CSCE 5543: Homework 1

April Walker
adw027@uark.edu

University of Arkansas,
Fayetteville, AR, 72701, USA

1 Problems from the Textbook

2.1.a - Proving the Addition Rule of Probability

Using axioms and set theory we can prove $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Consider first intuitively: if we find the the probability of either of two events occurring by individually adding up the probability of each event occurring ($P(A) + P(B)$) we have accidentally added the intersection of these events twice, since $(A \cap B) \subset A$ and $(A \cap B) \subset B$. This means we must subtract one of these two intersections, giving us the form above.

A Minimal Proof:

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(B \cap A^c) \qquad \text{(i)}$$
$$P(A) = P(A \cap B^c) + P(A \cap B)$$
$$P(B) = P(B \cap A^c) + P(B \cap A) \qquad \text{(ii)}$$

$$\therefore \quad P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \blacksquare$$

To Expand:

(ii)

$$\begin{aligned} B &= B \cap \Omega \\ &= B \cap (A \cup A^c) \\ &= (B \cap A) \cup (B \cap A^c) \\ \therefore \quad P(B) &= P(B \cap A) + P(B \cap A^c) \end{aligned} \qquad \text{(and similar for } A\text{)}$$

(i)

$$\begin{aligned} A \cup B &= (A \cup B) \cap \Omega \\ &= (A \cap B^c) \cup B \\ &= (A \cap B^c) \cup (B \cap A) \cup (B \cap A^c) \\ \therefore \quad P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(B \cap A^c) \end{aligned}$$

2.3 - Probability of Multiple Events Given Conditional Probability

Given the event "is-abbreviated" is A and event "three-letter-word" is TLW we are given the following information:

$$P(A|TLW) = 0.8$$
$$P(TLW) = 0.0003$$

We also know the following:

$$P(A \cap B) = P(A|B)P(B)$$

Using this equation in the context of our equation we find:

$$P(A \cap TLW) = 0.8 \cdot 0.0003 = 0.00024$$

2.8 - Confirming the Maximum of an Equation

This question references the following equations:

$$P(s|\mu_m) = m^i(1-m)^j \qquad (2.15)$$

$$P(\mu_m|s) = \frac{m^i(1-m)^j \cdot 6m(1-m)}{P(s)} \qquad (2.17)$$

Then we can show the maximum of equation (2.15) occurs at 0.8 by taking the derivative of the log of the equation with respect to m and setting that derivative to zero (that is find the MLE).

$$L(m|\mu_m) = f(s|\mu_m) = m^i(1-m)^j$$

$$l(m|\mu_m) = \log(L(m|\mu_m)) = i\log(m) + j\log(1-m)$$

$$\frac{\delta l}{\delta m} = \frac{i}{m} - \frac{j}{1-m} = 0$$

$$\frac{i}{m} = \frac{j}{1-m}$$

$$\frac{1-m}{m} = \frac{j}{i}$$

$$\hat{m} = \frac{i}{j+i}$$

Thus when $i = 8$ and $j = 2$ we can say:

$$\arg\max_m P(s|\mu_m) = \frac{8}{2+8} = 0.8$$

To find the maximum of equation (2.17) we can simply do the same thing as with (2.15), but only consider the numerator. Since s is given, $P(s)$ is a real numerical value and will not impact the MLE. I will further show this below:

$$L(m|s) = f(\mu_m|s) = \frac{m^i(1-m)^j \cdot 6m(1-m)}{P(s)}$$

$$l(m|s) = i\log(m) + j\log(1-m) + \log(6) + \log(m) + \log(1-m) - \log(P(s))$$

$$\frac{\delta l}{\delta m} = \frac{i}{m} - \frac{j}{1-m} + \frac{1}{m} - \frac{1}{1-m} = 0$$

$$\frac{i+1}{m} = \frac{j+1}{1-m}$$

$$\hat{m} = \frac{i+1}{j+i+2}$$

$$\arg\max_m P(\mu_m|s) = \frac{8+1}{2+8+2} = 0.75$$

Now let us consider instead the following:

$$P(\mu_m) = 30m^2(1-m)^2$$

In the book, we had originally found $P(\mu_m|s)$ using Baye's. Now instead, we must say:

