# Broadcaster & Social Recommendation

Junru Li, Yimin Ma, Dianwen Jin, Huize Yan

Biomedical Engineering, SEAS

# Motivation



Live streaming App(part shown) in China
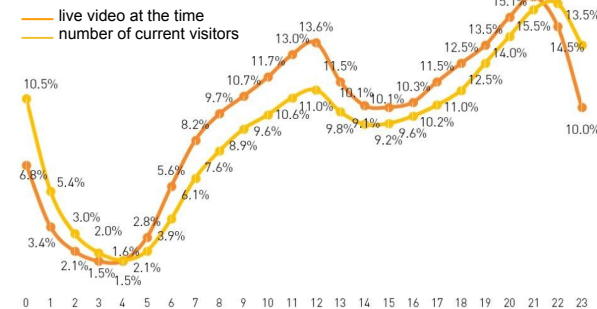


No. 1 and most active[1]
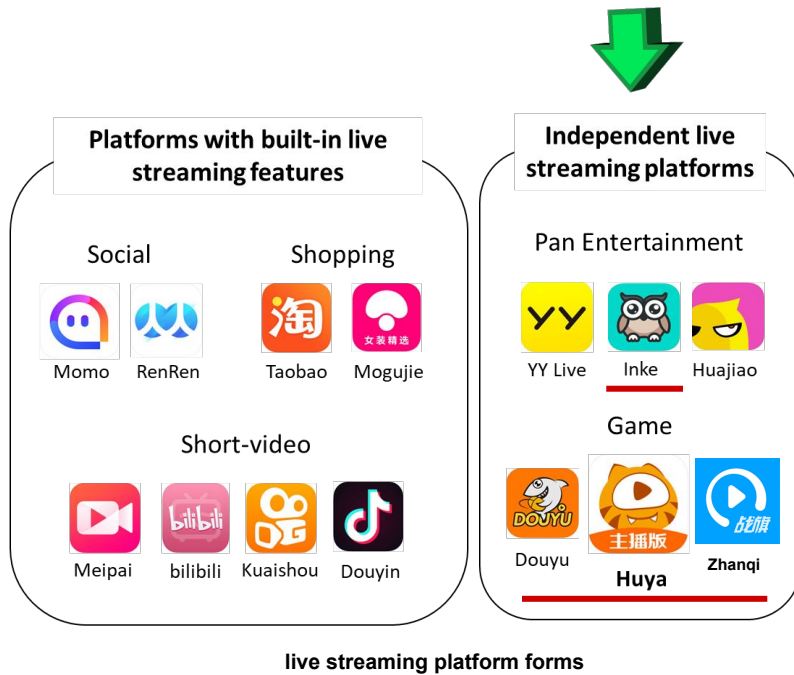game live streaming platform in China

86.7m
average
MAUs[2]

99mins
spent on app
per day per
active user[2]

2,600+
games
covered[3]

live video at the time
number of current visitors

**NEED:** provide recommendation based on viewers's likes

COLUMBIA ENGINEERING
The Fu Foundation School of Engineering and Applied Science

# Motivation

**Platforms with built-in live streaming features**

Social

Momo  RenRen

Shopping

Taobao  Mogujie

Short-video

Meipai  bilibili  Kuaishou  Douyin

**Independent live streaming platforms**

Pan Entertainment

YY Live  Inke  Huajiao

Game

Douyu  Huya  Zhanqi

**live streaming platform forms**

**Recommendations**

**3 similar streamer based on the category you like**

**#SOCIAL#
Find the one with similar liked streamers**

http://jl4939.pythonanywhere.com/

# Scrape Data

A big data service platform of streamers in China



A Python library for pulling data out of HTML and XML files

Scrape data of **2000** streamers.

| Rank | Streamer | Platform | Category | URL |
|---|---|---|---|---|
| 1 | kaerlol | 虎牙 | LOL | http://www.huya.com/kaerlol |
| 2 | xuxubaobao | 龙珠 | DNF | http://star.longzhu.com/xuxubaobao |
| 3 | 66666 | 熊猫 | PUBG | http://www.panda.tv/66666 |
| 4 | haddis | 虎牙 | 王者荣耀 | http://www.huya.com/haddis |
| 5 | 229346 | 斗鱼 | TRAVELING | https://www.douyu.com/229346 |
| 6 | dongxiaosa | 虎牙 | PUBG | http://www.huya.com/dongxiaosa |
| 7 | heigou | 虎牙 | PUBG | http://www.huya.com/heigou |
| 8 | xiaozhan | 虎牙 | PUBG | http://www.huya.com/xiaozhan |
| 9 | 24422 | 斗鱼 | PUBG | https://www.douyu.com/24422 |
| 10 | 44397 | 全民 | LOL | http://www.quanmin.tv/44397 |

