Exploring Taiwanese SNS PTT Travel Board: Text and Trend Analysis

Introduction

Japan is the most frequently visited country by Taiwanese travelers, even more so than China. Due to this frequent travel to Japan by Taiwanese people, I became curious whether Taiwanese travelers might start excluding popular tourist destinations (such as Tokyo and Kyoto) and begin exploring less popular ones.

Therefore, this study will analyze the Japan travel discussion board on Taiwan's community website PTT, which is similar to Reddit. I will observe if there is an increasing trend among PTT users towards visiting non-popular tourist spots. Additionally, I will use supervised classification methods like random forest and Naive Bayes, introduced in class, to classify text content about popular and non-popular tourist attractions. I will assess how well these models perform in correctly classifying the content. Finally, I will use a Semi-Supervised Model called Structural Topic Model (STM) to observe if it categorizes based on geographical regions or other criteria and attempt to categorize and label all travel experience articles accordingly.

Research Questions/ Methodology

According to the Visit Rate Ranking by Prefecture from Japan Tourism Statistics, I define "popular tourist destinations" as the following areas. I exclude Chiba Prefecture, ranked second, because I speculate its high visit rate is mainly due to Tokyo Disneyland and Narita International Airport, which are not typical tourist destinations:

- Tokyo
- Osaka
- Kyoto

Next, I will use Python to scrap the articles from the PTT Japan Travel forum regarding "experiences" from March 9, 2016, to June 18, 2024, and build an Excel database with the following variables:

- Title: Article title
- Type: If the article title contains keywords of our defined tourist destinations (Tokyo, Osaka, Kyoto), Type will be Y; otherwise, N.
- Time: Publication time of the article
- Like: Number of likes on the article
- Dislike: Number of dislikes on the article
- CommentNum: Number of comments on the article
- Content: Article content, initially stored in full for further processing

I will then use Excel Data and R language to address the following research questions:

Question 1 – Is there a increase in discussions about non-popular travel areas in Japan by time?

I will group articles by Type and calculate daily article publication counts to create a trend chart. Additionally, I will plot the proportion of articles discussing non-popular tourist areas against total daily publications to verify if there is indeed a yearly increase.

Question 2 – Can supervised classification models effectively distinguish content differences between discussions about popular and non-popular tourist destinations?

I will apply the random forest and Naive Bayes models introduced in class for classification. If the models perform below expectations, we will explore potential reasons for this.

Question 3 – Can differences in regional descriptions in articles be successfully classified using the designed STM?

I will use the Semi-Supervised Model Structural Topic Model (STM) to extract the most distinct classifications and attempt to label each topic ourselves. I will observe if these topics clearly categorize based on geographical regions.

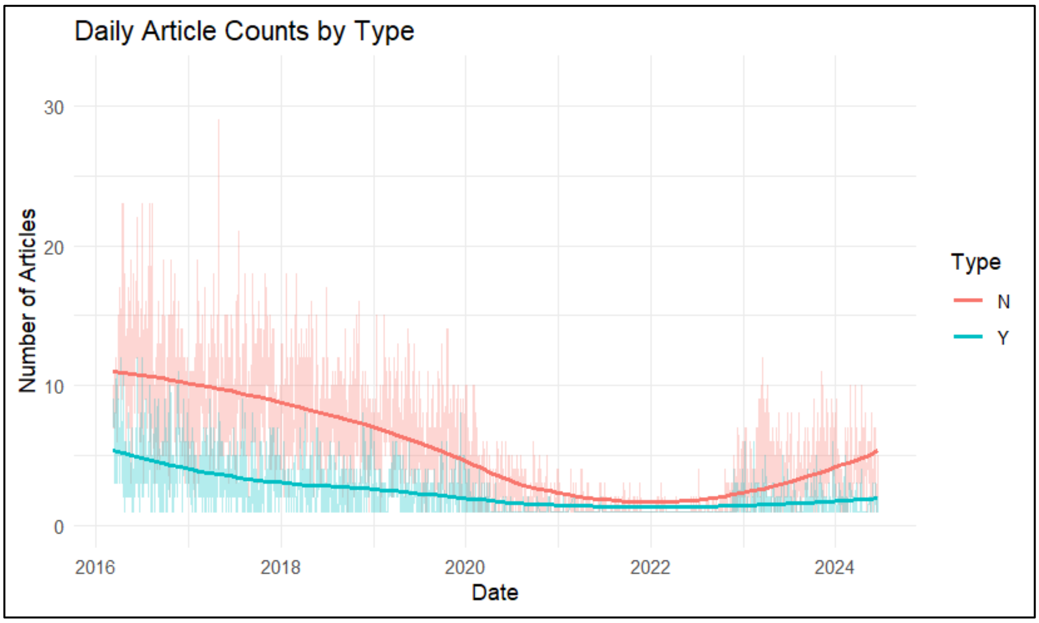
Result

I will now use the results from the R language analysis to answer the questions I listed earlier and provide insights and interpretations.

