Corpus analysis for Japanese locations

Cheolyeon BYUN
1M170020

Introduction to Text Mining and Digital Humanities
byuncheolyeon@akane.waseda.jp

2023/01/29

Abstract

This paper shows findings that were made through the use of the corpus work bench, analyzing the corpus 'TRIP_ADVISOR', see how the use of words regarding locations of Japan and other famous activities like visiting temples, onsens, and festivals, have changed over the years, and how it compares with actual tourist numbers of Japan.

1 Introduction

Data has been extracted from the corpus 'TRIP_ADVISOR', that are visualized with use of python and matplotlib. The theme of the data are different famous Japanese tourist locations, such as Tokyo, Kyoto, Osaka, Hokkaido, Yokohama, and Okinawa. These locations weren't chosen deliberately, but only chosen because those were the only ones that returned sufficient information for an academic paper. These data will be compared with actual tourist numbers of Japan at the end to see if we can observe a correlation.

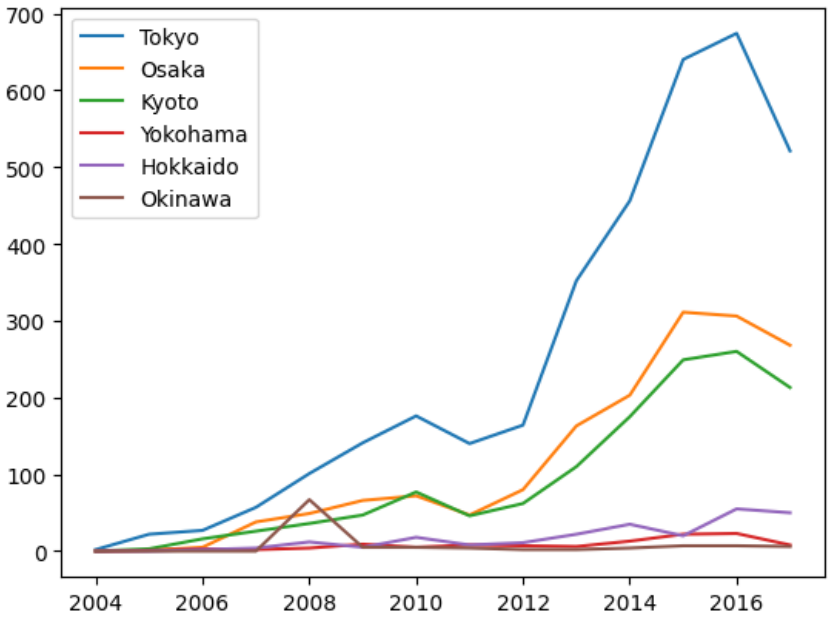
2 Data and Method

With the corpus workbench operating from the linux operating system, I was able to fetch data over the years about words and how frequently those words come up in the corpus for each year. From the year 2004 to 2017, data that was into a txt file format was extracted again to a pandas dataframe and was ultimately created into a csv file, shown below. The words that I've decided to use were the locations metioned above, with some extra words that are all famous Japanese things to do, such as visiting onsens, shrines, temples, ryokans, and festivals.

words/year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
tokyo	2	22	27	57	101	141	176	140	164	352	456	640	674	521
kyoto	0	3	16	26	36	47	77	46	62	110	175	249	260	213
osaka	0	1	5	38	49	66	72	47	80	163	203	311	306	268
yokohama	0	1	3	2	4	9	5	8	7	6	13	22	23	8
hokkaido	0	0	2	4	12	5	18	8	11	22	35	20	55	50
okinawa	0	0	0	0	67	5	5	4	2	2	4	7	7	6
onsen	1	7	14	43	44	87	99	83	116	152	175	262	370	283
ryokan	0	3	15	123	132	215	160	134	149	245	256	336	460	377
temple	0	39	16	145	167	260	231	189	179	394	374	498	668	613
shrine	0	28	12	72	80	116	83	68	65	142	146	184	244	214
festival	1	25	8	34	48	83	60	58	70	94	156	165	177	200

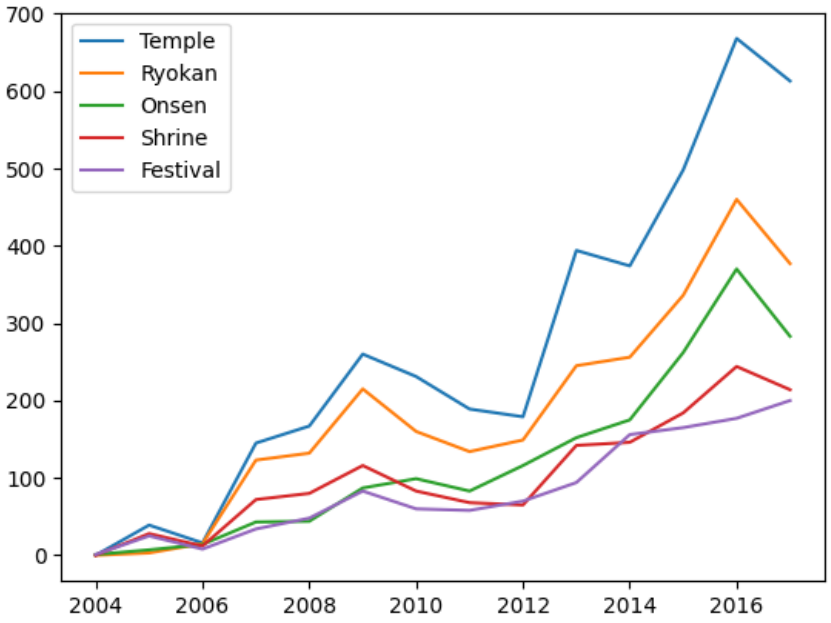
3 Results

The results are divided into two different charts, plotted with the help of matplotlib.