**Check integrity:**
missing values; non-sensical variables…...

| 1668 | live?anchorid=68891368 | 企鹅电竞 | 绝地求生 全军出击 |
| 1677 | live?anchorid=2527039 | 企鹅电竞 | 绝地求生 全军出击 |
| 1687 | 155883 | 龙珠 | 全军出击 |
| 1722 | 222530 | 斗鱼 | 军事游戏 |
| 1756 | 2478985 | 斗鱼 | 全军出击 |

Code:

```python
import requests
from bs4 import BeautifulSoup
import csv
import time
from selenium import webdriver
import pandas as pd

# 使用Chrome浏览器创建浏览器对象
options = webdriver.ChromeOptions()
options.add_argument('--headless')

# Initialize the driver
driver = webdriver.Chrome(chrome_options=options)

# 使用get方法加载页面
driver.get("http://live.chinaz.com/anchor.html?platform=all&gameType=all&ti

# 展开页面，提取n个数据
for i in range(1,8):
    # class="zhubo_btn"是下一页按钮，click()是模拟点击
    driver.find_element_by_class_name('zhubo_btn').click()
    # 缓3s响应时间
    time.sleep(5)

soup = BeautifulSoup(driver.page_source, "lxml")

# Name & Platform
artist_name_list = soup.find_all(class_='info_avator')
artist_name = []
artist_platform = []

#从info_avator抓出了一堆文字，挑选span里面的是主播名，i是平台名
for m in range(len(artist_name_list)):
    try:
        artist_name.append(artist_name_list[m].find('span').contents[0])
        artist_platform.append(artist_name_list[m].find('i').contents[0])
    except:
        print(m+1)  # 出错数据的排名
        artist_name.append('Error Name')
        artist_platform.append('Error Platform')
```

Columbia ENGINEERING
The Fu Foundation School of Engineering and Applied Science

# Set up Database

Set up database to store users' infomation

```
mysql> DESCRIBE Users;
+----------+---------------+------+-----+---------+----------------+
| Field    | Type          | Null | Key | Default | Extra          |
+----------+---------------+------+-----+---------+----------------+
| uid      | int(11)       | NO   | PRI | NULL    | auto_increment |
| username | varchar(20)   | YES  |     | NULL    |                |
| password | varchar(50)   | YES  |     | NULL    |                |
| email    | varchar(50)   | YES  |     | NULL    |                |
| settings | varchar(500)  | YES  |     | NULL    |                |
+----------+---------------+------+-----+---------+----------------+
```

Code:

```
mysql> CREATE TABLE users (
    -> uid INT(11) AUTO_INCREMENT PRIMARY KEY,
    -> username VARCHAR(20),
    -> password VARCHAR(50),
    -> email VARCHAR(50),
    -> setting VARCHAR(32500));
Query OK, 0 rows affected, 1 warning (0.20 sec)

mysql> SHOW TABLES
    -> ;
+---------------------------------+
| Tables_in_DianwenJin$DataSci_final |
+---------------------------------+
| Users                           |
| users                           |
+---------------------------------+
2 rows in set (0.01 sec)
```

# Storage & Retrieval of Data

Functions related to database:
(1)Storage: register; add 3 favorite streamer
(2)Retrieval: login; show users' profile; social