Question 1

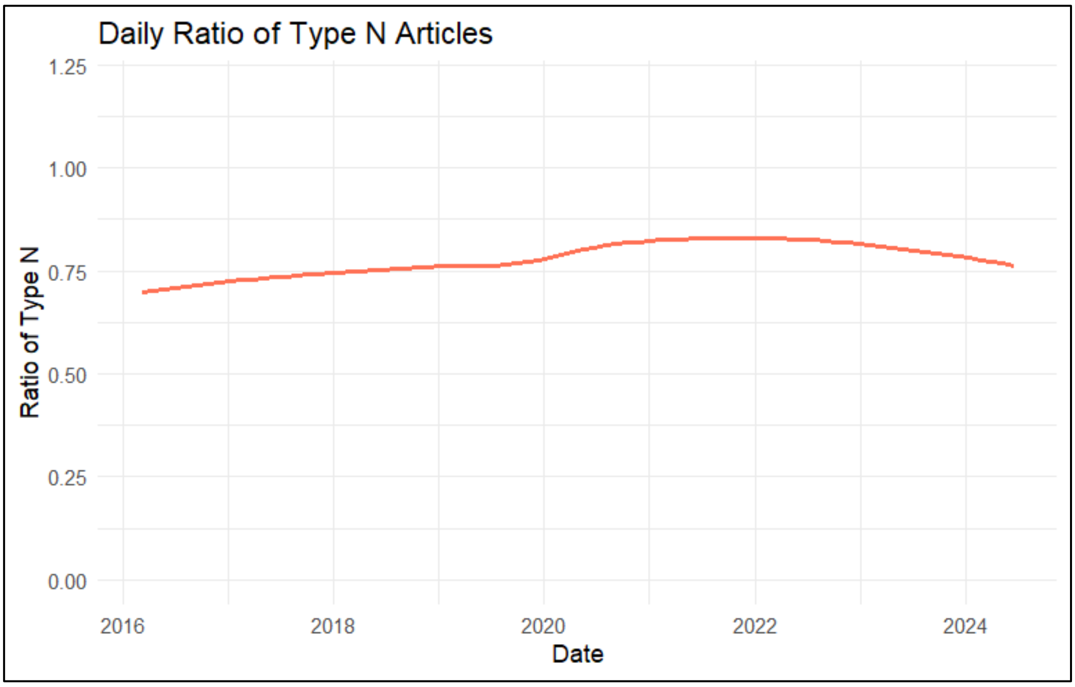
First, let's look at Figure 1, the word frequency chart. We can observe three insights:

1. **During COVID-19, the number of posts dropped significantly**
   The Japanese government started banning tourist visas on April 3, 2020, and gradually reopened for tourism in 2022. On October 11, 2022, they removed the daily entry limit and resumed visa-free travel for eligible countries (including Taiwan). We see that during the no-tourism period (2020-2022), the number of posts dropped significantly.
2. **The number of posts about non-popular tourist spots is greater than about popular tourist spots**
   According to the trend line, we can see that the number of posts about non-popular areas has always been higher than the number of posts about popular areas.
3. **There is a downward trend before COVID-19 and an upward trend after the pandemic**
   According to the trend line, we see a decline in both types of posts before the pandemic. After the pandemic, the trend rises for both, with a noticeable increase in posts about non-popular tourist spots.



▲ Figure 1

Next, we look at Figure 2. We can see that posts about non-popular tourist areas have been consistently high, showing an upward trend, although there was a slight drop in 2023. Overall, comparing the ratio from early 2016 to the latest data in June 2024, we see an increase in discussions about non-popular tourist areas because the initial ratio is lower than the most recent data although it is slight.



▲ Figure 2

Question 2

Since the original data is too large, I only used 0.5% of it to run the models. I used 80% as training data and 20% as testing data, using random forest and Naive Bayes models.

