**Classify Lego sets based on color themes and forecast the color dependency**

CAS 771 Project Proposal

Lai Wei

400145770

Introduction

Lego, one of the world's leading manufacturers of play materials, devotes to manufacturing the interlocking by bricks. In February 2015, Lego replaced Ferrari as Brand Finance's "world's most powerful brand". Nowadays, more and more customers choose to buy Lego sets online instead of going to the store.

Motivation

How do customers choose to buy their favourite Lego sets online? How does Lego firm classify customers’ choices? According to the Lego official website (Figure 1), it is clear that there are two primary categories: Themes and Interests. More secondary categories are also shown below the primary ones, such as Star Wars (Themes) and Animals (Interests). As customers, we can find that the choices are a little bit limited. Can customers make a better choice based on any other categories?

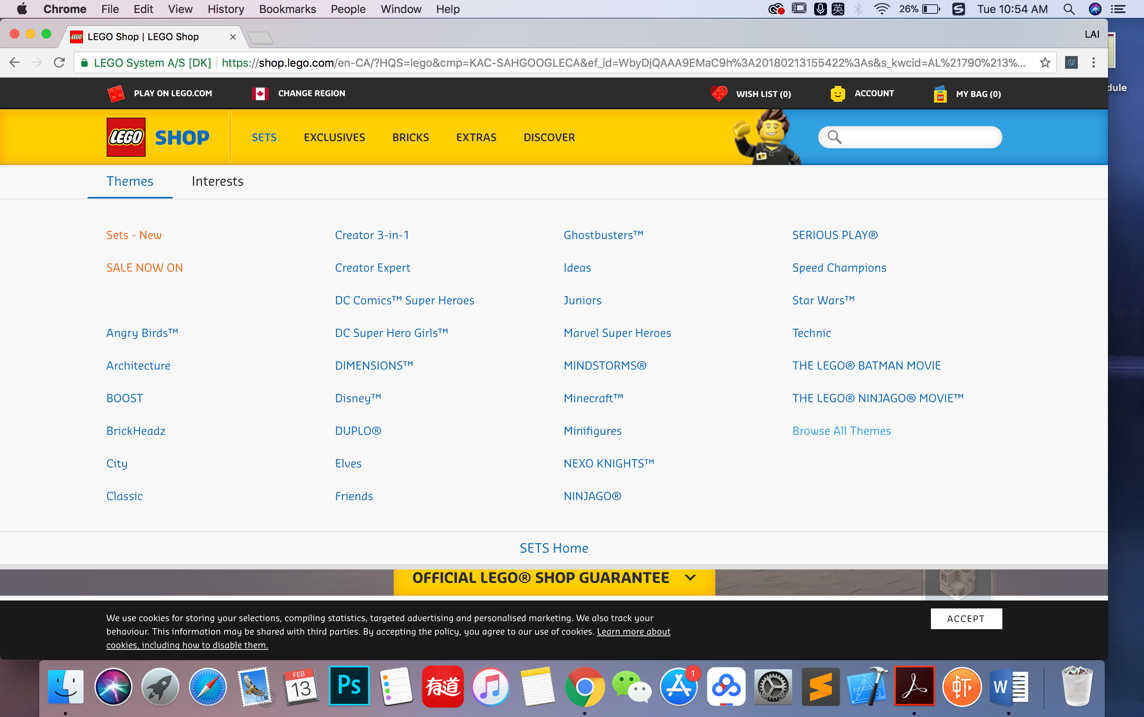


Figure 1 Lego official website

For each Lego brick, according to my knowledge, there are two dominant features which are the shape and the color respectively (shown in the Figure 2). But as we can see, the current selections that the Lego official website provides only focus on the shape of the Lego sets, ignoring the importance of the color, so what I plan to do is to classify Lego sets based on color themes to expand methods of choosing.

