Birju Patel

Dr. Michael Yoder

September 17, 2023

Homework 1 Report

1.3.

“juliet”

Term-Document Matrix

Rank	Synonym	Similarity Score
1	romeo	0.9899494936611666
2	capulet	0.9899494936611665
3	pump	0.9899494936611665
4	laura	0.9899494936611665
5	pitcher	0.9899494936611665
6	behoveful	0.9899494936611665
7	hurdle	0.9899494936611665
8	capulets	0.9899494936611665
9	petrucio	0.9899494936611665
10	heartless	0.9899494936611665

Term-Context Matrix

Rank	Synonym	Similarity Score
1	lucius	0.787796461449417
2	gloucester	0.7818803163235146
3	servants	0.7717159769405332
4	warwick	0.7701625274368058
5	nurse	0.7590373891930818
6	antonio	0.7531868738827995
7	paris	0.7519236427794577
8	brutus	0.7510764834961863
9	clifford	0.7494451562895195
10	buckingham	0.7492511026378086

“war”

Term-Document Matrix

Rank	Synonym	Similarity Score
1	our	0.8761797978447218
2	retire	0.8744977289562926
3	field	0.8559274552577374
4	blood	0.8453736038474777

5	wars	0.840769575013834
6	their	0.8322435799410592
7	revolt	0.8317673637529124
8	we	0.8288559072294951
9	drums	0.8267726515762036
10	arm	0.8253905634380312

Term-Context Matrix

Rank	Synonym	Similarity Score
1	france	0.9186011346212705
2	state	0.9104586835320344
3	course	0.9084921305841427
4	kings	0.9047033596576401
5	justice	0.8978439420834826
6	nature	0.8932543791781989
7	manner	0.8907479583045452
8	taste	0.8906995444907473
9	glory	0.8873752187730591
10	blood	0.8824951602046457

“soldier”

Term-Document Matrix

Rank	Synonym	Similarity Score
1	wars	0.8674692400629195
2	camp	0.8668600384816458
3	preparation	0.8413565752388104
4	edges	0.8227270096133007
5	swerving	0.8126053915961521
6	pleach	0.8126053915961521
7	prescript	0.8126053915961521
8	undid	0.8126053915961521
9	squares	0.8126053915961521
10	meeter	0.8126053915961521

Term-Context Matrix

Rank	Synonym	Similarity Score
1	fool	0.9332786557710041
2	dog	0.9280552851631676
3	gentleman	0.9225195474718277
4	man	0.9214376105340338
5	woman	0.9206079338228043
6	like	0.9179819895250835
7	traitor	0.9150058772096371
8	beggar	0.9132366078890773

9	beast	0.8947914853463569
10	little	0.8946892204601253

“king”

Term-Document Matrix

Rank	Synonym	Similarity Score
1	sovereign	0.8723889898823937
2	title	0.8722882623156409
3	seat	0.8624239750266073
4	subject	0.86240703404077
5	kingdom	0.8589420110465151
6	royal	0.8581872075414547
7	lords	0.8475107931059153
8	london	0.8386619739210417
9	liege	0.8382979584256983
10	sceptre	0.8335583390738851

Term-Context Matrix

Rank	Synonym	Similarity Score
1	people	0.9424616904127386
2	french	0.9359298167882195
3	devil	0.9347399890373561
4	queen	0.9332422031283539
5	dauphin	0.9318944640823045
6	prince	0.9314062641484058
7	of	0.9276054362264482
8	world	0.9268048810933017
9	duke	0.9213327188835964
10	next	0.9210089124576452

“jester”

Term-Document Matrix

Rank	Synonym	Similarity Score
1	bestowed	0.9354143466934853
2	europa	0.8451542547285167
3	congregation	0.8451542547285167
4	sexton	0.806946584785929
5	watchman	0.8017837257372731
6	dreamed	0.8017837257372731
7	fencer	0.8017837257372731
8	frugal	0.8017837257372731

9	panders	0.8017837257372731
10	negotiate	0.8017837257372731

Term-Context Matrix

Rank	Synonym	Similarity Score
1	devil	0.7153285434259797
2	emperor	0.7003684674445881
3	dauphin	0.6969992683520221
4	palace	0.6934805019581182
5	count	0.6932170174324136
6	jew	0.693065532733913
7	deer	0.6907492665935286
8	king	0.689529495003695
9	lion	0.6854813942357647
10	matter	0.6854448792634026

In our term-document matrix, we have rows of word vectors that are 36 items long. This corresponds to the 36 Shakespeare plays in our corpus. Shakespeare wrote across a variety of genres and explored many themes, so his plays give us a broad overview of the English language. I think this is enough information for us to get an idea of the sentiment and mood that each word conveys.

