ニコニコデータセットの分析

原田玲央

と

バイクでのログを用いた分析

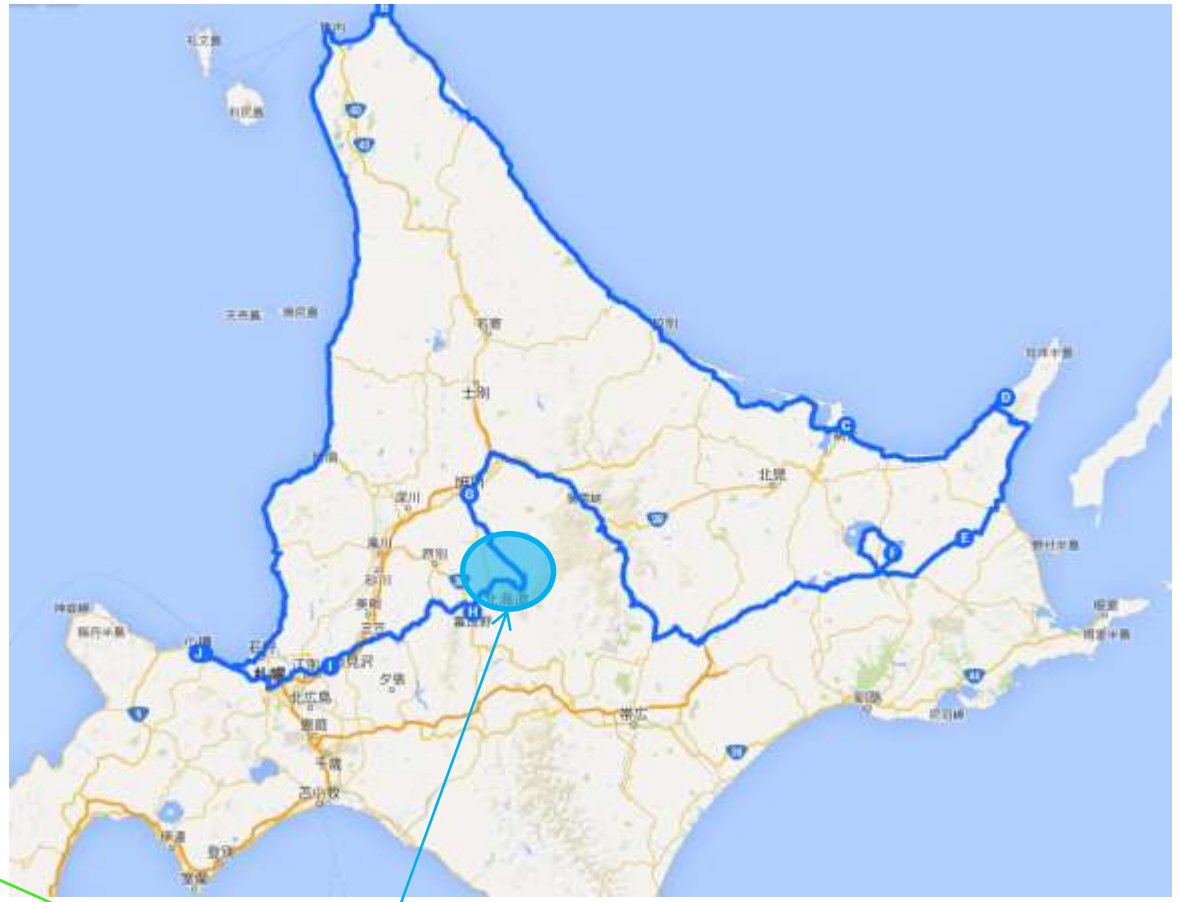
北海道行ってきました！！

せっかくだしログでもとれないかな、、、

このアプリ使えそう、、、

x,y,z軸に対する加速度を取得できて、csvファイルとして出力してくれる

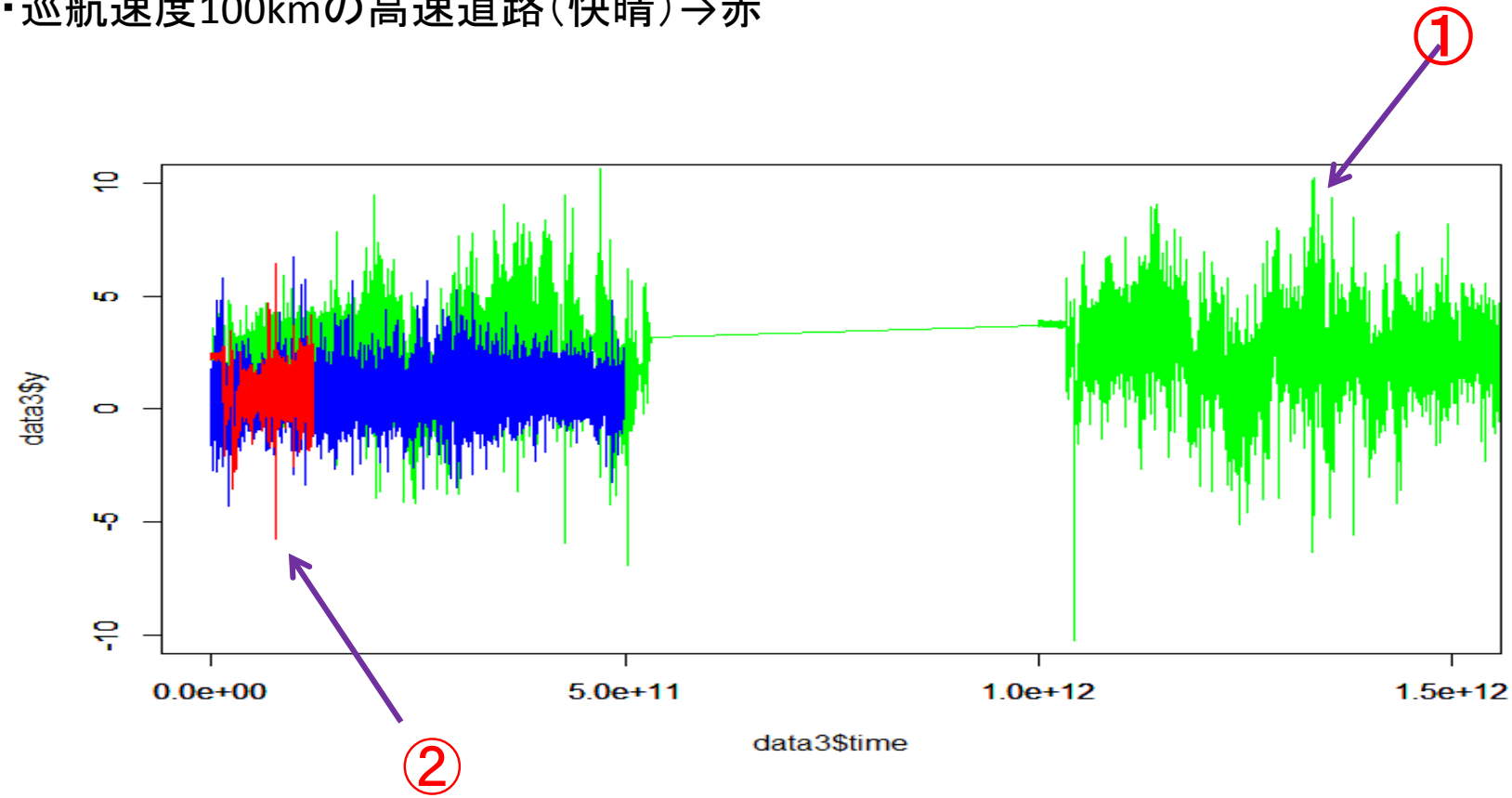
スマートフォンをハンドル部分につけて左右への傾きを計測



3ヶ所でデータを取ってきました！

分析結果

・巡航速度60km前後の２車線の峠（快晴）→緑
・巡航速度30km～40km前後の一本道の峠（濃霧）→青
・巡航速度100kmの高速道路（快晴）→赤



※約15万行のcsv

①巡航速度が比較的速い峠

→左右へのグラフの変動が大きい

②高速道路の合流地点（車線変更）

→加速度がかかる

つまり、

車線変更を頻繁にする車両はこのような波形が多く見られるのでは？

２．ニコニコデータセットを用いた分析

