Wordwordword Analysis Report

By Liang Tao

Orginal Data: *wordwordword.xlsx*

Main analysis steps:

1. Calculate each tweet sentiment by using AFINN-111.txt
2. Calculate the total term for each words
3. Calculate every keyword sentiment (sum the sentiment of each tweet that includes the keyword)
4. Filter the keyword by use the function(|keyword_sentiment| >= 100), so main attention can focus on the important keywords
5. Next, we are trying to find out why the keywords are important. For each important keyword, calculate the words and also times also are in the tweets at the same times
6. Import the result to Gephi to find out the community
7. Divide the whole important keywords into communities and analysis respectively.

Here are some interesting results.

The terms of some keywords

(Whole table please see: Healthcare Project/Wordwordword_analysis /Result/ 2_count_term.txt)

doctor 18898	choice 1021	mom 594	childrens 442
hospital 13526	death 940	sticker 593	call 432
medical 6807	days 915	unlocked 592	shit 430
abortion 2335	lol 829	countdown 590	care 430
life 1831	angelina 779	kermit 590	health 426
doctors 1324	season 765	people 587	bad 419
prison 1202	finale 723	jolie 574	office 414
convicted 1194	ca 716	told 557	gonna 411
day 1194	#getglue 715	baby 544	getting 406
time 1048	via 696	news 544	help 400
eye 475	gosnell 670	penalty 519	hes 397
dr 455	love 665	thats 516	waiting 386
	watching 648	hate 512	feel 446
	tomorrow 638	school 494	center 443
	im 632	ill 494	
	home 620	pa 493	
	philadelphia 604	please 488	

The sentiment of some keywords

(Whole table please see: Healthcare Project/Wordwordword_analysis /Result/ 3_theme_sentiment.xlsx)

doctor -12501	choice 912	mom -347	childrens 404
hospital -4123	death -4760	kermit -2652	call -181
medical 1194	days -166	people -434	shit -2721
abortion -9375	lol 2440	jolie 712	care 969
life -4660	angelina 829	told -258	health 183
doctors -636	ca -403	baby -228	bad -1674
prison -3874	via -775	news -737	office -242
convicted -4546	gosnell -2980	penalty -2840	gonna -352
day 125	love 2682	thats -177	getting -105
time -319	tomorrow -136	hate -2423	help 915
eye -252	im -604	ill -1290	waiting -267
dr -371	philadelphia -2411	pa -1418	feel -338
		please 840	center 148

