Web scraping

Digital Kultur

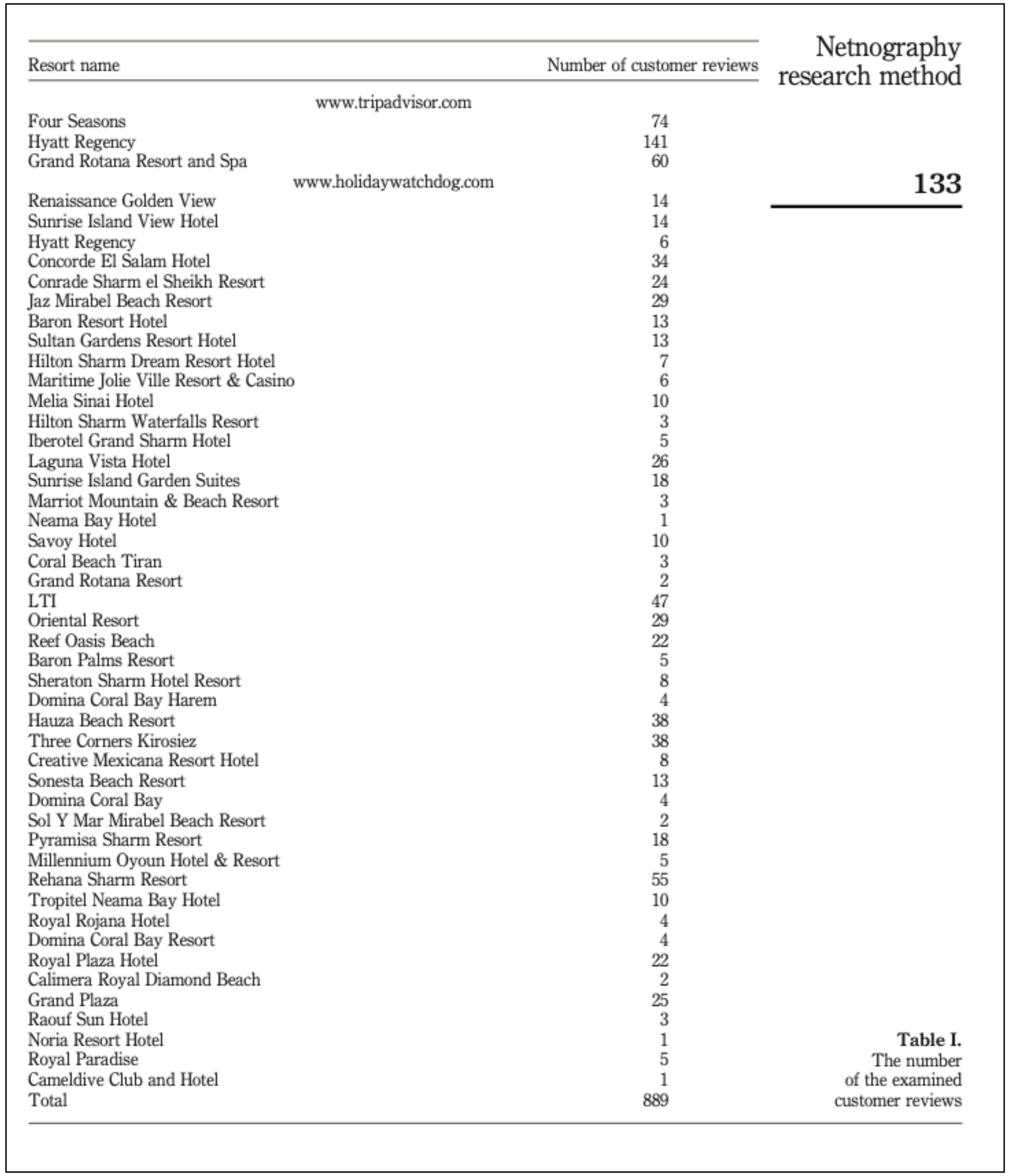
NIFR 2024 - https://github.com/nicklasdean/web-scraping-examples

Agenda

Digital Kultur

- Netnography - an example
- Web scraping configuration
- Web Scraping - static web pages
    - Exporting the results
- Web Scraping - dynamic web pages

