10-605 Assignment 3: Streaming Phrase Finding

Namit Katariya (andrew id: nkatariy)

February 13, 2013

QUESTION 1

- Top 20 phrases in apple dataset:

Phrase	Total Score	Phrasiness	Informativeness
the apple	1.2228536165218358	1.2432898744503482	-0.02043625792851254
an apple	0.8944406445408146	0.9141930611561296	-0.01975241661531503
apple computer	0.5454707350876162	0.5153269508174773	0.030143784270138994
apple pie	0.42511924266490203	0.4257813820316012	-6.62139366699174E-4
apple juice	0.3845523229817979	0.37532016939839324	0.009232153583404656
apple tree	0.3203031473549426	0.3300593266213714	-0.00975617926642878
and apple	0.2884501779931379	0.2829551017519293	0.005495076241208589
of apple	0.276889927504093	0.27666665778246463	2.2326972162837186E-4
apple menu	0.2674185989083051	0.23121463931942213	0.03620395958888299
apple and	0.25651541202909567	0.25677993318787395	-2.6452115877830017E-4
apple trees	0.2531338387783706	0.2597533431308324	-0.00661950435246185
apple macintosh	0.25075948131678205	0.2329743957619985	0.017785085554783565
apple cider	0.1987315888283459	0.19266771838529745	0.006063870443048458
apple ii	0.17676004988663813	0.18228225479836765	-0.005522204911729537
crab apple	0.13840079569011662	0.1357397183059859	0.0026610773841307125
big apple	0.13082483408143342	0.1291819135789967	0.001642920502436726
apple orchard	0.1050118516363894	0.10709009117147242	-0.002078239535083016
apple event	0.09766629161175833	0.06441261462073067	0.033253676991027666
apple of	0.09081973265221893	0.09665921413558815	-0.00583948148336922
with apple	0.08768841716919668	0.08568149612676443	0.0020069210424322444

- Top 20 phrases in the Full dataset:

Phrase	Total Score	Phrasiness	Informativeness
of the	0.015740232192325256	0.017813394986453294	-0.0020731627941280384
in the	0.010734501314750868	0.011247332134259502	-5.128308195086337E-4
on the	0.004977961433819259	0.005036476548505788	-5.851511468652945E-5
it is	0.004822895986738756	0.005143167904779404	-3.2027191804064867E-4
to be	0.0047495731325583315	0.0049241633342931275	-1.7459020173479645E-4
new york	0.004669992497639615	0.004677253397120053	-7.2608994804381016E-6
can be	0.003910800205769494	0.003863083241497727	4.771696427176755E-5
to the	0.0032016026888147735	0.003586433512796285	-3.8483082398151145E-4
et al	0.0031318276472495797	0.0028986416726793996	2.331859745701799E-4
have been	0.0029801576977533527	0.0030971464502816902	-1.1698875252833747E-4
as a	0.0029629052146228	0.0029140852510580262	4.88199635647739E-5
united states	0.002921780355416181	0.0029851908240693765	-6.341046865319557E-5
it was	0.002869638745334329	0.002966842862802441	-9.720411746811221E-5
from the	0.0028344451647436394	0.0029243777769915325	-8.993261224789303E-5
at the	0.002808246862117229	0.00279281567299676 3	1.543118912046588E-5
may be	0.0028042632082229246	0.0029639021899455172	-1.5963898172259266E-4
has been	0.002723665789608882	0.0028400505478043975	-1.1638475819551547E-4
such as	0.0026773995951420975	0.0024950322720301305	1.8236732311196704E-4
for the	0.0022182162223921163	0.002417832998790889	-1.9961677639877262E-4
the same	0.0022143770961817457	0.002322425397183739	-1.08048301001993E-4

QUESTION 2.A

Apple dataset : *apple computer*, *apple menu* and *apple macintosh* are three of the top phrases so one could conclude that Apple, the company must have been on the rise. We know that this was, in fact, true. *big apple* is also one of the top which is understandable since New York is the center of a lot of activity. Most of the remaining phrases are simply words that occur together (i.e are valid commonly used phrases) e.g.apple juice, apple cider, apple trees etc I don't think any conclusions can be drawn from those.

Full dataset : We do not get any meaningful insights since most of the phrases are frequently occurring words (stopwords) that usually appear together. New York and United States are the only two phrases that do not belong to this category but it's hard to draw any conclusions from them.

