STAT 7008 Project Report

Data analysis of LEGO idea product

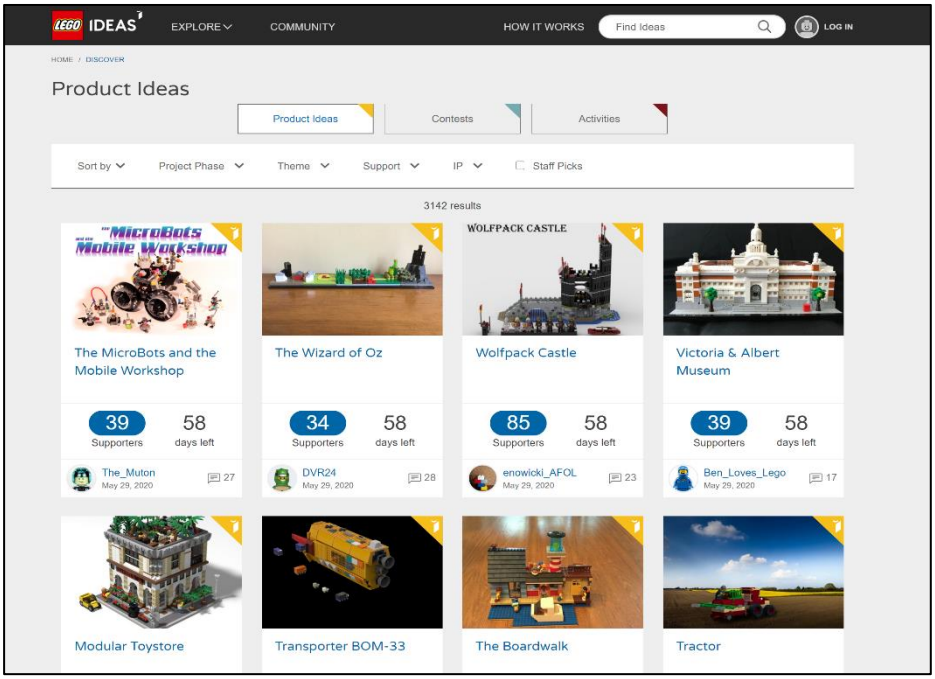
3035676389 WANG Yao

1. Introduction

Background:

The basic products in Lego company are the different types of bricks. People can combine them by their imagination to create different objects. The idea of the combination is very important. The Lego company has its own website for customers to share their new creations and vote for their favorite products. For these products have many focuses will be official produced for customers. But the customers have the feedback, the products recently they produced and featured are not the popular ones.

The website is like:



Not only the support bottom, but also many types of information, for example the comments in this website. We need to analyze them to give an opinion that which product is much more famous and what kind of style is now popular.

Object:

Analysis 1: according to the ideas and comments, find the most popular idea of product.
Analysis 2: according to the customers' comments, find the popular theme of Lego.
Analysis 3: According to the featured data, check whether the featured product is the popular product.

Brief conclusion summary:

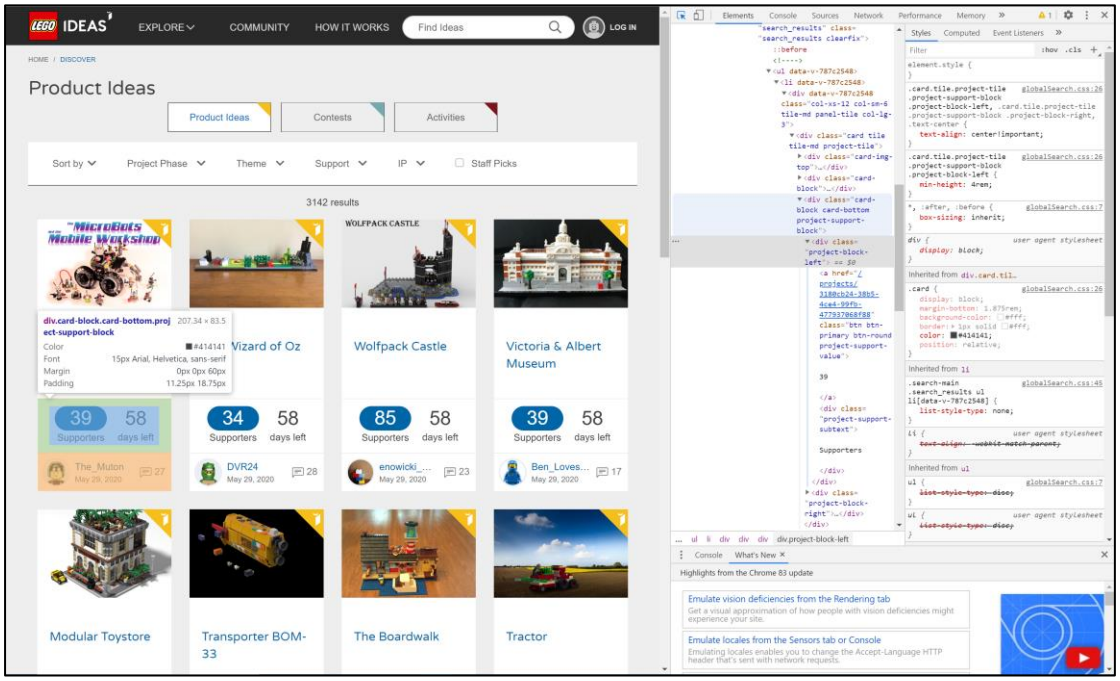
After analyzing the data from scraped three different resources. The featured products by the official are not the popular ones.

Brief problems summary:

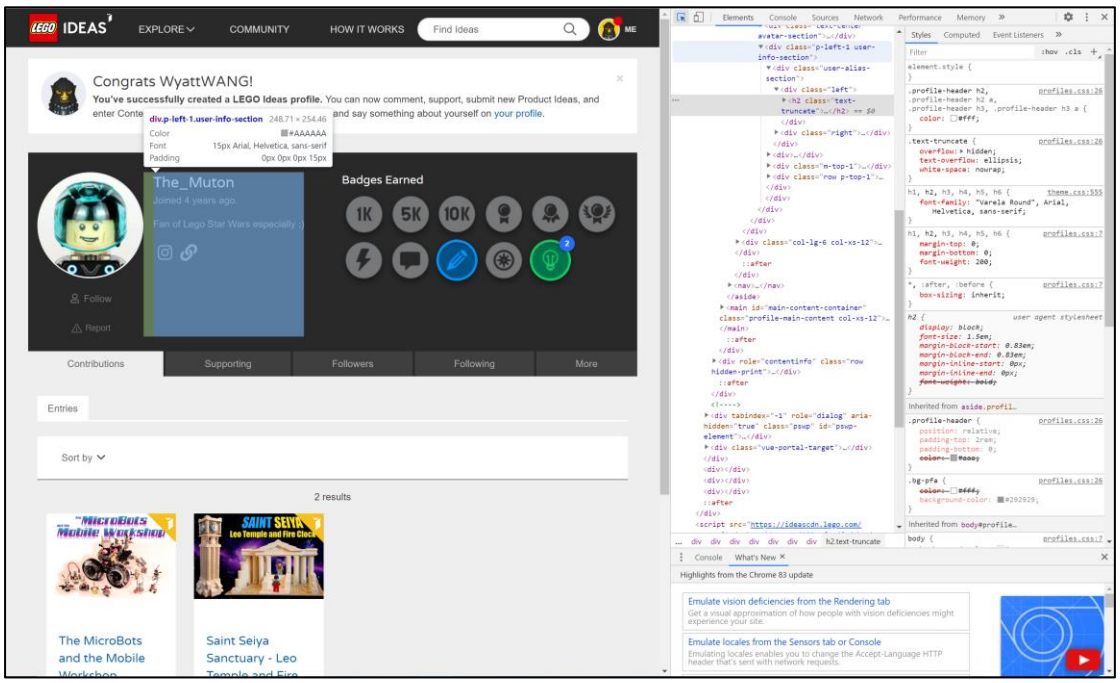
1. The datasets are limited, the Lego company has the access rules.
2. The limitation of the analysis package in python, when processing the comments of the customers.

