参赛队员姓名: 吴世悠

中学: 华南师范大学附属中学

省份: 广东省

国家/地区: 中国

指导教师姓名: 陈海兰

论文题目: The Role of Big Data in Smart Tourism: A Case Study of Understanding Museum Visiting Behavior

The Role of Big Data in Smart Tourism: A Case Study of Understanding Museum Visiting Behavior

Shiyou Wu
The Affiliated High School of South China Normal University

Abstract: We show the role of big data in smart tourism by proposing a case study of understanding museum visiting behavior. Using the obtained dataset of museum entries in the city of Florence, Italy, we provide insights from two different perspectives, namely time-series analysis and museum embedding model. Our results indicate that industry with economic significance such as tourism may benefit from economic modeling with the big data.
Key word: Smart tourism, Big data, Time-series, Word embeddings

1. Introduction

Over the past decades, tourism has been continuously growing to become one of the most important economic sectors in the world. According to the World Tourism Organization (UNWTO), by 2018 international tourist arrivals reach a total of 1,322 million and generates over 1.6 trillion dollars revenue globally [1]. As such, the business value of tourism even surpasses that of oil exports, food products or automobiles and represent significant impact on international commerce, especially for developing countries.

As modern tourism is closely linked with economic development for many countries, it not only creates jobs and services needed for tourists that benefit the local economy, but also contributes to the cultural understandings between tourists and residents. In order to further promote the tourism quality and its spillover impact on different industry from agriculture to telecommunications, effective management strategy and policy needs to be established and evaluated upon the core component of tourism, i.e., tourists.

However, traditional research studies on tourists and their behavior largely rely on qualitative analysis. For example, tourism department typically issue surveys to collect feedbacks from tourists regarding their perceived quality and satisfaction [2] . Furthermore, tourism department prefer using summary statistics by year or season to report the key tourism indicators. Both of these methods cannot directly reflect individual tourist behavior and preferences, which may limit the policymaker's strategic decision process from a micro-level scope.

Thanks to the recent advances in Information Technology, data of large scale has been collected to measure and understand population-level behavior at fine granularity. Certainly, the boom in big data also substantially change the tourism management. Li et al. (2018) performed the first comprehensive review on use of big data in tourism so far and they found that area still lacks the systematic understanding on the role of big data in smart tourism [3].

In this research, we aim to propose a case study of using big data in tourism area. Through the

obtained dataset with respect to transactions of museum entries in a touristic city of Florence, Italy, we propose two different perspectives: 1) we organize the museum entries as a time-series data at hourly level and adopt a state-of-the-art forecasting model for museum visit volumes; 2) we generate museum visiting as sequences at individual level and borrowed the word embedding model from the natural language processing task to yield latent representations of museums. Our contributions are mainly three folds. First, we show a data-driven case study in tourism that generates insights from different perspectives on the same dataset. Second, we apply a time-series forecasting model with hourly visit data, which is not common in tourism studies [3]. Third, our museum embedding model leverages the rich co-visitation patterns across tourists and generate the "semantic closeness" between museums beyond the spatial distance. We believe this work may not only provide interesting insights on tourist museum visitation patterns, but also shed lights on using big data in an area with economic significance.

2. Background and Data

We obtained the dataset of FirezeCard about museum entry records in the city of Florence, Italy. Florence is the capital city of region of Tuscany, and considered as the birthplace of the Renaissance with world-level cultural and artistic heritages across the city, such as the Uffizi Gallery, Palazzo Pitti, Piazza del Duomo, to name just a few. The tourism is the most significant industry in Florence with 13 million tourist per year [4] that often leads to the overcrowding problems.

However, Florence is seeking a better class of tourist to share its besieged medieval treasures, as the Mayor Dario Nardella said "No museum visit, just a photo from the square, the bus back and then on to Venice... We don't want tourists like that." [5]. As a result, Florence started the program of FirenzeCard (http://www.firenzecard.it), a tourism promotion initiative that allows tourists to visit all museums in Florence in 72 hours with only 72 euros. The rationale behind this strategy is to promote tourists to visit more museums and stay longer in the Florence. For example, tourists typically only visit the most famous museums within one day and leave the city. This poses challenges to the local economy as these tourists will overcrowd a small number of museums and do not fully explore the rich cultural heritage. Also, they tend to stay for a short period of time and thus contribute to the tourism revenues less than expected.