| uid | username | password | email | settings |
|-----|----------|----------|-------|----------|
| 2 | test1 | $5$rounds=535000$.Iq7d6c848aSnWNL$yG8QsBmxYtqgeFfaiwfpQquZpUVsI.sc4b5YAkXVgA/ | test1@gmail.com | 47,15,31 |
| 3 | test2 | $5$rounds=535000$WJiAKWkg1CNDBN8K$uTTiqUrtwCkOtyN.rLmiKuDlazjHjO1v4/yJpojaQEC | test2@gmail.com | 175,225,221 |
| 5 | test3 | $5$rounds=535000$PWEzA8kd1ElTwqP6$mE3zoHyeQo2FoXCzpRf2a4Cv2Ap2OyRuEAgMrO5V707 | test3@gmail.com | 31,47,15 |
| 6 | hy2523 | $5$rounds=535000$./YmUVFdKLW8pq1S$LdasMuoGs9wQCxI/Hhqu2h51XghkkSzYyYWz9zCBfP6 | hy2523@columbia.edu | NULL |
| 7 | yhz1 | $5$rounds=535000$JENesZCDKzM0/UMM$DPi2TF7P.jubOdBVmxGONfsfEGWupRQ31tkmuke13ZD | yhz1@gmai.com | NULL |
| 8 | yhz2 | $5$rounds=535000$RHlOrSPj6Pn7wyzI$v9OjY99jN29pkqP2XcJ2N9bNxiKKYWH6Pp7QUJj7G36 | yhz2@gmai.com | NULL |
| 9 | yhz3 | $5$rounds=535000$HeRuSDzMZPMKMV7E$mZPGRX2Ogwu8YpOImpjmJRroRu3g6/DsRDgjj.Ofom/ | yhz3@gmail.com | NULL |
| 10 | diandian1 | $5$rounds=535000$aiAfp9jj1iJcOZqL$0XLVvLXQUM6FsOfvRbsvDHvxKboExYq4XDFjSCqYt0B | diandian1@163.com | 1014,680,4 |
| 11 | diandian2 | $5$rounds=535000$gkdvX/2AEpW1x/xC$dvQOOwrTY0/.T6q3dzaHUlArvWTAiIOI.Ox9hGkO0iO | diandian2@163.com | 505,870,1015 |
| 12 | try1 | $5$rounds=535000$nllAkkHfoEmh9OHG$eYRfr76A/ZbHg1cJOMBWFSj.JeMiqvDdH87KPELuEU6 | try1@163.com | NULL |
| 13 | diandian3 | $5$rounds=535000$l6lbQmI5545xzz2lE$qHT5i7cvEOgJX7PzHDnMjWCvhEuef99e7lhXu7Wlxw6 | diandian3@163.com | NULL |
| 14 | diandian4 | $5$rounds=535000$dGspmEVbmlAWXhiP$scduQlKskLnYk3tiiGwVeU/JmF5bIXyHsnxFe2YzKzA | diandian4@163.com | NULL |
| 15 | try2 | $5$rounds=535000$w9Yn7rw8vJ3xfiDU$FaDxmDJx390KTLMIENwtyC71Qls.MbYc79TN6Nm4o86 | try2@163.com | NULL |
| 16 | try3 | $5$rounds=535000$.FgkBsLI/NkwLDzI$ZtOaNOky4ZBI0Cb2.b53W28CkjahifQFGS6vnK3p.tA | try3@163.com | NULL |
| 17 | jdw1 | $5$rounds=535000$SRZEVlzEawKMKu6k$b7CAFAmPn3.8tVxwJm.q7cTAsGc7Z3ArD2n9Qlmgfw0 | jdw1@163.com | NULL |
| 18 | try4 | $5$rounds=535000$jLirjojKNNb.xEis$6rjPdd3KOWAYqSdX11Q7r/wphzdQXgN8C5uygsuke.5 | try4@163.com | NULL |
| 19 | test5 | $5$rounds=535000$SGy6LwWsHGf49M/.$NWkON5DJFvY1IqoArOwzl5UUWRPqKoA606K/8dG2480 | test5@gmail.com | 15,31,47 |
| 20 | test6 | $5$rounds=535000$ArtHRfHgNcWAutpH$BErXO3AO5DfmrOQfKKt5kKtiLRYxrWC34CV65bnTjH8/ | test6@gmail.com | 47,15,31 |
| 21 | 123123 | $5$rounds=535000$3Ztt6cj5QjaTbRpo$tSlubw23owC9SwozaAI19LACipTMGBnX0xp5ejkeP1. | 123123@163.com | 864,1184,11 |
| 22 | wanshabi | $5$rounds=535000$VoPHR7p8UGqknreE$Z.2/me6mg525EcuPmEJm5cW3xLGoDD/gfdP13HVpGi3 | wanzhihe@msu.edu | NULL |

Code of connection:

```python
import MySQLdb

def connection():
    conn = MySQLdb.connect(host="jl4939.mysql.pythonanywhere-services.com",...
                           user="jl4939", passwd="apma4990", db="jl4939$apma4990")
    c = conn.cursor()

    return c, conn
```