2. Data resources

idea products page: we can scrap the data from the website. It totally has more than 1600 pages, and every page share the same structures.
From this page we can get the name of the idea, the number of the supports. The number of the comments. The authors' name.



author page (log-in required):
More detailed information about the authors. We can get the id of the authors and relevant ideas.



Data collections method:

1. Scrapping from the websites.
2. Aggregate the data from different resources, drop NA data.
3. Store the file in a csv file in local.

I need to setup a loop for all the pages. Because the limitation of the access, I cannot scrap all the data and then store them into csv file. I have to store them page by page. (more than 1400 pages in total for idea products)

.

The dataFrame: (Only 200/1600 data is scrapped from the website.)

```
In [286]: file
Out[286]:
                                                     0   ...          15
0                 The MicroBots and the Mobile Workshop  ...  2020/05/26
1                                      The Wizard of Oz  ...  2020/05/26
2                                       Wolfpack Castle  ...  2020/05/26
3                              Victoria & Albert Museum  ...  2020/05/26
4                                      Modular Toystore  ...  2020/05/26
...                                                  ...  ...         ...
2155                The Vintage Hello Kitty Brickheadz  ...  2020/05/26
2156                    Large Scale LEGO Octopus Piece  ...  2020/05/26
2157  Nereu Ramos Palace [ National Congress of Braz...  ...  2020/05/26
2158                                       Soda Machine  ...  2020/05/26
2159                          Concept Fuego GT V-12  ...  2020/05/26

[2160 rows x 16 columns]
```

One row of the dataframe:

```
In [288]: data.iloc[0]
Out[288]:
p_n                              The MicroBots and the Mobile Workshop
p_url              https://ideas.lego.com/projects/3180cb24-38b5-...
p_content          Here are the MicroBots!These little bots just ...
p_type                                                       project
author                                                     The_Muton
a_url                        https://ideas.lego.com/profile/The_Muton
a_id                             cf24ce1f-02a0-4b8b-83e9-7cc8203c6c7c
countOfSupport                                                   37
daysLeft                                                         59
commend                                                     comment
commend_num                                                      27
view                                                           view
view_num                                                        138
featured                                                          0
published_date                                           2020-05-29
collect_date                                             2020/05/26
Name: 0, dtype: object
```

The explanation of the columns in the Appendix

3. Analysis

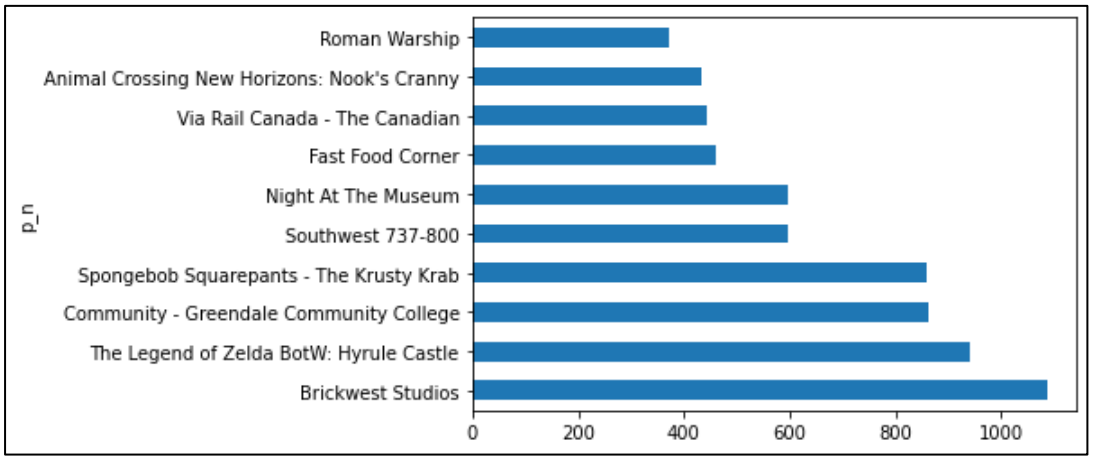
Analysis 1: find the most popular idea of product.

