# Massive Data の収集・分析手法を用いた観光イメージ分析

## ――宮島に関する Trip Advisor[1] の英文 Reviews を事例に――

金　　徳　謙

（受付　2018年 10 月 31 日）

## Ⅰ　は　じ　め　に

抜井（2012）は，ICT の発達により，過去 1 位だった観光情報の収集方法が「家族・友人の話」から「インターネット」にとって変わったとしている。このような観光情報の収集方法の移行のながれは現在もつづき，従来の受け入れ側が提供する情報を収集する方法からさらに進み，現在は，ウェブサイトはもちろん，情報利用者の多くが参加もし，情報を提供するようになった。多種多様の SNS が質の高い情報として収集対象とされ，今日の観光において情報収集の対象として欠かせない存在になっている。

SNS には投稿するひと[2] が提供する情報（どこで何をしていたのかなどの経験）をはじめ，観光地の様子などがリアルタイムで投稿されるため，他の観光者に信頼できる有益な情報として利用されている。このため，SNS に対する観光者からの注目度は高まる傾向にある。

しかし，若山（2016）が「このような情報を整理し有効活用する技術の研究は十分には進んでいない」と指摘しているとおり，観光分野において

1）　Trip Advisor は2000年に設立された，観光地，レストラン，ホテルなどのレビューやホテルの予約，その他観光に関連する各種情報を提供するアメリカの企業である。現在，世界中で 3 億1,500万人を超える利用者をもつ世界最大の観光情報提供サイトである。日本を訪れる多くの外国人観光者もこのサイトを利用し，各種情報の収集やレビューの掲載を行っている。

2）　本稿ではレビュー投稿者すべてを観光者と見なす。

も情報の収集や活用の技術を応用する研究は始まったばかりといえ，日本では抜井（2012），金（2013及び2015），馬場ら（2017）などがみられる程度である。このため，さらなる研究の蓄積が必要とされる。他方で，外国の研究に目を向けると，とくに英文研究誌において諸外国からの研究が多くみられる。Xiang et al.（2010）は，オンライン上の観光情報において，SNS は観光情報の提供だけでなく，観光研究においても重要な役割を果たすことになるとし，SNS に注目する必要性を指摘している。Amaro et al.（2016）は，旅行計画を立てる際，SNS の影響は大きいと指摘した上で，旅行者と検索情報の関係を分析している。Kim et al.（2017，韓国語文献）は，SNS が提供する情報の質が観光地のイメージに影響することを明らかにし，SNS の適切な利用がマーケティング戦略の策定などに役立つと指摘している。初期の研究は今後の観光研究における必要性が注目された SNS を取り上げるものが多くみられたが，近年 SNS の活用方法や研究の有効性の検証などの研究へと拡大した。応用先や分析手法など，さらに専門的な研究も加わり，SNS を取り上げる研究は多岐にわたっている。このように観光分野における SNS 研究が重要視され，研究の進展が確認できる。

これらの研究は SNS への投稿内容に着目した研究や，いつ・どこで投稿したかなどの場所と時間に着目した研究など，研究内容は多岐にわたる。分析対象も，従来の文字だけの分析にとどまらず，画像（主に写真）や動画を取り上げるなど，拡大している（金；2015など）。若山（2016）は，SNS を代表するツイッター投稿文を分析することで位置情報が分かるため，①話題の場所の最新状態の確認，②投稿したユーザーの趣味嗜好や属性に応じた観光情報の提供，③訪れた観光地の順路特定などができるが，場所情報が含まれているのは全体の0.18％に過ぎず，投稿文言から場所を推定する研究が盛んであるとしている。

SNS の普及は，情報量の増加を意味し，観光業者にとっては，大量の SNS データを収集・分析することで観光者のニーズをより正確に把握することができるため，精度の高い需要予測や満足度の高いサービスの提供に

つながるメリットがある。SNS に投稿されたデータの内容を分析する代表的な手法はテキスト分析であり，近年，とくに SNS の大量の文字（ビッグ）データから有益な情報を抽出する手法として利用されている。しかし他方で，増加する SNS データを効率よく収集するためにはプログラミングについての知識が求められ，それがサプライヤー[3] 側の SNS 分析の普及を妨げる高いハードルとなっている。このため，とくに日本における観光研究ではアンケートなど従来の手法による研究がほとんどで，真の観光者のニーズ解明に課題が多く残っている。