Consider the results produced for the word “war”. The term-document matrix provides us with synonyms that have the same sentiment as the target word. These are words which evoke the same mood as the word “war” and intuitively belong together, such as “blood” and “drums”. By contrast, the term-context matrix provides us with words that would be used in a similar context. So, for instance, one would declare “war” on a “state”, and the war would be prosecuted by a “king” in order to achieve “justice”. The term-context matrix provides more information on word usage, but the term-document matrix provides more intuitive results. However, in the case of “jester”, the term-document matrix provided synonyms that made no sense, while the term-context matrix provided words that did fit. This is likely because “jester” appears with a very low frequency and appears across many plays with varied themes.

When I expanded the window size from 4 to 5 and ran the program again, I saw no change to the results provided by the term-context matrix. However, when I shrank the window from 4 to 1, the results shifted. I began to see words that had no connection to “war” appear on the list with high similarity scores. This is likely because with a small window size, there was not enough context for the algorithm to provide meaningful results. I did not see “war” on the list of results, which is to be expected since I assumed that a word cannot co-occur with itself.

1.4

“juliet”

tf-idf Matrix

Rank	Synonym	Similarity Score
1	procures	0.9611235625180706
2	benedicite	0.9611235625180706
3	ghostly	0.9380710616293596
4	capulet	0.8748623439660793
5	pump	0.8748623439660791
6	laura	0.8748623439660791
7	pitcher	0.8748623439660791
8	behoveful	0.8748623439660791
9	hurdle	0.8748623439660791
10	capulets	0.8748623439660791

PPMI Matrix

Rank	Synonym	Similarity Score
1	capulet	0.17047704421722754
2	vauntingly	0.149494730192251
3	barnardine	0.1408585432561904
4	provost	0.13702927994670755
5	montague	0.13489512115273272
6	mercutio	0.12996388934657044
7	stricken	0.12335620337836761
8	tybalt	0.12191570453282441
9	ruminat	0.11756596255144802
10	katarina	0.11683190214071515

“war”

tf-idf Matrix

Rank	Synonym	Similarity Score
1	field	0.9211833638630113
2	wars	0.9190362070287909
3	sword	0.9042630641102515
4	strike	0.9035892578236605
5	ours	0.9016379812546453
6	strong	0.9006737071706502
7	noble	0.8991704765725292
8	proud	0.899139417499278
9	revolt	0.8980745001906689
10	yield	0.894789216123648

PPMI Matrix

Rank	Synonym	Similarity Score
1	investing	0.09106280488564122

2	wars	0.08995711364966252
3	unexecuted	0.0881596537655478
4	curlish	0.07788555514928075
5	muskets	0.07762864190869034
6	of	0.07727323370293915
7	their	0.07557572271959145
8	butcherly	0.07486202623148497
9	scarecrow	0.07421746030624776
10	unscarr	0.07411446631527918

“soldier”

tf-idf Matrix

Rank	Synonym	Similarity Score
1	sword	0.887730140001851
2	war	0.87868305223274
3	valiant	0.8764477017737677
4	spoke	0.8717738631100879
5	send	0.8708232193388834
6	given	0.8642505253634335
7	already	0.8635562798737787
8	seem	0.8632476814547339
9	worthy	0.8630801037927566
10	itself	0.8628378618292695

PPMI Matrix

Rank	Synonym	Similarity Score
1	skitless	0.16437252281841086
2	statesman	0.12764143573032238
3	pursueth	0.12257132859943476
4	linguist	0.110557179494151
5	flask	0.10213972627787726
6	abler	0.09261114471154086
7	lot	0.08623277592119938
8	van	0.0856839297329014
9	scour	0.08495921365175152
10	scholar	0.08482930549772982

“king”

tf-idf Matrix

Rank	Synonym	Similarity Score
1	royal	0.9526816549204372

2	crown	0.9366026731979491
3	majesty	0.9348720224515122
4	arms	0.9289309924356932
5	queen	0.9196828330828575
6	gracious	0.9115408052441484
7	lords	0.9110150653223522
8	sovereign	0.9091543179938315
9	highness	0.9075962825555225
10	days	0.9044861479512178

PPMI Matrix

Rank	Synonym	Similarity Score
1	henry	0.22008680463253716
2	richard	0.15522462986780128
3	lewis	0.15010934134465592
4	queen	0.1435288245979196
5	enter	0.1409576109324997
6	edward	0.14081458763350585
7	the	0.13346405138231454
8	attendants	0.132619808808069
9	flourish	0.13234282556761956
10	viii	0.12350227085489662

“jester”

tf-idf Matrix

Rank	Synonym	Similarity Score
1	bestowed	0.9119005598537532
2	europa	0.7982244569370549
3	congregation	0.7982244569370549
4	blazon	0.7972625740429038
5	watchman	0.7785391634049672
6	dreamed	0.7785391634049672
7	fencer	0.7785391634049672
8	frugal	0.7785391634049672
9	panders	0.7785391634049672
10	negotiate	0.7785391634049672

PPMI Matrix

Rank	Synonym	Similarity Score
1	yorick	0.2938451595960623
2	proposing	0.23698137961300225
3	benign	0.2349582242990229
4	bookmates	0.234954713776372
5	misdoubts	0.23083717526448666

6	unmindful	0.22840987142690206
7	calydon	0.22301553806284768
8	fleming	0.22130654176191233
9	trothed	0.21720709360708024
10	stupefy	0.21710521281013673

The tf-idf matrix produces than the term-document matrix because it reduces the weight of filler words. For instance, when finding synonyms for “war”, the tf-idf matrix gave the filler word “ours” a rank of 4, while the term-document matrix gave it a rank of 1. However, the PPMI matrix does not remove these filler words. For instance, it includes “of” and “their” as synonyms for “war”.