Method:

1. Combined the forms,
2. Count the number of the supports and sort by value
3. Plot the graph.

Result:

the Brickwest Studios is most popular.



Analysis 2, find the popular theme of Lego.

Method:

1. Process the data from the p_n and p_content to single word.
2. Filter the words, for example 'In','a','this'…
3. Count the frequency of the words, and sort by value.
4. Draw the word cloud graph.

Result:

the House and truck are popular theme. This result is in line with the analysis 1. We can find the types of product in analysis 1 are all in analysis2.

```
In [285]: wrc
Out[285]:
[('Brick', 14),
 ('Park', 14),
 ('Spaceship', 14),
 ('Modern', 14),
 ('Mech', 14),
 ('Adventure', 14),
 ('Robot', 14),
 ('Dragon', 14),
 ('Beach', 14),
 ('Temple', 15),
 ('Ship', 15),
 ('Mini', 15),
 ('Hot', 16),
 ('Vintage', 16),
 ('Station', 16),
 ('Zelda', 16),
 ('Medieval', 21),
 ('Set', 21),
 ('Of', 22),
 ('Island', 22),
 ('Tower', 23),
 ('Legend', 27),
 ('Train', 28),
 ('Castle', 31),
 ('Brickheadz', 33),
 ('Space', 37),
 ('City', 42),
 ('Car', 45),
 ('Truck', 52),
 ('House', 78)]
```

Analysis 3: whether the product featured is popular.

Column features: 0 means not featured by the official.

Out[299]:

p_n	commend_num	featured
Brickwest Studios	1086	0
The Legend of Zelda BotW: Hyrule Castle	942	0
Community - Greendale Community College	863	0
Spongebob Squarepants - The Krusty Krab	860	0
Southwest 737-800	597	0
Night At The Museum	596	0
Fast Food Corner	461	0
Via Rail Canada - The Canadian	444	0
Animal Crossing New Horizons: Nook's Cranny	432	0
Roman Warship	372	0

Result:
From the dataset, we can see the featured products are not the popular products.

4. Problems

1. The website limits the time of access. The problems will occur.

```
SSLError: HTTPSConnectionPool(host='ideas.lego.com', port=443): Max retries
exceeded with url: /search/global_search/ideas?
sort=most_recent&query=&idea_phase=idea_gathering_support&idea_phase=idea_a
chieved_support&idea_phase=idea_in_review&idea_phase=idea_idea_approved&ide
a_phase=idea_idea_not_approved&idea_phase=idea_on_shelves&idea_phase=idea_e
xpired_ideas&time=1590833458069 (Caused by SSLError(SSLError("bad
handshake: Error([('SSL routines', 'tls_process_server_certificate',
'certificate verify failed')])")))
```

2. The package to judge the comments is very limited. Some people will convey the opposite idea on purpose. Therefore it will influence the results.

5. Conclusion

The featured products by the official are not the popular ones.
We can get the real commends by aggregating the comments page. After we sort them, we can find the top 10 products with most commends are not featured by the official. Therefore, as the customers said, the featured system use a different way, and cannot reflect the demand of the customers.
Suggestions: The official can aggregate the comments and modify the featured system by the analysis results. And manufact the new product according to the results.

6. Appendix

The explanation of the columns in the dataFrame:

p_n: The name of the product
p_url: the url of the product
p_content: the details description of the product
p_type: the type of the product
author: the name of the author
a_url: the url of the author
a_id: the system uid of the author, can be used to follow the author.
countOfSupport: the number of the supporters
daysLeft: the voting day left
commend: whether the product has the commend
commend_num: the number of the commends
view: whether the product has been viewed by customers
view_num: the number of viewers.
featured: whether the product is suggested by the official group, where 0 is not featured,1 is featured.
published_date: the date of publishing this idea
collect_date: the date of collecting the data

7. References

1. The most recent idea website of LEGO: https://ideas.lego.com/search/global_search/ideas?idea_phase=idea_gathering_support&query=&sort=most_recent
2. Matplotlib, NumPy, pandas, request packages documentations.