そこで，本研究では①外国人観光者が抱く観光地についてのイメージを明らかにすること，②SNS からの大量なデータの収集や分析に用いる手法の有効性の検証，③SNS からのデータ収集や分析に必要なプログラミング知識がなくても SNS データの収集や分析に利用できるコードの作成・開示を目的として，世界最大の観光情報提供サイト Trip Advisor に掲載されている広島県廿日市宮島に関する全てのレビューを収集しテキスト分析を行う。

## Ⅱ　データの収集

### 1　収集対象

本研究では，世界最大の観光情報提供サイト Trip Advisor に広島県の全観光スポットの中，レビューがもっとも多く投稿されている宮島に関するすべてのレビューを収集，分析する。

レビューを収集するサイトは Trip Advisor[4] のトップページから‘Miyajima’で検索した結果のトップ画面である。データ収集の対象は図

3)　一般的には観光客に必要なサービスの提供を事業とする側の総称であるが，本稿ではこれに加え，自治体など観光客のニーズや行動，意識などに強い関心をもつ広義の事業者を含むものとした。

4)　本稿では Trip Advisor 日本語サイトではなく，https://www.tripadvisor.com/（英語サイト）をもとにしている。

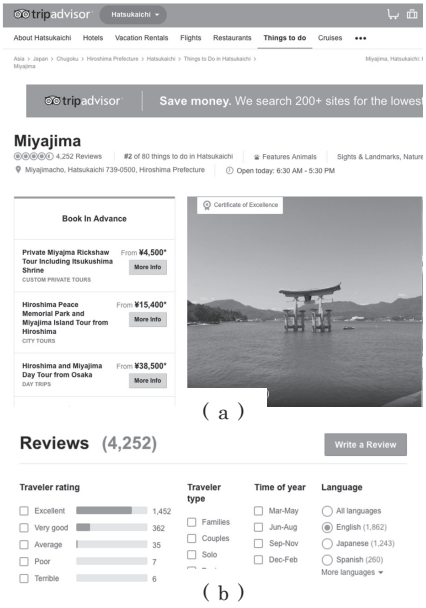
1-a で確認できるとおり，レビュー件数は全4,252件で，日本語（1,243件）や英語（1,862件）の他にも中国語，韓国語，スペイン語など様々な外国語によるレビューが掲載されている[5)]。今回はその中で，英語のレビュー（図 1-b 参照）を全てを収集し，分析に用いる。

## 2　収集内容・方法

### (1)　収集内容

本研究で取り上げる SNS（インターネット上で提供されるデータを含む）のような大量データは，人手により 1 件ずつ収集することはほぼ不可能で

**図 1　tripadvisor.com での “Miyajima” 検索結果**




5)　レビューは，多くの利用者がリアルタイムで掲載するため，掲載件数もリアルタイムに変化する。本研究に用いたデータの収集は2018年10月29日に行っているため，閲覧時のレビュー件数と異なることもあり得る。本研究におけるレビュー件数（掲載件数）は図 1 に示した件数をさす。

ある。そのため，本稿では Web Scraping 手法[6] を用いて Trip Advisor に掲載されているレビューを収集する。収集する項目はレビューに焦点を当てるため，「レビュー」に限定した。今回収集するレビューは検索のトップ画面（図 1-a 及び図 1-b 参照）に示されている全1,862件である。

(2) 収集方法

Web Scraping 手法によるデータ収集には様々な方法があるが，いずれの方法も多少のコード作成（プログラミング）が必要となる。今回は，コード作成が比較的簡単とされる“R”言語（以下，R と記す）を用いて行う。

なお，本稿で用いるデータの収集及び分析方法が，プログラミングについての知識がない研究者にも利用可能な研究手法となることを期待し，データの収集から分析までのコードを開示することにする。

初めに，レビュー収集のために必要な Package を R にロードした上で Trip Advisor のトップページから Miyajima を検索し，掲載されているレビューの件数などを確認して収集に必要なコード作成を過程別に区分する。