ニコニコデータセットって、、、

- 国立情報学研究所のダウンロードサービスより(株)ドワンゴが提供
- 2007年3月～2012年11月に投稿された約830万件のデータ
- 動画のメタデータ（12GB）
- コメントデータ（300GB）

データの量が膨大すぎる(;´д`)

ということで一部だけ利用しました

とりあえず中身を見てみる

{”video_id”:”sm1001”,”thread_id”:1179132……}

⇒{”key”:”value”,…….}

※key=video_id,thread_id,upload_time,length,size_high,view_counter,tag,etc…

(´･ω･`).oO(　フムフム、、、なるほど、、　)

このデータを使って何か分析してみよう

Ⅰ.カテゴリータグの出現確率

Ⅱ.タグの数と再生数の関係

Ⅲ.タグを用いたアソシエーション分析

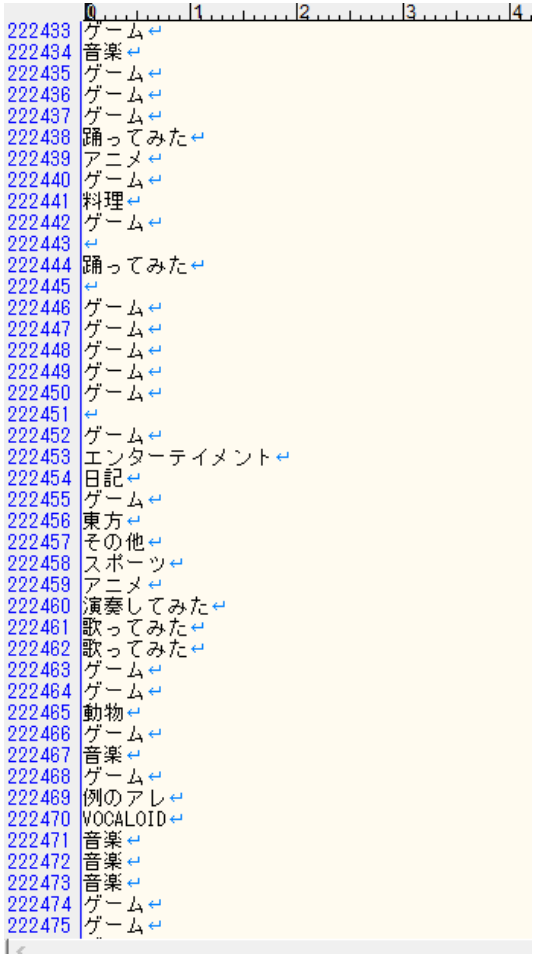
調べてみた!!

