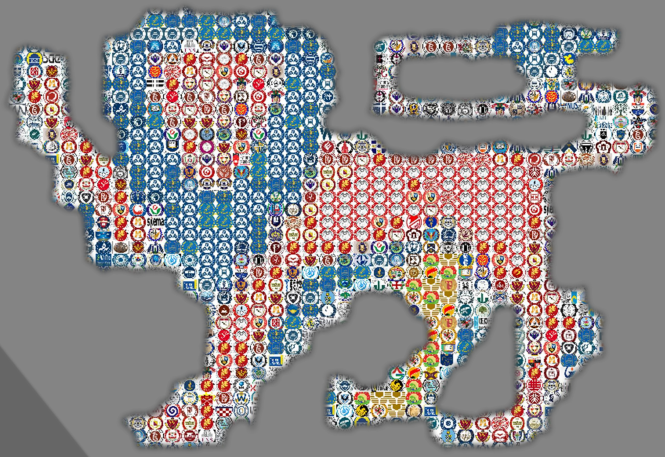
Community Detection for University Badges

Ge Song, Qin Yu, Tang Longfei, Zhan Mingwei

Team 01

SWS3001
Summer 2019

FIND YOUR TWIN-UNIVERSITY



INTRODUCTON

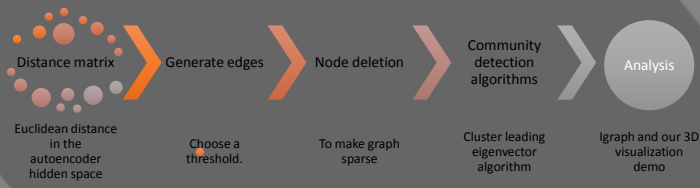
TOPIC: CD for university badges
OBJECT:
MOTIVATION: many badges look alike
PURPOSE: analysis the similarity and dissimilarity of university badges over different countries and within each country

DATASET & METHOD

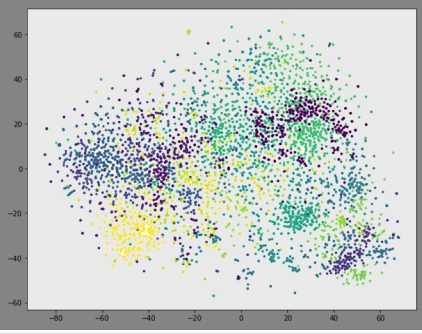
DATASET
4000+ badges from 12 countries

METHOD
AUTOENCODER to extract features
MATRIC of distance computed
perform **Girvan-Newman**

WORKFLOW

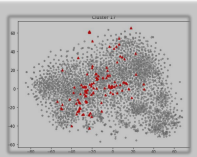


Clustering Result



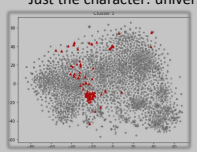
RESULTS OBTAINED

Badges contains the same icon tends to be clustered together.
All the medical school has a icon of snap twisted on a cane.
Also, we find that the nodes connects green community (medical schools) and orange community are the medical schools that its icon is reshaped from the original snake cane.



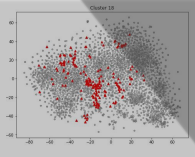
Japan : 13%

Just the character: university



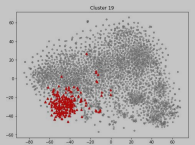
UK : 17%

Shield and a hat



Germany : 53%

Kin, Queen, Goddess



Indian : 24%

Rising Sun from sea and ocear

‘NUTRITION FACTS’ ON YOUR DAILY NEWS FEEDS

COMMUNITY DETECTION ON NEWS TITLES

Gao Ruifeng Pi Haowen Xiang Yunfan Yin Xiaolan

SWS3001
Summer 2019
CLUSTER2-02

WHAT ARE WE GOING TO DO?

Classification: find the relationship between different news published by one news media.

Comparison: Compare the similarities and differences among components of news from different medias.

Time sequence: Tracking trends in the news community reported by the same media in time.

WHAT WE HAVE LEARNED

- Data tells stories.
- Conduct different news media portraits to illustrate their role positioning and starting point.

Extends:

- News recommendation.

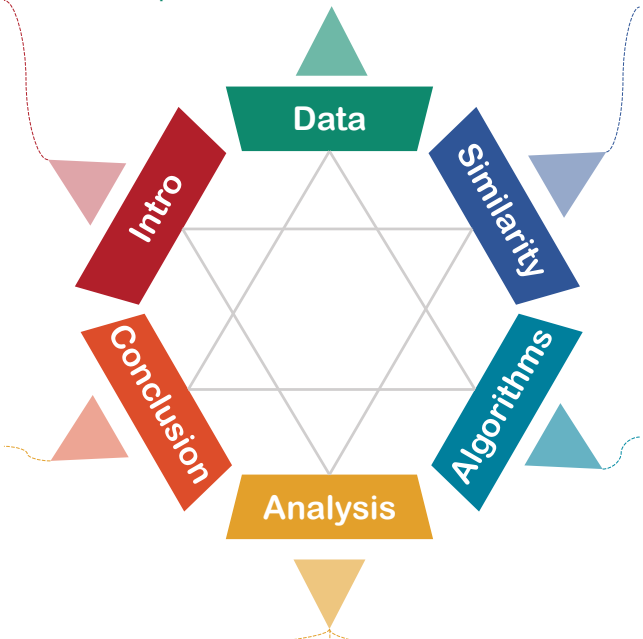
WHAT DATASET DO WE USE?

今日头条 你关心的 才是头条

搜狐新闻

the answer company THOMSON REUTERS

Classification: Jinri Toutiao, 2000 headlines.
Comparison: Jinri Toutiao, Sohu, 2000 each.
Time sequence: Reuters news, 2000 headlines.



HOW WE GET SIMILARITY ?