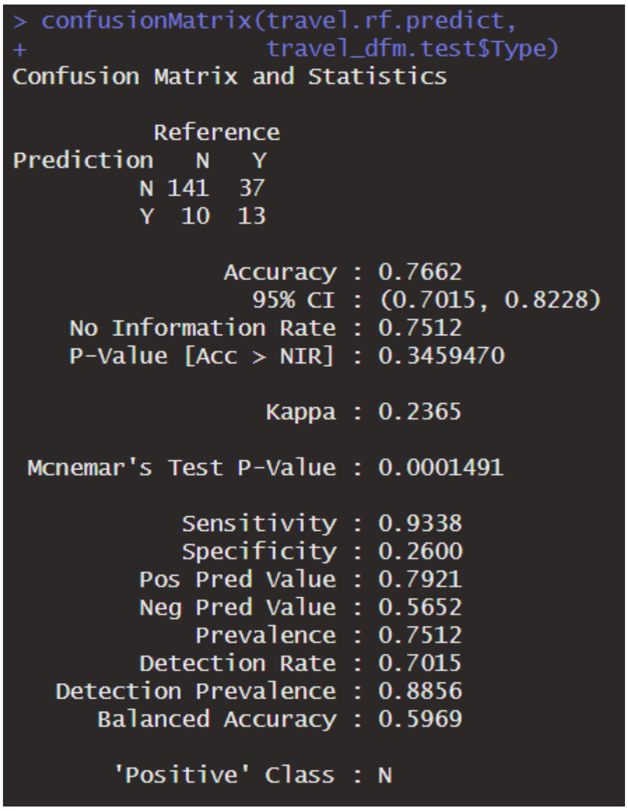
Figure 3 and Figure 4 show our results. Both models have an accuracy of about 70%. However, in the random forest model, the Kappa value is very low, close to random, and the specificity is also below 0.5. In contrast, Naive Bayes performs better, with a Kappa value close to moderate and specificity around 60%. Overall, Naive Bayes performs better than random forest, but both models have weaker performance in classifying Type Y.

There may be three reasons for these classification errors:

1. Model Choice:
   It's possible that these two models are not suitable for this text, and other models

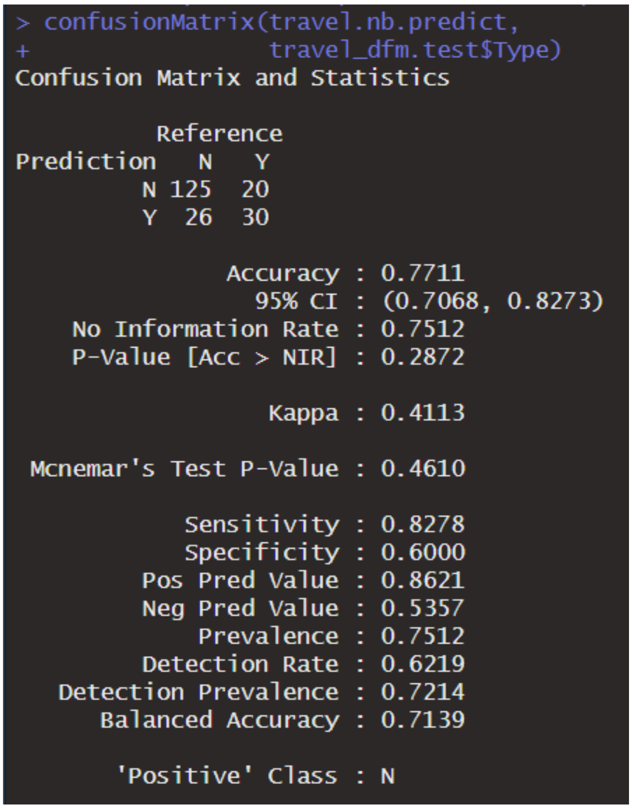
may be needed for better prediction.