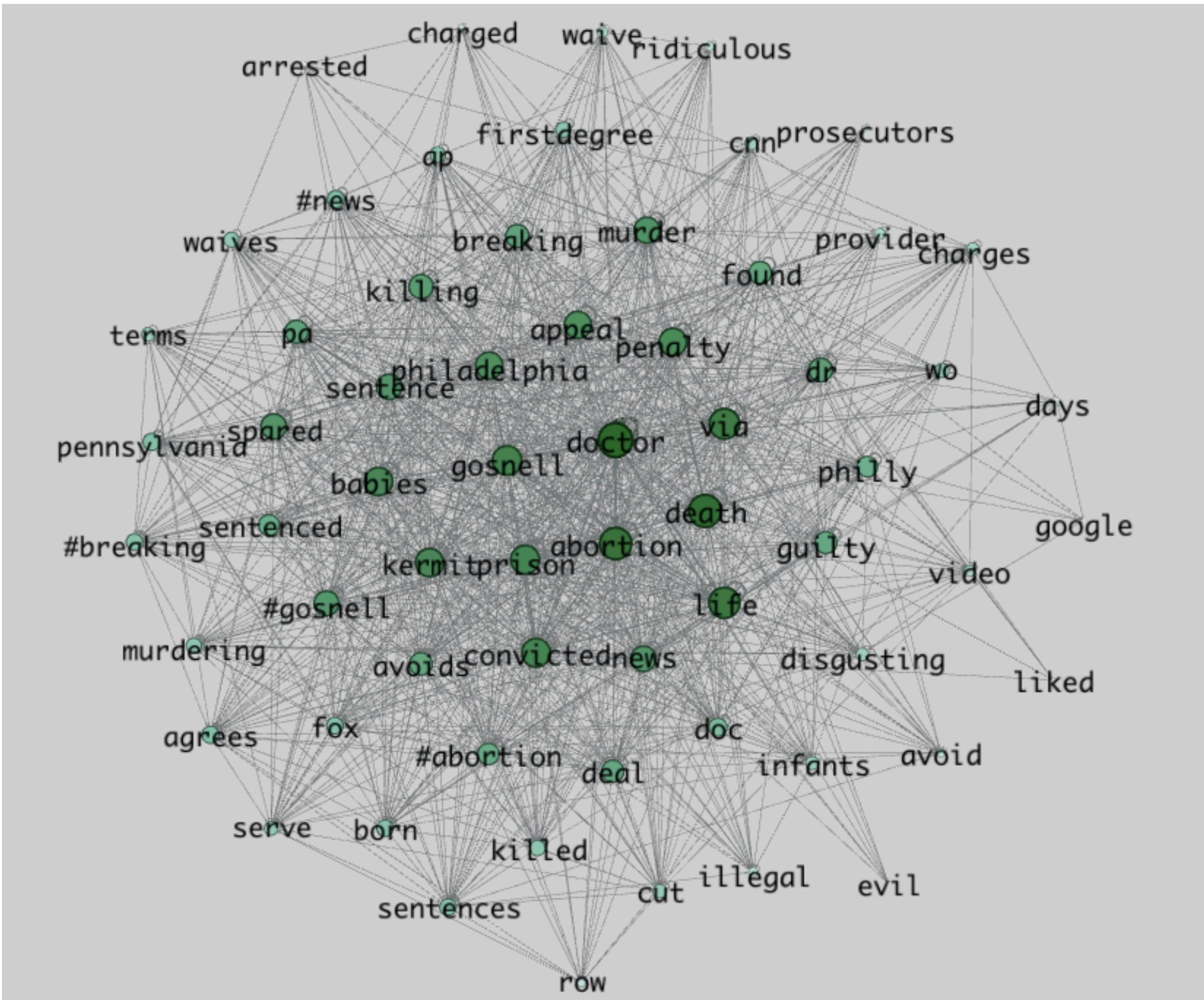
The concurrence of some keywords

(Whole table please see: Healthcare Project/Wordwordword_analysis /Result/ 5_concurrence_keywords.csv)

doctor follow 67	doctor leave 57	doctor makes 134	doctor sore 87
doctor hate 278	doctor team 74	doctor thats 275	doctor broke 23
doctor row 45	doctor brave 7	doctor teenage 5	doctor baby 256
doctor tv 143	doctor tired 32	doctor via 356	doctor glad 31
doctor disgusting 14	doctor sigh 15	doctor love 381	doctor worst 38
doctor #angelinajolie 6	doctor video 181	doctor win 39	doctor join 15
doctor tf 21	doctor giving 42	doctor fake 11	doctor niggas 30
doctor sorry 85	doctor waiting 169	doctor eye 451	doctor car 23
doctor worth 15	doctor suspect 2	doctor positive 17	doctor sucks 30
doctor sent 31	doctor appeal 221	doctor hospital 87	doctor beautiful 39
doctor risk 11	doctor told 426	doctor live 108	doctor crazy 56
doctor fat 30	doctor angelina 53	doctor call 311	doctor embarrassing 63
doctor choice 31	doctor sexy 24	doctor calm 27	doctor awesome 47
doctor fan 45	doctor strong 10	doctor #ologycb 23	doctor share 12
doctor awful 17	doctor prosecutors 57	doctor injury 21	doctor fuck 178
doctor courageous 1	doctor healthy 41	doctor tell 243	doctor sought 3
doctor condition 16	doctor study 121	doctor holy 19	doctor rape 5
doctor cool 56	doctor guilty 139	doctor haha 131	doctor native 129
doctor die 66	doctor jamb 87	doctor award 4	doctor awkward 34
doctor shock 15	doctor weird 50	doctor hurt 55	doctor breaking 57