$$P(\mu_m|s) = \frac{P(s|\mu_m)P(\mu_m)}{P(s)}$$

$$P(\mu_m|s) = \frac{m^i(1-m)^j \cdot 30m^2(1-m)^2}{P(s)}$$

Skipping some trivial steps, we can instead say:

$$l(m|s) = i\log(m) + j\log(1-m) + \log(30) + 2\log(m) + 2\log(1-m) - \log(P(s))$$

$$\frac{\delta l}{\delta m} = \frac{i}{m} - \frac{j}{1-m} + \frac{2}{m} - \frac{2}{1-m} = 0$$

$$\frac{i+2}{m} = \frac{j+2}{1-m}$$

$$\hat{m} = \frac{i+2}{j+i+4}$$

$$\arg\max_m P(\mu_m|s) = \frac{8+2}{2+8+4} = \frac{10}{14} \approx 0.7143$$

2 Collocations in Amazon Review Corpus

2.1 Introduction

This assignment explores bigram collocation detection. Overall I found my methodology produced many collocations, however my more minimal processing technique performed poorly under the conditions of the assignment. My choice to use Mutual Information as a method to detect true collocations didn't work incredibly due to it's bais towards rare word occurances, however with some more intense preprocessing higher PMI generally coincided with true collocations when looking at the top 100 bigrams by frequency. In order to utlize the PMI as a method to find the most likely collocation bigrams, more stringent standards on minimum word occurrence would likely be needed.

2.2 Methodology and Results

For this assignment I used Python and the `nltk` library for preprocessing. My initial and more "official" method for doing this assignment involved little preprocessing. The methodology linked on your website was used with a minimal addition to remove words with only one occurrence. Preprocessing took approximately 30 seconds. Once cleaned, our corpus contained 2,764,330 words.

The bigrams were added to a hash-table with the value being their frequency occurrence. The process to find the bigrams took approximately 2.86 seconds. The process found 626,469 unique bigrams.
Looking at bigrams simply by frequency, none of the top 100 were actually collocations (at least as I defined it, I'll later argue on potential one). This was rather expected, since stop words were not removed and part-of-speech was also not considered.

In order to get more fruitful results I used the Pointwise Mutual Information (PMI) method. This methodology is very straightforward to implement, however it has a strong bias towards rarely occurring words. This methodology would thus be more fruitful with a larger corpus or more stringent standards on minimum word occurrence. To meet the standard outlined by the assignment, the word occurrence minimum was not increased.

The PMI for a bigram can be calculated as follows:

$$PMI(w_1, w_2) = \log\left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right)$$

Where $P(w_1, w_2)$ refers to the probability of the bigram occurring (that is the first and second word occuring together), $P(w_1)$ refers to the probability of the first word occurring, and $P(w_2)$ refers to the probability of the second word occurring. This probability can be calculated by counting up the occurances of each bigram or word and then dividing by the total number of words in the corpus N. Thus we can say:

$$PMI(w_1, w_2) = \log\left(\frac{C(w_1, w_2)/N}{(C(w_1)/N)(C(w_2)/N)}\right) = \log\left(\frac{N \cdot C(w_1, w_2)}{(C(w_1))(C(w_2))}\right)$$

Where $C(w)$ refers to the count. This final equation is what was used within my code.

Calculating the PMI for each bigram took about 58 seconds. While the top 100 bigrams by frequency were not collocations, the PMI did seem to point out the words that were more legitimately related rather than simply grouped stop words. Bigrams like "it a", "it to", and "that the" had values incredible close to zero (predicting independence) while other bigrams like "this game", "a few", and "have been" scored higher. With a minor stretch one could even say the highest performing bigram in the top 100 list "have been" with a PMI of 4.01 truly is a collocation as in certain contexts it refers to a star passed their prime.

Sorting my hashtable by PMI, I had somewhat more luck. If I had considered names as collocations, approximately half of the top 100 bigrams via PMI would have met the criteria. Discluding potential names, 16 of the 100 were collocations. Even so, these were more along the lines of titles and locations. Due to the bias of PMI, the new top 100 even picked up some bigrams in other languages.