The dataset contains about 360,000 individual museum entries since its activation about when and which museum has been visited by over 51,000 tourists between June 2016 and September 2016. This period is commonly considered as the peak season and may well represent the tourist visit patterns. Figure 1 shows the number of monthly visit and we see a clear trend that there are more visitor in summer time (July and August).

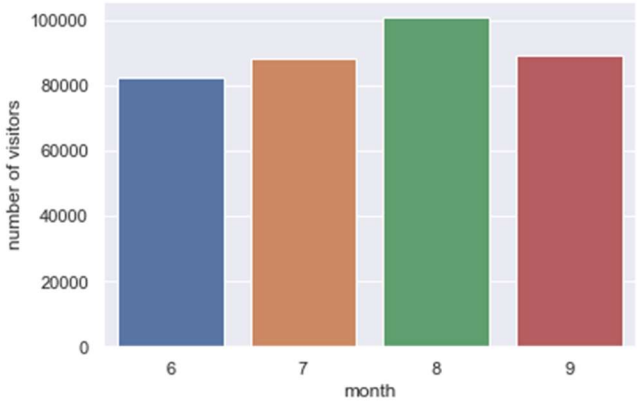

Figure 1: Number of museum visitors by months

Moreover, we aggregate the visitor volume by museums and obtain the popularity of each museum, as shown in Figure 2 with the top 10 most popular museums. We find that Battistero di San Giovanni (Florence Baptistery) is the most popular museum, followed by Uffizi Gallery, Galleria dell'Accademia, etc. This is largely consistent with the inherent popularity of museums.

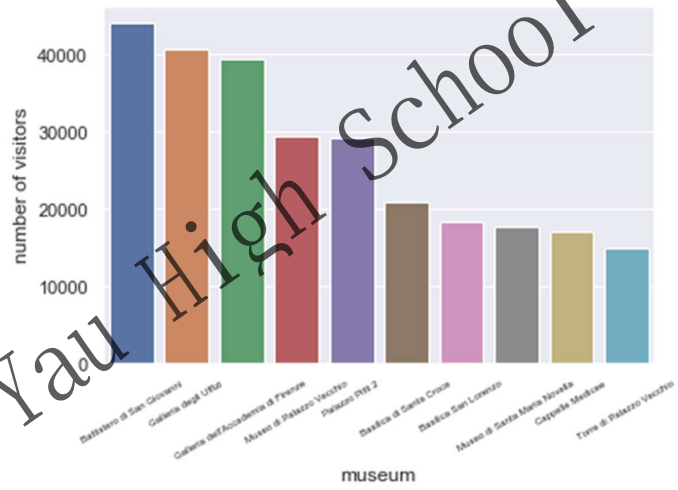

Figure 2: Top 10 museums in terms of popularity

We are also interested in knowing how many museums each cardholder has visited. Figure 3 shows that on average cardholders visit 6 museums and most visit between 1 and 10, given the 72-hour constraints. Interestingly, we find that about 20% of visitors have explored more than 10 museums. This would generate sequences of museums visits with varying length, leading extra heterogeneity of tourists' museum preferences.

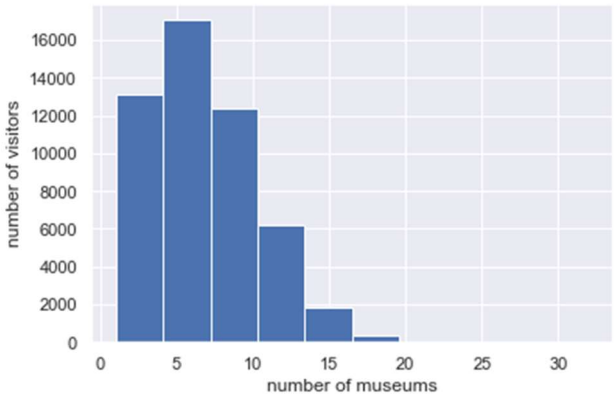

Figure 3: Histogram of number of museums

The explorative data analysis provides us with insights to reorganize the data and to apply different methods from two perspectives. More specifically, we first aggregate the visit volume at hourly level and perform time-series analysis. Then, we generate the museum visit sequences and consider these sequences as "sentences" where each museum is a word. As such, we can perform embedding model to find out latent representation of each museum.

3. Models

3.1 Forecasting Tourist Museum Visiting Volume

Tourism is largely affected by seasons, and Southern European countries such as Italy are especially popular during summer, presumably summer vacation. Thus, the number of tourist visits to Florence sites is likely to follow a predictable trend. Moreover, the number of visits can be seen as periodic both in weeks and in a single day, hence we need a model that strongly reflects such seasonal patterns.