Motivation for papier

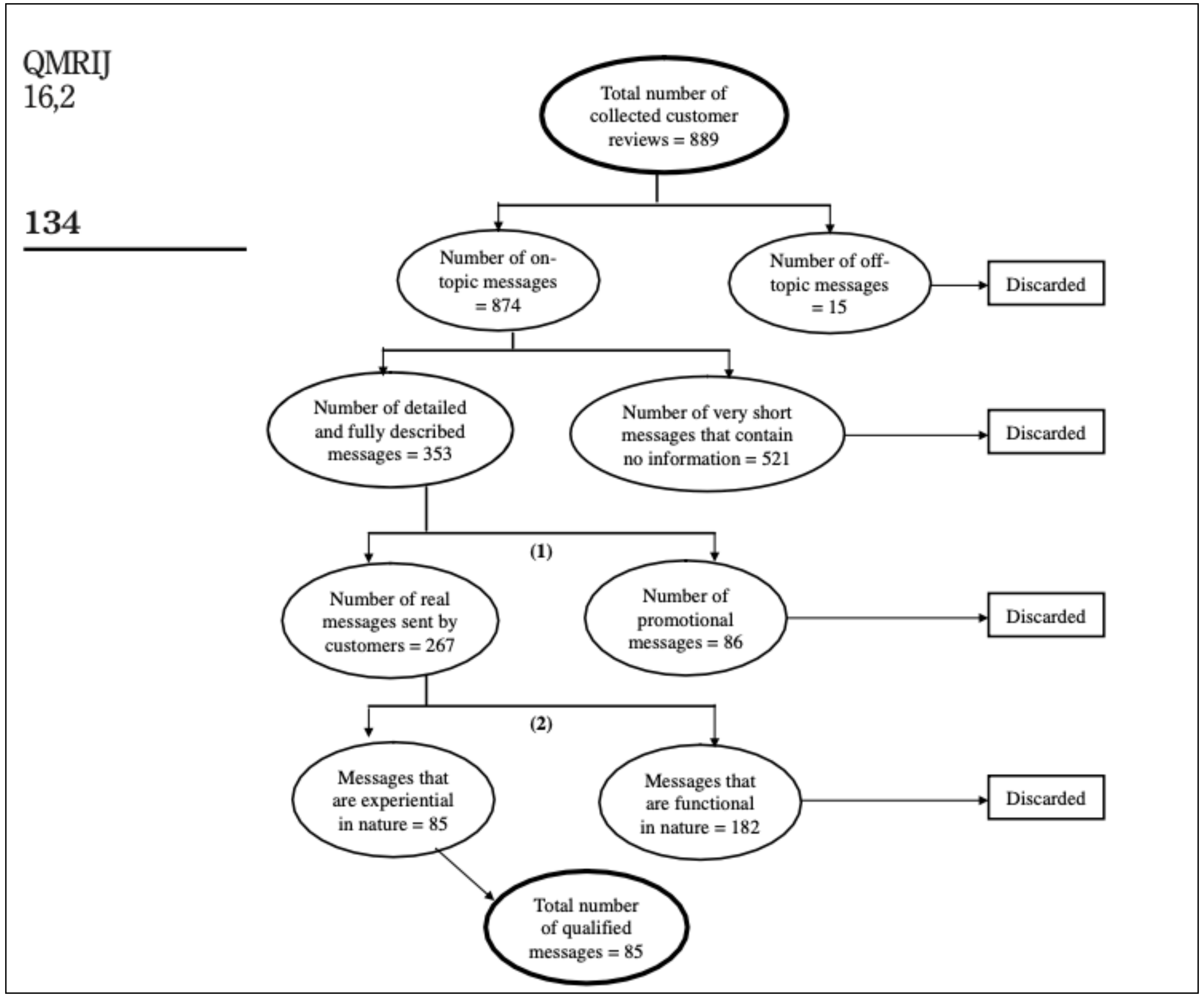


Netnography research method

133

Resort name	Number of customer reviews
www.tripadvisor.com	
Four Seasons	74
Hyatt Regency	141
Grand Rotana Resort and Spa	60
www.holidaywatchdog.com	
Renaissance Golden View	14
Sunrise Island View Hotel	14
Hyatt Regency	6
Concorde El Salam Hotel	34
Conrade Sharm el Sheikh Resort	24
Jaz Mirabel Beach Resort	29
Baron Resort Hotel	13
Sultan Gardens Resort Hotel	13
Hilton Sharm Dream Resort Hotel	7
Maritime Jolie Ville Resort & Casino	6
Melia Sinai Hotel	10
Hilton Sharm Waterfalls Resort	3
Iberotel Grand Sharm Hotel	5
Laguna Vista Hotel	26
Sunrise Island Garden Suites	18
Marriot Mountain & Beach Resort	3
Neama Bay Hotel	1
Savoy Hotel	10
Coral Beach Tiran	3
Grand Rotana Resort	2
LTI	47
Oriental Resort	29
Reef Oasis Beach	22
Baron Palms Resort	5
Sheraton Sharm Hotel Resort	8
Domina Coral Bay Harem	4
Hauza Beach Resort	38
Three Corners Kirosiez	38
Creative Mexicana Resort Hotel	8
Sonesta Beach Resort	13
Domina Coral Bay	4
Sol Y Mar Mirabel Beach Resort	2
Pyramisa Sharm Resort	18
Millennium Oyoun Hotel & Resort	5
Rehana Sharm Resort	55
Tropitel Neama Bay Hotel	10
Royal Rojana Hotel	4
Domina Coral Bay Resort	4
Royal Plaza Hotel	22
Calimera Royal Diamond Beach	2
Grand Plaza	25
Raouf Sun Hotel	3
Noria Resort Hotel	1
Royal Paradise	5
Cameldive Club and Hotel	1
Total	889

Table I. The number of the examined customer reviews



QMRIJ 16,2

134

Total number of collected customer reviews = 889

Number of on-topic messages = 874

Number of off-topic messages = 15

Discarded

Number of detailed and fully described messages = 353

Number of very short messages that contain no information = 521

Discarded

(1)

Number of real messages sent by customers = 267

Number of promotional messages = 86

Discarded

(2)

Messages that are experiential in nature = 85

Messages that are functional in nature = 182

Discarded

Total number of qualified messages = 85

Rageh, A., Melewar, T. C., & Woodside, A. (2013). Using netnography research method to reveal the underlying dimensions of the customer/tourist experience. *Qualitative Market Research An International Journal*, *16*(2), 126–149. doi:10.1108/13522751311317558

Netnography in detail

Insiders & Devotees

Kozinets highlights devotees and insiders as the most enthusiastic, actively involved and sophisticated users and thus as the most important data sources for researchers.

Bowler, G. M. (2010). Netnography: A Method Specifically Designed to Study Cultures and Communities Online. The Qualitative Report, 15(5), 1270-1275. https://doi.org/10.46743/2160-3715/2010.1341

