Interpretable Model for Classifying Best Seasons of Travel Destination

1. Project Description

During the project idea brainstorming sessions, we wanted our project to be as relevant to our daily lives as possible. For the past two years, the pandemic of COVID-19 has undoubtedly put a stop to world's travel, which has caused a great impact on the world's economy and travel industries such as airlines, hotels, cruises, etc.

As the vaccination rates steadily increase, more and more countries around the world have decided to fully reopen and to be ready to welcome the travelers and tourists from around the world for the first time since March 2020, and some of them, such as Singapore, South Korea, Australia, New Zealand, etc, were even the ones that used to impose the most strict travel bans and quarantine regulations back then.

As a matter of fact, one of our team members Ben proposed this idea, and he already took initiatives by starting traveling to Europe during the spring break. While we can't be more excited about all the new potential travel destinations that we can explore, Ben has convinced us, and we think it would be a great idea that we implement a project that can provide us with more ideal travel recommendations with automatic web crawling and language processing that we've been learning and practicing in this class.
The topic of this project is determining the best season to travel to a certain location. To make the determination, textual information about locations would be crawled from the internet, and predictions would be made based on the information. The system can output the best season to travel given a location as input.

- Data: Textual data about locations would be scraped from informational websites such as Wikipedia or travel review websites such as Tripadvisor.
- Method: Textual data scraped from the internet would be preprocessed for the analysis. A set of parameters would be constructed for each season, and the determination of best season would be computed through correlation between the data and season parameters.
- Evaluation: Widely-accepted opinions about the best season to travel to a location would be used as the correct answer. The location-season combinations would be collected by searches such as "best summer destinations" on the internet. The correct answers would be compared against model output of this project.

2. Related Work (References)

To help us evaluate our project's scope and plan our approach more holistically, our team has dived into researching existing works that are related to our subject. Our research is based on two main aspects, including the current decision process of the best season to visit a destination used by travelers and travel recommendation publications, and the existing usage of textual data correlation in projects related to subject classification.

To understand the decision process, our team has chosen a summary article about the best season to visit all regions of the world with a map illustration [1]. Rather than using the generalized classification of seasons as a

guideline for our own classification model, we used this map as a way to understand the key factors that contribute to seasons' suitability for destinations. An example would be having the term "Northern Lights" as a parameter for winter and summer destinations, as it is generally the best to visit Nordic destinations in the summer months or the winter months. Another article with detailed descriptions of activities in specific destinations, named *A destination for every season*, is used to extract additional words that could represent seasons [2]. An article that discusses the best seasons to visit the world's most visited destinations is used to construct the correct answer key discussed above [3].

[1] https://scarletscribs.wordpress.com/2018/01/07/this-awesome-map-tells-you-the-best-time-to-visit-every-destination-in-the-world/
[2] https://bragpacker.com/a-destination-for-every-season-a-quick-guide-to-picking-the-best-travel-destinations-all-year-long/
[3] https://www.farandwide.com/s/best-time-to-travel-around-the-world-0df967e053fa4328

For the classification algorithm, basic searches about the ways to classify objects would frequently lead to the answer of text vectorization, which is the underlying implementation of our group's proposed approach [4]. A description article of text classification in Python provides an example of classifying the ethical characteristics of tweets [5]. The data preparation and feature engineering sections of the article have been reviewed by our group's members to inspire the implementation plan of our model. Another article about document classification suggests a similar approach that our team is planning on taking [6]. In the subsection of *defining keywords*, the practice of collecting a list of keywords, similar to the parameters in our project proposal, for each subject category is introduced. An example would be "protein" and "calories" for the category of nutrition.

3. Data Collection, Annotation, and Data Samples

Since our project goal is to find the best seasons for traveling to some fantastic places, our team plans to get the data about destinations and season information crawled from the internet searching in regards to the data collection. The project would require us to first give a list of destinations that our team is interested in traveling to, which can either be countries, regions, or cities. To be specific, the target destinations should be sorted logically, with all cities tagged with their corresponding country or region. Then the data annotation would require us to identify a list of keywords related to climate and weather for the targeted destinations, such as "temperature", "Fahrenheit degrees", "Celsius degrees", "weather", "climate" and so on. These data can be caught through historical temperature reports and professional databases, but it would be easier if we look for them in the textual introduction of the target destinations. Moreover, special events or seasonal sights should also be considered, including any festival and celebrations, flower season or snow season, or even concerts and special exhibitions in the museums if applicable. Such events are so meaningful and attractive that the visitors may come to visit the place mainly for them. The feature of this type of data would contribute in recognizing when is the best time period to travel to this place so as to reach the highest value, and this is based on multiple purposes or metrics in regards to different needs from the individual travelers. Some data samples for this purpose could be "festival", "celebration", "limited time", "annual", "best time", "best season" and so on. Some slogans like "if you miss it, you will regret it" may also be helpful for identification and annotation, so we would include samples like "miss" and "regret" or some words in this same logic as well. These

data may be found in the written-up city guide, tourists' experiences or blogs, or textual suggestions from the locals. Some advertisements from the traveling agents would also be a good source for the data collection and worthy of crawling.