Similarity:

Using BERT for sentence vectors. Calculate distance by cosine distance:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Matrix:

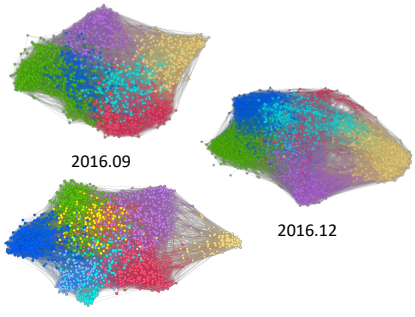
- Setting Threshold: Bad performance.
- k-NN: Keep the highest k edges, set the remaining to 0.
  - Works better. Forms large, tidy communities.

THE CD ALGORITHM?

Louvain: A greedy algorithm that maximize the MODULARITY Q.

$$\begin{aligned} Q &= \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\sum_i \delta(c_v, i)\delta(c_w, i) \\ &= \sum_i \left[\frac{1}{2m}\sum_{vw} A_{vw}\,\delta(c_v, i)\delta(c_w, i) \right. \\ &\qquad \left. - \frac{1}{2m}\sum_v k_v\,\delta(c_v, i)\frac{1}{2m}\sum_w k_w \delta(c_w, i)\right] \\ &= \sum_i (e_{ii} - a_i^2). \end{aligned}$$

TIME SEQUENCE: REUTERS

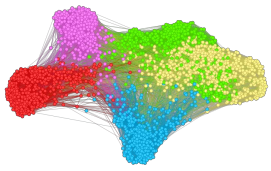


		2016/9	2016/12	2017/3	2017/6	2017/9
	The Middle East	√	√	√	√	√
	public safety	√	√	√	√	
	Large enterprise	√	√	√	√	√
	America, politic	√	√	√	√	√
	Asia-Pacific, politic	√	√	√		√
others	Big event				[1][2]	[3][4][5]

[1]2017.06 UK PM
[2[2017.06 Trump, Paris Agreement
[3]2017.09 Merkel, Germany
[4]2017.09 American hurricane
[5]2017.09 Myanmar ethnic conflict

- When conflicts became violent in 12/16, community of Middle-East news becomes larger.
  When situation there rested in 03/17, the community shrinks a lot.

COMPARISON: JINTRI TOUTIAO & SOHU



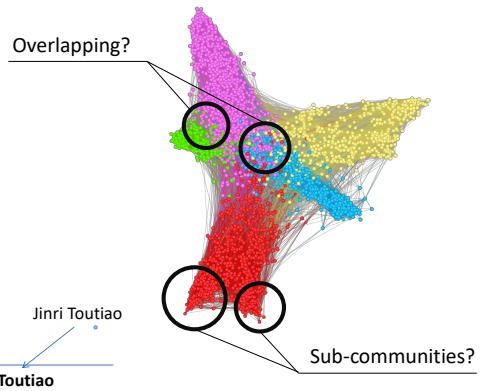
Sohu

Jinri Toutiao

		Sohu	Jinri Toutiao
	Sports	More kinds of sports	Mainly basketball and soccer
	Entertainment	Official	casual
	?	**Policy, education**	**family affairs**
	Global situation	Statement	Questions
	Finance	/	/

- The two graphs both have same four communities.
- Jinri Toutiao has a community of family affairs and Sohu has policy and education part.
- Toutiao prefers to report news more casually and asks readers questions. Sohu is more official and objective.

CLASSIFICATION: JINRI TOUTIAO



- Sports are mainly divided into two small communities—— basketball and football.
- Family affairs overlaps with the entertainment.
  - contains many family affairs and daily life of celebrities.
  - there are also films related to family relationship.
- The middle area have the same structure in their sentences. They are all like “What do you think of something”.

DIGGING INTO MOVIE POSTERS

NUS National University of Singapore

Team3 Chen Yiyan, Wu Longfei, Yan Shiqin, Zhang Yanshu

SWS3001

CLUSTER2-03

Introduction

·Explore common design style of movie posters.

·Find out potential relation between movie content and poster with regard to composition, colour,

·The object of this study is movie posters, supplemented by their corresponding movie information

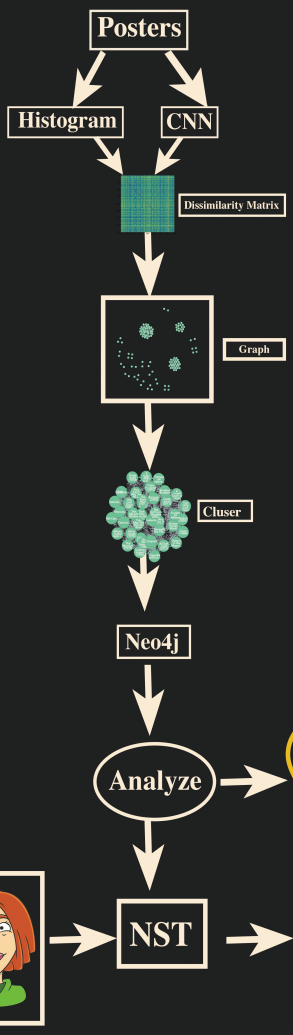
Methods adopted

·Two methods are combined to calculate the dissimilarity between two images: calculation based on histogram and convolusional neural network.