Partially out of curiosity, I decided to revisit my methodology with better preprocessing. I further used `nltk` to filter out bigrams with stop words and only allow bigrams which had the form "noun noun" or "adjective noun". Rerunning my hash-table through this filter my top 100 by frequency looked much better. 50 of my bigrams were collocations, although most were related to PC parts or game titles and similar. Even so collocations like "video game/s", "roller coaster", "story line", "heavy duty", "air filter", and more were picked up. However, sorting by PMI my results were practically identical to before.

While this methodology clearly outperformed my minimal preprocessing, the total process took approximately 200 seconds to complete. For our corpus this time is negligible, but in some situations with a larger dataset it may be more beneficial to go with a "good enough" solution that avoids a time consuming process.

2.3 Final Statements

Overall, PMI performed somewhat poorly when used as the absolute teller of collocations. However, my methodology still produced many collocations, although the most successfully were found with the help of additional preprocessing. In the future, harsher limits on word rarity would need to be utilized to produce more adequate results, however I felt that was outside the specific directions of this assignment since we could only remove absolutely unique tokens from the corpus. A more hypothesis-based testing method, such as χ^2 might prove more useful.

On the next page, I document the top 100 bigrams found utilizing various methods and the PMI and frequency for each (with the exception of the top 100 from the alternate method sorted by PMI, since it closely mirrors the top 100 sorted bby PMI of the original method). Bigrams I believed to be collocations were indexed with an asterisk (*), but note I generally did not consider names to be collocations although they clearly are closely related.

