Classifying Hotel Topics and Opinions from Online Guest Review Comments

By:
N. Nobel

Master Thesis
VU University Amsterdam
March, 2013
Classifying Hotel Topics and Opinions from Online Guest Review Comments

Author:
Nicolaas Nobel

Supervisors:
Mark Hoogendoorn
Piek Vossen
Sandjai Bhulai

VU University Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam

Booking.com B.V.
Herengracht 597
1017 CE Amsterdam

March, 2013
PREFACE

The final part of the Master Business Analytics consists of a six months internship at an organisation by choice. During this internship students are required to use their theoretical knowledge of business mathematics and informatics to support the organization with data-driven decision making, optimisation of business processes and data analysis. The deliverables generally consists of a final thesis along with a data analysis, program or simulation. The research for and creation of this thesis took place at the content department of Booking.com from September 2012 until March 2013. During this period a text mining tool is developed and a business focused data analysis is performed.

In the first and foremost place I would like to thank Vladimir Sterngold for his on-going guidance and support, giving me insight into the hotel business and providing me detailed information about the Booking.com website. In the second place I would like to acknowledge my special gratitude to my supervisors from the VU University, Mark Hoogendoorn and Piek Vossen, giving me direction and constructive comments on my thesis. In the third place I would like to thank Sandjai Bhulai, who accepted to take on the task of second reader. In the fourth place, I would like to use the opportunity to thank all the guest review associates that helped me develop a high quality labelled dataset for this project. Finally, I would like to express my grateful appreciations to my colleagues from the business analytics team, content and IT department for their collegiality and discussions that helped me write this thesis.

Nicolaas Nobel

Amsterdam, March 21, 2013
ABSTRACT

[Classified]
Contents

PREFACE .............................................................................................................................. 5
ABSTRACT ............................................................................................................................ 6

1 Introduction ...................................................................................................................... Error! Bookmark not defined.

2 Background ..................................................................................................................... Error! Bookmark not defined.
   2.1 Booking.com’s content department .......................................................................... Error! Bookmark not defined.
   2.2 Online guest reviews ............................................................................................... Error! Bookmark not defined.
   2.3 ................................................................................................................................. Error! Bookmark not defined.
   2.4 ................................................................................................................................. Error! Bookmark not defined.
   2.5 .................................................................................................................................. Error! Bookmark not defined.
   2.6 Used Algorithms ...................................................................................................... Error! Bookmark not defined.
   2.7 Conclusion ............................................................................................................... Error! Bookmark not defined.

3 Related work & text mining concepts .............................................................................. Error! Bookmark not defined.
   3.1 Data Mining and Machine learning ........................................................................ Error! Bookmark not defined.
   3.2 Text Mining – Pre-processing texts ........................................................................ Error! Bookmark not defined.
   3.3 Classifiers for text classification .............................................................................. Error! Bookmark not defined.
   3.4 Opinion Mining ....................................................................................................... Error! Bookmark not defined.
   3.5 Performance measurement and results from other studies ...................................... Error! Bookmark not defined.

4 Proposed approach ......................................................................................................... Error! Bookmark not defined.
   4.1 Department needs and resources ............................................................................. Error! Bookmark not defined.
   4.2 Approach summary .................................................................................................. Error! Bookmark not defined.
   4.3 Data collection ........................................................................................................ Error! Bookmark not defined.
   4.4 Text mining system ................................................................................................. Error! Bookmark not defined.
      4.4.1 Implementation requirements ............................................................................ Error! Bookmark not defined.
      4.4.2 Functionalities .................................................................................................. Error! Bookmark not defined.
   4.5 Data Analysis .......................................................................................................... Error! Bookmark not defined.

5 Experimental set-up ........................................................................................................ Error! Bookmark not defined.

6 Results ............................................................................................................................. Error! Bookmark not defined.
   6.1 Experiment 1 .............................................................................................................. Error! Bookmark not defined.
   6.2 Experiment 2 .............................................................................................................. Error! Bookmark not defined.
   6.3 Experiment 3 .............................................................................................................. Error! Bookmark not defined.
   6.4 Experiment 4 .............................................................................................................. Error! Bookmark not defined.
   6.5 Data analysis results ................................................................................................. Error! Bookmark not defined.
7 Discussion........................................................................................................ Error! Bookmark not defined.

7.1 Overall performance...................................................................................... Error! Bookmark not defined.

7.2 Conclusions based on the data analysis....................................................... Error! Bookmark not defined.

7.3 Future work.................................................................................................... Error! Bookmark not defined.

8 Summary.......................................................................................................... Error! Bookmark not defined.

9 Bibliography..................................................................................................... 9

10 Appendix......................................................................................................... Error! Bookmark not defined.

10.1 ...................................................................................................................... Error! Bookmark not defined.

10.2 ...................................................................................................................... Error! Bookmark not defined.

10.3 ...................................................................................................................... Error! Bookmark not defined.

10.4 ...................................................................................................................... Error! Bookmark not defined.
Bibliography

   http://www.booking.com/general.en-gb.html?aid=303948&dcid=1&label=bookings-naam-3jHr6EsDyn1Uq9ttdGmKKws17149403061%3ApI%3Atha%3Ap1%3Ap2260%2C000%3Aac%3Aap1t1%3Aneg&lang=en-gb&sid=64fd0b6cb562f018c6cc9132ec9cd39f&tmpl=docs%2Fabout.


https://wiki.booking.com/display/INT/Content+Request+4+-
+Customer+Experience+Tab+in+Extranet+%28gettext%29.

32. Worditout. [Online] [Cited: 08 02 2013.]

33. IBM.com. [Online] [Cited: 15 02 2013.]

34. Online reputation management is hot — but is it ethical? Hoffman, T. s.l. : Computerworld, 2008.


37. Mayock, P. *Measuring the impact of online reviews on rate*. [Online]


hotelnewsnow, 16 08 2012. [Cited: 20 02 2013.]


43. *From Data Mining to Knowledge Discovery in Databases*. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. s.l. : American Association for Artificial Intelligence, 1996.


48. Norouzi, M, Fleet, D. J. and Salakhutdinov, R. Hamming Distance Metric Learning. s.l. : Departments of Computer Sciencey and Statistics University of Toronto.


55. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Joachims, T.


59. Techniques for Improving the Performance of Naive Bayes for Text Classification. Schneider, K. M.


64. *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?* Manning, C. D. s.l.: Departments of Linguistics and Computer Science.


74. CPAN. [Online] [Cited: 02 02 2013.] http://www.cpan.org/.