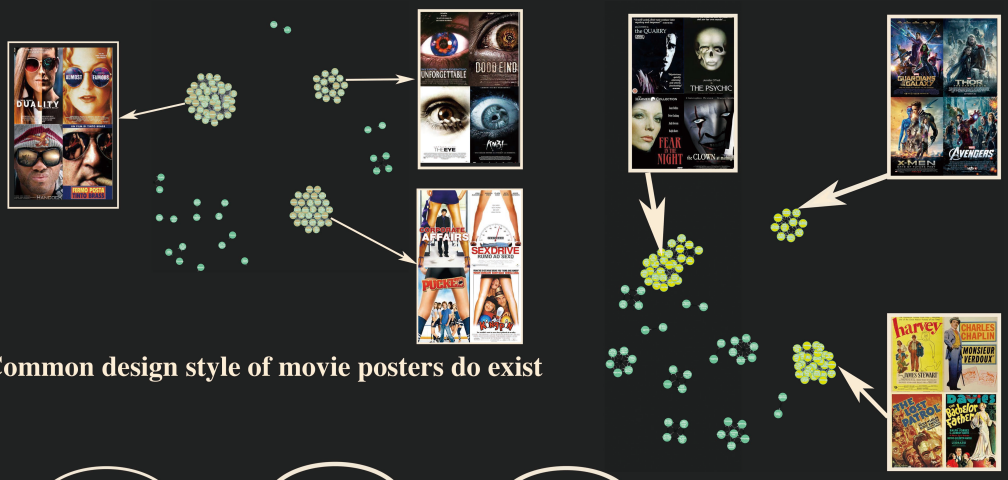
·A few different community detection algorithms(Greedy Modularity, Asyn Lpa, Girvan Newman, K Clique, etc) are implemented.

·Neural Style Transfer algorithm is used to generate new posters according to style posters.

Workflow

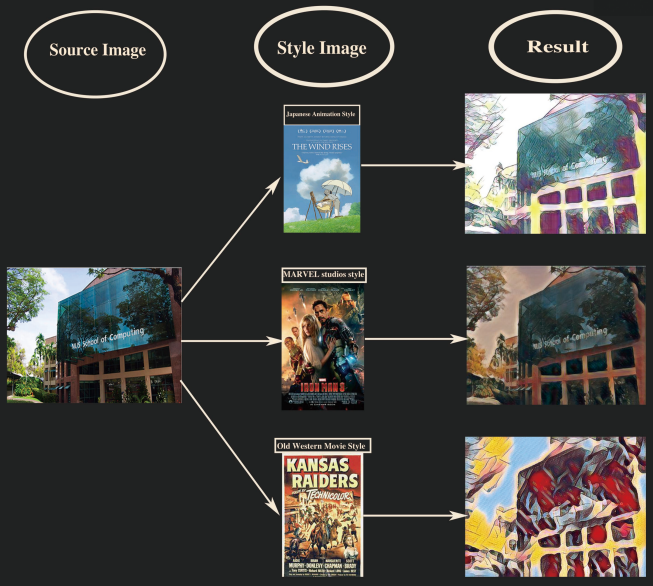


Results



Common design style of movie posters do exist

Posters of the same genre, series, period of producing time share common design style.



After analyzing the most representative posters of a certain theme, new poster can be generated with the help of Neural Style Transfer algorithm.

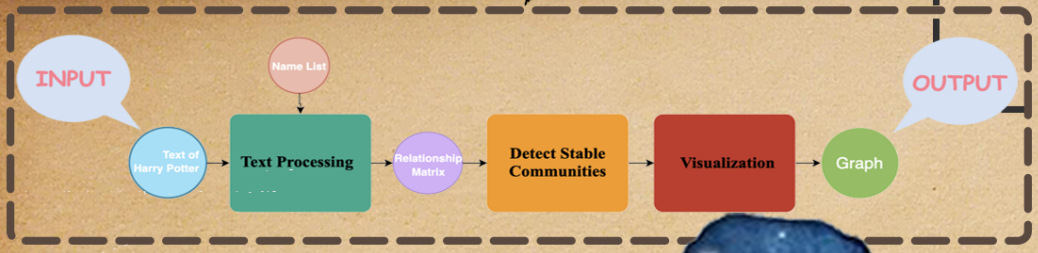
NUS National University of Singapore
Liu Xiaotong, Liu Xinyu, Peng Sen, Zhang Hanyi

SWS3001 CLUSTER2-04

Introduction

Harry Potter & communities detection.
We analyzed the communities to find some hidden relationships.
This method can also help you to read other books, like *A Song of Ice and Fire*..

Workflow

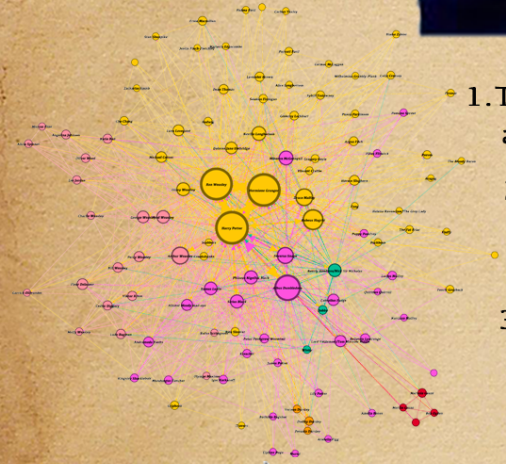


Data & Method Used

Data: Original text content of Harry Potter
Method:
1. Natural Language Processing.
2. Modularity Optimization.
3. Other community algorithms like fast-greedy for comparing results.
4. Removing main characters for better communities.

CAN DATA PERFORM MAGIC?

Result Analysis



1. The Weasleys have a Quidditch tradition
2. The Malfoys and the Longbottoms have strong connections.
3. Snape and Wormtail are in the intersection of Death Eaters and Order of Phoenix. One of them is a spy, the other one is a traitor.
4. Dumbledore is the only person that Grindelwald has a strong connection with. It is revealed in *Fantastic Beasts* that they were in love.