Ⅰ.カテゴリータグの出現確率

カテゴリータグ

エンターテインメント、音楽、歌ってみた、演奏してみた、踊ってみた、VOCALOID、ニコニコインディース
動物、料理、自然、旅行、車載動画、スポーツ、ニコニコ動画講座、歴史
政治
科学、ニコニコ技術部、ニコニコ手芸部、作ってみた
アニメ、ゲーム、アイドルマスター、東方、ラジオ、描いてみた
例のアレ、日記、その他
R-18
ファッション

まずデータの加工



```
222433 ゲーム↵
222434 音楽↵
222435 ゲーム↵
222436 ゲーム↵
222437 ゲーム↵
222438 踊ってみた↵
222439 アニメ↵
222440 ゲーム↵
222441 料理↵
222442 ゲーム↵
222443 ↵
222444 踊ってみた↵
222445 ↵
222446 ゲーム↵
222447 ゲーム↵
222448 ゲーム↵
222449 ゲーム↵
222450 ゲーム↵
222451 ↵
222452 ゲーム↵
222453 エンターテイメント↵
222454 日記↵
222455 ゲーム↵
222456 東方↵
222457 その他↵
222458 スポーツ↵
222459 アニメ↵
222460 演奏してみた↵
222461 歌ってみた↵
222462 歌ってみた↵
222463 ゲーム↵
222464 ゲーム↵
222465 動物↵
222466 ゲーム↵
222467 音楽↵
222468 ゲーム↵
222469 例のアレ↵
222470 VOCALOID↵
222471 音楽↵
222472 音楽↵
222473 音楽↵
222474 ゲーム↵
222475 ゲーム↵
```

Json形式のファイル

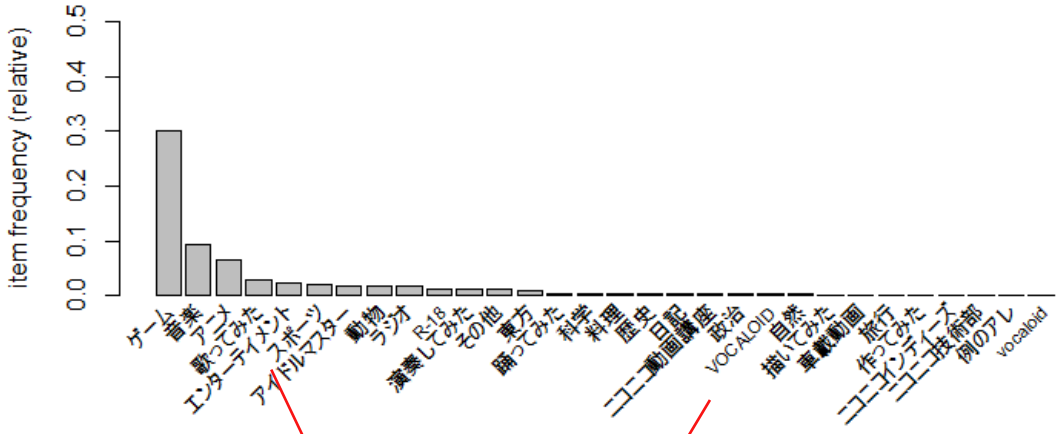
↓

カテゴリタグだけを抽出

↓

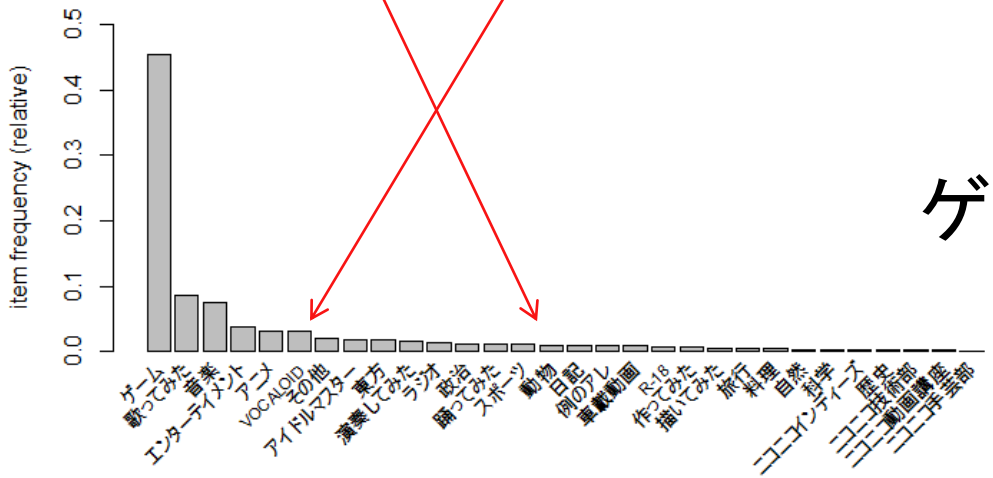
RでitemFrequencyPlot(data)