Storage & Retrieval code:

```python
############################## Profile Page ##############################
# -----------------
@app.route('/profile/')
@login_required
def profile_page():
    try:
        c, conn = connection()
        profile_username = str(session['username'])
        c.execute("SELECT * FROM users WHERE username='{0}'".format(str(session['username'])))
        profile_data = c.fetchall()

        profile_email = profile_data[0][3]
        profile_settings = profile_data[0][4]
        c.close()
        conn.close()

        return render_template("profile.html", profile_username=profile_username, ...
                        profile_email=profile_email, profile_settings=profile_settings)

    except Exception as e:
        return str(e)

######################## Add 3 Liked Streamer Page  ########################
# -----------------
@app.route('/add_streamer/', methods=['POST'])
@login_required
def add_streamer_page():
    try:
        temp_streamer_rank = []
        temp_streamer_rank.append(str(request.form['streamer_rank_1']))
        temp_streamer_rank.append(str(request.form['streamer_rank_2']))
        temp_streamer_rank.append(str(request.form['streamer_rank_3']))
        streamer_rank = list(set(temp_streamer_rank))
        sql_data = ','.join(streamer_rank)
        c, conn = connection()
        c.execute("UPDATE users SET settings='{0}' WHERE username='{1}';".format(sql_data, str(session['username'])))
        conn.commit()
        flash('Successfully added!')
        c.close()
        conn.close()

        return redirect(url_for('test'))
    except Exception as e:
        return str(e)
```

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

# Recommendation

## Recommend Streamer

| | Rank | Streamer | Platform | Category | URL |
|---|---|---|---|---|---|
| 0 | 1 | kaerlol | 虎牙 | LOL | http://www.huya.com/kaerlol |
| 1 | 2 | xuxubaobao | 龙珠 | DNF | http://star.longzhu.com/xuxubaobao |
| 2 | 3 | 66666 | 熊猫 | PUBG | http://www.panda.tv/66666 |
| 3 | 4 | haddis | 虎牙 | 王者荣耀 | http://www.huya.com/haddis |
| 4 | 5 | 229346 | 斗鱼 | TRAVELING | https://www.douyu.com/229346 |
| 5 | 6 | dongxiaosa | 虎牙 | PUBG | http://www.huya.com/dongxiaosa |

| Rank | Category | P | U | B | G |
|---|---|---|---|---|---|
| 1 | LOL | 0 | 0 | 0 | 0 |
| 2 | DNF | 0 | 0 | 0 | 0 |
| 3 | PUBG | 1 | 1 | 1 | 1 |
| 4 | 王者荣耀 | 0 | 0 | 0 | 0 |
| 5 | TRAVELING | 0 | 0 | 0 | 1 |

```python
def set_matrix(df, REF_VAR):
    for s in REF_VAR: df[s] = pd.Series([0 for _ in range(len(df))])
    colonnes = ['Category']
    for categorie in colonnes:
        for index, row in df.iterrows():
            if pd.isnull(row[categorie]): continue
            for s in row[categorie]:
                if s in REF_VAR: df.set_value(index, s, 1)
    return df

df_new = set_matrix(df,variables)
df_new
```

# Recommendation

## Recommend Streamer

```
nbrs = NearestNeighbors(n_neighbors=31, algorithm='auto', metric='euclidean').fit(X)
#nbrs.kneighbors(X)
xtest = df_new.iloc[id_entry].as_matrix(variables)
xtest = xtest.reshape(1, -1)
distances, indices = nbrs.kneighbors(xtest)
nbrs.kneighbors(xtest)

listed_streamer = indices[0][:]
listed_streamer
```

KNN

```
[3, '66666', '熊猫', 'PUBG', 'http://www.panda.tv/66666'],
[26, '20360', '斗鱼', 'PUBG', 'https://www.douyu.com/20360'],
[78, '10066', '熊猫', 'PUBG', 'http://www.panda.tv/10066'],
[98, '199211', '全民', 'PUBG', 'http://www.quanmin.tv/199211'],
[99, '80000', '熊猫', 'PUBG', 'http://www.panda.tv/80000'],
[190, '339610', '斗鱼', 'PUBG', 'https://www.douyu.com/339610'],
[342, 'lol77', '虎牙', 'PUBG', 'http://www.huya.com/lol77'],
[346, '687423', '斗鱼', 'PUBG', 'https://www.douyu.com/687423'],
[424, '652734', '熊猫', 'PUBG', 'http://www.panda.tv/652734'],
[505, '101367', '斗鱼', 'PUBG', 'https://www.douyu.com/101367']]
```