To forecast the tourist museum visiting volume, we aggregate the number of museum entries at the hour level. As such, we organize the data into the time-series format to indicate the number of visits over time. We decide to apply the Facebook Prophet, the state-of-the-art time-series analysis model [5]. The Prophet model provides much more straightforward to create a reasonable, accurate forecast based on an additive model, where the predicted output value is the sum of several components, such as trend, seasonality, and holiday effect, respectively denoted as g(t), s(t), and h(t). The additive model is shown as

$$Y(t) = g(t) + s(t) + h(t)$$

Prophet provides two types trend models, logistic model and linear model, respectively. In the logistic model, it is assumed that the data being studied behave like an ecosystem where saturation is a key factor, and the model fits the data with logistic model given by sigmoid function $g(t) = \frac{C(t)}{1+\exp(-(k+\mathbf{a}(t)^\intercal\boldsymbol{\delta})(t-(m+\mathbf{a}(t)^\intercal\boldsymbol{\gamma})))}$. to reflect the characteristic that the total number will not exceed the carrying capacity C(t) and grow more slowly when the number approaches the carrying capacity. On the other hand, the linear model is targeted at data without a definite upper bound. The model fits the data with a linear regression function

$$g(t) = (k + \mathbf{a}(t)^\intercal \boldsymbol{\delta})t + (m + \mathbf{a}(t)^\intercal \boldsymbol{\gamma}).$$

given by , a function of time, "where k is the growth rate, δ has the rate adjustments, m is the offset parameter, and γ_j is set to $-s_j\delta_j$ to make the function continuous", where changepoints s_j could be specified using known dates of growth-altering events[5]. In addition to trend, seasonality is modeled by a Fourier Series to accurately reflect the periodic characteristic of the data. The Fourier Series

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

provides smooth seasonal effects, and works with P=7 to feature the period of week and P=365.25 to feature the period of a year. More importantly, Prophet model includes holiday effect to avoid the noise of special dates where the data may appear different from usual, which in other models might be treated as outliers. The holiday effect is modeled by a linear regression model

These components add up to fit the training data and make prediction, and reliable prediction can be attained considering the model's accurate reflection of changepoints and seasonality and low susceptibility to holiday effects. Since Prophet is more acceptive of the given data, it is expected that Prophet will perform better than traditional models that tackle time series data. To analyze the Florence's tourist data, for which seasonality and trend are significant, Prophet model will be used in this paper to make predictions, and cross validation on the most fit result will indicate success level.

3.2 Museum2Vec Model

Aside from analyzing the number of Firenze Card holders' visits to Florence sites and predicting number of visits in the future, we seek for another key component reflecting the tourism pattern in Florence—the underlying connection between different tourist attractions. To be specific, tourists' visiting sequences indicate some sort of invisible similarity between sites. For instance, tourists' preferences in a certain type of site, such as Renaissance-themed galleries, determine their visiting sequences. Therefore, by analyzing their behaviors, we will be able to identify most similar and least similar sites. The aim is to understand how the forty sites covered by Firenze Card are correlated from the perspective of tourists, in a way that might not appear obvious from a normal point of view. The information indicated by tourists' visiting sequences can help tourism management agencies comprehend the underlying connections between tourist attractions and predict visitors' decisions and next destination on individual level.

Echo this, we adapt the word embedding model to our scenario to measure the "semantic closeness" between the museums as the Museum2Vec. We choose the Continuous bag-of-words, or CBOW, as it is a main type of word embedding model, and a constructed CBOW architecture predicts the current word based on adjacent words, or context [6]. The training of a CBOW model requires context and current word and generates similarity between different words based on proximity in sentences. Similarity is shown as dense representations of words in a vector space, known as word vectors, and is converted into cosine value of the angle between the two compared word vectors using the formula

$$\text{Similarity}(A, B) = \cos\theta = A \cdot B / (\text{norm}(A) \cdot \text{norm}(B))$$

In short, similarity between words is indicated by proximity in sentences and shown in the form of space vectors in the Word2Vec model. Because similarity of tourist attraction sites is indicated by proximity and simultaneous presence in visiting sequences, the model can be applied to tackle the Firenze Card data, where each site is viewed as a unique word in the CBOW model and each visiting sequence is viewed as a sentence, based on which the dense vector representations of sites are constructed.

