REDDIT

Finding Structure in Comment Networks

STAT 578 Project

23.04.2018

Shriyak. Sridhar

MS in Statistics, 2016-18

[–] jeffgerickson CS prof 3 points 5 days ago

Degrees don't land jobs. Skills and accomplishments land jobs.

permalink embed save report reply

[–] icecoolsushobhan CEE/PhD 2 points 2 days ago

I didn't see the sun for such a long time in February, I was afraid that it was gone and NASA didn't tell us about it.

permalink embed save report reply

classes in spring, internship in summer, more classes the next fall.)

- The UIUC program probably has more flexibility than you realize. In my role as the program's advisor, I've been rather accepting about substitution of newer, more data science and machine learning focused courses.

Somewhat coincidentally, I've had meetings with some of the leaders of that Duke program, and I like a lot of what they plan to do. It would be hard to say that one program is "better" than the other, but there are a lot of differences, mainly based on the department that runs the program. At UIUC your instructors would be statisticians. (Maybe some computer scientists.) At Duke they would come from a wide variety of departments since the degree is offered by SSRI and IID.

There's also the usual differences between the schools as a whole to consider. Man, that Duke campus is nice, and the weather is a little nicer in Durham. (Champaign has been interesting this spring...) Obviously UIUC has a much lower price tag.

permalink embed save report reply

Hey, everyone: I'll be taking your questions online today. Ask yours here: OFA.BO/gBof44

-bo

I am Barack Obama, President of the United States -- AMA •...

Hi, I'm Barack Obama, President of the United States. Ask me anything. I'll be taking your questions for half an hour starting at about 4:30...

reddit.com

Is that the map with the rancor? I thought battlefront 2 had that map. Gamorrean guards came out and murdered both teams in epic fashion.

THAT 1ST BOSS FIGHT WAS INCREDIBLE

Best I've played in a long time

I compiled a list of all the software in this thread that got a 1000+ score (in order from top to bottom), along with a short description of each.

Over 1000 upvotes:

1. Google Maps: Navigation app - https://www.google.com/maps
2. Blender: 3D modeling software - https://www.blender.org/
3. VLC: Video player - https://www.videolan.org/index.html
4. The Windows Snipping Tool: Screen capture tool -

Will there be Jakarta fried rice that can beat Cravings's?

permalink embed save report reply

I know a kid who got 100s on the first two quizzes and a 22 on this last one so yeah...

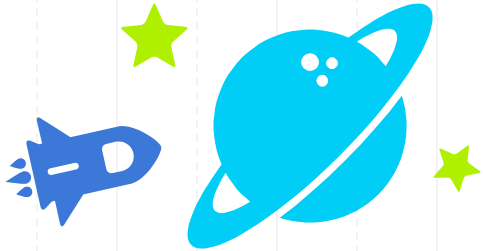
permalink embed save

The St. Louis Cardinals lost 9-4 to the New York Mets, starting their season at 0-1, and destroying all hope of a perfect season.

The Cubs have started their season at 1-0, a 100% win rate.

Analysis

- Subreddit networks
- Topic Modelling
- Clustering communities

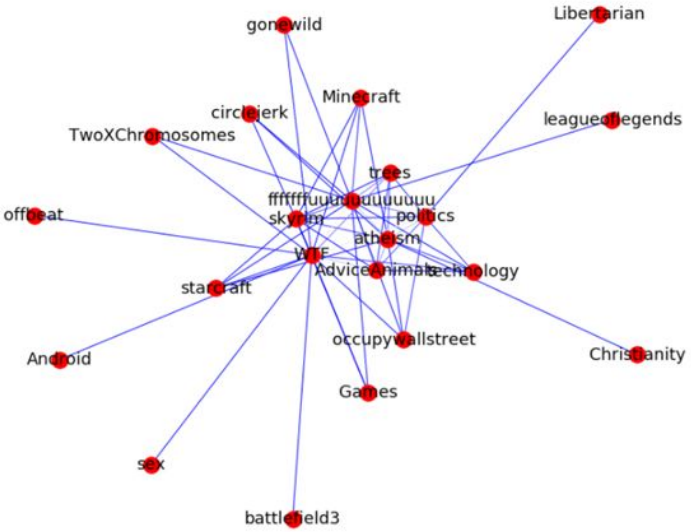


SubReddit Networks

Subreddits are forums dedicated to specific topics. The interactions between these forums are mapped using comments.