Top 100 Freq	Freq	PMI	Top 100 (PMI)	PMI	Top 100 Alt (Freq)	Freq	PMI	Top 100 Alt (PMI)
of the	16294	1.565842423	everett vinsonsan	14.13916167	*final fantasy	927	7.193362432	everett vinsonsan
in the	10234	1.556412021	vinsonsan antonio	14.13916167	great game	722	2.037976488	monte carlo
this game	9247	3.08766059	*monte carlo	14.13916167	long time	511	4.144542436	kinetic hc
is a	6747	1.644964011	somerset maugham	14.13916167	first time	483	3.406578135	hugh lofting
the game	6632	1.492655642	garcia marquez	14.13916167	best game	466	2.306247684	somerset maugham
it is	6575	1.744725273	kay summersby	14.13916167	*resident evil	459	7.953388626	garcia marquez
and the	6422	0.2836021156	galen rowell	14.13916167	*video game	423	3.876570366	kay summersby
this book	6104	3.209339694	grandi isole	14.13916167	*story line	397	5.209118393	galen rowell
to the	6087	0.3744110261	croque monsieur	14.13916167	circus life	378	4.66027574	grandi isole
if you	6026	3.776546108	*taj mahal	14.13916167	*nursing home	374	7.691745538	croque monsieur
this is	6016	2.020786324	*dalai lama	14.13916167	great book	350	1.85080357	taj mahal
i have	5637	2.381678826	anastasia romanov	14.13916167	*main character	348	5.954646808	dalai lama
on the	5237	1.552295073	algun modo	14.13916167	old man	347	5.302104352	anastasia romanov
for the	5101	1.069919784	tholly tregolis	14.13916167	next book	345	3.561801079	algun modo
it was	4767	2.216124531	tregolis jud	14.13916167	sara gruen	340	7.986746562	tregolis jud
and i	4569	0.698947442	jud trudy	14.13916167	*game play	312	1.885682438	jud trudy
i was	4446	2.00449358	trudy paynter	14.13916167	*game boy	303	4.118055185	trudy paynter
with the	4283	1.24625658	horace verity	14.13916167	great product	295	3.254759743	horace verity
is the	4068	0.3669030243	marjorie morningstar	14.13916167	great story	283	2.404660432	marjorie morningstar
to be	3989	2.497234078	bearl barbor	14.13916167	*super mario	270	6.693911911	bearl barbor
the best	3924	2.639809151	*iwo jima	14.13916167	great read	261	2.256519551	iwo jima
one of	3881	2.583993815	valles marineris	14.13916167	good game	259	1.223100511	areo hotah
you can	3805	3.367861087	areo hotah	14.13916167	*star wars	245	7.643676176	mance rayder
the book	3685	1.442053735	mance rayder	14.13916167	fun game	242	1.951471408	arkks zzzts
and it	3494	0.5726188808	arkks zzzts	14.13916167	*video games	241	4.708444	zzzts dizarss
game is	3456	2.003260788	zzzts dizarss	14.13916167	*street fighter	230	8.382954975	dizarss razakss
in a	3310	1.199759203	dizarss razakss	14.13916167	fantasy vii	220	7.37817317	tibetian foothills
the story	3191	2.064460675	*tibetian foothills	14.13916167	good book	219	1.592263246	banjoe kazooi
for a	3035	1.322809112	banjoe kazooi	14.13916167	robert jordan	219	8.279736526	sadf asdf
a great	2919	2.409741276	sadf asdf	14.13916167	*tomb raider	218	9.304225767	asdf asf
the first	2784	2.17900721	asdf asf	14.13916167	*character development	214	6.418239359	liu kang
that i	2753	1.208204722	liu kang	14.13916167	*single player	212	6.770956242	quan chi
from the	2691	1.643163608	quan chi	14.13916167	many times	209	4.189521422	nitrus brio
all the	2656	1.362143693	tani loje	14.13916167	*memory card	207	7.69978294	tani loje
i would	2636	2.562614613	mooney tls	14.13916167	good read	199	2.195625069	mooney tls
have to	2592	1.615879879	spongebob squarepants	14.13916167	much fun	198	3.318558892	friederike knabe
you have	2588	2.3146011	jules verne	14.13916167	many years	194	3.616792405	timothy zahn
with a	2577	1.510343316	friederike knabe	13.73369656	jacob jankowski	192	7.22481689	swan slimline
i had	2543	2.613479983	snidely whiplash	13.73369656	replay value	192	8.184297146	bagmitten racket
to get	2539	2.474961461	khaled hosseini	13.73369656	many people	189	3.687481601	debo decir
but i	2494	1.555969723	valar morghulis	13.73369656	*roller coaster	189	8.920023366	kwik tek
i am	2462	3.461569833	timothy zahn	13.73369656	k n	186	9.14978035	lao tzu
was a	2430	1.416713966	helly hansen	13.73369656	*main characters	183	4.498735258	poppy eyebright
of a	2416	0.4292732646	*bagmitten racket	13.73369656	george martin	180	6.776877173	christoph waltz
at the	2378	1.495548853	debo decir	13.73369656	quot quot	176	3.946100041	mitch albom
when i	2259	2.348604718	kwik tek	13.73369656	*sonic adventure	168	6.983033015	gail cooke
but it	2243	1.591814721	*kinetic hc	13.73369656	best books	166	3.358406904	danielle steele
a lot	2196	3.563466928	lao tzu	13.73369656	little bit	163	4.425352851	eladio andres
there are	2154	3.752027298	hugh lofting	13.73369656	*tom clancy	163	8.284196454	alta cocina
as a	2142	1.516878649	poppy eyebright	13.73369656	*red mars	161	6.759476118	arabian peninsula
that the	2057	0.1609815468	christoph waltz	13.73369656	good quality	159	3.831750773	ronald reagan
a good	2047	2.265273649	mitch albom	13.73369656	young man	158	5.833982402	bella poldarki
there is	1994	2.642574761	gail cooke	13.73369656	*civil war	158	7.55491051	caitlin kiernan
i do	1950	2.041394812	danielle steele	13.73369656	great job	157	3.736027002	clive barker
a little	1950	3.079415307	eladio andres	13.73369656	*bell tolls	157	9.580414535	sookie stackhouse
in this	1946	1.159084686	alta cocina	13.73369656	*soul calibur	156	8.996455962	valar morghulis
have a	1915	1.318378383	ronald reagan	13.73369656	several times	154	5.215051048	irc tibick
the characters	1913	1.802342145	bella poldarki	13.73369656	good story	151	1.986814472	peirce brosnan
you are	1881	2.094993192	caitlin kiernan	13.73369656	great price	146	3.233519661	whitewater rapids
to play	1867	2.644551517	clive barker	13.73369656	john clark	139	8.070393953	seung mina
the same	1825	2.738434408	sookie stackhouse	13.73369656	*great depression	138	4.368899982	catfish maw
it i	1772	0.02694039455	irc tibick	13.73369656	*depression era	137	7.668594427	stemme motorglider
a few	1768	3.339918732	peirce brosnan	13.73369656	*expansion pack	137	7.904062969	nar shaddaa
of this	1763	0.6046745763	*whitewater rapids	13.73369656	*crash bandicoot	137	8.606153117	ingrid bergman
and a	1740	-0.2501278639	nitrus brio	13.73369656	good product	135	2.683380266	mack bolan
in my	1729	1.986766946	seung mina	13.73369656	*battle system	134	5.369231333	khal drogo
i could	1727	2.645387061	*catfish maw	13.73369656	*heavy duty	132	8.075047309	barkhang monastery
that it	1712	0.8750943631	stemme motorglider	13.73369656	new characters	130	2.887565049	saya sangat
this one	1711	2.053897501	nar shaddaa	13.73369656	*world war	129	4.764935614	saya beli
i did	1708	2.526778526	ingrid bergman	13.44601449	*mario kart	127	7.78997301	dutton lainson
it a	1706	0.005316934977	mack bolan	13.44601449	*story lines	126	4.927498724	woo hoo
out of	1704	2.24449311	khal drogo	13.44601449	*theme park	124	7.395335016	costa rica
is not	1703	1.75250641	janos slynt	13.44601449	best part	123	3.819600264	alan bunkel
i bought	1702	3.107790136	barkhang monastery	13.44601449	great deal	121	4.139318986	kurt cobain
it to	1697	-0.00518318795	swan slimline	13.44601449	*sound effects	120	7.151636179	andre marty
so i	1681	1.613434013	saya sangat	13.44601449	*book club	118	4.147593182	genghis khan
as the	1669	0.4952488187	saya beli	13.44601449	*benzini brothers	118	9.078689697	puedo jugar
want to	1655	3.1198772	desde hace	13.44601449	boy color	118	7.128244741	raytheon beechjet
to read	1655	2.107616789	dutton lainson	13.44601449	great graphics	117	2.265301099	graphite interiors