Since the aim is to understand the connections between sites, the visiting sequences are crucial, serving as the fundamental indications of similarities. When two sites are simultaneously present in many visiting sequences, it can be determined that they are closely connected to some extent. Likewise, when two sites are consecutive in many visiting sequences, they probably have a high-level connection. Out of the consideration that similarity is indicated by proximity and simultaneous presence in visiting sequences, word2vec model under natural language processing is taken into accounts because the criterion of similarity in the tourism data resembles the criterion in word2vec model.

4. Result Analyses

4.1 Introduction to museum visiting data

The data used in this paper records the number of uses of Firenze Cards at the Entrance of forty tourist attraction sites in Florence, ranging from 8:00 am to 21:00 daily and spanning across June, 2016 to September, 2016. The raw data contains over 360,000 detailed records over the four months.

Among the forty sites covered by Firenze Card, Battistero di San Giovanni was the most popular site over the four months and in each month too.

Museum Name	Visits
Battistero di San Giovanni	11951
Galleria degli Uffizi	10924
Galleria dell'Accademia di Firenze	10137
Museo di Palazzo Vecchio	8985
Palazzo Pitti 2 - Giardino di Boboli, Museo degli Argenti, Museo delle Porcellan	8250
Basilica di Santa Croce	6040
Museo di Santa Maria Novella	5625
Basilica San Lorenzo	5122
Cappelle Medicee	4748
Torre di Palazzo Vecchio	4169
Museo Galileo	3599
Palazzo Medici Riccardi	3560
Museo Nazionale del Bargello	3459
Museo di San Marco	2207
Biblioteca Medicea Laurenziana	2045

Figure 4: Top 10 visits in August

Museum Name	Visits
Battistero di San Giovanni	44047
Galleria degli Uffizi	40622
Galleria dell'Accademia di Firenze	39364
Museo di Palazzo Vecchio	29403
Palazzo Pitti 2 – Giardino di Boboli, Museo degli Argenti, Museo delle Porcellan	29142
Basilica di Santa Croce	20908
Basilica San Lorenzo	18260
Museo di Santa Maria Novella	17715
Cappelle Medicee	17144
Torre di Palazzo Vecchio	14996
Museo Galileo	13675
Museo Nazionale del Bargello	13301
Palazzo Medici Riccardi	12188
Museo di San Marco	8051
Biblioteca Medicea Laurenziana	6753

Figure 5: Top 10 visits from June to September

In order to simplify the full names of the museums, each one of them is assigned an index number from 0 to 39.

Index	Museum	Index	Museum
0:	'Basilica San Lorenzo',	20:	'Museo Novecento',
1:	'Basilica di Santa Croce',	21:	'Museo Stefano Bardini',
2:	'Battistero di San Giovanni',	22:	'Museo Stibbert',
3:	'Biblioteca Medicea Laurenziana',	23:	'Museo degli Innocenti',
4:	'Cappella Brancacci',	24:	'Museo del Calcio',
5:	'Cappelle Medicee',	25:	"Museo dell'Opificio delle Pietre Dure",
6:	'Casa Buonarroti',	26:	'Museo di Antropologia',
7:	'Fondazione Scienza e Tecnica – Planetario',	27:	'Museo di Geologia',
8:	'Galleria degli Uffizi',	28:	'Museo di Mineralogia',
9:	"Galleria dell'Accademia di Firenze",	29:	'Museo di Palazzo Davanzati',
10:	'La Specola',	30:	'Museo di Palazzo Vecchio',
11:	'Musei Civici Fiesole',	31:	'Museo di Preistoria',
12:	'Museo Archeologico Nazionale di Firenze',	32:	'Museo di San Marco',
13:	'Museo Casa Dante',	33:	'Museo di Santa Maria Novella',
14:	'Museo Ebraico',	34:	'Orto Botanico',
15:	'Museo Ferragamo',	35:	'Palazzo Medici Riccardi',
16:	'Museo Galileo',	36:	'Palazzo Pitti 2 – Giardino di Boboli, Museo degli Argenti
17:	'Museo Horne',	37:	'Palazzo Pitti Cumulativo',
18:	'Museo Marini',	38:	'Torre di Palazzo Vecchio',
19:	'Museo Nazionale del Bargello',	39:	'Villa Bardini'}

Figure 6: Museum names with indices

4.2 Forecasting Result of Tourist Museum Visiting Volume

To fit the Firenze Card data, we use the prophet model under prophet packet in Python. Changepoint prior and seasonality prior are the most significant hyper-parameters in the Facebook Prophet model because they determine the model's fitness to trend and seasonality, and the prior concerning holiday effect is of less significance because only three days were slightly affected, thus no extra regressors are needed. The domain of these two hyperparameters are within (0, 1), and cross validation will be employed to find the root mean square error (abbreviated as RMSE in the following text) and determine the best set of hyperparameters.