One of the communities

(Whole figures please see: Healthcare Project/Wordwordword_analysis /Result/ 7_community_figure)



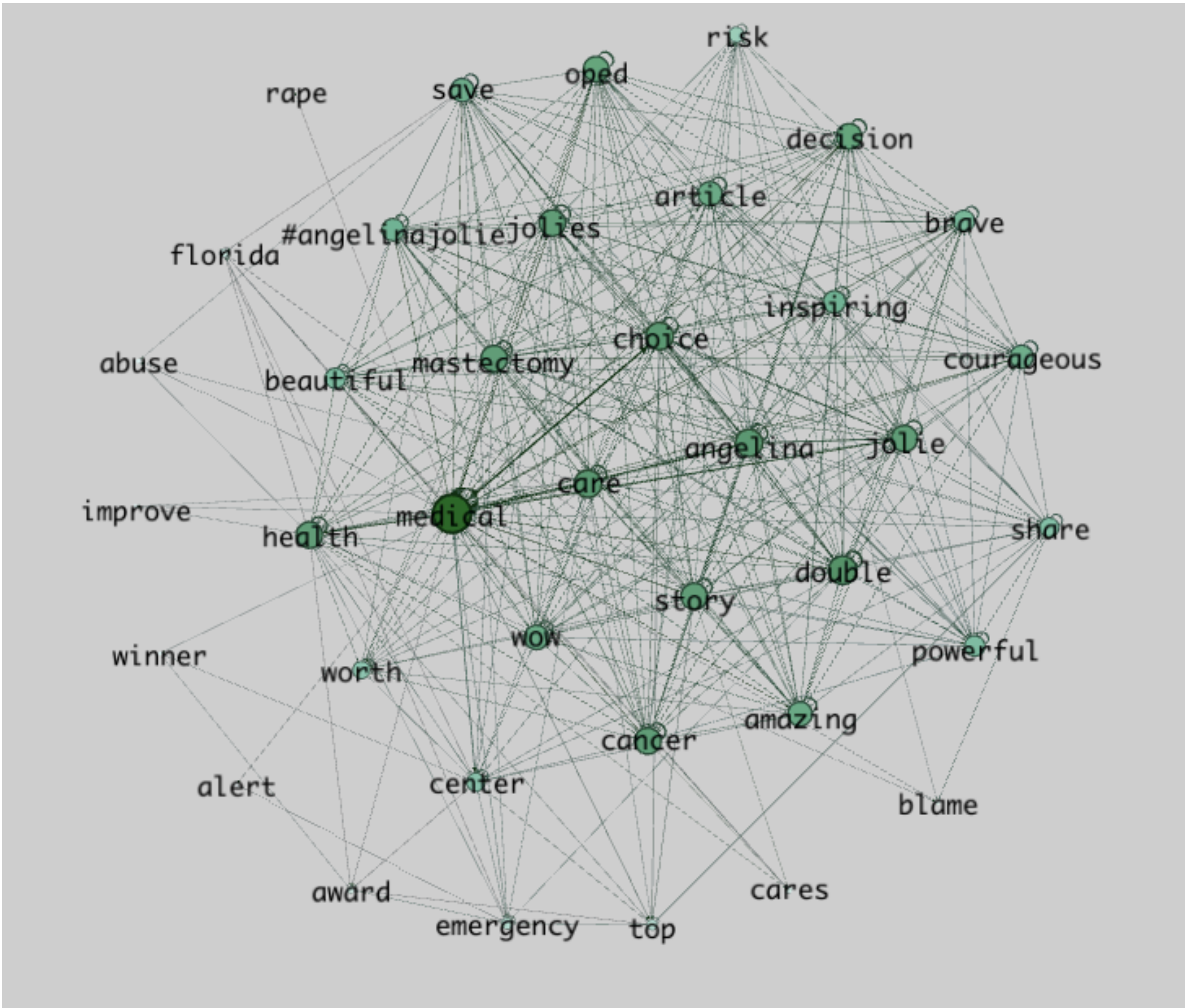
The bigger the node, the larger the degree.