～2008/01



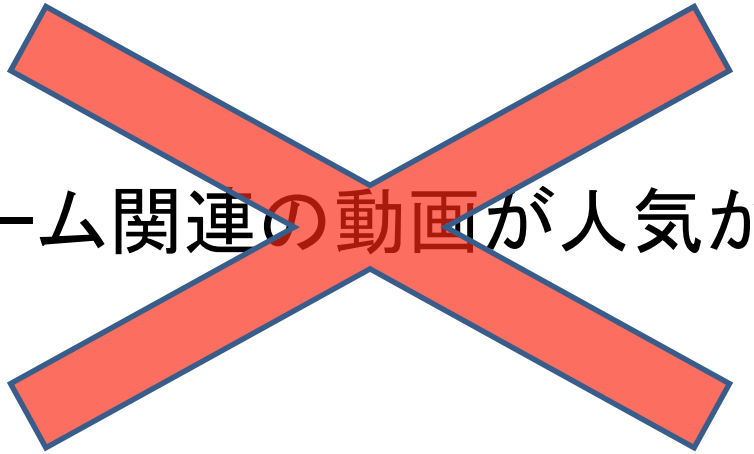
・ゲームタグが多く出現
・VOCALOIDタグが増加
・スポーツタグが減少

2012/06～2012/11

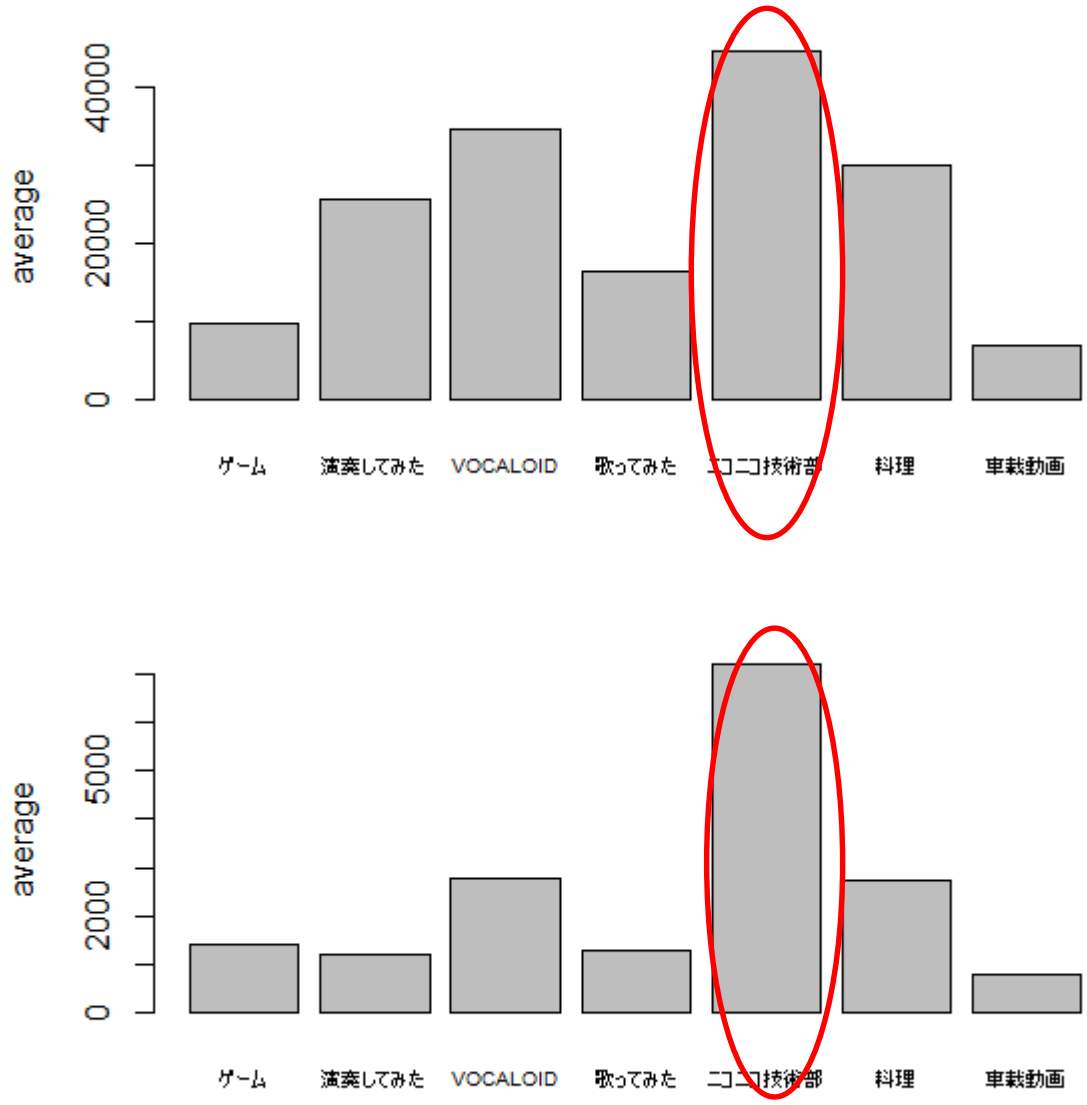


ゲームタグが3割超！！

(´･ω･`).oO(　ゲーム関連の動画が人気があるのか、、)