The PPMI matrix uses Bayesian statistics to find synonyms, giving the word that has the highest probability of co-occurring with the target word, in comparison with the probability of it just randomly occurring. This is seen in its output when given the target “king”. Since “king” is a title, it is likely to appear next to names like “henry” and “richard". However, the term-context matrix does not use probability, it uses the shape of the co-occurrence vector to determine similarity. Therefore, it will miss this connection, because names are not used in the same context as titles.

2

“lesbian”

PPMI Matrix

Rank	Synonym	Similarity Score
1	embracing	0.2263154120184383
2	newlywed	0.19155620187715394
3	pansexual	0.18327637671635066
4	sharing	0.1725244302557183
5	romantic	0.13046849157498241
6	loving	0.12995486715799986
7	shares	0.1266324798842623
8	share	0.12312443885150082
9	meet	0.12002616189161075
10	date	0.11913755475609156

“gay”

PPMI Matrix

Rank	Synonym	Similarity Score
1	pride	0.27399783533814526

2	marriage	0.22845142680721686
3	veterans	0.22375743690827987
4	protesters	0.2101103218181274
5	rights	0.20687007344067143
6	protest	0.20674857609073727
7	protesting	0.20024344319426834
8	couple	0.17424954393159486
9	parade	0.17312953843473522
10	protester	0.16813920647239533

“black”

PPMI Matrix

Rank	Synonym	Similarity Score
1	white	0.5918175563989874
2	wearing	0.5163888472364351
3	and	0.48927315593517695
4	blue	0.4775602302469326
5	red	0.47406336909263735
6	brown	0.4480652363531139
7	gray	0.41616032120479773
8	shirt	0.39757811280328925
9	purple	0.3952237113643676
10	pink	0.3931365332150112

“white”

PPMI Matrix

Rank	Synonym	Similarity Score
1	black	0.5918175563989874
2	red	0.5118148118585463
3	blue	0.5087344293471318
4	wearing	0.4765089005245807
5	and	0.4730161306646712
6	brown	0.43085362099101476
7	pink	0.4209114269728297
8	yellow	0.41531422322377265
9	gray	0.3980688773440444
10	purple	0.3931080261907618

“latino”

PPMI Matrix

Rank	Synonym	Similarity Score
1	indians	0.13833990115780614
2	kimonos	0.12765918002644638
3	entertain	0.12087925239090025
4	feminine	0.11922581621281736
5	lasso	0.11273609847956312
6	bulls	0.11079909060866888
7	nicely	0.11015190433451183
8	elementary	0.10942732502207964
9	african	0.10727128906701511
10	steer	0.10549564692324231

“lesbian”

- A woman shares her umbrella with her lesbian lover. (first order similarity)
- Seventeen pansexual lovers cavort beneath a gigantic coverlet. The women are lesbian lovers. (second order similarity)
- Two women embracing. The two young women are lesbian. (second order similarity)
- A newlywed couple sharing a kiss under a structure. A lesbian couple kiss while walking down a busy street. (second order similarity)

“gay”

- The retired gay couple are dressed. (first order similarity)
- Some men are marching in a gay pride parade. (first order similarity)
- Two men are protesting the ruling on gay marriages. (first order similarity)

The word “lesbian” appears in a vastly different context than the word “gay” in this corpus. “Gay” is used usually relating to men, as shown in the above examples. The word is also used in the context of the politics of gay rights. The word “lesbian” is naturally used to talk about women, but the sentences convey a different theme, often talking about love and marriage. For instance, it is used in the same context as the word “newlyweds”, with both of these sentences are depicting scenes of romantic love. The usage reflects heteronormative and patriarchal values, which assume that women are meant to value love and romance, and that politics is the domain of men. It is possible that representational harm is being done here. For instance, the importance of lesbian activists to the success of the gay rights movement is being erased, since most of the examples in the corpus involving gay pride protests mention only men.

“black”

- An older man dressed in black and blue is working alone at a printing press. (first order context)
- A surfer is wearing a black shirt. (first order context)

“white”

- A man is wearing a white shirt with the band's name on it as he crowd surfs. (first order context)
- A man in a white suit and no helmet is on a yellow dirt bike. (first order context)

The words “black” and “white” are mainly used in the corpus to refer to the color of objects. Since this usage is neutral, there is minimal risk of representational harm.