0.005004625	0.005059079	-5.45E-05
it	is	0.004844729	0.005163353	-3.19E-04
to	be	0.004772374	0.00494443	-1.72E-04
new	york	0.004687369	0.004693502	-6.13E-06
can	be	0.003926743	0.003877402	4.93E-05
to	the	0.003228446	0.003608138	-3.80E-04
et	al	0.00314301	0.002908514	2.34E-04
have	been	0.002992145	0.003108425	-1.16E-04
as	a	0.002977785	0.002926659	5.11E-05
united	states	0.002932418	0.002995389	-6.30E-05
it	was	0.002882509	0.002978492	-9.60E-05
from	the	0.002850234	0.002937803	-8.76E-05
at	the	0.002823642	0.002805642	1.80E-05
may	be	0.002815883	0.00297488	-1.59E-04
has	been	0.002734426	0.002850264	-1.16E-04
such	as	0.002688202	0.002504307	1.84E-04
for	the	0.002233539	0.002430603	-1.97E-04
the	same	0.002224628	0.002331717	-1.07E-04

Apple data set:

the apple	1.230975	1.251297	-0.020321
an apple	0.900258	0.919985	-0.019727
apple computer	0.548942	0.518554	0.030388
apple pie	0.427808	0.428426	-0.000618
apple juice	0.386985	0.377655	0.009330
apple tree	0.322353	0.332125	-0.009771
and apple	0.290373	0.284788	0.005584
of apple	0.278744	0.278463	0.000282
apple menu	0.269106	0.232661	0.036446
apple and	0.258231	0.258445	-0.000214
apple trees	0.254754	0.261379	-0.006625
apple macintosh	0.252340	0.234423	0.017916
apple cider	0.199980	0.193860	0.006120
apple ii	0.177902	0.183431	-0.005529
crab apple	0.139275	0.136583	0.002692
big apple	0.131667	0.129996	0.001671
apple orchard	0.105678	0.107756	-0.002078
apple event	0.098277	0.064819	0.033458
apple of	0.091442	0.097293	-0.005852
with apple	0.088271	0.086235	0.002036

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)

Most of them in the full data data are stop words combination. They are just the middle of the sentence to be got. Not much information could be inferred from this article about the trends of the 90s.

In apple result, “apple computer” ranks the third after “the apple” and “an apple”. It shows that Apple computer has drawn great attention since 1990s. Also we can see that Apple computer, Apple menu and Apple macintosh all have higher scores in informative scores.
Also food made from apple also get higher attention since apple pie, apple juice are also listed here.

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)

Yes, for full set size:
I think most of them should be removed from the result set. All I get are the combinations of common stop words in the texts. They appear here simply because they appear far too many

times both in background corpus and foreground corpus. We can also see that both the infor score and phrase score are near to zero or small than zero. This could not provide any further information.

For apple set:
of apple, apple and, apple of, with apple, etc, Those are all apple and a “stop word”. Those should not appear in the word list since they don’t provide much information but because they have appeared here so many times.

4. How could you improve upon the total score proposed by Tomokiyo and Hurst?
Remove stop words to build a more accurate language model.
Give the score a weight combination.
Also, we could manually train the model parameters of the two. By adjusting the combination factor to finally determine a possible results.

5. Part 1. Answer the questions below (5 points):
(a) What are the entries in eventCounts.dat associated with the words “toast", “likes", and “steak"?

toast c[Y=breakfast] 3
likes c[Y=breakfast]2
likes c[Y=dinner] 2
steak c[Y=dinner] 2

(b) What are the entries in words.dat associated with the words “toast", “likes", and “steak"?

toast C[Y=breakfast] =3
likes C[Y=breakfast]=2, C[Y=dinner] 2
steak C[Y=dinner] 2

(c) 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 //~crt to id 1
ordered //~crt to id 1
eggs //~crt to id 1
and //~crt to id 1
toast //~crt to id 1

(d) What is the output of answerWordCountRequests on the test corpus?
id1 ~crt for Jane C[Y=breakfast] =1
id1 ~crt for and C[Y=breakfast] = 1, C[Y=dinner] = 1
id1 ~crt for eggs 0

id1 ~crt for ordered 0
id1 ~crt for toast C[Y=breakfast] =3

(e) What is the input to testNBUsingRequests?

id1 // Jane ordered eggs and toast
id1 ~crt for Jane C[Y=breakfast] =1
id1 ~crt for and C[Y=breakfast] = 1, C[Y=dinner] = 1
id1 ~crt for eggs 0
id1 ~crt for ordered 0
id1 ~crt for toast C[Y=breakfast] =3

Part 2. 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):

(a) The number of integers that are stored in eventCounts.dat.

V to KV (a word may appear in one to many corpus)

(b) The number of key-value pairs that are stored in eventCounts.dat.

The same as (a)

(c) The number of integers that are stored in words.dat.

The same as (a)

(d) The number of key-values pairs that are stored in words.dat.

V

(e) The number of key-value pairs output by requestWordCounts.

N (suppose there is only one test id)

(f) The number of key-value pairs read as input by answerWordCountRequests.

N+V (suppose there is only one test id)

(g) The number of key-value pairs produced as output by answerWordCountRequests.

N (suppose N tokens are included in V)

6. I do not receive any help from anybody.