RMSE		Hyperparameter "Seasonality Prior"				
		0.005	0.007	0.01	0.02	0.05
Hyperparameter "Changepoint Prior"	0.005	0.738	0.591 (T)	0.554 (T)	0.540 (T)	0.538 (T)
	0.01	0.737	0.599	0.565 (S)	0.547 (S)	0.545 (S)
	0.02	0.741	0.602	0.563 (T)	0.551 (T)	0. (S)

T refers to unrealistic trend prediction

S refers to unrealistic daily seasonality prediction

Chart 1: Average RMSE of each set of examined hyperparameters

According to Chart 1, if we only focus on the average RMSE, we can draw a conclusion that the best set of hyperparameters is (0.01, 0.05). However, not all parameters accurately reflect the realistic trend and daily seasonality. Since summer is the most popular season for tourism, the monthly trend should increase until October and starts descending in September, typically shown in figure 7. Therefore, multiple sets of results are invalidated for not matching the trend in reality despite relatively low RMSE.

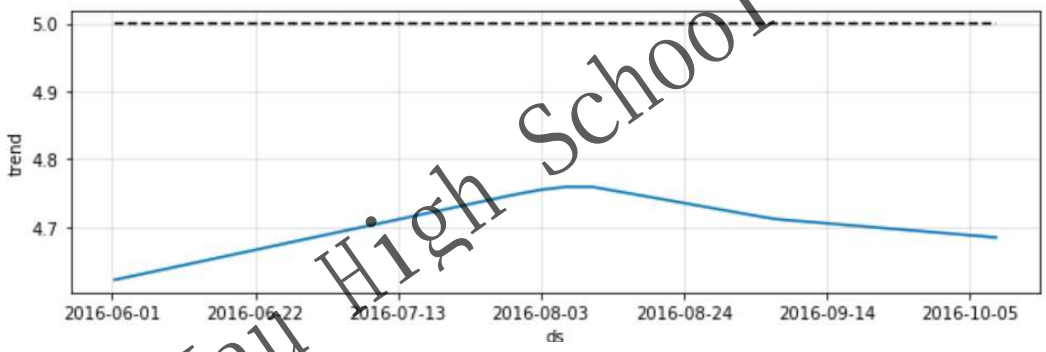

Figure 7: Trend over months

Similarly, the daily seasonality graph is an approximate reflection of the rate of change of the number of tourist visits from hour to hour. Since all sites are closed from 22:00pm to 8:00am, the number of visits during this period is constantly zero. Therefore, the rate of change of this period should be close to zero, hence we can eliminate results that demonstrate large fluctuations shown in figure 8, where the seasonality prior equals 0.1, and retain the normal ones, as exemplified in figure 9, where seasonality prior equals 0.005.