SubReddit Network - 2009

- One Central Cluster
- Many isolated outer-nodes

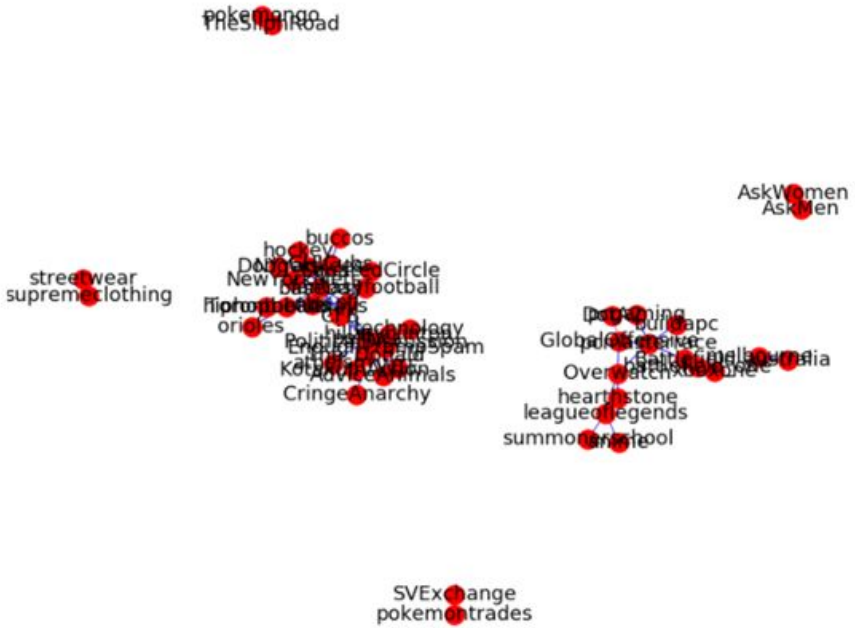


Important themes:

- Gaming
- Religion
- Technology
- Politics

SubReddit Network - 2016

- Multiple clusters
- Outer nodes are connected
- Inner nodes are hard to identify

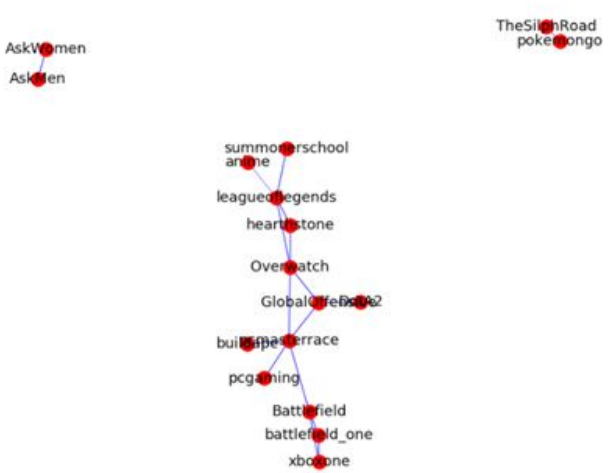


Important themes:

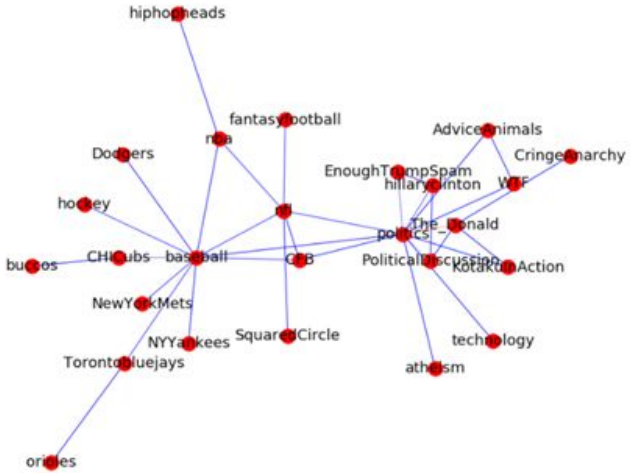
- Gaming
- Politics
- Sports
- Fashion, tech, advice, memes

SubReddit Network - 2016

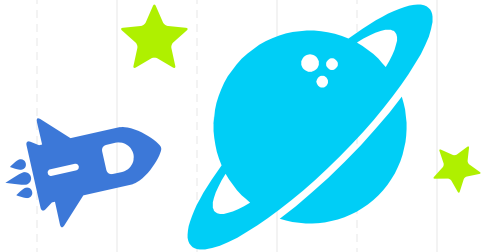
Focussing on the central nodes



Redditors love gaming, 2009-2016!