・データ収集のための準備過程[7]

レビューを収集するために R と組み合わせて利用する各種パッケージをロードする準備段階で，詳しくは次のコード文から確認できる。

---

6) インターネット上の HP などは簡単な情報伝達の場合に使われる HTML 型式や複雑な構造で多くの情報を効率よく提供できる XML 型式などで提供されている。近年は後者の型式による情報提供が多くみられる。提供されている情報が構造化されているため，効率よく収集することができる。本研究では，インターネットに提供される情報を収集する Web Scraping 手法を用いて Trip Advisor のレビューを収集する。

7) グレー枠内はコード実行箇所で，白枠内の‘##’で始まる箇所はコードの実行結果である。

・データ収集に必要なPackageをロードする

```
library(rvest)
```

```
## Loading required package: xml2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
```

・データ収集の過程

ここでは，SNS に掲載されているすべてのレビューを収集するために必要な準備（全体のレビューの量の確認及びレビューを自動で一括収集するためのコード作成）を行う。

次に，収集したレビューから空白や改行など，分析に不要な部分を削除するクリーニング作業を行い，テキストデータのみを抽出する。また，複数項にわたり掲載されている全てのレビューを一括収集するため，URL（情報が掲載されているインターネット上の住所）を分割するコードを作成し，プログラミングにより自動作業でデータを取り込むための準備を行う。

・データ取得のため、宮島検索のトップページからデータの取得確認を行い、　２ページ以降すべてのページからデータを取得するため、ページ設定ができるように準備

```
#最初のページを取得
url <- "https://www.tripadvisor.com/Attraction_Review-g1022438-d1161271-Reviews-Miyajima-Hatsukaichi_Hiroshima_Prefecture_Chugoku.html"

# Get the title and reviews from above url.
trip.adv <- read_html(url)
trip.rev <- html_nodes(trip.adv,"p.partial_entry") %>% html_text()
trip.rev <- str_remove_all(trip.rev,"\n")
trip.rev <- str_remove_all(trip.rev," ") %>% as.data.frame()

####Get Data from after 2page of Miyajima in Trip Advisor.
# urlの取得。複数ページを取得するため、変化するページ数を記入する箇所を前後して
# 2つに分けて、間にページ数を挟み、urlを組み立てる。
url2 <- "https://www.tripadvisor.com/Attraction_Review-g1022438-d1161271-Reviews-or"
url3 <- "0-Miyajima-Hatsukaichi_Hiroshima_Prefecture_Chugoku.html"
```

・データ取得のため、宮島検索のトップページからデータの取得確認を行い、　2ページ以降すべてのページからデータを取得するため、ページ設定ができるように準備

```
#最初のページを取得
url <- "https://www.tripadvisor.com/Attraction_Review-g1022438-d1161271-Reviews-Miyajima-Hatsukaichi_Hiroshima_Prefecture_Chugoku.html"

# Get the title and reviews from above url.
trip.adv <- read_html(url)
trip.rev <- html_nodes(trip.adv,"p.partial_entry") %>% html_text()
trip.rev <- str_remove_all(trip.rev,"\n")
trip.rev <- str_remove_all(trip.rev," ") %>% as.data.frame()

####Get Data from after 2page of Miyajima in Trip Advisor.
# urlの取得。複数ページを取得するため、変化するページ数を記入する箇所を前後して
# 2つに分けて、間にページ数を挟み、urlを組み立てる。
url2 <- "https://www.tripadvisor.com/Attraction_Review-g1022438-d1161271-Reviews-or"
url3 <- "0-Miyajima-Hatsukaichi_Hiroshima_Prefecture_Chugoku.html"
```

取得するすべてのURLをもとに自動作業でデータを取得するため，コードを作成した。繰り返す頃数は検索トップ画面（図 1-a）で確認した187頃である。そのため，コード文には次のとおり，187－1の186回を指定した上でデータ取得を進めていく。