you will	1639	2.664339676		cecil demille	13.44601449		good condition	116	4.288920684	rican rico
easy to	1602	3.040891493		*costa rica	13.44601449		favorite game	116	2.307056475	upc barcode
the graphics	1575	2.10229523		alan bunkel	13.44601449		*sega dreamcast	115	5.834828912	alluminum sleve
of my	1555	1.425048256		kurt cobain	13.44601449		*easy read	114	2.664345485	conectado todo
but the	1520	0.3050116509		andre marty	13.44601449		*science fiction	113	7.647127145	cotek cotek
about the	1509	1.205766539		lieutenant berrendo	13.44601449		*coaster tycoon	112	8.865419232	sangat ini
i can	1507	1.730277109		genghis khan	13.44601449		*real life	111	4.112300008	beli ini
game i	1504	0.7646580092		zacky tholly	13.44601449		*time period	111	4.651602568	gh gh
that you	1489	1.305007155		*frappe snowland	13.44601449		*circus train	110	5.400062919	gj kg
into the	1485	1.805817317		*raytheon beechjet	13.44601449		*good job	109	3.581450179	excelentes excelentes
it has	1483	2.01382956		*puerto rican	13.44601449		good thing	108	2.826590692	crispy peking
lot of	1483	3.372475612		*upc barcode	13.44601449		*donkey kong	108	9.908044719	womp womp
on a	1482	1.062053288		*vroom vroom	13.44601449		*air filter	107	6.434288951	impreza sti
on my	1467	2.488292985		*alluminum sleve	13.44601449		stop reading	107	4.97382993	ba kup
have been	1446	4.006526633		conectado todo	13.44601449		several years	106	4.343284616	unchecked hedonism
and you	1431	0.249415486		cotek cotek	13.44601449		many characters	105	2.52471282	mawkish sentimentalism
for my	1414	1.995432981		sangat ini	13.44601449		great condition	102	3.949982206	ecclesiastes ecclesiastes
the most	1413	1.901957924		beli ini	13.44601449		awesome game	102	2.285622767	hustle bustle
they are	1410	3.247473752		gh gh	13.44601449		great games	102	1.475373042	ralph hammy
book i	1392	1.22430533		elisabeth welch	13.44601449		*high quality	100	5.583031831	urchin pompey
the circus	1392	2.10884289		excelentes excelentes	13.44601449		*hard time	100	2.798876875	psuedonym lesieg