2. High Text Diversity:
   Using `textstat_lexdiv()`, we found that 97% of the texts have a diversity above 0.5. High text diversity might make accurate predictions more difficult, as the high variability makes it hard to distinguish between Type Y and N.
3. Too Less Training Samples:
   Because the original data is very large, I only used a small portion for training and prediction. The small data size might also affect the model's accuracy.



```
> confusionMatrix(travel.rf.predict,
+                 travel_dfm.test$Type)
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 141  37
         Y  10  13

               Accuracy : 0.7662
                 95% CI : (0.7015, 0.8228)
    No Information Rate : 0.7512
    P-Value [Acc > NIR] : 0.3459470

                  Kappa : 0.2365

 Mcnemar's Test P-Value : 0.0001491

            Sensitivity : 0.9338
            Specificity : 0.2600
         Pos Pred Value : 0.7921
         Neg Pred Value : 0.5652
             Prevalence : 0.7512
         Detection Rate : 0.7015
   Detection Prevalence : 0.8856
      Balanced Accuracy : 0.5969

       'Positive' Class : N
```

▲ Figure 3 Random Forest Outcome



```
> confusionMatrix(travel.nb.predict,
+                 travel_dfm.test$Type)
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 125  20
         Y  26  30

               Accuracy : 0.7711
                 95% CI : (0.7068, 0.8273)
    No Information Rate : 0.7512
    P-Value [Acc > NIR] : 0.2872

                  Kappa : 0.4113

 Mcnemar's Test P-Value : 0.4610

            Sensitivity : 0.8278
            Specificity : 0.6000
         Pos Pred Value : 0.8621
         Neg Pred Value : 0.5357
             Prevalence : 0.7512
         Detection Rate : 0.6219
   Detection Prevalence : 0.7214
      Balanced Accuracy : 0.7139

       'Positive' Class : N
```

▲ Figure 4 Naive Bayes Outcome

Question 3

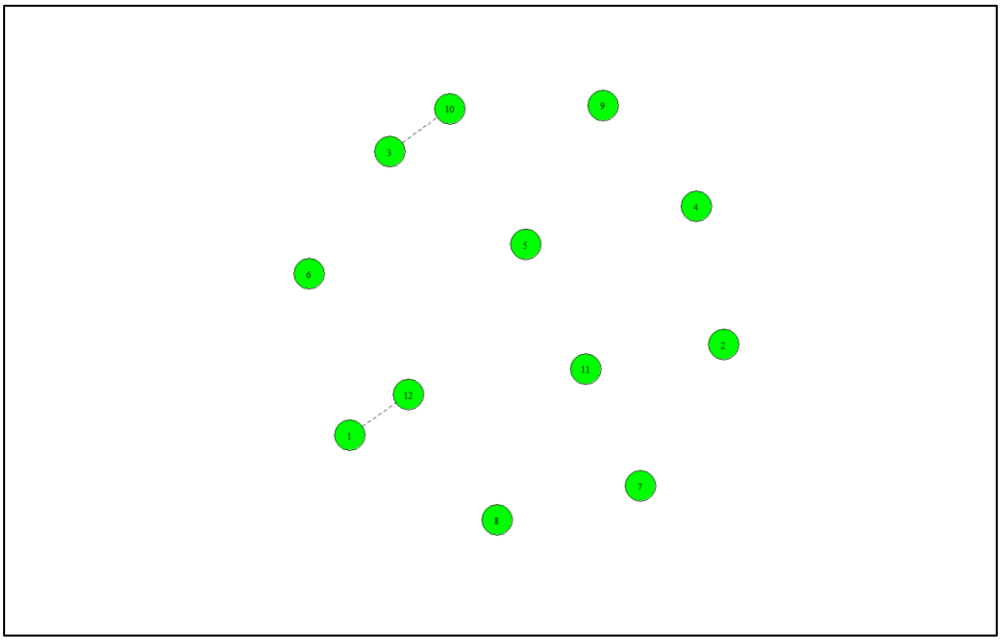
Can differences in regional descriptions in articles be successfully classified using the designed STM?

I set several different STM models using different formulas and finally selected the following model:

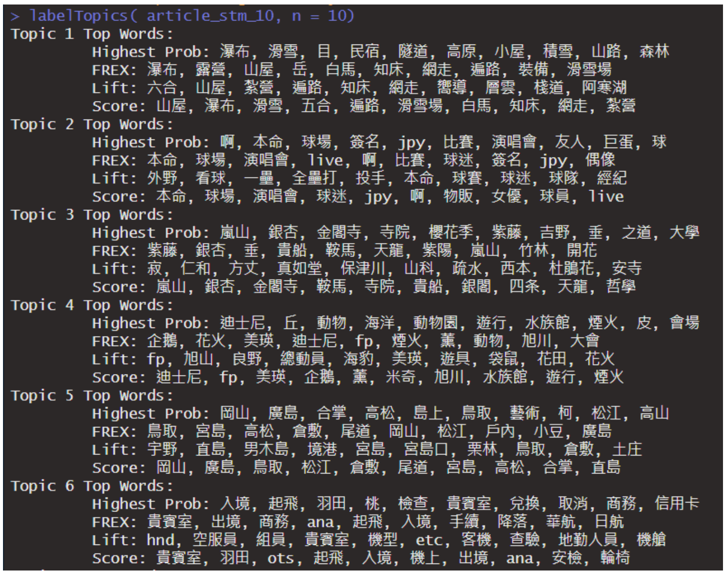
- k = 12

- prevalence =~ number of comments + Type + s (Time)