4. Methods and Evaluation

Methods

The first step is the Dataset Preparation step which includes the process of loading a dataset and performing basic pre-processing. First, we need four documents that contain keywords for each season. Therefore, we will build a crawler to crawl the top N results from Google web search. Since Google can determine the relevance that a website holds for any given search term, we will not waste time on ranking the relevance of pages. We will just crew the articles related to the queries like "travel expectations for spring", processing the pages for finding our keywords in further steps. At this point, we form a list of keywords that will help us define whether a city relates to a given season. Also, we need another crawler, taking trip advisor website as the root, crawling the reviews for our input cities.

Next step is building our model. We can simply use the rule-based features. We will use tf-idf with to implement the vector space model, which enables us to gives us a way to associate each word in a document with a number that represents how relevant each word is in that document. The highest scoring words of a document are the most relevant to that document, and therefore they can be considered keywords for that document. The model will score each crawled review document by the similarity measure with keywords documents.

Alternatively, we can use unsupervised classification. Since categories (four seasons) are defined, annotating a large number of training documents takes much time and effort. By applying NLP to understand the context of words, the model scans the existing crawled review documents and find similarities between documents. Then, it divides the set of similar documents into a cluster. This cluster will contain records with content that theoretically falls into the same category.

Evaluation

One of the evaluations of the project will be measured by comparing the season result we get from the model's output to our labeled data collection. The labeled data collection represents the actual popular season, which is mainly determined by two resources: Trip-advisor and google. At this stage we use labeled data containing the authoritative answers to test the model. The labeled data is crawled on these resources and the output should be whether a period of time or an exact season and stored in a document. These results are considered the gold standard when evaluating the accuracy of our model.

Based on the former data training process, we have already computed the similarity scores between the review contents for our targeted location and season-specific query documents. Then we use the labeled data collection to calculate the macro-averaged recall and macro-average-precision for our model. We are aiming to find the query document for a season with high precision and low recall because higher precision means the model that season-related query document retrieves more relevant instances within the whole relevant pool, and also indicates that a higher percentage of our output is considered relevant. The lower recall usually infer that the model with that season-related query document eliminates more irrelevant results and returns fewer irrelevant documents. Higher precision often comes with lower recall. Therefore, since precision and recall are inversely related, as the precision values decrease, the recall values should increase. In this case we compare the result we generated and the season whose query document with the

highest macro-averaged recall and the closest macro-average-precision.

We will also include other resources like information on Kaggle, where we can easily incorporate the existing evaluation results and other evaluation methods written in python. We could also use Bayes' Theorem to evaluate whether our training model is suitable. Bayes' Theorem is defined as the probability of an event happening given another event occurs. If we consider the season as one event and the possible season characteristics as another event, we can compute the conditional probability of a season given the existing characteristics. In this case, the assumption here is that a season can always find a characteristic to match with.

References

Darwitan, Posted by Andrew. "This Awesome Map Summarizes the Best Time to Visit Every Country in the World." *Scarlet Scribbles*, 9 Nov. 2021, https://scarletscribs.wordpress.com/2018/01/07/this-awesome-map-tells-you-the-best-time-to-visit-every-destination-in-the-world/.

Bragpacker. "A Destination for Every Season: Guide to Picking the Best Travel Destinations All Year Long." *Bragpacker*, https://bragpacker.com/a-destination-for-every-season-a-quick-guide-to-picking-the-best-travel-destinations-all-year-long/.

"Best Times to Travel to the World's Most-Visited Cities." *Far & Wide*, https://www.farandwide.com/s/best-time-to-travel-around-the-world-0df967e053fa4328.

BlackBeansBlackBeans 12333 bronze badges, and Nicolas MartinNicolas Martin 94011 silver badge1010 bronze badges. "How to Classify Objects from a Description in Natural Language." *Data Science Stack Exchange*, 1 June 1969, https://datascience.stackexchange.com/questions/98308/how-to-classify-objects-from-a-description-in-natural-language.

"A Comprehensive Guide to Understand and Implement Text Classification in Python." *Analytics Vidhya*, 26 July 2019, https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/.

Editor. "Document Classification with Machine Learning: Computer Vision, OCR, NLP, and Other Techniques." *AltexSoft*, AltexSoft, 17 Nov. 2021, https://www.altexsoft.com/blog/document-classification/.