**・レビュー数の確認と取得するページ数を指定し、レビューデータの取得と保存を自動で行う　ようにコードを用意する**

```
trip.rev.all <- NULL

# 最後のレビューページが187なので、for文を186 (187-1) までと指定。
for(i in 1:186){
  # urlをfor文と組み合わせて作り直す
  url_f <- paste0(url2,i,url3)

  #各ページからレビューデータを抽出、取得してデータフレームに変換。
  trip.adv_f <- read_html(url_f)
  trip.rev_f <- html_nodes(trip.adv_f,"p.partial_entry") %>% html_text() %>% as.data.frame()

  # 各ページで取得したデータをバインドしてファイルに書き込む。
  # そのあと、全部をrbindしてまとめる。
  trip.rev.all <- rbind(trip.rev.all,trip.rev_f)

  #次のデータ取得のためにtrip.rev_fを空けてから、指定回数まで繰り返す。
  trip.rev_f <- NULL
}
```

```
# 最初のページのデータと2ページ以降のデータをまとめ、フィールド名を変更する。
trip.rev.all <- rbind(trip.rev.all,trip.rev)
colnames(trip.rev.all) <- c("review")
#最初の１０件を確認する
head(trip.rev.all,2)
```

ここまでのコード文の実行には1,862件にのぼるレビューを収集するため，多少時間がかかるが，それにより全てのレビューが収集され，テキスト文のみを残すクリーニング作業まで終了する。収集したレビューの内容確認のため，最初の２件（１番目と２番目のレビュー）と最後の２件

（1,861番目と1,862番目のレビュー）を確認したところ，次のとおり内容が表示され，レビューが正しく収集されていることが確認できた。

```
##
                                                                    review
## 1                  If you do Hiroshima, follow with a happy place. This island is miraculous and magical. You are surrounded b
y natural beauty but cuddled at the same time by cultural wonders. The shrines will knock you out. Run away from the ch
eesy village, put on comfy...More
## 2 Whoever's advertising this place is not doing a good enough job. The island is much more lively and developed than
I'd imagined. The staff at our hostel recommended staying overnight, when many of the tourists have left. The floating t
orii and itsukushima shrine are, of...More
```

```
#最後の10件を確認する
tail(trip.rev.all,2)
```

```
##                                                                                                                      re
view
## 1861       WecaughttheelectrictramfromHiroshimathenaJRFerrytotheisland.TheferrypullsupdirectlybehindthefloatingTorigatewhichenticesyoutovisitit.MakesureyoulookuphightideandlowtidetimestogetthebestMore
## 1862 Mustbedone.Theislandissimplyfantasticandtheltsukushimashrineisunique.ItisbesttotaketheferrybeforeBam,thenyoucantakenicephotosoftheshrinewithoutmanypeople.ThehiketoMt.MiesenisbeautifulandtheviewMore
```

最後は次章の分析に用いるため，収集したレビューを csv 形式のファイル（ファイル名：miyajima_revieweng.csv）として保存する過程である。コード文は次のとおりである[8)]。

・取得したデータを分析に用いるため、保存する

```
# 最後にShift-JIS形式にencodingして、csv形式に出力する。
write.csv(trip.rev.all,"miyajima_review_eng.csv",fileEncoding = "cp932")
```

## Ⅲ　分　　析

近年，テキストマイニングは SNS のレビューなど大量の文字データから有益な情報を抽出する手法として注目されている。本章では，前章で示し

8）R は Windows 及び Mac OS を含む Unix 系などマルチ OS 対応ソフトである。そのため，使っているパソコン環境（OS）に合わせて文字コードを指定する必要がある。文字コードの指定は，Unix 系 OS には utf-8を，Windows 系 OS には Shift-JIS（cp-932）を指定する必要がある。文字コードが異なると，文字化けにより意味不明の文字となる。本稿では，Windows OS での利用に合わせ，文字コードを cp932と指定した。

た手法を用いて収集したレビューを，次の２つのテキストマイニング手法を用いて分析していく。

まず，レビューをもとに形態素分析を行い，宮島がどのような形態素で表現されているのかを分析する。次に，感情分析（Sentiment Analysis）を行い，宮島に対するイメージを抽出する。

分析は，データの収集と分析を１つのパッケージで行える点，及び必要なコードの作成が他のコンピュータ言語より優しい点，の２点から有用であると考えられるため，データ収集と同じ R を用いることとし，コード文を作成して分析を進めていく。

