Vacation Vision: A holiday recommendation engine

Peter Moore

2018-10-26

SELECT @@SERVERNAME

1 records

|  |
| --- |
| 612a03e6bdaa |



University of Dundee

Course: AC53011

Studentid: 170000983

Supervisor: Dr Karen Petrie

A thesis submitted in fulfillment of the requirements for the degree of MSc Data Science (School of Computing)

**Vacation Vision: A recommendation engine for the travel industry using twitter-based sentiment analysis; image detection and collaborative filtering** \*\*\*

## Executive Summary

This project…

## Declaration

“I declare that the special study described in this dissertation has been carried out and the dissertation composed by me, and that the dissertation has not been accepted in fulfilment of the requirements of any other degree or professional qualification.”

Peter Moore \*\*\*

Peter Moore

## Certificate

I certify that Peter Moore has satisfied the conditions of the Ordinance and Regulations and is qualified to submit this dissertation in application for the degree of Master of Science.

Dr Karen Petrie

## Acknowledgements

I would like to thank my supervisor, Dr Karen Petrie for their guidance throughout the project.

I would also like to thank Professor Mark Whitehorn, Professor Annalu Waller and Mr Andy Cobley fortheir extra-curricular support.

I would like to thank my family for their support. Especially those my partner and daughter who had to live with me whilst writing this.

## Vacation Vision: A recommendation engine for the travel industry using twitter-based sentiment analysis; image detection and collaborative filtering

## Introduction

The aim of this project is to recommend exotic holiday destinations for people. “Exotic” in this case meaning places they may never have thought of going but that would suit them. These would be recommended via an analysis of their own holiday snaps per their Twitter feed and a comparison with their peers.

## Approach

## Workflow

## Background

### Twitter

Twitter is a …

As per Finin et al. ([2010](#ref-finin2010annotating))

Totally worth invoking Crisp-DM at this point.

### Contraints

* Twitter’s move to 280 characters in late 2017[[1]](#footnote-34) leads t o a lack of establised literature on the new longer tweet length and the vast majority of thepreviouswork discussed here pre-dates this. It would be interesting to review Twitter with a view to the linguistic effect of this change (for example has it led to less abbreviations). Since this project is not concerned directly with tweet lenght and expands abbreviations, this is not considered deleterious to the research.
* The Cheng, Caverlee, and Lee ([2010](#ref-cheng2010you)) dataset is US-based potentially introducing bias to the control corpus
* Furthermore, the analyses are English-language analyses although the methodology should be scalable to other languages

## Ethics statement

All tweets collected were available publicly and collected using the (public) twitter API. Twitter’s international terms of service were observed (<https://twitter.com/en/tos#intlTerms>).

## Conclusion

## Reflection

This represents merely the beginning of what can be achieved by mining twitter data in the travel sector. Ancillary to this project, but fascinating overall, is the processing of real-time tweets. For example Odlum and Yoon ([2015](#ref-odlum2015can)) use real-time twitter information to monitor the spread of the Ebola disease virus. Methodologies such as this could be adapted for data subjects as their vacations approach, for safety (for example political unrest), for packing the suitcase (for example forthcoming extreme weather events) or for fun (for example local sports teams reaching finals).

This study is not longitudinal, it does not examine the impact of longer-period trends. For example, the relative decline in attraction of the USA and rise in attraction of Oceania over the last fifty years Ballantyne et al. ([2009](#ref-ballantyne2009trends)).

Peter Moore

2018-10-26

## Appendices

### Appendix 2: Travel Agency Twitter Feeds

Skills Matrix

|  |  |  |
| --- | --- | --- |
| Agency | TwitterFullURL | TwitterUsername |
| Audley Travel | <https://twitter.com/audleytravel> | audleytravel |
| Turquoise travel | <https://twitter.com/TurquoiseUK> | TurquoiseUK |
| STA Travel | <http://twitter.com/STATravel_UK> | STATravel\_UK |

### Appendix 3: Twitter Pseudonimisation Algorithm

def fromtextExtractNameObjects(text):  
# import nltk  
# # adapted from https://tim.mcnamara.nz/post/2650550090/extracting-names-with-6-lines-of-python-code  
# for sent in nltk.sent\_tokenize(text):  
# for chunk in nltk.ne\_chunk(nltk.pos\_tag(nltk.word\_tokenize(sent))):  
# if hasattr(chunk, 'label'):  
# print(chunk.label(), ' '.join(c[0] for c in chunk.leaves()))  
  
 return text

### Appendix X

Matrix of skills used and whether or not I had prior[[2]](#footnote-47) experience.

Skills Matrix

|  |  |
| --- | --- |
| Skill | Prior exerience |
| SQL | Y |
| Docker | N |
| Python | N |
| R | N |
| Power BI | Y |
| ggplot | N |
| nltk | N |
| Weka | N |
| MongoDB | N |
| Tweepy | N |
| Tensorflow | N |
| Keras | N |

## References

Ballantyne, Roy, Jan Packer, Megan Axelsen, and others. 2009. “Trends in Tourism Research.” *Annals of Tourism Research* 36 (1). Elsevier Science: 149–52.

Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. 2010. “You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users.” In *Proceedings of the 19th Acm International Conference on Information and Knowledge Management*, 759–68. ACM.

Finin, Tim, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. “Annotating Named Entities in Twitter Data with Crowdsourcing.” In *Proceedings of the Naacl Hlt 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 80–88. Association for Computational Linguistics.

Odlum, Michelle, and Sunmoo Yoon. 2015. “What Can We Learn About the Ebola Outbreak from Tweets?” *American Journal of Infection Control* 43 (6). Elsevier: 563–71.

1. <https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html> [↑](#footnote-ref-34)
2. “Prior” meaning before 1st January 2017 [↑](#footnote-ref-47)