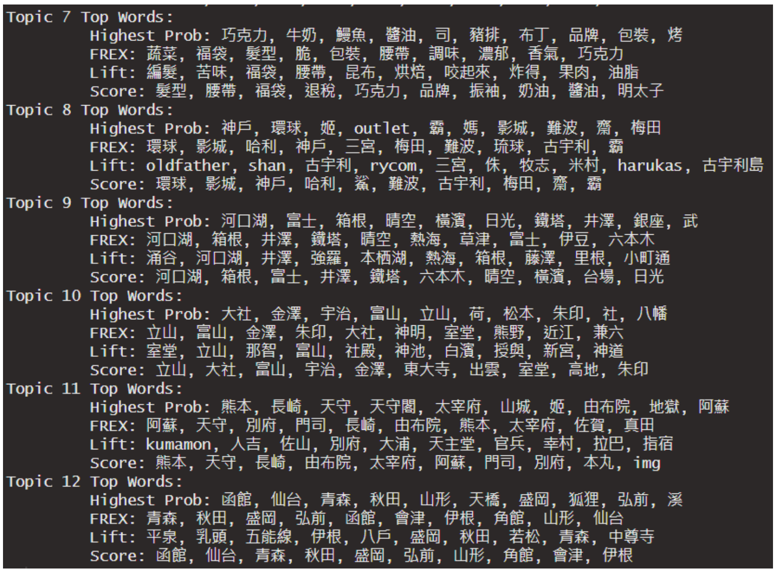
This model performed better compared to others. The correlation plot (Figure 5) also showed clear clustering effects.