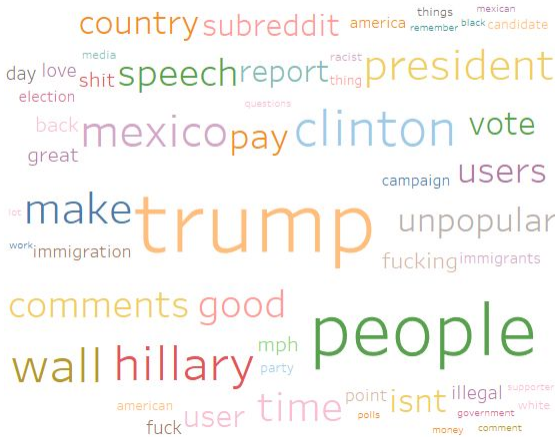
Sports connects and divides Reddit!



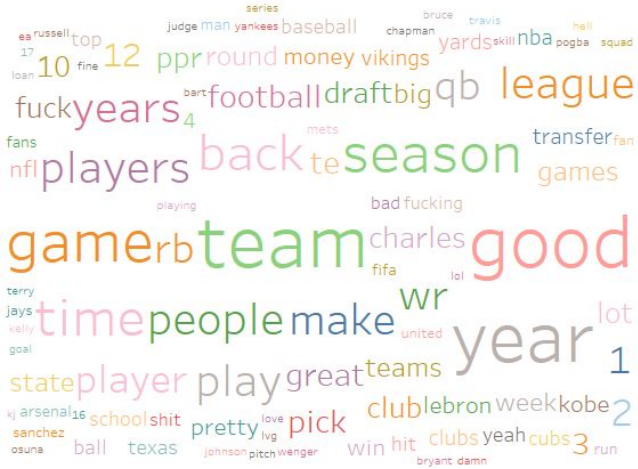
Topic Modelling

Topic modelling finds an abstract *topic* (i.e. group of words) from a collection that best represents the information.