Netnography in detail

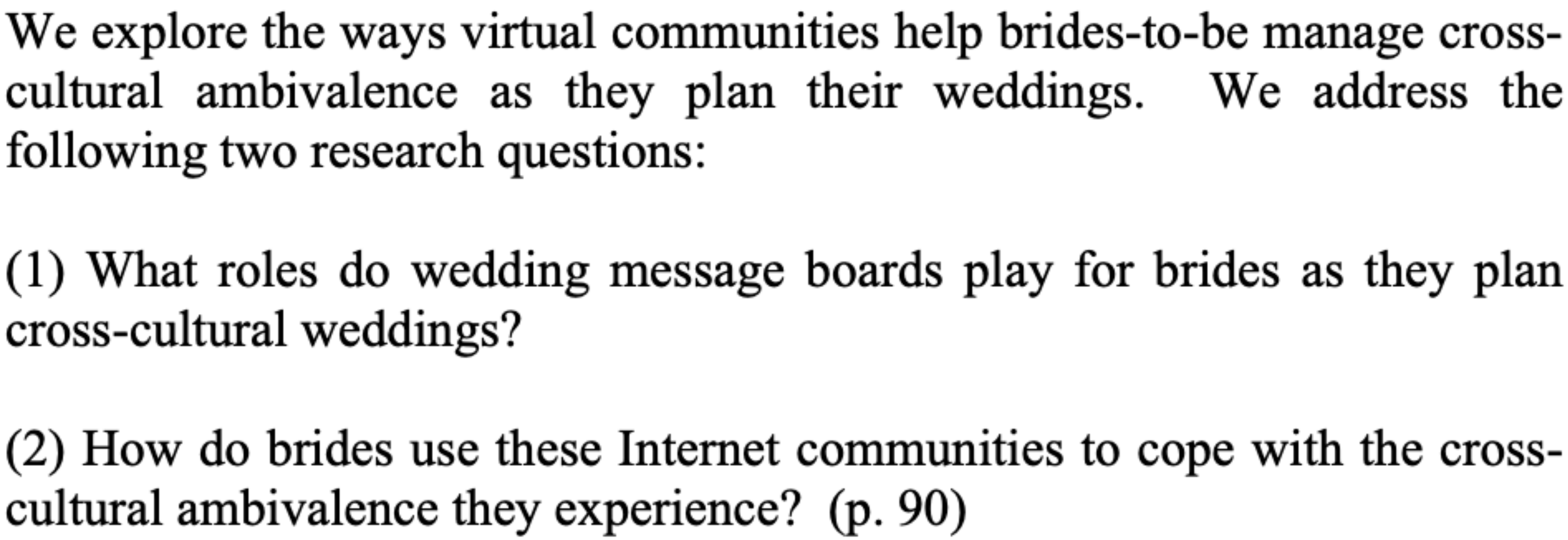
Strategy

- Ask one or two central questions followed by no more than seven related sub-questions.
- Relate the central question to the specific qualitative strategy of inquiry.
- Begin the research questions with the words “what” or “how” to convey an open-ended and emergent research design.
- Focus on a single phenomenon or concept.
- Use exploratory verbs such as “discover”, “understand”, “explore”, “describe”, or “report”.
- Use open-ended questions.
- Specify the participants and the research site for study.

Bowler, G. M. (2010). Netnography: A Method Specifically Designed to Study Cultures and Communities Online. The Qualitative Report, 15(5), 1270-1275. https://doi.org/10.46743/2160-3715/2010.1341

Netnography in detail

Research questions



We explore the ways virtual communities help brides-to-be manage cross-cultural ambivalence as they plan their weddings. We address the following two research questions:

(1) What roles do wedding message boards play for brides as they plan cross-cultural weddings?

(2) How do brides use these Internet communities to cope with the cross-cultural ambivalence they experience? (p. 90)

Bowler, G. M. (2010). Netnography: A Method Specifically Designed to Study Cultures and Communities Online. The Qualitative Report, 15(5), 1270-1275. https://doi.org/10.46743/2160-3715/2010.1341

Web scraping

Motivation

- To engage with and capture data that is accessible in the browser/web but has no API by building a scraper (bot)
- Lot of solutions exists (plugins, add-ons, IDE's)
  - Often times - a custom solution is necessary, except for very simple cases
- We will be using a popular framework Selenium