Death Eaters

Order of Phoenix

DEPRESSIVE COMMUNITY DETECTION AND ANALYSIS

SWS3001
CLUSTER2-05

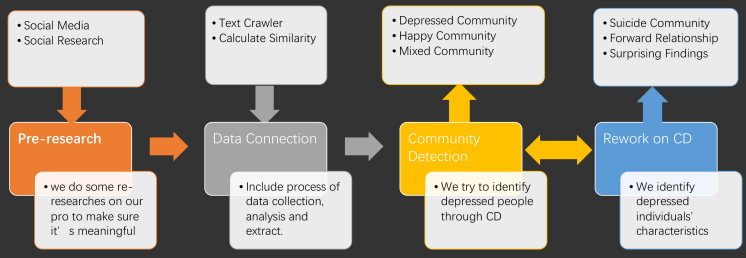
DENG YANGTAO, RUAN PENGHUI, TANG HANGYUN, WANG YUJIA

INTRODUCTION

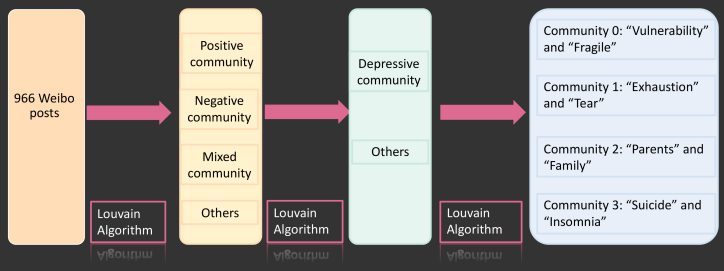
-Depression has been a increasing potential killer among normal people
-A common way to express -> Online social media
-Depressive people -> Possible discover by posts

We mainly focus on Weibo posts using NLP and CD algorithms to find out potential depressive groups.

WORKFLOW



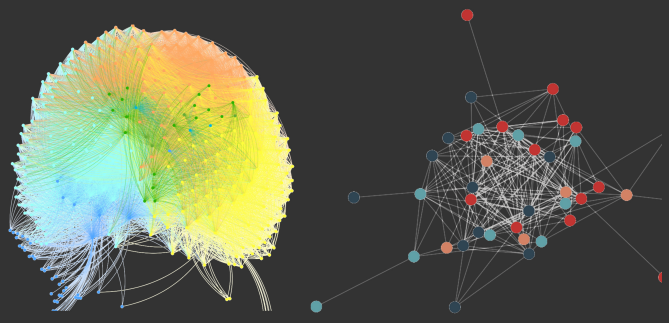
COMMUNITY DETECTION



The threshold to cut off unnecessary edges is 0.05.
We tried Girvan-Newman, Louvain,
Fast unfolding of communities in large networks algorithms to do community detection and compared them. Louvain behaves the best and the form of communities makes sense.

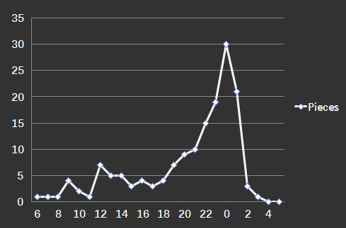
ANALYSIS

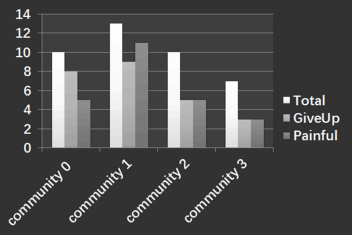
A positive community, a negative community and a mixed community in the left graph. Further community detection results on negative community including depressive population is shown in the right graph.



MORE RESULTS

Time analysis show that midnight is the time that depressive posts happened the most.
48 depressive patients are divided into 4 communities. Their posts have something in common : they all use words like " Death" "Medicine" " Torture"





CONCLUSIONS

After community detection with user's Weibo posts, we can predict whether a person have depressive tendency or not in some degree.

THE GALLERY OF CALLIGRAPHY

Team 06 Guo Minghao, Niu Yibo, Wu Yi, Yu Lin

SWS3001
CLUSTER 2--06

Introduction

- Chinese calligraphy has been developing along with Chinese history. After a long time of evolution, it now consists of five types of fonts: seal, clerical, cursive, regular and semi-cursive.

Five styles of the Chinese character "永"

Objectives

- Explore the connections between different styles of calligraphy.
- Analyze and validate the evolution of calligraphy.
- Analyze the relationships between calligraphers.

Details of Data

- Collect Chinese calligraphy images from the Internet.
- Each class contains 10 images of one style and one calligrapher.
- Crop the images to the same size (resolution: 440*440) and implement the binarization.
- Prepare two training datasets. One contains 47 classes, and the other is enlarged to 74 classes.
- Below is an example of one class:

Regular characters by Wang Xizhi, in Jin dynasty

Overall Workflow of Study

Collect data and form two datasets.

↓

Use Siamese neural network to generate dissimilarity matrix.

↓

Apply different community detection algorithms to detect community.

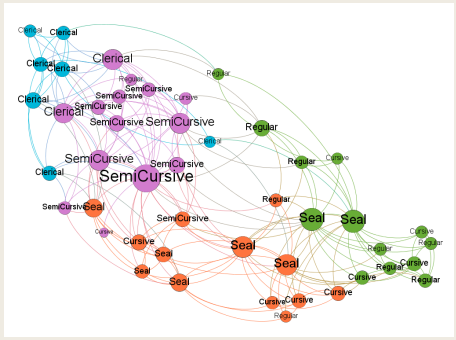
↓

Do the visualization and analyze the results.

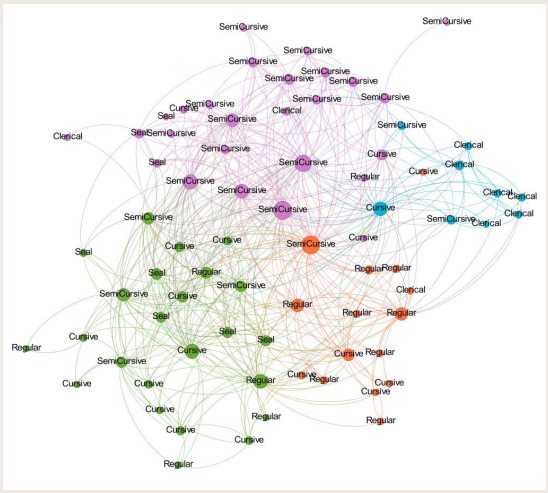
Community Detection

- Try three community detection algorithms: Girvan-Newman, Markov Clustering, and Louvain Algorithm.
- The parameters tuning depends on the modularity and the distribution of the communities.
- After comparison, we eventually chose the Louvain Algorithm.