▲ Figure 5 Correlation Map



```
> labelTopics( article_stm_10, n = 10)
Topic 1 Top Words:
 	 Highest Prob: 瀑布, 滑雪, 目, 民宿, 隧道, 高原, 小屋, 積雪, 山路, 森林
 	 FREX: 瀑布, 露營, 山屋, 岳, 白馬, 知床, 網走, 遍路, 裝備, 滑雪場
 	 Lift: 六合, 山屋, 紮營, 遍路, 知床, 網走, 嚮導, 層雲, 棧道, 阿寒湖
 	 Score: 山屋, 瀑布, 滑雪, 五合, 遍路, 滑雪場, 白馬, 知床, 網走, 紮營
Topic 2 Top Words:
 	 Highest Prob: 啊, 本命, 球場, 簽名, jpy, 比賽, 演唱會, 友人, 巨蛋, 球
 	 FREX: 本命, 球場, 演唱會, live, 啊, 比賽, 球迷, 簽名, jpy, 偶像
 	 Lift: 外野, 看球, 一壘, 全壘打, 投手, 本命, 球賽, 球迷, 球隊, 經紀
 	 Score: 本命, 球場, 演唱會, 球迷, jpy, 啊, 物販, 女優, 球員, live
Topic 3 Top Words:
 	 Highest Prob: 嵐山, 銀杏, 金閣寺, 寺院, 櫻花季, 紫藤, 吉野, 垂, 之道, 大學
 	 FREX: 紫藤, 銀杏, 垂, 貴船, 鞍馬, 天龍, 紫陽, 嵐山, 竹林, 開花
 	 Lift: 寂, 仁和, 方丈, 真如堂, 保津川, 山科, 疏水, 西本, 杜鵑花, 安寺
 	 Score: 嵐山, 銀杏, 金閣寺, 鞍馬, 寺院, 貴船, 銀閣, 四条, 天龍, 哲學
Topic 4 Top Words:
 	 Highest Prob: 迪士尼, 丘, 動物, 海洋, 動物園, 遊行, 水族館, 煙火, 皮, 會場
 	 FREX: 企鵝, 花火, 美瑛, 迪士尼, fp, 煙火, 薰, 動物, 旭川, 大會
 	 Lift: fp, 旭山, 良野, 總動員, 海豹, 美瑛, 遊具, 袋鼠, 花田, 花火
 	 Score: 迪士尼, fp, 美瑛, 企鵝, 薰, 米奇, 旭川, 水族館, 遊行, 煙火
Topic 5 Top Words:
 	 Highest Prob: 岡山, 廣島, 合掌, 高松, 島上, 鳥取, 藝術, 柯, 松江, 高山
 	 FREX: 鳥取, 宮島, 高松, 倉敷, 尾道, 岡山, 松江, 戶內, 小豆, 廣島
 	 Lift: 宇野, 直島, 男木島, 境港, 宮島, 宮島口, 栗林, 鳥取, 倉敷, 土庄
 	 Score: 岡山, 廣島, 鳥取, 松江, 倉敷, 尾道, 宮島, 高松, 合掌, 直島
Topic 6 Top Words:
 	 Highest Prob: 入境, 起飛, 羽田, 桃, 檢查, 貴賓室, 兌換, 取消, 商務, 信用卡
 	 FREX: 貴賓室, 出境, 商務, ana, 起飛, 入境, 手續, 降落, 華航, 日航
 	 Lift: hnd, 空服員, 組員, 貴賓室, 機型, etc, 客機, 查驗, 地勤人員, 機艙
 	 Score: 貴賓室, 羽田, ots, 起飛, 入境, 機上, 出境, ana, 安檢, 輪椅
```



```
Topic 7 Top Words:
 	 Highest Prob: 巧克力, 牛奶, 鰻魚, 醬油, 司, 豬排, 布丁, 品牌, 包裝, 烤
 	 FREX: 蔬菜, 福袋, 髮型, 脆, 包裝, 腰帶, 調味, 濃郁, 香氣, 巧克力
 	 Lift: 編髮, 苦味, 福袋, 腰帶, 昆布, 烘焙, 咬起來, 炸得, 果肉, 油脂
 	 Score: 髮型, 腰帶, 福袋, 退稅, 巧克力, 品牌, 振袖, 奶油, 醬油, 明太子
Topic 8 Top Words:
 	 Highest Prob: 神戶, 環球, 姬, outlet, 霸, 媽, 影城, 難波, 齋, 梅田
 	 FREX: 環球, 影城, 哈利, 神戶, 三宮, 梅田, 難波, 琉球, 古宇利, 霸
 	 Lift: oldfather, shan, 古宇利, rycom, 三宮, 侏, 牧志, 米村, harukas, 古宇利島
 	 Score: 環球, 影城, 神戶, 哈利, 鯊, 難波, 古宇利, 梅田, 齋, 霸
Topic 9 Top Words:
 	 Highest Prob: 河口湖, 富士, 箱根, 晴空, 橫濱, 日光, 鐵塔, 井澤, 銀座, 武
 	 FREX: 河口湖, 箱根, 井澤, 鐵塔, 晴空, 熱海, 草津, 富士, 伊豆, 六本木
 	 Lift: 涌谷, 河口湖, 井澤, 強羅, 本栖湖, 熱海, 箱根, 藤澤, 里根, 小町通
 	 Score: 河口湖, 箱根, 富士, 井澤, 鐵塔, 六本木, 晴空, 橫濱, 台場, 日光
Topic 10 Top Words:
 	 Highest Prob: 大社, 金澤, 宇治, 富山, 立山, 荷, 松本, 朱印, 社, 八幡
 	 FREX: 立山, 富山, 金澤, 朱印, 大社, 神明, 室堂, 熊野, 近江, 兼六
 	 Lift: 室堂, 立山, 那智, 富山, 社殿, 神池, 白濱, 授與, 新宮, 神道
 	 Score: 立山, 大社, 富山, 宇治, 金澤, 東大寺, 出雲, 室堂, 高地, 朱印
Topic 11 Top Words:
 	 Highest Prob: 熊本, 長崎, 天守, 天守閣, 太宰府, 山城, 姬, 由布院, 地獄, 阿蘇
 	 FREX: 阿蘇, 天守, 別府, 門司, 長崎, 由布院, 熊本, 太宰府, 佐賀, 真田
 	 Lift: kumamon, 人吉, 佐山, 別府, 大浦, 天主堂, 官兵, 幸村, 拉巴, 指宿
 	 Score: 熊本, 天守, 長崎, 由布院, 太宰府, 阿蘇, 門司, 別府, 本丸, img
Topic 12 Top Words:
 	 Highest Prob: 函館, 仙台, 青森, 秋田, 山形, 天橋, 盛岡, 狐狸, 弘前, 溪
 	 FREX: 青森, 秋田, 盛岡, 弘前, 函館, 會津, 伊根, 角館, 山形, 仙台
 	 Lift: 平泉, 乳頭, 五能線, 伊根, 八戶, 盛岡, 秋田, 若松, 青森, 中尊寺
 	 Score: 函館, 仙台, 青森, 秋田, 盛岡, 弘前, 山形, 角館, 會津, 伊根
```