Word Clouds - Politics vs Sports



Politics

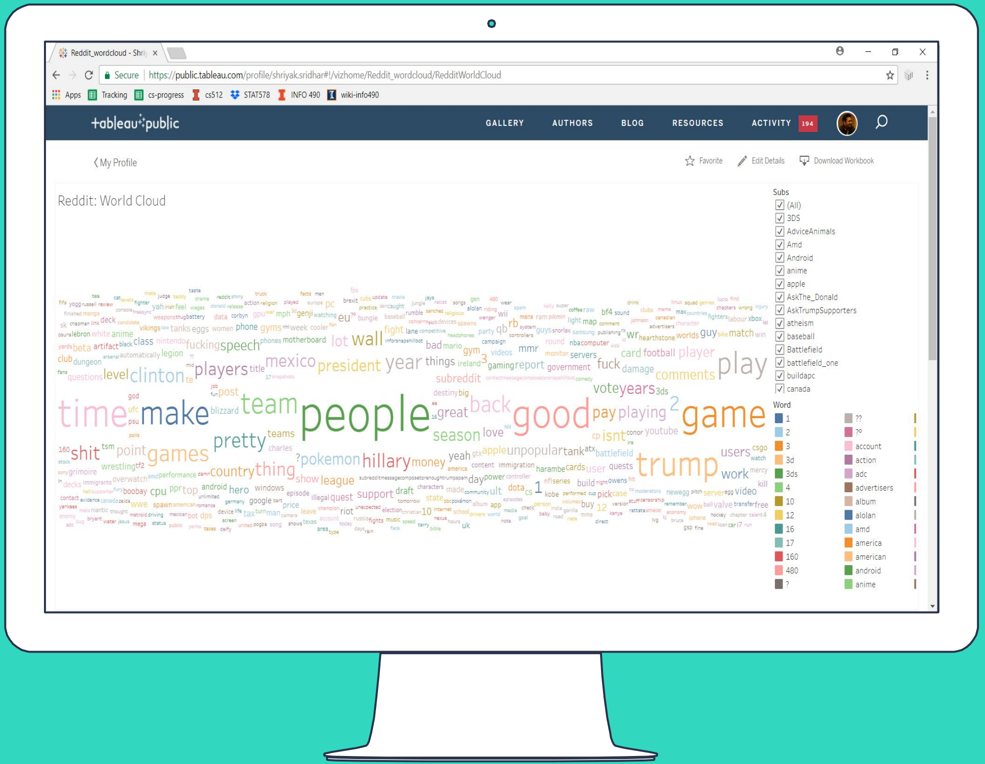


Sports

PUBLISHED ONLINE:

https://tabsoft.co/2qNUapE

Explore the word-clouds for different Reddit communities



Topic Modelling

PROCESS TEXT
Remove punctuation, “stop words”, make lower case.

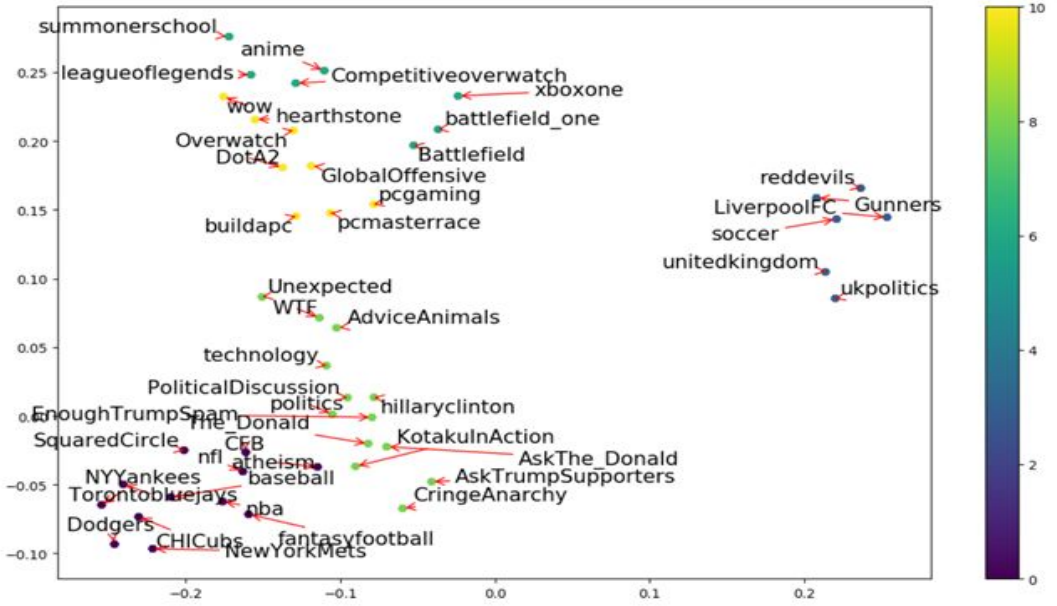
WORD COUNTS
Find word frequency, inverse doc. frequency

GENERATE MODELS
Based on corpus, word cluster and similarities

Topic Models

Topic 1 - *politics*		Topic 2-*gaming*		Topic 3-*sports*		Topic 4-*other*	
people	0.005711	good	0.005944	games	0.003602	lol	0.008317
trump	0.00455	play	0.005336	post	0.003126	f*ck	0.005966
make	0.002577	game	0.005128	play	0.002927	god	0.004031
money	0.002179	team	0.004651	time	0.002909	awesome	0.003942
clinton	0.001757	wow	0.003218	yeah	0.002899	man	0.003672
country	0.001668	kill	0.002631	work	0.002608	damn	0.003447
vote	0.001668	damage	0.002561	years	0.002349	guy	0.003331

Clusters Formed - 2016



Clear clusters form:

- Politics and memes
- American sports
- Gaming
- European Cluster
  - UK
  - British teams