ある期間におけるタグごとの平均再生回数



この2つの期間においては共に「ニコニコ技術部」が人気？

投稿する側

→ゲーム関連の動画をよくうp

視聴者側

→(※この期間においては)「ニコニコ技術部」の動画をよく再生

（＝おもしろい動画が多い）

Ⅱ.タグの数と再生数の関係

- タグは1つの動画に対して最大１１個
  （うちカテゴリタグは3個まで）
- キーワードを含むタグ検索が可能

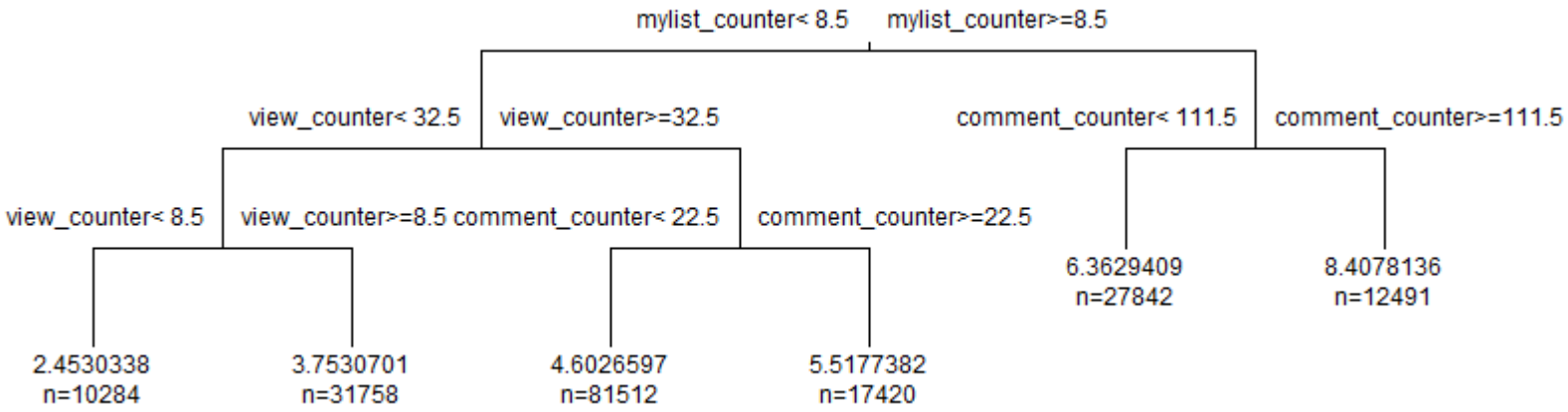
↓

タグが多いほど盛り上がっている動画なのでは？

まずデータの加工

	A	B	C	D	E	F	G	H	I	J	K	L
1	video_id	thread_id	upload_time	title	movie_type	size_low	size_high	mylist_counter	length	view_counter	comment_counter	tag_counter
2	sm19247769	1351682087	2012-10-31T20:14:47+09:00	夕暮れ時に聴	mp4	0	334397195	4	6370	285	3	7
3	sm19190025	1351008005	2012-10-24T01:00:04+09:00	ゴールデンボ	mp4	0	330772456	105	6068	4794	819	9
4	sm19056468	1349519736	2012-10-06T19:35:36+09:00	【ラテール】【1	mp4	0	329617200	6	6052	392	2	4
5	sm19009272	1348966550	2012-09-30T09:55:50+09:00	HITS! THE T(	mp4	0	314243418	6	5758	959	3	5
6	sm19136152	1350398270	2012-10-16T23:37:50+09:00	ゴールデンボ	mp4	0	307575580	100	5643	4119	304	10
7	sm19222191	1351398493	2012-10-28T13:28:13+09:00	10/28朝礼拝	mp4	0	301622420	0	5528	3	0	6
8	sm19030839	1349186606	2012-10-02T23:03:26+09:00	ゴールデンボ	mp4	0	295198172	109	5416	11931	333	11
9	sm19180689	1350912067	2012-10-22T22:21:07+09:00	【2005年】永	mp4	0	287891470	13	5278	1284	49	1
10	sm19224809	1351417903	2012-10-28T18:51:43+09:00	10/28夕礼拝	mp4	0	285215448	0	5227	7	0	6
11	sm19082678	1349791487	2012-10-09T23:04:47+09:00	GARO ZERO	mp4	0	282847214	3	5856	82	55	4
12	sm19083775	1349796714	2012-10-10T00:31:54+09:00	ゴールデンボ	mp4	0	280324231	127	5145	9968	467	11
13	sm19023945	1349099321	2012-10-01T22:48:41+09:00	悪魔のQP ウ	mp4	0	280178882	65	5155	11829	87	10
14	sm19167701	1350774072	2012-10-21T08:01:12+09:00	HITS! THE T(	mp4	0	279514768	3	5122	627	5	6
15	sm19055425	1349511144	2012-10-06T17:12:24+09:00	MADOKA MA	mp4	0	278997793	37	5164	574	13	4
16	sm19171937	1350819360	2012-10-21T20:36:00+09:00	10/21 夕礼拝	mp4	0	278162582	0	5099	3	0	6
17	sm19135257	1350393578	2012-10-16T22:19:38+09:00	映画史上最悪	mp4	0	272720968	22	4998	774	71	10
18	sm19121900	1350224647	2012-10-14T23:24:07+09:00	【走行音】ホー	mp4	0	272057736	2	5025	85	3	5
19	sm19169045	1350792945	2012-10-21T13:15:45+09:00	10/21 朝礼拝	mp4	0	263064774	0	4822	2	0	5
20	sm19238090	1351568249	2012-10-30T12:37:29+09:00	うちらの時間	mp4	0	262028267	0	8060	18	0	3
21	sm19127648	1350303905	2012-10-15T21:25:05+09:00	ふぃあ通 20	mp4	0	260418835	6	5341	199	2	5
22	sm19098172	1349997416	2012-10-12T08:16:56+09:00	【うんこちゃん	mp4	0	252414169	11	4665	2493	22	1
23	sm19039298	1349288965	2012-10-04T03:29:25+09:00	LAST EXILE	mp4	0	249814624	14	4669	359	2	3
24	sm19237958	1351565832	2012-10-30T11:57:12+09:00	うちらの時間	mp4	0	246462591	0	7583	1	0	4
25	sm19115627	1350183938	2012-10-14T12:05:38+09:00	HITS! THE T(	mp4	0	242030175	1	4435	329	4	6
26	sm19225207	1351421737	2012-10-28T19:55:37+09:00	スフィアのオ	mp4	0	238194567	18	6120	463	20	2
27	sm19018052	1349019648	2012-10-01T00:40:47+09:00	切ない・癒し系	mp4	0	237465340	22	4525	633	6	3

- Jsonからメタデータ抽出＋タグの数をカウント
- csvとして出力
- rpartで決定木(目的変数＝tag_counter)