▲ Figure 6 / Figure 7 label Topic outcome

Figure 6 and Figure 7 are the final results. I tried to focus on the words under "FREX" and labeled the topics accordingly:

1. Nature Travel & Hokkaido

- ♦ FREX: Waterfall, Camping, Mountain Hut, Take, Hakuba, Shiretoko, Abashiri, Pilgrimage, Gear, Ski Resort
- ♦ This topic frequently mentions nature spots and Hokkaido place names, so I named it "Nature Travel & Hokkaido".

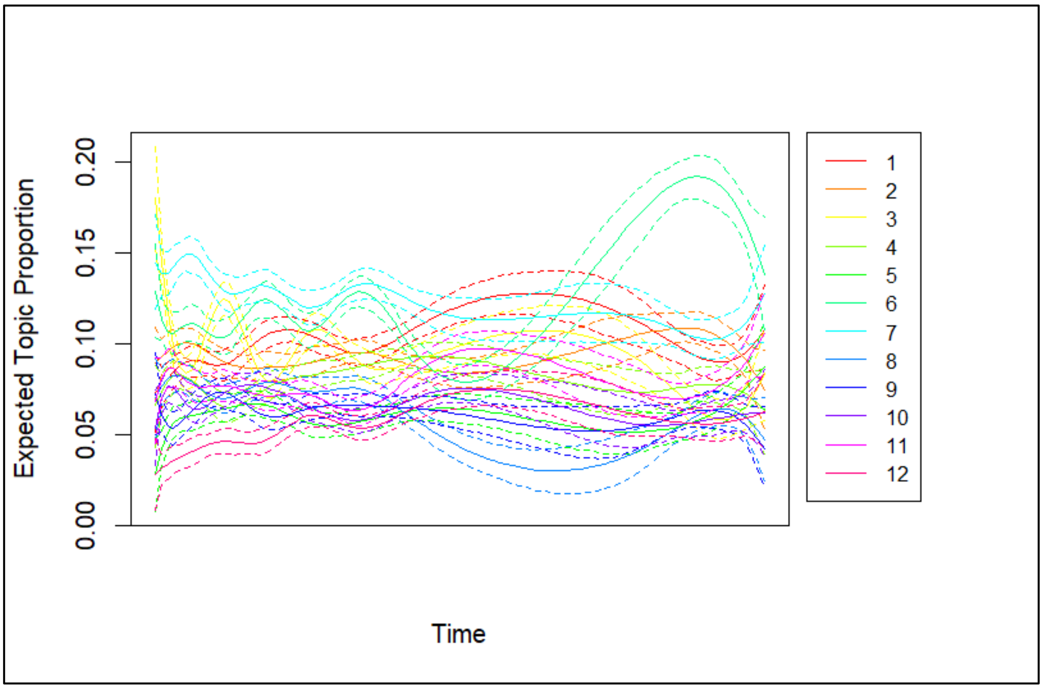
2. **Sports & Entertainment**
   - ♦ FREX: Favorite, Ballpark, Concert, Live, Ah, Game, Fan, Autograph, JPY, Idol
   - ♦ This topic includes words related to sports games, concerts, and fan activities, so it’s labeled "Sports & Entertainment".
3. **Kyoto Flower Viewing**
   - ♦ FREX: Wisteria, Ginkgo, Hanging, Kibune, Kurama, Tenryu, Hydrangea, Arashiyama, Bamboo Forest, Blooming
   - ♦ This topic has many words related to Kyoto places and flowers, so it os labeled "Kyoto Flower Viewing".
4. **Theme Parks**
   - ♦ FREX: Penguin, Fireworks, Biei, Disney, FP, Fireworks, Lavender, Animal, Asahikawa, Conference
   - ♦ This category includes words about theme parks like Disney or zoos.
5. **Shikoku & Chugoku Region**
   - ♦ FREX: Tottori, Miyajima, Takamatsu, Kurashiki, Onomichi, Okayama, Matsue, Indoor, Shodoshima, Hiroshima
   - ♦ This topic has many place names from Shikoku and Chugoku regions, so I labeled it "Shikoku & Chugoku Region".
6. **Airports & Airlines**
   - ♦ FREX: Lounge, Departure, Business, ANA, Takeoff, Immigration, Procedures, Landing, China Airlines, Japan Airlines
   - ♦ This topic includes many airport-related terms and airline names, so it’s labeled " Airports & Airlines".
7. **Food & Shopping**
   - ♦ FREX: Vegetables, Lucky Bag, Hairstyle, Crisp, Packaging, Belt, Seasoning, Rich, Aroma, Chocolate
   - ♦ This topic revolves around food descriptions and shopping experiences, so it is named "Food & Shopping".
8. **Osaka & Okinawa**
   - ♦ FREX: Universal, Movie City, Harry, Kobe, Sannomiya, Umeda, Namba, Ryukyu, Kouri, Boss
   - ♦ This topic has many Osaka and nearby Kobe area place names, as well as Okinawa names, so it is named "Osaka & Okinawa".