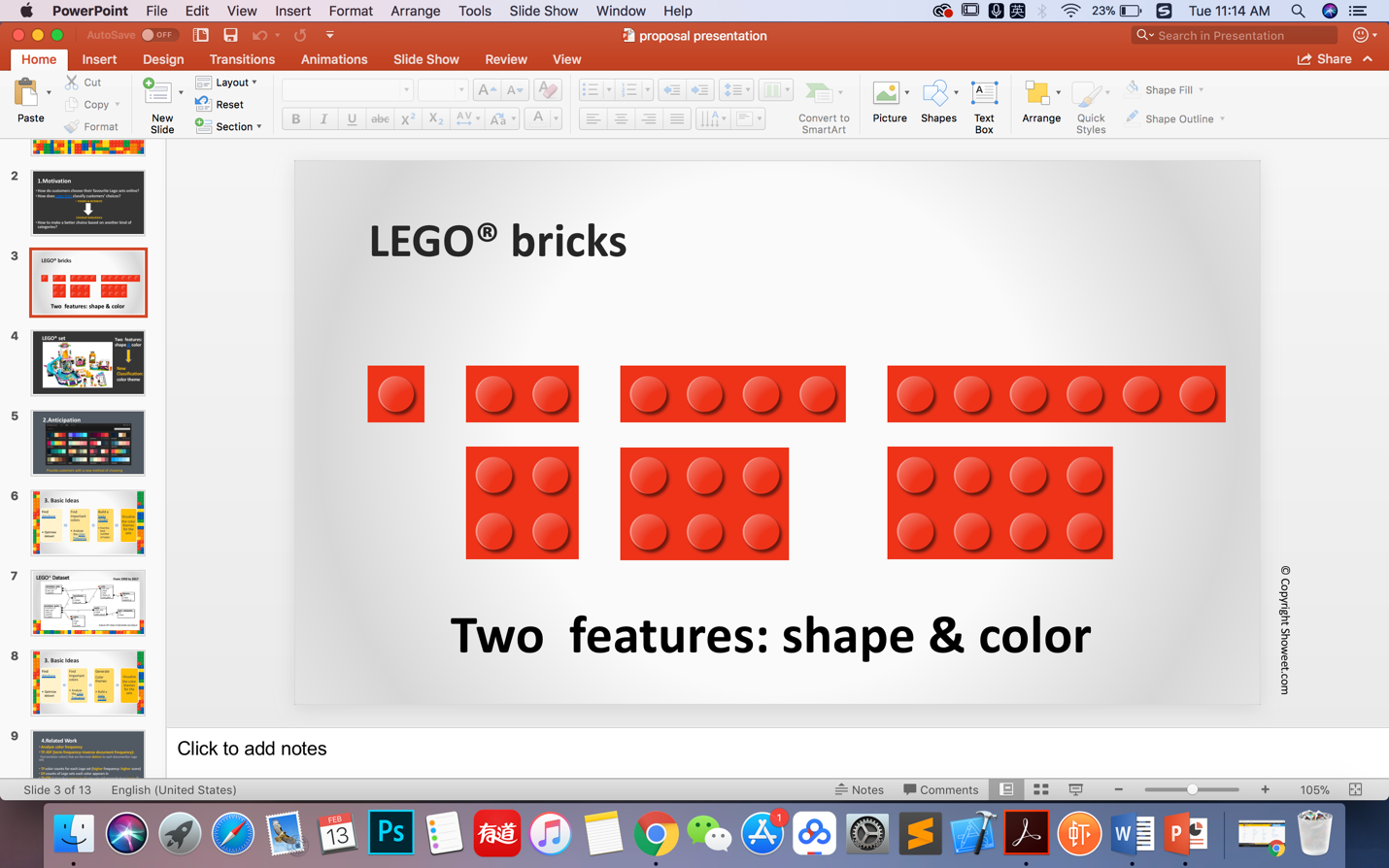


Figure 2 Lego brick features: shape and color

Dataset

There is an official Lego API – rebrickable, where I can get all Lego set information from 1950 till now. The dataset consists of eight CSV files (Figure 3) including themes, sets, colors and so on. In order to get the useful data for my project, I should do some optimization for the dataset.

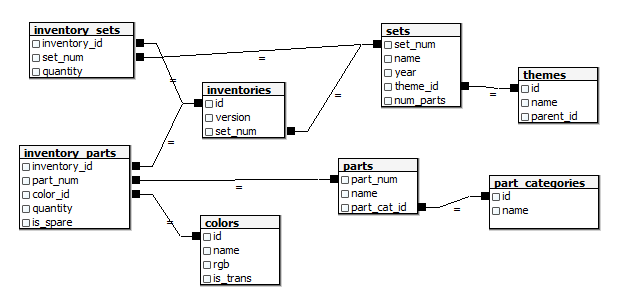


Figure 3 The schema of Lego dataset

Basic Ideas

**1.Color frequency analysis**

In order to find the most important color in the Lego sets (which color can define each Lego set), I plan to use TF-IDF (term frequency-inverse document frequency) method to analyze the color frequency. TF-IDF is a popular method in texting mining which aims to find words that are most distinct in the documents. In this case, words should be colors and documents are the Lego sets, which is a good fit. To be more specific, TF is color counts for each Lego set. For example, if the white color appears three times in a Lego set, the TF score of the white color is three. DF is counts of Lego sets each color appears in, which means that if the white color appears in five different Lego sets, DF score should be five for the white. If the white color appears in all the Lego sets, the TF-IDF score of it will be lower than those colors which only appear in one Lego set.

According to the results of the analysis, I think I can get more detailed information about the color in Lego sets, which can help me build the model for the color themes.

**2.Color theme generation**

Topic model is a good fit to implement the color themes for the Lego sets. Nathanael Aff in his project which is also related to color themes used five model validation methods to compare which one is the best, and what I try to do is to explore other validation methods to make the model more efficient. Furthermore, I plan to use python as my programming language because there are various python-based libraries for building topic models I can leverage.

**3.Color dependency analysis and forecast**

Not only try to find the relationship between the color and Lego sets, but I also plan to do the exploration about the colors themselves. The color dependency is another important relation in the Lego sets, so based on the color dependency analysis, I can do the prediction of the color dependency as well. For instance, if the white color appears in a Lego set, I can know any other colors which probably appear in the same Lego set.

**4.Visualization**

After the color analysis, visualization is also a very important part in my project. I plan to visualize the final result of the best model validation method to provide customers with a new way of choosing Lego sets. The figure below shows my anticipation of the visualization of the color themes of Lego sets.

Conclusion

Topic model is a good fit in the color theme analysis, but the exploration of finding more efficient model validation methods can keep going. Furthermore, the color dependency analysis and the visualization are also my core parts of my project.

References

1.Wikipedia. *Lego* from:

<https://en.wikipedia.org/wiki/Lego>

2.Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng.*Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey* Retrived on 12 Nov 2017 from:

<https://arxiv.org/abs/1711.04305>

3.Rebrickable. *Lego API* from:

<https://rebrickable.com/>

4.Nathanael Aff.*Finding Lego color themes with topic models* Retrived in October from:

<https://www.kaggle.com/nateaff/finding-lego-color-themes-with-topic-models/notebook>

5.Adobe Color CC. *most popular color themes*. from:

<https://color.adobe.com/explore/?filter=most-popular&time=month>

6. Wikipedia. *Topic Model* from:

<https://en.wikipedia.org/wiki/Topic_model>

7. Wikipedia. *TF-IDF* from:

<https://en.wikipedia.org/wiki/Tf%E2%80%93idf>