Frequency of Japanese Locations

This here is a line chart of the data. Obviously the word 'Tokyo' came up the most, next to Osaka, Kyoto, Yokohama, Hokkaido, and Okinawa that came in last. It seems like numbers associated with Tokyo can be a huge outlier, and the data in general is not balanced at all among all locations.

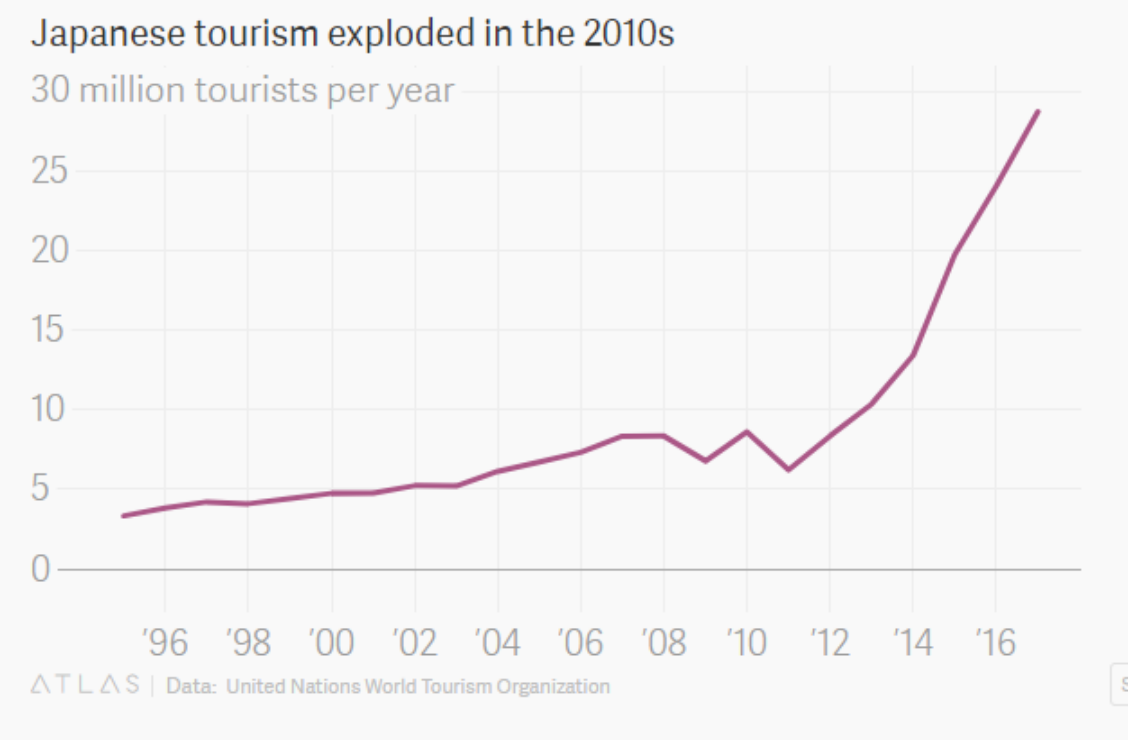


Frequency of Famous Japanese Activities

Next is a graph made from data collected from Japanese activities that most tourists do. The word Temple had the highest frequency, followed up by Ryokan, Onsen, Shrine, then Festival. As opposed to the data from our first graph, the data seems a more balanced, and no data strays too far from all the other data.

4 Conclusion and Limitations

The WorldEconomicForum (2018) states that, tourist numbers kept increasing until the year 2019 when there were nearly 32 million tourists that visited Japan.



Tourist Numbers Before Covid

As evident by the actual statistics regarding the number of tourists, the graph is very analagous to the two other graphs regrading the frequency of words that appeared on the TRIP_ADVISOR corpus, which in turn represents the amount it appeared on the trip travel forum from tripadvisor.com. However, as tempting as it is to conclude that this correlation must have soemthing to do with the increasing interest in Japan by people around the world, more due-diligence is required, and there could be more factors that is playing a role here. For example, the data for frequency of words were extracted from a forum, and the words could have came up more in the forum simply because the tripadviosr website and the forum themselves were gaining more popularity, and might not reflect the global trend of more people wanting to travel to Japan. The only logical conclusion one can make from these results is that, we can only infer that the people who uses that forum had a growing interest in Japan and traveling there, but the upward trend of the data coinciding with actual tourist numbers in Japan is probably nothing more than a mere coincidence.

5 Code

`https://github.com/OWO-hue/CQP-Text-Mining`

References

WorldEconomicForum (2018). The number of tourists that visit japan has increased by more than 20 million in five years.