HW3 – Report
Daniel Ribeiro Silva (drsilva)

Question 1

Top 20 Phrases for apple dataset:

Phrase	Score	Phraseness	Informativeness
the apple	1.230975068	1.25129651	-0.020321442
an apple	0.900257662	0.919985061	-0.019727399
apple computer	0.548942001	0.518554253	0.030387748
apple pie	0.427807519	0.428425611	-6.18E-04
apple juice	0.386985076	0.377655105	9.33E-03
apple tree	0.322353388	0.332124555	-0.009771167
and apple	0.290372684	0.284788262	0.005584422
of apple	0.278744215	0.278462638	2.82E-04
apple menu	0.269106458	0.23266057	3.64E-02
apple and	0.258231142	0.258444941	-2.14E-04
apple trees	0.25475374	0.261378783	-6.63E-03
apple macintosh	0.252339609	0.234423213	0.017916395
apple cider	0.199980386	0.193860474	0.006119913
apple ii	0.177901787	0.183430504	-0.005528717
crab apple	0.139275303	0.136583166	0.002692137
big apple	0.13166656	0.12999564	0.00167092
apple orchard	0.10567812	0.107756168	-0.002078048
apple event	0.098277012	0.064818791	0.033458221
apple of	0.091441662	0.097293313	-0.00585165
with apple	0.088270976	0.086235388	0.002035589

Top 20 phrases for the full dataset:

Phrase	Score	Phraseness	Informativeness
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

Question 2
For the full data it is hard to observe any trend, since most of them are composed by stop words, but for the apple dataset we see a some trend given by the appearance of the Apple products, such as apple computer. Since the background corpus is from a time where that did not exist, the Apple-related phrases have a very high informativeness score. Nevertheless, since the Informativeness score is orders of magnitude smaller than the phraseness, that trend is not sufficiently highlighted. A score normalization of both P and I would make that more clear.

Question 3
There are many phrases that are high ranked and should not be. The most common examples are phrases made up of stop words such as “of the” or “in the”. They are very high ranked because they have a high phraseness score. They are common language constructions, but with no real meaning (with a very low informativeness score). But since the phraseness score is so large, that effect dominates the final score. That is quite dramatic for the large dataset. Among the top 20 phrases, only 2 of them are actual phrases (“new york” and “united states”). Here we see the importance of removing stop words and/or balancing/scaling both scores

Question 4

A big problem with Tomokiyo and Hurst final score is that phraseness and informativeness can have different orders of magnitude. That is actually what happened in our dataset. In our particular example, the phraseness value was many orders of magnitude bigger than the informativeness values. As a consequence, adding them up to create the final score was basically equivalent to only conseidering the phraseness score. One suggestion would be to normalize both the scores. Ideally we would use a linear scale, but since it is a streaming algorithm , we are unable to know beforehand what are the min and max values of both the informativeness and the phraseness. One possible way to make it work with streaming is to use a logistic function. Instead of using I and P for the score, we would use $\sigma(I)$ and $\sigma(P)$. So one possible score is $score = \sigma(P) + \sigma(I)$ Another second possible score is $score = \sigma(P) * \sigma(I)$.

Question 5

Part 1

a.
   - X=toast^Y=breakfast 3
   - X=likes^Y=breakfast 2
   - X=likes^Y=dinner 2
   - X=steak^Y=dinner 2

b.
   - Toast C[w^ Y=breakfast]=3
   - Likes C[w^ Y=breakfast]=2, C[w^ Y=dinner]=2
   - Steak C[w^ Y=dinner]=2

c.
   - Jane // ~ctr to id1
   - ordered // ~ctr to id1
   - eggs // ~ctr to id1
   - and // ~ctr to id1
   - toast // ~ctr to id1

d.
   - id1 ~ctr for Jane is C[w^ Y=breakfast]=1
   - id1 ~ctr for and is C[w^ Y=breakfast]=1, C[w^ Y=dinner]=1
   - id1 ~ctr for toast is C[w^ Y=breakfast]=3

e.
   - id1 Jane ordered eggs and toast
   - ~ctr for Jane is C[w^ Y=breakfast]=1
   - ~ctr for and is C[w^ Y=breakfast]=1, C[w^ Y=dinner]=1
   - ~ctr for toast is C[w^ Y=breakfast]=3

Part 2

a. V*K
b. V*K
c. V*K
d. V
e. N
f. V+N

g. N

Question 6
I have discussed some ideas about the pipeline with Poorna and Vinay, but nothing very deep really.