Visualization & Analysis



Graph of the 47 classes by Louvain algorithm



Graph of the 74 classes by Louvain algorithm

- Each color represents one community. Both datasets are clustered into 4 communities through Louvain algorithm.
- As expected, vertices of the same styles tend to be in the same community.
- Even some classes belong to the same style, they can be clustered into different communities because of historical factors, such as evolution and revolution.
- Classes of different styles may be clustered into the same community. For instance, the evolution of the seal script in Qing dynasty made the style more similar to the cursive script. Therefore, many seal script fonts are clustered with the cursive script fonts.
- The newly added classes in the 74 classes can influence the clustering of the 47 classes. For example, since Chu Suiliang, a calligrapher, had studied the regular calligraphy of Ouyang Xun and Wang Xizhi, when the regular script of Ouyang Xun was added, the regular script of Chu Suiliang moved to the community with those two calligraphers.

Conclusion

By utilizing community detection to study calligraphy history, we can gain new perspectives to classify calligraphy instead of traditional classification.

NUS National University of Singapore

STEAMCODE

Steam COmunity DEtection

Jia Lianchen, Liu Yifan, Tian Runxin, Zheng Cuiyu

SWS3001
CLUSTER-07

WHAT A GAME POSTER NEED?

ATTRACTION!

INTRODUCTION

SteamCODE (Steam COmmunity DEtection) stands for “Detecting Communities of Steam Game Posters”.

Steam, one of the most famous PC game platform, serves a huge amount of games.

From Communities of Game Posters, we aim to provide a system and try to answer three questions.

1. IF YOU WERE A LEADER OF ONE COMPANY, WHETHER YOUR GAME POSTER IS ATTRACTIVE ENOUGH?
2. IF YOU WERE A PLAYER, WHICH UPCOMING GAME WOULD HAVE BETTER QUALITY?
3. IF YOU WERE A SENIOR PLAYER, DO YOU WANT TO DISCOVER MORE INTERESTING STORIES?

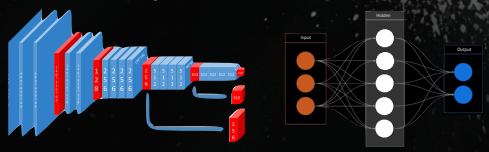
WORKFLOW

DATA

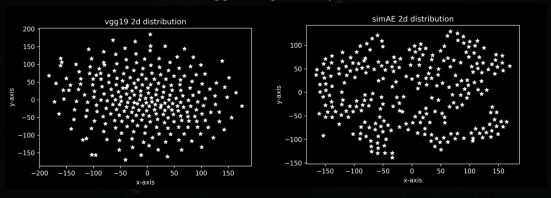
- Tiny poster of steam games with size (460*215px)
- Game info from steamdb.info

DISTANCE MATRIX

- **Extract Feature Vector.** We tried pre-trained both vgg19 (left) and Simple AutoEncoder (right) to get the feature vector matrix.
  INPUT: 460x215 image -> OUTPUT: 1x4306 vector.



- **Dimension Reduction.** Apply Principal Component Analysis to reduce the number of dimensions to a reasonable amount (eg.50).
  OUTPUT: 2-d data (left: vgg19, right: simple AutoEncoder)

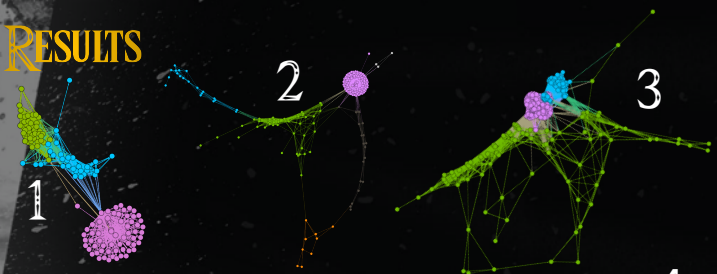


- **Compute Distance Matrix.**

COMMUNITY DETECTION

- **Construct and Refine the graph.** Set threshold to weight and degree to delete edges and nodes to get better CD result.
- **Apply CD algorithms.** Try different CD algorithm and get the most suitable one.
  FastUnfolding Algorithm perform best on our case.

RESULTS



ID	Model	Dimension	Modularity	Clusters
1	simAE		0.041	5
2	simAE		0.076	11
3	simAE		0.089	13
4	Vgg19			
5	Vgg19			

INSIGHTS

1. WHAT WE KNOW WE KNOW

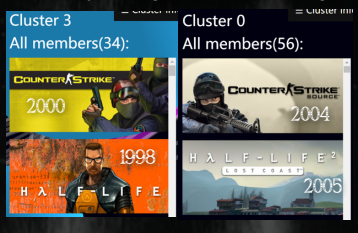
- **Series games** belong to the same (closer) community generally.
- Most action games in the same community.

2. WHAT WE KNOW WE DON’T KNOW

- Images in the same community appear similar in tone. (simpleAE)
- Images in the same community appear similar in shape. (vgg19)
  Which means simple AutoEncoder may extract more features about tone, and vgg19 may extract more about shape.

3. SOMETHING INTERESTING?

- **The style changes within a series of games may reflect interesting stories.**



- Counter Strike comes from Half-Life.
- Mike Harrington (author of Half-Life) left Valve and travel all over the world at 2000.
- Counter Strike is more popular for Half-Life to mirror the success.
- They both changes to source engine after 2000.