QUESTION 2.B

Yes. The top phrases are all often occurring words. So we could ignore phrases that contain stopwords / both the words are stopwords. These words almost always occur together so their phrasiness score is very high.

QUESTION 2.C

We could weight the informativeness higher. Second, as mentioned in 2b, we could ignore phrases containing stopwords.

BONUS QUESTION

I implemented the stopwords idea mentioned above. I wrote a script that uses the nltk stopwords corpus to remove phrases both of whose words are phrases. The following are the results I obtained.

Apple dataset

Phrase	Total Score	Phrasiness	Informativeness
apple computer	0.5454707350876162	0.5153269508174773	0.030143784270138994
apple pie	0.42511924266490203	0.4257813820316012	-6.62139366699174E-4
apple juice	0.3845523229817979	0.37532016939839324	0.009232153583404656
apple tree	0.3203031473549426	0.3300593266213714	-0.00975617926642878
apple menu	0.2674185989083051	0.23121463931942213	0.03620395958888299
apple trees	0.2531338387783706	0.2597533431308324	-0.00661950435246185
apple macintosh	0.25075948131678205	0.2329743957619985	0.017785085554783565
apple cider	0.1987315888283459	0.19266771838529745	0.006063870443048458
apple ii	0.17676004988663813	0.18228225479836765	-0.005522204911729537
crab apple	0.13840079569011662	0.1357397183059859	0.0026610773841307125
big apple	0.13082483408143342	0.1291819135789967	0.001642920502436726
apple orchard	0.1050118516363894	0.10709009117147242	-0.002078239535083016
apple event	0.09766629161175833	0.06441261462073067	0.033253676991027666
apple butter	0.0688542339353877	0.06927867734905391	-4.244434136662066E-4
apple computers	0.06857986620449702	0.06506008964898637	0.0035197765555106493
apple orchards	0.06747071413895256	0.06854275255126634	-0.0010720384123137847
golden apple	0.05520249070200211	0.05700665302209527	-0.0018041623200931588
apple slices	0.053124515326728364	0.052688086979418604	4.3642834730976205E-4
apple events	0.04998160155210139	0.03404379132823658	0.015937810223864816
red apple	0.04875258938756106	0.049330272120867706	-5.776827333066506E-4

Full dataset

Phrase	Total Score	Phrasiness	Informativeness
new york	0.004669992497639615	0.004677253397120053	-7.2608994804381016E-6
et al	0.0031318276472495797	0.0028986416726793996	2.331859745701799E-4
united states	0.002921780355416181	0.0029851908240693765	-6.341046865319557E-5
university press	9.994306440311385E-4	9.091222142710631E-4	9.030842976007531E-5
see also	8.898648674514287E-4	8.427731170118197E-4	4.709175043960893E-5
health care	7.046305584573879E-4	5.968230409758352E-4	1.0780751748155269E-4
los angeles	6.155616361130264E-4	6.06685698518084E-4	8.875937594942451E-6
even though	5.780372551168027E-4	5.695084820442275E-4	8.52877307257522E-6
san francisco	5.538273337247622E-4	5.477530950510836E-4	6.074238673678554E-6
world war	4.4443649090537027E-4	4.479087967733365E-4	-3.4723058679662084E-6
years ago	4.232104320867491E-4	4.2183609622501784E-4	1.374335861731284E-6
dialog box	4.068341335925972E-4	2.832297250213754E-4	1.2360440857122175E-4
de la	3.993099952899138E-4	4.0811482156959843E-4	-8.804826279684638E-6
high school	3.7183405095419917E-4	3.713363459881821E-4	4.9770496601706E-7
mental health	3.336227193714332E-4	3.226538885632896E-4	1.0968830808143552E-5
supreme court	3.134821450707294E-4	3.206201961664764E-4	-7.138051095747015E-6
nineteenth century	3.1166884591307063E-4	3.141040764103935E-4	-2.435230497322834E-6
human beings	3.098892911758977E-4	2.999964782128753E-4	9.892812963022406E-6
ve got	3.049002938036539E-4	2.887283943076438E-4	1.6171899496010094E-5
north carolina	3.0275110510455465E-4	3.007156246969150 3E-4	2.0354804076396403E-6