## 1　形態素分析

本節では，レビューにどのような形態素がどの程度使われているのかを分析する，いわゆる形態素別の頻度分析を行う。そのため，分析に不要な意味を持たない形態素[9)] を取り除き，意味をもつ形態素のみ抽出した上でレビューでの出現頻度を分析し，グラフ化（図２参照）していく。

まず，分析に必要なパッケージを読み込む過程である。

**・取得したデータを頻度分析に用いるため、データ形式を整え、分析を準備する過程**

```
#データ分析（頻度分析）に必要なPackageをロードする
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

9)　分析に不要な形態素とは，‘the’，‘that’，‘this’，‘of’，‘for’，‘a’ などのような英語独特な語で，特別な意味をもっているものではないが，文中に多く使われている。これらの形態素を含めて分析すると，本来の意図とは異なる結果になってしまうため，分析の前に取り除く必要がある。本研究では，この類の形態素を集めたファ ↗

次に，前章で収集，保存した1,862件のレビューを読み込み，分析に不要な意味を持たない形態素を取り除き意味をもつ形態素のみを抽出した。続いて，有意味な形態素がレビューで使われた頻度を調べるため，頻度分析を行った。

この結果，使われていた形態素は3,907種で，もっとも多く出現した形態素はisland 1,177回であり，その他 ferry 824回，hiroshima 786回であった。

```
library(tidytext)
library(tidyr)
library(stringr)
library(ggplot2)

trip <- read.csv ("/Users/kimutoku/Desktop/miyajima_review_eng.csv",
                  header = TRUE,
                  stringsAsFactors = FALSE)
trip_df <- data_frame(line = 1:1862, text = trip$review)


#↑分析に不要な'the'、'a'、'of'、'for'などを取り除く過程

#英文の分析に不要な表現を取り外すためのパッケージを呼び込む
data(stop_words)


#ファイル内の単語の出現頻度を調べ、出現順に並び替える過程

trip_freq <- trip_df %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)
```

```
## Joining, by = "word"
```

続いて，レビューに用いられた形態素から上位10位までを抽出し，出現頻度を調べた。出現頻度は，次のコードの実行結果から確認できる。

```
# 出現頻度が高い上位10位の単語の出現頻度を確認
head(trip_freq, 10)
```

```
## # A tibble: 10 x 2
##    word        n
##    <chr>     <int>
##  1 island    1177
##  2 ferry     824
##  3 hiroshima   786
##  4 miyajima    768
##  5 day       632
##  6 jr        452
##  7 beautiful   418
##  8 trip      397
##  9 deer      387
## 10 shrine     379
```

↘ イル（stop words）を参考に，レビューの中から不要な形態素を取り除き，分析する。

次に，出現頻度が多い形態素（200回以上出現）を抽出した。その結果は次のコードの実行結果に示したとおり，出現頻度がもっとも多い island 1,177回からもっとも少ない walk 211回まで，20個の形態素が検出された。

```
# 頻度分析の結果をグラフで表示し、分析に用いた200回を点線で示す過程
# 出現頻度をもとに分析し、グラフ化する
trip_freq %>%
  # 出現頻度200回以上のワードのみ抽出する過程
  filter(n > 200) %>%