GitPals: Mining Communities in GitHub

SWS3001
CLUSTER2-08

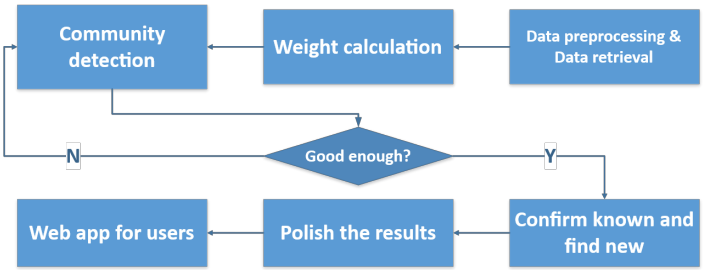
Wang Ziqin, Wang Zhiyi, Wu Qianwei, and Yin Junlin
https://gitpals.wangziq.in

Introduction

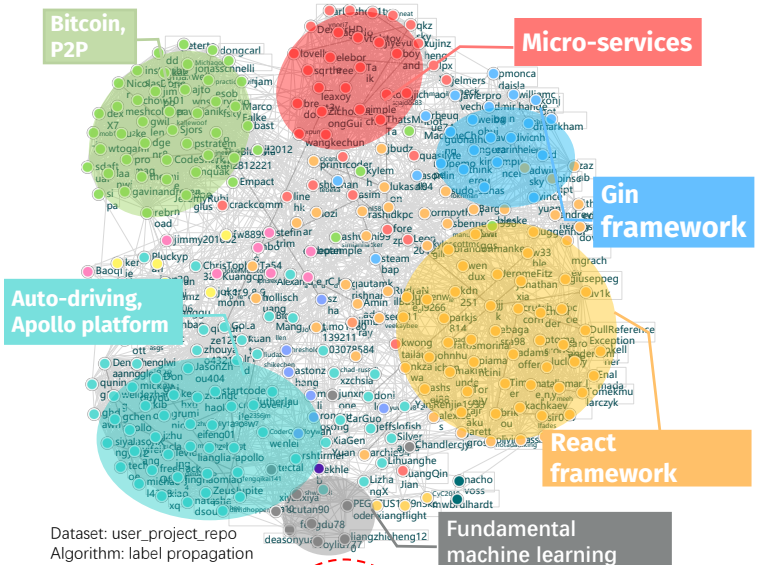
GitHub is a platform for code sharing and collaborative programming, where there are many naturally formed interest communities that we would like to analyse.

- **Goals:**
  - Detect real-world involvement for skill evaluation
  - Find out emerging interest groups
  - Discover hidden user relationships
  - Repository/User recommendation system (Future)
- **Algorithms**: Markov Clustering (MCL), Label Propagation (LP), Girvan-Newman(GN) and K-clique community detection algorithms.
- **Dataset**: Graph data retrieved via GitHub Developer API

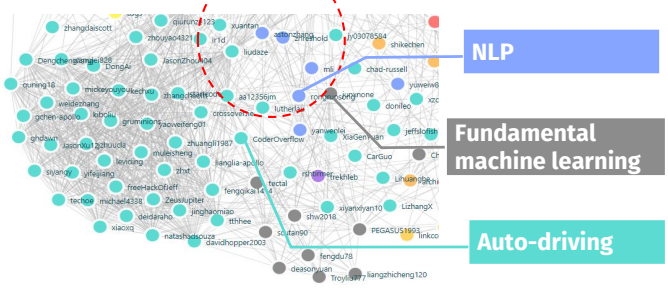
Workflow of Study



Results Obtained



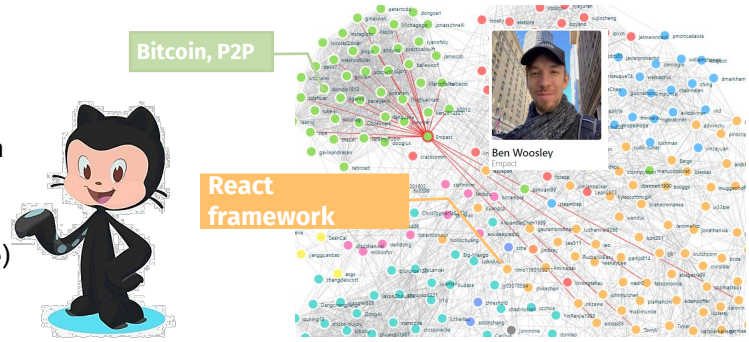
Dataset: user_project_repo
Algorithm: label propagation



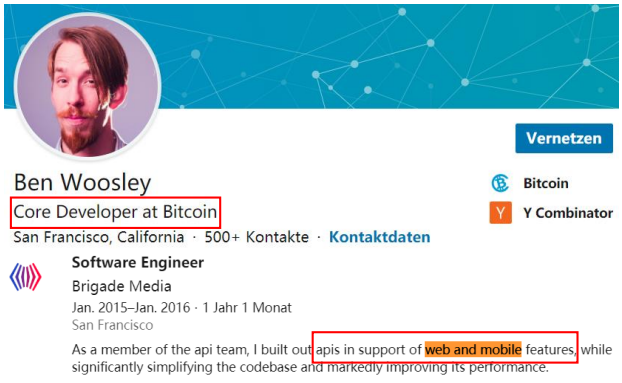
- The three topics are close to each other.
- They are all sub-fields of AI!

Interesting insights

Case 1

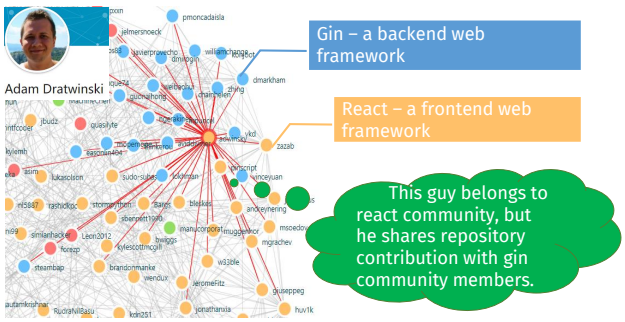


This Bitcoin guy has connection with react framework community. This suggests that he may also work on react-related software development.