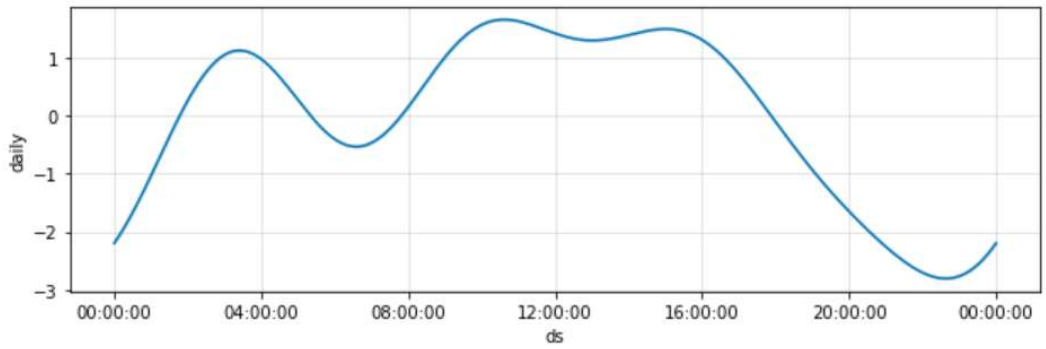

Figure 8: Daily seasonality with seasonality prior = 0.1

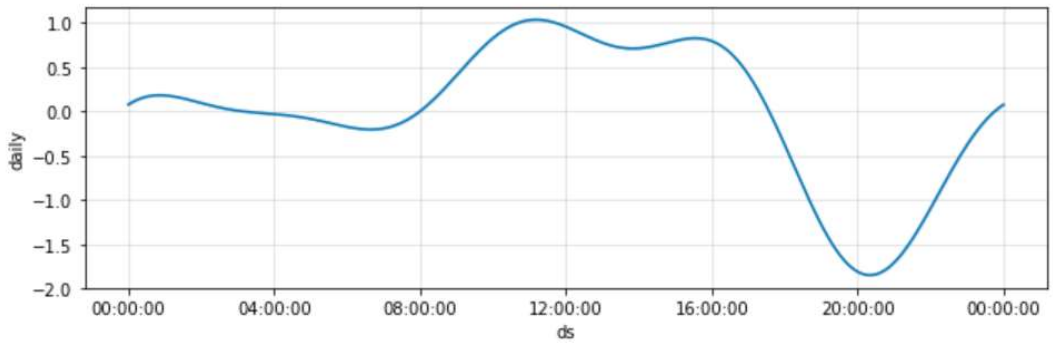

Figure 9: Daily seasonality with seasonality prior = 0.005

Taking all these factors into accounts, we obtain a modified version of the result that the best set of hyperparameters is “changepoint prior” equal to 0.01 and “seasonality prior” equal to 0.007, where the average RMSE is 0.599. The corresponding fit result is shown in Figure 10.

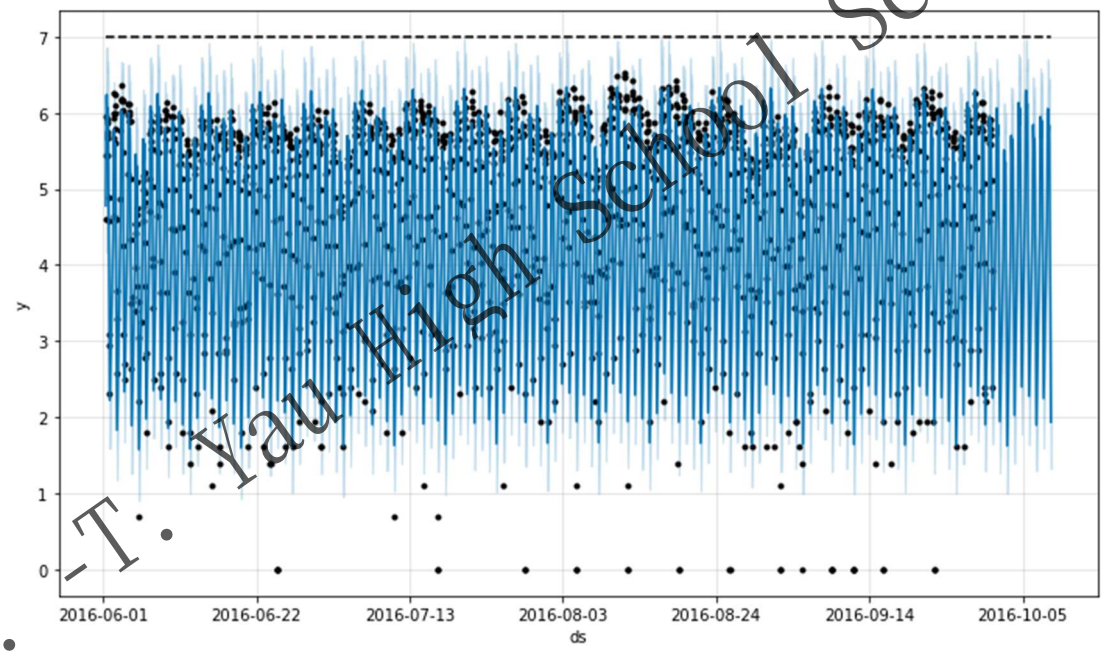

Figure 10: Fitting result with the best set of hyperparameters

4.3 Individual Tourist Data Insight Result with Museum2Vec

The most important priors involved in the word embedding model in gensim packet of Python are size and window. Size determines the dimension of the space vectors, and window determines the number of preceding and succeeding units being used to predict the current unit. There are no principal restrictions on the domain of these two hyperparameters, yet the size hyperparameter is usually less than the number of vectors produced.