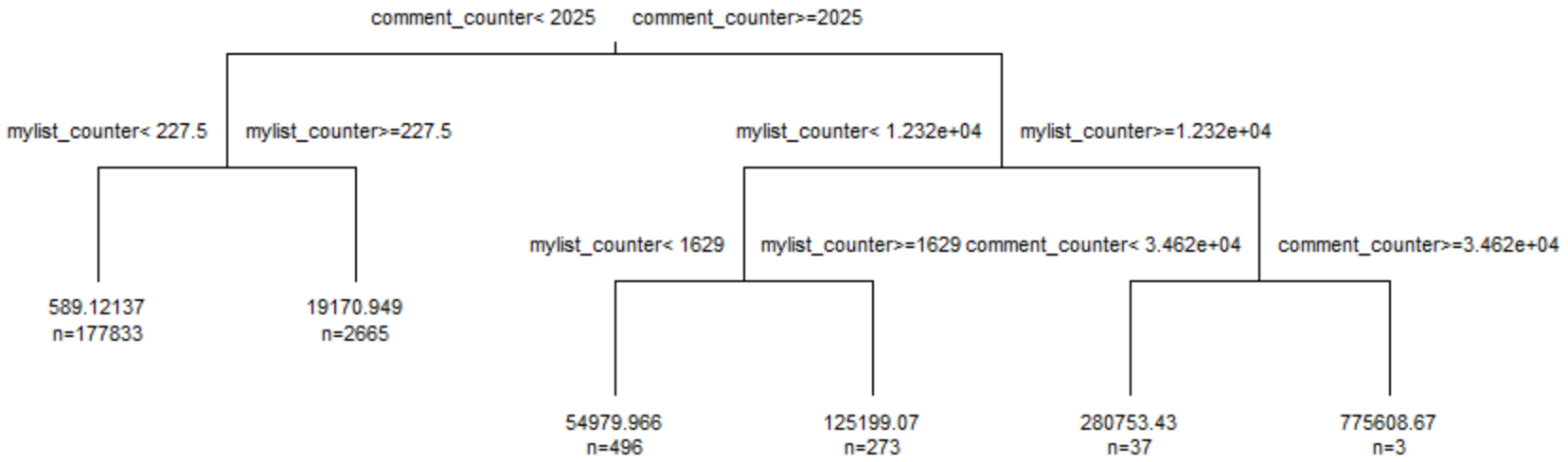
・分岐基準はマイリス登録数、再生回数、コメント数
・マイリス登録が多くコメントも多い→タグの数も多い
・マイリス登録が少なく、再生数も少ない→タグの数も少ない

しかし、
分岐基準となる値が小さい→少しでも見られている動画はタグの数が少ないことはない？

つまり、

盛り上がる動画はタグの数が多くつけられる

逆は？？



目的変数＝view_counter

分岐基準にtag_counterが出てこない

↓

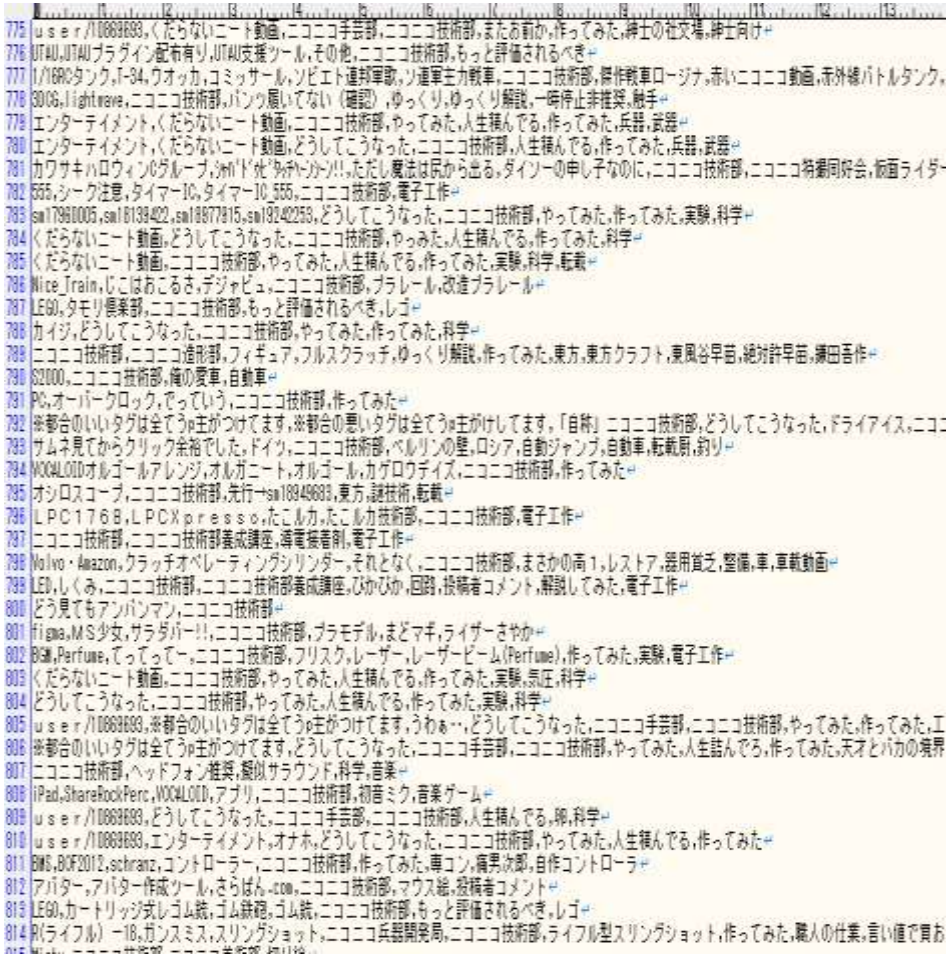
再生数が多いのはタグの数によるとは限らない

Ⅲ.タグを用いたアソシエーション分析

タグの間にどのような関連性があるのかを調べてみた

条件
2012年8月、9月に投稿されたデータ(35万件)
その期間のうち「ニコニコ技術部」タグを含む（818件）

まずデータを加工



Json形式のファイル

↓

条件に適したタグを抽出

（1行につき1件の

トランザクションデータ）

↓

>read.transactions(data)