If node “doctor” is larger than node “google”, it means that the tweets contain “doctor ” have more other keywords than the tweets contain “google”.

The thicker the edge, the higher the concurrence.

If edge “doctor life” is larger than node “google life”, it means that there are more tweets contain “doctor” and “life” than the tweets contain “google” and “life”.

Another community



Some high concurrence keywords

(Whole table please see: Healthcare Project/Wordwordword_analysis /Result/ Concurrent_words.csv)

Keyword1	Keyword2	Times	Percentage in Keyword1	Percentage in Keyword2
doctor	hate	278	0.01	0.54
doctor	tv	143	0.01	0.74
doctor	video	181	0.01	0.51
doctor	waiting	169	0.01	0.44
doctor	told	426	0.02	0.76
doctor	study	121	0.01	0.52
doctor	guilty	139	0.01	**0.89**
doctor	makes	134	0.01	0.59
doctor	thats	275	0.01	0.53
doctor	via	356	0.02	0.51
doctor	love	381	0.02	0.57
doctor	eye	451	0.02	**0.95**
doctor	live	108	0.01	0.56
doctor	tell	243	0.01	0.65
doctor	haha	131	0.01	0.49
doctor	fuck	178	0.01	0.5
doctor	native	129	0.01	0.92
doctor	wrong	120	0.01	0.55
doctor	help	139	0.01	0.35
doctor	office	122	0.01	0.29
doctor	qualified	124	0.01	**0.89**
doctor	found	155	0.01	0.57
doctor	yeah	138	0.01	0.55
doctor	god	106	0.01	0.45
doctor	feel	240	0.01	0.54
doctor	tomorrow	261	0.01	0.41
doctor	little	147	0.01	0.41
doctor	murder	332	0.02	0.93
doctor	doctors	152	0.01	0.11
doctor	sentenced	183	0.01	**1**
doctor	play	190	0.01	0.78
doctor	knee	116	0.01	0.74
doctor	medical	112	0.01	0.02
doctor	left	124	0.01	0.37
doctor	penalty	490	0.03	**0.94**
doctor	hope	129	0.01	0.34
doctor	stop	116	0.01	0.52
doctor	dr	259	0.01	0.57

doctor	kermit	589	0.03	**1**
doctor	please	148	0.01	0.3
doctor	wo	137	0.01	0.62
doctor	killing	299	0.02	**1.07**
doctor	spared	275	0.01	**0.99**
doctor	news	366	0.02	0.67
doctor	ill	214	0.01	0.43
doctor	pa	532	0.03	**1.08**
doctor	life	1636	0.09	0.89
doctor	look	159	0.01	0.55
doctor	im	328	0.02	0.52
doctor	id	122	0.01	0.55
doctor	pennsylvania	212	0.01	**1.09**
doctor	mom	233	0.01	0.39
doctor	sentence	251	0.01	**1.01**
doctor	prison	1232	0.07	**1.02**
doctor	real	112	0.01	0.54
doctor	gonna	184	0.01	0.45
doctor	people	220	0.01	0.37
doctor	fucking	137	0.01	0.44
doctor	gosnell	635	0.03	0.95
doctor	fox	101	0.01	0.89
doctor	getting	190	0.01	0.47
doctor	wanna	151	0.01	0.48
doctor	seeing	134	0.01	0.61
doctor	whats	110	0.01	0.53
doctor	philadelphia	698	0.04	**1.16**

Explanation to why some percentage is larger than one:

In my code, I count the times every two words are in the same tweet. For example, if one tweet is like this:

" A likes B but B dislikes A"

Here I count <A, B> four times in this tweet. But there are only two A. If I divide <A,B>/A, the result is 2. That's why some value of percentage is larger to 1.