Configuration + Hello World

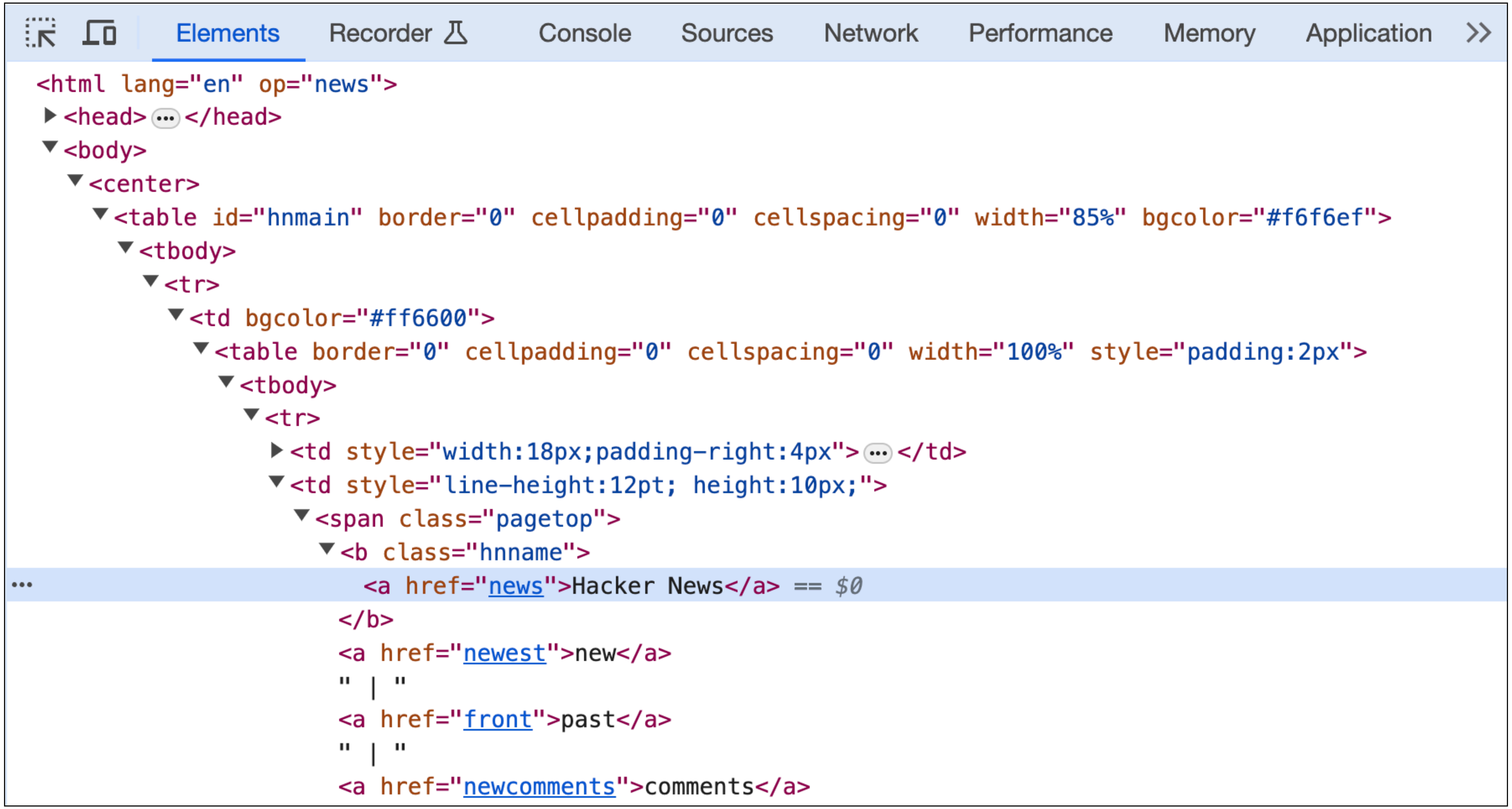
Selecting elements

Using the Selenium WebDriver

```
//Returns a single WebElement with HTML id = 35
driver.findElement(By.id("35"));

//Returns a list of Elements with HTML class = "row"
driver.findElements(By.className("row"));

//Returns a list of elements with the HTML tag <ul></ul>
driver.findElements(By.tagName("ul"));

//Returns a single element with the xpath query
driver.findElement(By.xpath("//*[@id=\"39587344\"]/td[3]/span/a"));
```

Deeply nested elements

Example - The Hackernews Title



```
<html lang="en" op="news">
 ▶<head> ⋯ </head>
 ▼<body>
   ▼<center>
     ▼<table id="hnmain" border="0" cellpadding="0" cellspacing="0" width="85%" bgcolor="#f6f6ef">
       ▼<tbody>
         ▼<tr>
           ▼<td bgcolor="#ff6600">
             ▼<table border="0" cellpadding="0" cellspacing="0" width="100%" style="padding:2px">
               ▼<tbody>
                 ▼<tr>
                   ▶<td style="width:18px;padding-right:4px"> ⋯ </td>
                   ▼<td style="line-height:12pt; height:10px;">
                     ▼<span class="pagetop">
                       ▼<b class="hnname">
⋯                           <a href="news">Hacker News</a> == $0
                         </b>
                         <a href="newest">new</a>
                         " | "
                         <a href="front">past</a>
                         " | "
                         <a href="newcomments">comments</a>
```

Exercises: Pairs in DK2 groups