The Gallery degli Uffizi is the landmark and second most popular site in Florence. Over August 2016, the data shows sites most similar to and most different from Uffizi. Each of the forty sites

covered by Firenze Card is given an index from 0 to 39, respectively.

hyper parameter "window"	museum (most similar)	Distance to "8"	Visits Difference (/month)	museum (most different)	Distance to "8"	Visits Difference (/month)
5	"30"," 2", "38"	2 minutes	2000	"3", "11", "37"	10 minutes	9000
10	"2", "38", "36"	7 minutes	1000	"20", "3", "37"	12 minutes	10000
20	"19","1", "16"	3 minutes	8000	"11", "27", "5"	60 minutes	10000

Chart 2: First round of results with hyperparameter "window" based on August dataset

Chart 2 records the result generated from the most popular month, August, featuring three different models. The distance between two compared sites is measured by Google Map's estimation of the time required to walk from one site to the other, which reflects the convenience to commute. The third column indicates the distance from Galleria degli Uffizi to the most similar site, which is the leftmost site in the "museum (most similar)" column, while sixth column indicates the required traveling time from Uffizi to the most different site, which is the leftmost site in the "museum (most different)" column. Visits difference is measured by the absolute value of the difference between the number of visits to Uffizi and the number of visits to the compared site.

As is shown in the chart, the result closest to reality is given by window size equal to 5, on accounts of the highest similarity and highest difference. Considering the top similarity results in the three models, respectively Museo di Palazzo Vecchio (30), Battistero di San Giovanni, and Museo Nazionale del Bargello (19). Bargello Museum is the first eliminated because the dramatic popularity difference with Uffizi. In the other two competitors, Museo di Palazzo Vecchio (30) is the closest site to Uffizi, requiring only two minutes' walk, and despite San Giovanni (2)'s slightly higher similar in terms of visit, Museo di Palazzo Vecchio (30) can be deemed as the most similar site because San Giovanni raises the difficulty of commute. In fact, the result is rational and conforms to intuition since Uffizi is the second most popular site in August while Museo di Palazzo Vecchio is the fourth most popular, and the two are geographical close to one another. Moreover, the most different sites generated by the three models are not significantly distinguished. Therefore, the first model, with the hyperparameter window equal to 5, is the fittest model among the three.

Taking a step further, we feature the models with hyperparameter "window" close to 5, and their results are shown in chart 3.

Hyper Parameter "Window"	Museum (most similar)	Distance to "8"	Visits Difference (/month)	Museum (most different)	Distance to "8"	Visits Difference (/month)
3	"19","30","2"	3 min	8000	"3", "37", "0"	10 min	9000
4	"30","2","36"	2 min	2000	"3", "37", "0"	10 min	9000
5	"30","2","38"	2 min	2000	"3", "11", "37"	10 min	9000
6	"30", "38", "2"	2 min	2000	"3", "37", "0"	10 min	9000
7	"2", "30","38"	7 min	1000	"11", "5", "0"	60 min	10000

Chart 3: Second round of results with hyperparameter "window" based on August dataset

As Chart 3 illustrates, while the window is equal to 4, 5, or 6, identical results of the "Museum

(most similar)" and "Museum (most different)" are generated. The difference varies at the second most similar and the second most different sites, therefore further tests are required. Chart 4 features the results generated from the dataset from June to September, offering a more comprehensive insight.

Hyper Parameter "Window"	Museum (most similar)	Distance to "8"	Visits Difference (/month)	Museum (most different)	Distance to "8"	Visits Difference (/month)
4	"30","38","2"	2 min	1.1×10^4	"3", "37", "0"	10 min	3.5×10^4
5	"2", "30", "38"	7 min	4×10^3	"11", "3", "21"	60 min	4×10^4
6	"30", "2", "38"	2 min	1.1×10^4	"3", "11", "37"	10 min	3.5×10^4

Chart 4: Results with hyperparameter "window" based on dataset of four months

The model whose window is equal to 5 generates a different result on the data set over the course of four months. Inconsistency suggests that the model with window value equal to 5 is not as stable as the other two candidates. Also, it has been shown previously that Uffizi ("8") is more similar to site 30 than it is to site 2, therefore the model with hyperparameter window equal to 5 gets eliminated.

As for second row and fourth row in Chart 4, identical results are obtained for the most similar and most different site. Therefore, we need to compare the second most similar and second most different tourist attractions. We can observe that the model whose window is equal to 6 performs better by comparing "2" with "38" and comparing "11" with "37". According to Chart 5, "2" differs from the Uffizi by approximately 4000 visits over four months, thus bearing higher resemblance with Uffizi than "38" does. Another conclusion drawn from Chart 5 is that "11" has a higher difference level from Uffizi than "37" in addition to its significant geographic isolation from Uffizi, which is reflected by Google Map's estimation. Overall, the model for which the hyperparameter window is equal to 6 performs the best among all three candidates.