X

| Rank | Category | P | U | B | G |
|------|----------|---|---|---|---|
| 1 | LOL | 0 | 0 | 0 | 0 |
| 2 | DNF | 0 | 0 | 0 | 0 |
| 3 | PUBG | 1 | 1 | 1 | 1 |
| 4 | 王者荣耀 | 0 | 0 | 0 | 0 |
| 5 | TRAVELING | 0 | 0 | 0 | 1 |

# Recommendation

## User System and Social

# Recommendation

## User System and Social

**User/User based collaborative filtering**

- **0-1 Matrix**
- **Compute User/User Cosine similarity**

```python
# Generate the "0-1" Matrix and Compute Cosine Similarity
df = pd.DataFrame(0, index=list_all_username, columns=item)
length_of_users = 0

for i in list_all_username:
    for index, row in df.iterrows():
        if user_liked_list[length_of_users] != None:
            for s in user_liked_list[length_of_users].split(','):
                if s in row:
                    df.set_value(i, s, 1)
    length_of_users = length_of_users+1
# Compute User to User consine distance
user_user = 1-pairwise_distances(df, metric="cosine")
df_users = pd.DataFrame(user_user, columns=df.index, index=df.index)
```

In [30]: `df_users.head()`

Out[30]:

| | test1 | test2 | test3 | test4 | test5 | test6 | test7 | test8 | test9 | test10 | ... | test91 | test92 | test93 | test94 | test95 | test96 | test97 | test98 | test9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **test1** | 1.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| **test2** | 0.0 | 1.000000 | 0.333333 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.333333 | 0.0 | ... | 0.333333 | 0.0 | 0.0 | 0.333333 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| **test3** | 0.0 | 0.333333 | 1.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.333333 | 0.0 | ... | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| **test4** | 0.0 | 0.000000 | 0.000000 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| **test5** | 0.0 | 0.000000 | 0.000000 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |

Columbia ENGINEERING
The Fu Foundation School of Engineering and Applied Science

# Recommendation

## User System and Social

Because we do not have enough amount of data in our SQL database. So we randomly generate some user data to verify our method.

```
In [29]:  print(user_liked_list[5])
          print(user_liked_list[17])

          50,98,46
          50,98,71
```
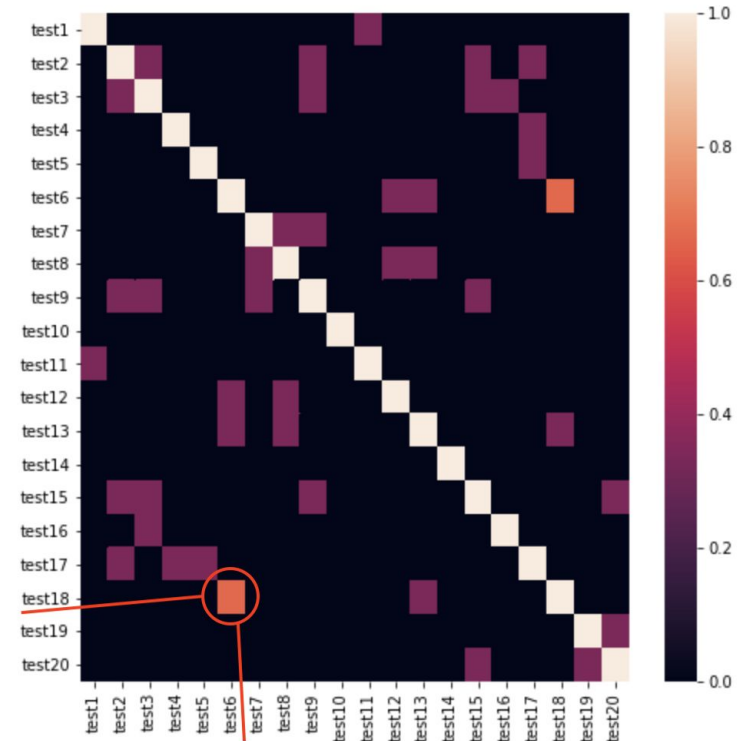
# Recommendation