CONCLUSION

Diverse topics

As the site matured from 2009 to today, it has become more diverse

Variety

There is a clear difference in topics discussed in various communities

Potential

There is a potential for recommender system

Communities

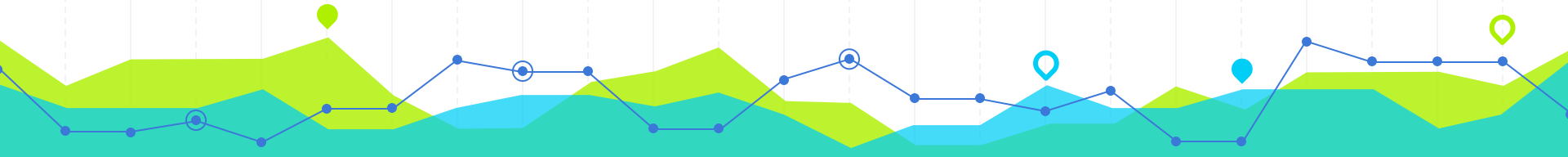
There are clear communities that appear (politics, gaming, sports)

Cliques

We can classify which cluster a comment belongs

Difficulties

Predicting comment score is very difficult!

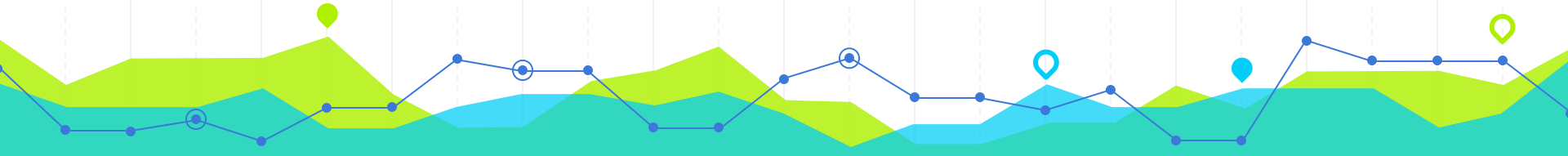


THANKS!

Any questions?

Special thanks to the people who curated and released this awesome dataset for free:

- /u/Stuck_In_the_Matrix
- /u/fhoffa



INTERESTING AND UNEXPECTED FINDINGS

Gaming

The popularity, size of gaming subreddits and their persistence with time. These are especially true for multiplayer games

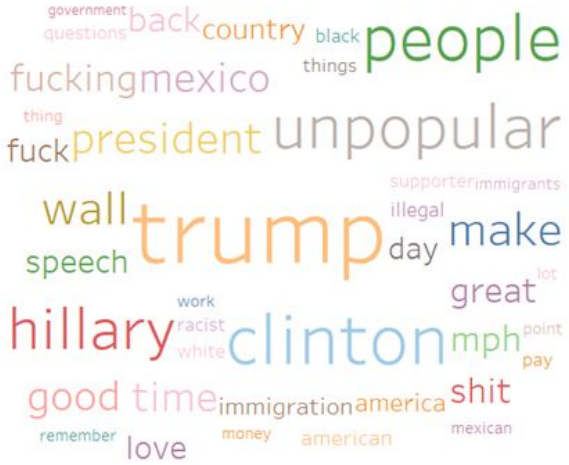
Trump makes folks angry

There is a marked negative sentiment in Donald Trump related subreddits and a lot of use of profanity, name calling

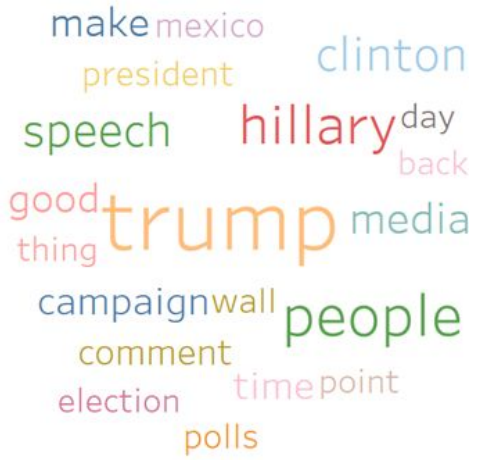
GMM >>> Kmeans

Despite using the Elbow method and ideal k, K-means has extremely poor performance in identifying clusters.

Word Clouds - Trump vs Clinton



Trump



Clinton

Gaussian Mixture Models - Elbow method and contour plot