	Museum with second highest similarity to Uffizi		Museum with second highest difference from Uffizi		Benchmark
Hyper-parameter	Hyperparameter "window" = 6	Hyperparameter "window" = 4	Hyperparameter "window" = 6	Hyperparameter "window" = 4	
Museum and index	Battistero di San Giovanni ("2")	Torre di Palazzo Vecchio ("38")	Musei Civici Fiesole ("11")	Palazzo Pitti Cumulativo ("37")	Galleria Degli Uffizi ("8")
Visits	44047	29403	655	1593	40622

Chart 5: Comparison between results of hyperparameter "window" = 4 and results of hyperparameter "window" = 6

By applying the Museum2Vec algorithm, we find out in chart 6 that tourists are more likely to visit the top three most similar sites to the Galleria degli Uffizi, respectively Museo di Palazzo Vecchio ("30"), Battistero di San Giovanni ("2"), Torre di Palazzo Vecchio ("38"), rather than

geographically closest Museo Galileo ("16"), which is usually thought to be visited by Uffizi's vistiors. This proves our algorithm successful by revealing similarity that cannot be traced from common sense.

Museums		Visits to Uffizi and the below museum	Percentage (Visits to the below museum and Uffizi / Visits to Uffizi)
Museums most favored by Uffizi visitors	Museo di Palazzo Vecchio ("30")	24318	60%
	Battistero di San Giovanni ("2")	35578	86%
	Torre di Palazzo Vecchio ("38")	12473	31%
Museum geographically closest to Uffizi	Museo Galileo ("16")	11431	28%

Chart 6: Comparison between museums most favored by Uffizi visitors and museum geographically closest to Uffizi

4. Conclusion

In industry with economic significance such as tourism, the use of big data may help to generate insights on individual tourist behavior at a fine granularity. In this work, we aim to provide a case study of using big data in the tourism, i.e., understanding the tourist museum visiting behavior in the city of Florence, Italy. We use the dataset of museum entries and organize it into hourly museum visit volumes and individual museum visiting sequence. For the former, we perform a time-series analysis to forecast the museum visit volumes using a state-of-the-art forecasting model. We find that the museum visiting volumes can be decomposed into several predictable components, such as the trend, weekly, daily seasonality as well as the holiday effect. As such, we are able to forecast how many visitors to these museums at each hour, and may help planning the tourist traffic ahead of the time.

Moreover, we further propose a Museum2vec model that takes the individual visit patterns and infer the latent embeddings between the museums, similarly as the Word2vec model commonly used in the NLP tasks. We find that inherent similarity between museums based on the co-visitation patterns of tourist, and such similarity reveals certain semantic closeness that is well beyond that measured by the spatial distance. For example, museums that have strong cosine similarity in terms of their latent embeddings implies that these musuems are more likely to be visit together due to the tourists' preferences. This finding, however, can only be generated from the large-scale museum vising sequence at individual level.

We believe our work shows a case study on the role of big data in the smart tourism and how we could perform economic modeling from different perspectives by organizing the same dataset into the different format.

References

[1] The UNWTO World Tourism Barometer (June 2018), Volume 16, available at http://marketintelligence.unwto.org/content/unwto-world-tourism-barometer

[2] Chen and Chen (2010). Experience quality, perceived value, satisfaction and behavioral intentions for heritage tourists, *Tourism Management*, 31(1), 29-35.

[3] Li et al. (2018). Big data in tourism research: A literature review, *Tourism Management*, 68, 301-323

[4] Tourism in Florence, Wikipedia, https://en.wikipedia.org/wiki/Florence#Tourism

[5] Taylor and Letham (2018), Forecasting at Scale, *The American Statistician*, 72, 37-45.

[6] Mikolov et al. (2013), Distributed Representations of Words and Phrases and their Compositionality, Proceedings of NIPS, 3111-3119

Acknowledgements

I wish to express sincere gratitude to Mrs. Hailan Chen who guided me until this paper is finished.

I sincerely thank my parents for their support and help in life.

I also thank Yau Award for Science, for endowing me the precious opportunity to be involved in economic modeling in high school.

Without the help of my parents and tutor, I would not be able to complete my first academic paper.

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员： 吴世悠 **指导老师：** 陈海兰

2018 年 9 月 28 日