## User System and Social

**Model Score**
- **Model compared with random guess about the similiarity**
- **Model accuracy is higher than random guess**
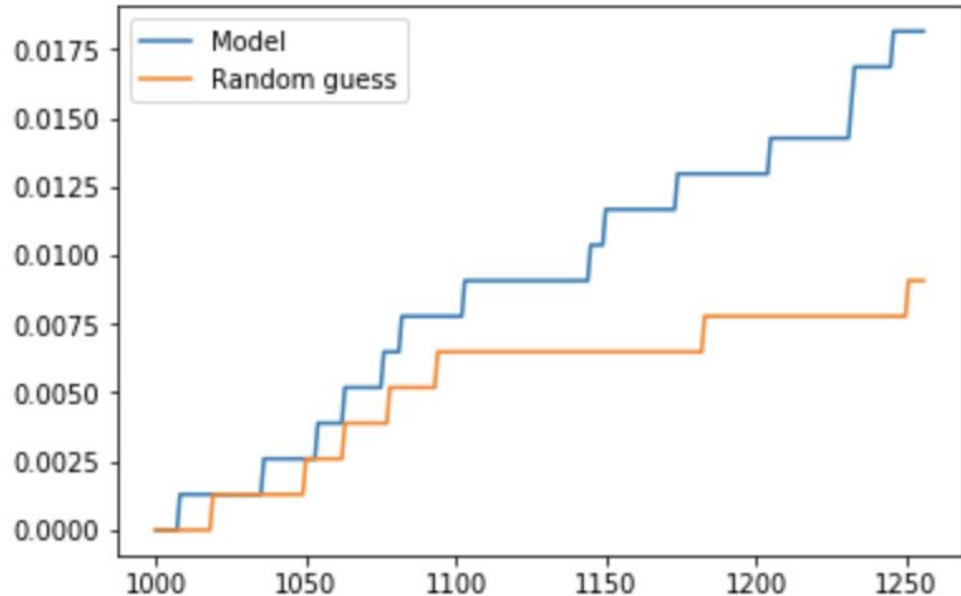
# Web Development

**1. Homepage**: show basic info of web

**3. Recommendation**: recommend suitable streamers



**2. Sign up & Login**: connect to database to read and write



**4. Social**: recommend the similar user

## Sign up & Login

# Web Development

## Jinja2 & HTML

### All Data

| Rank | Broadcaster | Platform | Category | URL |
|------|-------------|----------|----------|-----|
| 1 | 卡尔 | 虎牙 | 英雄联盟 | http://www.huya.com/kaerlol |
| 2 | 旭旭宝宝 | 龙珠 | 地下城与勇士 | http://star.longzhu.com/xuxubaobao |
| 3 | 刘杀鸡 | 熊猫 | 绝地求生 | http://www.panda.tv/66666 |
| 4 | 哈迪斯 | 虎牙 | 王者荣耀 | http://www.huya.com/haddis |
| 5 | 高冷男神钱小佳 | 斗鱼 | 户外 | https://www.douyu.com/229346 |
| 6 | Style丶董小飒-90327 | 虎牙 | 绝地求生 | http://www.huya.com/dongxiaosa |
| 7 | 黑店百地 | 虎牙 | 绝地求生 | http://www.huya.com/heigou |
| 8 | 小战 | 虎牙 | 绝地求生 | http://www.huya.com/xiaozhan |
| 9 | pigff | 斗鱼 | 绝地求生 | https://www.douyu.com/24422 |
| 10 | 英雄联盟职业联赛 | 全民 | 英雄联盟 | http://www.quanmin.tv/44397 |

```html
<div class="w3-container">
<table>
  <tr>
    <th>Rank</th>
    <th>Broadcaster</th>
    <th>Platform</th>
    <th>Category</th>
    <th>URL</th>
  </tr>

  {% for i in range(1950) %}
    <tr>
    <td>{{ data_matrix[0][i] }}</td>
    <td>{{ data_matrix[1][i] }}</td>
    <td>{{ data_matrix[2][i] }}</td>
    <td>{{ data_matrix[3][i] }}</td>
    <td><a href="{{ data_matrix[4][i] }}">{{ data_matrix[4][i] }}</a></td>
    </tr>
  {% endfor %}


</table>
</div>
```

Columbia Engineering
The Fu Foundation School of Engineering and Applied Science

# Thanks for listening!
# &
# Demo Time

Junru Li, Yimin Ma, Dianwen Jin, Huize Yan
Biomedical Engineering, SEAS