>apriori

>inspect

分析結果

```
>library(arules)
>ap<-apriori(data,parameter=list(maxlen=4,support=0.02,confidence=0.7,ext=T))
>inspect(head(sort(ap,by="lift"),n=50))
```

---

1 {やってみた, 科学, 作ってみた} => {どうしてこうなった}

support	confidence	lhs.support	lift
0.02	0.89	0.02	17.020807834

2 {やってみた, 科学} => {どうしてこうなった}

0.02	0.85	0.025	16.305647841

3 {ニコニコ技術部, やってみた, 科学} => {どうしてこうなった}

0.02	0.85	0.025	16.305647841

4 {ニコニコ動画講座} => {ニコニコ技術部養成講座}

0.03	0.81	0.039	16.210365854

5 {やってみた, 作ってみた} => {どうしてこうなった}

0.037	0.83	0.045	15.938403520

やってみた、作ってみた⇒どうしてこうなった

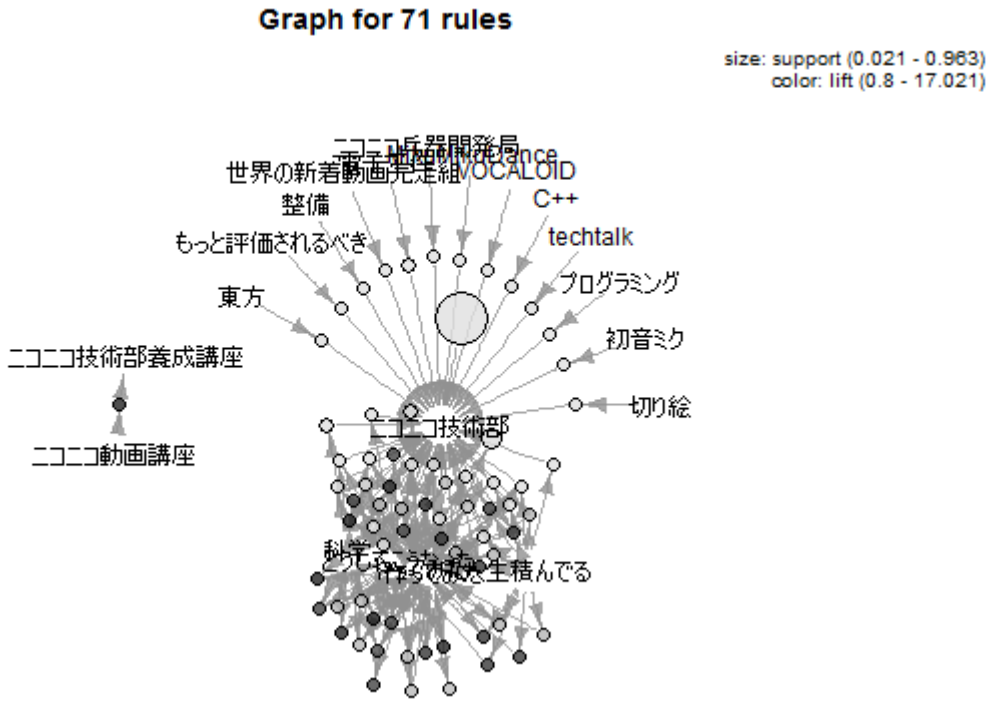
というルールが多く出現

↓

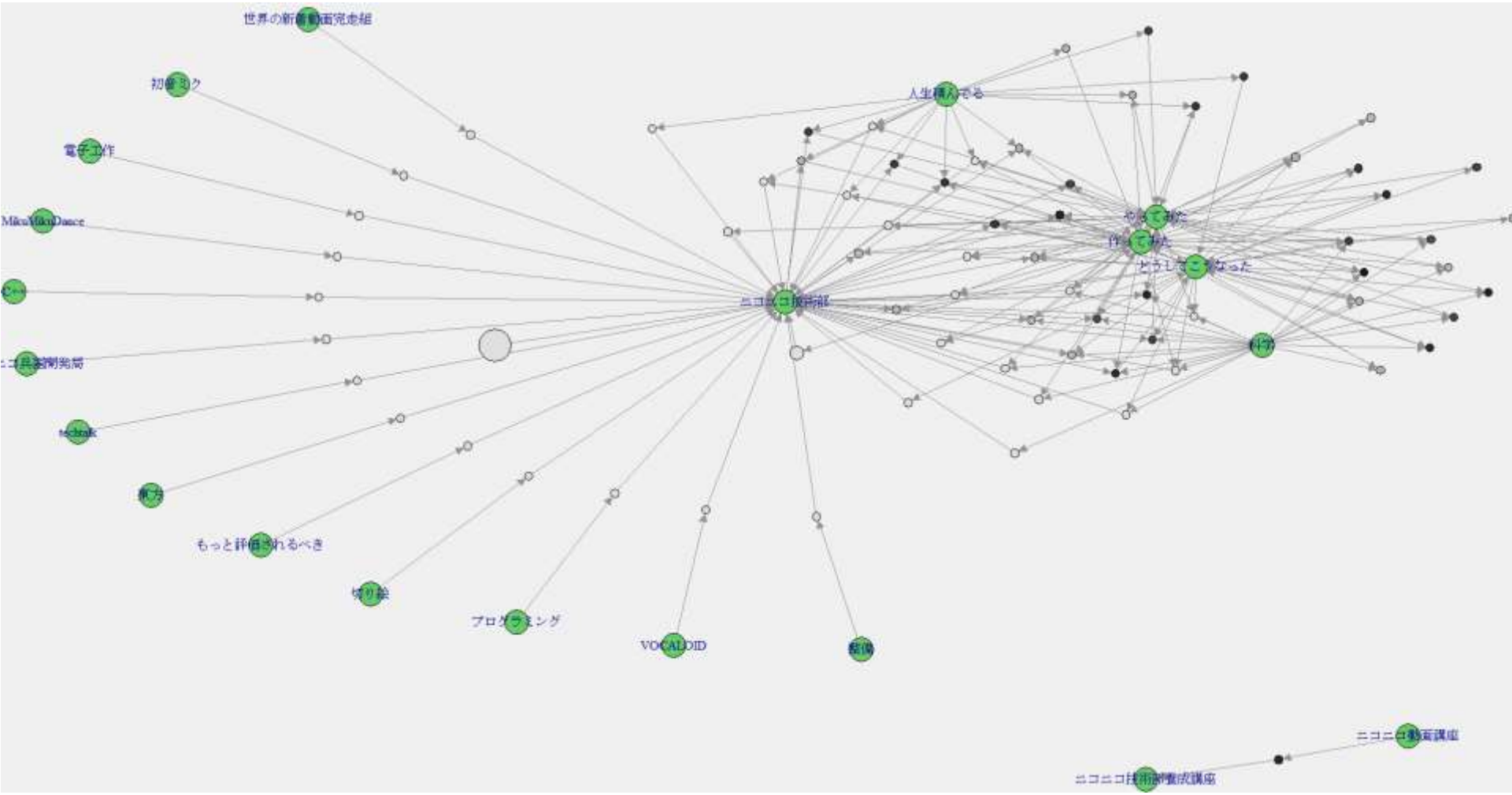
「すごい」というより「面白い」という要素が強そう

可視化すると、、、

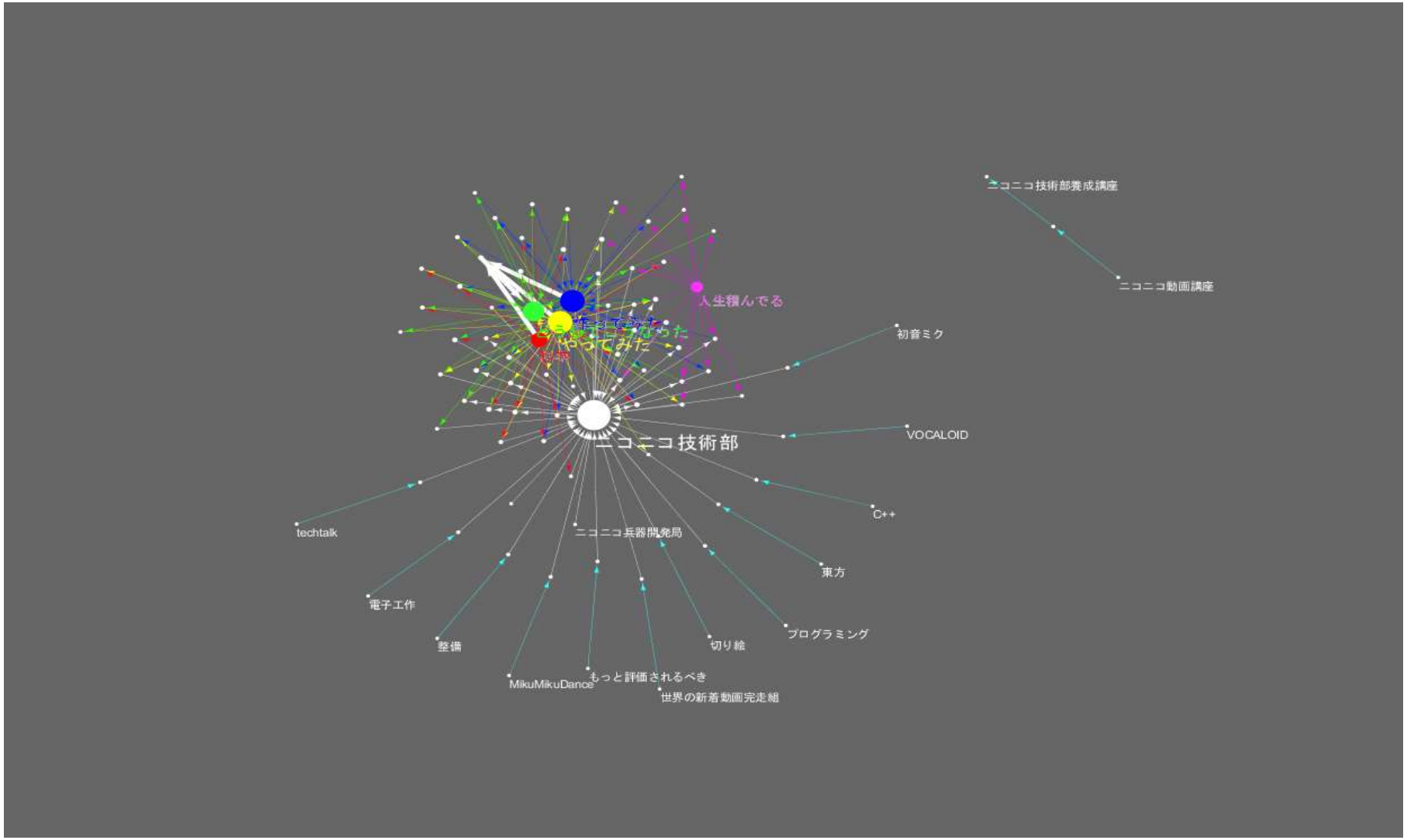
```
>library(igraph)
>plot(ap,method='graph',control=list(type='items',layout=layout.fruchterman.reingold))
```



Plotにinteractive=TRUEを追記するとちょっと見やすくなる



Cytoscapeで可視化



わからないということが、わかった

反省点

- コメントデータを使えなかった

　→関連性の薄いタグが複数あったら「タグ理解」のコメントが出現するのでは？

　→wwwが複数あったら8888がある可能性が高いのでは？

ご清聴ありがとうございました