Our investigation with LinkedIn indicates that our results is basically consistent with the reality.

Case 2



"I was a co-founder and main developer of the company's web framework." *— Adam Dratwinski's self introduction*

"Adam is a superb software developer with deep knowledge not only in Ruby, but in complete stack."

— Armin Pasalic, Senior Software Engineer worked with Adam

Conclusions & Lessons Learnt

Conclusions:

- Detected communities have real-life connection
- Bridge users of communities often have multiple skills or interests

Lessons Learnt:

- Computational thinking is critical to this project
- The better data set and visualisation, the better results

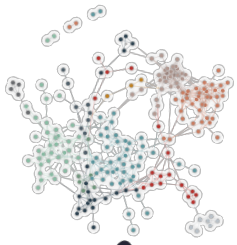
YouTube

SWS3001 Cluster2-9

DEEP IN TUBE

Team9

Wang Yisu
Li Yichen
Han Mingji
Yang Jiangling



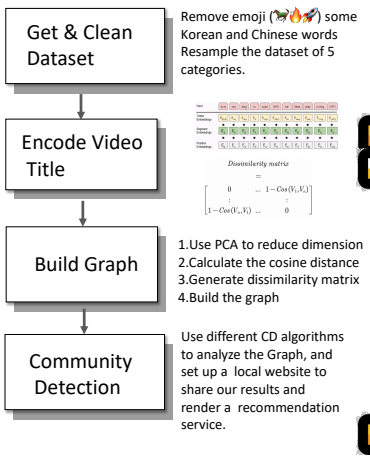
INTRODUCTION

When you search the videos in YouTube, you simply choose to watch what attract you. But if you dig deeper into YouTube, you will find more than you can imagine.
In this project, we will work on the video titles and discover the community in videos. And we hope to discover new insights based on the analysis of video titles.

Method

DATASET: YouTube dataset is collected from YouTube US of 5 categories.
NPL: We use BERT, a state of the art pretrained
Adjacent Matrix: Cosine Distance
CD Algorithm: Girvan-Newman, Hierarchy Clustering, Louvain algorithms

WORKFLOW



RESULTS & INSIGHTS

INSIGHTS 1

Videos about Donald Trump are always adjacent to some videos titles involved “High tax”, ”Tax Reform”, and "Google".

INSIGHTS 2

A video about a winter Olympics athlete is in community3(music), we find out that she is called beauty queen in Olympics and has a lot of intersection with the entertainment industry.

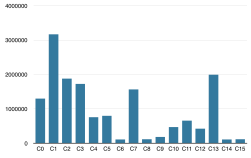
INSIGHTS 3

It’s a boundary case between sports and news. Conor McGregor, a super super star in UFC, who is known as ‘zuipao’ in China. This is the video about his appearance in court. So it become the boundary case.

INSIGHTS 4

Community 1: Interrogative questions beginning with How, What, Is.....which have the highest average views among all the group.

- Can you guess these WWE Superstars?
- Do You Hear Yanny or Laurel?
- How Trump makes extreme things look normal
- Is the Surface Book 2 a MacBook Pro Killer?



INSIGHTS 5

Most of these videos from community1 are uploaded by Vox, who has their own styles to name the videos.

1 Why female condoms are so hard to find	Vox
1 Why Wednesday's Super Blue Blood Moon will be a sight to behold	CBS Evening News
1 How this military explosive is poisoning American soil	Vox
1 How smart is today's artificial intelligence?	Vox
1 Why the Oscars love method actors	Vox
1 How Trump makes extreme things look normal	Vox
1 Which Countries Are About To Collapse?	NowThis World
1 Why does the Hawaii Kilauea volcano keep erupting?	CBC News
1 Can you guess these WWE Superstars?	WWE
1 It's not you. Commuting is bad for your health.	Vox
1 Why Stradivarius violins are worth millions	Vox

YouTube

BOOKTOMOVIE

Liu Yihong, Peng Jiawei, Tian Fengrui, Ye Qihao

website:

SWS3001
CLUSTER2-10

INTRODUCTION & OBJECTIVE

➢ **Introduction**:

- ***classify movies*** according to the audience's preferences on ***books***.

➢ **Objective**:

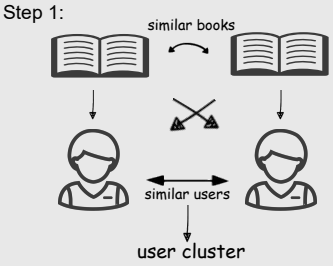
- Choose the **users of Douban** as our sample of movie audience and select the books they read as the representation of audiences' preferences;
- Run **Community Detection** on the ***User Graph*** with the books as reference, so as to classify the audience.
- Build the ***Movie Graph*** based on the preferences of each community.
- Observe and analyze the hidden relationship between different movies.

OVERVIEW & OVERALL WORKFLOW

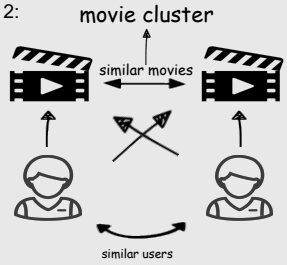
- **Input data:** 105 users' information, including their comments on the books and movies that they have read or watched
- **Labels:** general tags of books and movies
- **Workflow:**

DETAILS OF ALGORITHM & METHODS USED

Step 1:



Step 2:



Step 3:

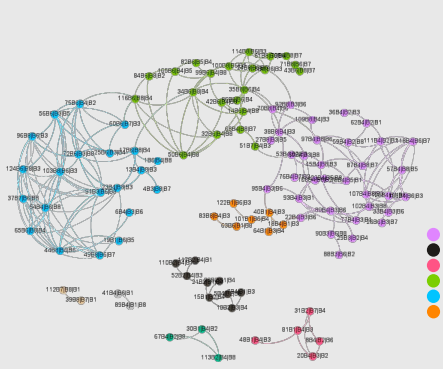
➢ Cosine similarity $$\text{similarity}(u,v) = \frac{|\min(N(u),N(v)|}{\sqrt{|N(u)||N(v)|}}$$

➢ Normalization $$x' = \frac{x - \min(X)}{\max(X) - \min(X)},\ x' \in [0,1]$$

Step 4:

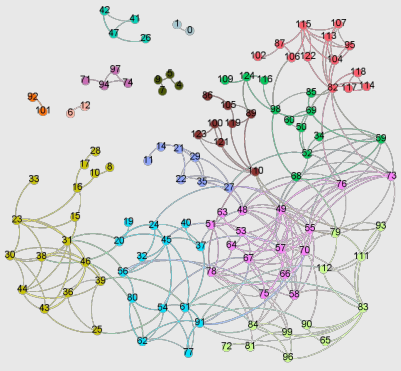
➢ Community detection:
- Girvan-Newman
- Louvain

RESULT OBTAINED & ANALYSIS



B: Book Categories
B1 = Literature
B2 = Popular
B3 = Culture
B4 = Philosophy
B5 = Military
B6 = Economic
B7 = Art
B8 = Science

C: User Clusters
C1: Economic
C2: Popular literature
C3: Philosophy
C4: Science
C5: Philosophy
C6: Literature



Family + Sci-Fi
Comedy + Romance C2+C3
High tense but not irritating C2
Classic
Suspense C4
Drama C6
Popular
Exciting Movie C4
Sci-Fi C5/C1
Popular Adventure
Popular comedy
Popular High rating
Life Philosophy Sense of Time

CONCLUSIONS&RECOMMENDATION

Conclusions

- Everyone loves happy ending movies - this is the common phenomenon of the Chinese audience. We all like a big reunion at the ending.
- The community that likes science books especially likes suspense.
- Those who reads philosophy and the economy like science fiction movies very much.
- People who prefer popular don't like stimulating movies, but they like movies with a tight narrative, such as Black Swan, Léon, You Are the Apple of My Eye.

Recommedation

- Like **economics books**: recommend sci-fi movies, such as The Butterfly Effect, "2012", Inception, we can also recommend movies of the compact plot or suspense type.

- Like **science books:** Recommendations should be careful, these people are strict, mainly recommend suspense, science fiction, stimulating plots, if it does not work ,we can also recommend comedy for insurance.
- Like **philosophy**: the same strict category, but they really like movies with philosophical insights and dramatic twists and turns, as well as science fiction movies.
- Like **literature**: Basically saying, they can watch everything. But slightly less like science fiction movies.

WHO CAN REPLACE CRISTIANO RONALDO?

SWS3001
CLUSTER2-11

Cao JiaPing, Liu Xiaoyu, Yang Junyi, Zhu Jiongli

INTRODUCTION

- A study over data of 160,000 football players around the world
- Detect invisible communities in players
- Different clustering methods tried
- Analysis over different subsets of the data

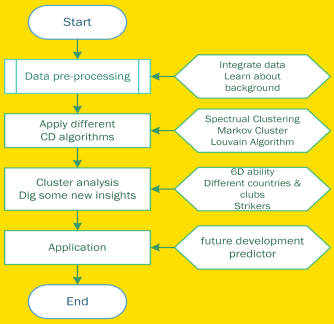
OBJECTIVE

- Verify what we already know about football players
- Discover new insights based on the graphs obtained from different subsets of data
- Compare the effects of different clustering methods

OVERVIEW

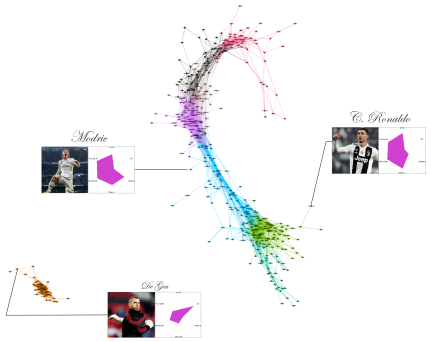
- Excellent players (with high value and wages) will form a cluster.
- Players with similar styles will be close to each other.
- Some players are so unique(excellent) that they will set themselves apart from a cluster.
- Some teams' degree of homogeneity will be relatively high, which leads to unbalanced performance on the field.

WORKFLOW



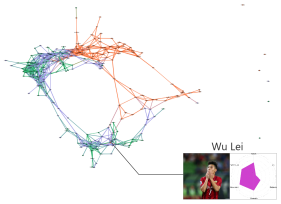
RESULTS

TOP 500 players:



players from different countries:

orange -- China
green -- Brazil
blue -- Germany

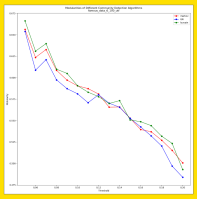


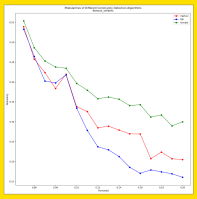
ALGORITHM COMPARISION

a. The performances of the three community detection algorithms on a relatively random and large data set (top salary data set) is about the same.

b. When the graph has some bridges, unique nodes, unbalanced number of edges among nodes, GN and markov are more likely to produce small communities in the cracks of some large communities. However, Louvain algorithm still results in balanced partitions.

c. As a conclusion, Louvain algorithms produce the best communities evaluated by modularity at most of the time.





CONCLUSION

a. Nobody can really replace Cristiano Ronaldo.

b. Clubs usually avoid homogeneity in each line.

c. It still takes a long way for Chinese players to be better.

d. If possible, Chelsea should sign Neymar to replace Hazard.

WELCOME TO VISIT OUR WEBSITE:
http://172.25.104.238:5000 (LAN)