9. **Tokyo & Surroundings**
    - FREX: Kawaguchi Lake, Hakone, Iizawa, Tower, Sky Tree, Atami, Kusatsu, Fuji, Izu, Roppongi
    - This category has many Tokyo places (like Sky Tree, Roppongi) and nearby areas (like Kawaguchi Lake, Yokohama), so it’s labeled "Tokyo & Surroundings".
10. **Hokuriku Region**
    - FREX: Tateyama, Toyama, Kanazawa, Seal Stamp, Taisha, Shinmei, Murodo, Kumano, Omi, Kenroku
    - This topic has many Hokuriku place names, so it is named "Hokuriku Region".
11. **Kyushu Region**
    - FREX: Aso, Tenshu, Beppu, Moji, Nagasaki, Yufuin, Kumamoto, Dazaifu, Saga, Sanada
    - This topic has many Kyushu place names, so it’s labeled "Kyushu Region".
12. **Tohoku Region**
    - FREX: Aomori, Akita, Morioka, Hirosaki, Hakodate, Aizu, Ine, Kakunodate, Yamagata, Sendai
    - This topic mainly includes Tohoku place names, so it is named "Tohoku Region".

In summary, we can see that using STM clustering, the content of the articles shows region-based themes. Additionally, through the content of STM, we can discover the following insights:

1. **Not All Topics Are Region-Based**For example, Topic 6 (Airports) and Topic 7 (Food & Shopping) do not have obvious location names.
2. **Connection Between Region and Travel Purpose**
   For instance, Topic 1 includes many nature spots (waterfalls, skiing, plateaus, mountain paths, etc.) and Hokkaido names. This may be because many visitors share their experiences of skiing or nature travel in Hokkaido, so these terms often appear together in Topic 1. This also suggests that many people go to Hokkaido for nature travel.
3. **Topic 6 (Airports) Shows a Peak Post-Pandemic**

▲ Using the `estimateEffect()` function, we can plot a figure estimating time effect on each topic, which shows in Figure 8. We see a clear peak in Topic 6. The reason of peak might be the adjustments or new services in airport and airline when Japan reopened after the pandemic, leading to increased discussions.



▲ Figure 8 Time Effect on each Topic

Research Limitations

There are still limitations in this study, with some biases that I currently cannot eliminate. Here list the points:

1. Issues with Type Classification

The research method used checks if the Title contains keywords to classify articles as popular or non-popular areas. However, classification errors can still occur. For example, the writer may have actually visited a popular tourist spot but did not mention the city name, or the Title might be "Don’t Want to Stay in Tokyo, So I Started a Journey," mentioning a popular city name but the content is about a non-popular spot. Manually checking labels is the most reliable solution, but it takes a lot of time.

2. Difficulties in Text Processing

Articles may contain both Traditional Chinese and Japanese Kanji, which have overlapping characters. This might require manual judgment. Although I tried to use a custom dictionary for proper nouns, the high diversity of vocabulary makes it hard to cover all terms, leading to possible incorrect word segmentation.

Conclusion

In this study, I set three main questions. We can summarize that non-popular tourist spots have slightly increased but not significantly. Article classification performs better with the Naive Bayes model, though there is still room for improvement. In the STM classification, only some topics are regional, while more are related to travel themes.

This study still has limitations. In the future, I plan to address these research limitations by using more complex methods for Type classification and better techniques for text processing.