  # 出現頻度が多い順に並び替える過程
  mutate(word = reorder(word, n)) #%>%
```

```
## # A tibble: 20 x 2
##    word        n
##    <fct>     <int>
##  1 island    1177
##  2 ferry      824
##  3 hiroshima  786
##  4 miyajima   768
##  5 day        632
##  6 jr         452
##  7 beautiful  418
##  8 trip       397
##  9 deer       387
## 10 shrine     379
## 11 visit      364
## 12 gate       343
## 13 train      325
## 14 japan      301
## 15 pass       271
## 16 torii      260
## 17 tide       250
## 18 ride       246
## 19 time       225
## 20 walk       211
```

200回以上の出現が検出された形態素をグラフ化すると図２のとおりであり，上位５位までの形態素の出現頻度がとくに多い。さらに，１位 island は２位 ferry との差が大きく，island の出現頻度の多さが際立つ。

最後に，レビューに使われた形態素を，出現頻度をもとに文字の大きさと表示場所を変え，分かりやすく可視化できる WordCloud 分析を実施した。その結果は図３に示したとおりであり，island，hiroshima，miyajima，ferry，beautiful，day などが中心に大きく表示され，これらの形態素の出現頻度の高さが分かる。

図 2　頻出形態素（200回以上）

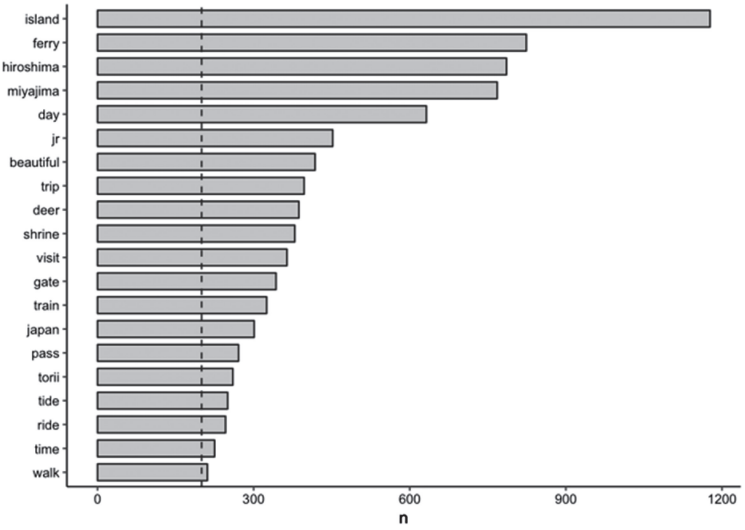



図 3　頻出形態素の可視化


## 2　感 情 分 析

前節の頻度分析に続き，本節ではレビューに使われた形態素がもつ意味を分析していくため，使われた形態素を positive と negative に区分し，一

定回数以上の出現頻度をもつ形態素を抽出し，宮島のイメージを検討していく。詳しい分析の手順は次に示すとおりである。

まず，分析に必要なパッケージを読み込んだ上で positive と negative に区分するための感情データを読み込み，レビューの形態素を区分していく。

```
# 必要なPackageをロードする過程
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidytext)
library(tidyr)

#頻度分析で抽出、保存したデータを分析のため読み出す過程
trip_freqency <- read.csv("0_trip_freq.csv",head = TRUE)
trip_freqency <- trip_freqency[,-1]

# "bing"を使い、表現を「ポジティブ」と「ネガティブ」に区分して
# ポジティブ−ネガティブで差をカウントする段階
trip_freq_sentiment <- trip_freqency %>%
  inner_join(get_sentiments("bing")) %>%
  spread(sentiment,n,fill = 0) %>%
  mutate(sentiment = positive - negative)
```

```
## Joining, by = "word"
```

```
## Warning: Column `word` joining factor and character vector, coercing into
## character vector
```

・positive 要因

ここではレビューの中から positive な形態素を抽出し，さらに出現頻度50回超の形態素をグラフ（図 4 参照）を用いて可視化した。

分析では positive なイメージをもつ形態素328種が検出されたが，その中に出現頻度 1 回の形態素が132種確認された。次に，形態素からイメージを分析するため，出現頻度50回超の形態素を抽出し，可視化を行った。結果は図 4 及び図 5 に示したとおりで，beautiful の出現頻度がもっとも多く，400回を超ている。

```
library(ggplot2)
# positive 表現の抽出・グラフ表示
posi <- trip_freq_sentiment %>% filter(positive > 50)
ggplot(data = posi, aes(word,positive)) +
 theme_classic() +
 geom_col(width = .7, fill = "gray80", colour = "black") +
 xlab(NULL) +
 # グラフを90度回転させる
 coord_flip() +
 geom_hline (yintercept = 50, lty = 2, colour = "red") +
 geom_hline (yintercept = 100, lty = 2, colour = "red") +
 geom_hline (yintercept = 150, lty = 2, colour = "red") +
 geom_hline (yintercept = 200, lty = 2, colour = "red")
```

図 4　ポジティブ表現（頻度50回以上）

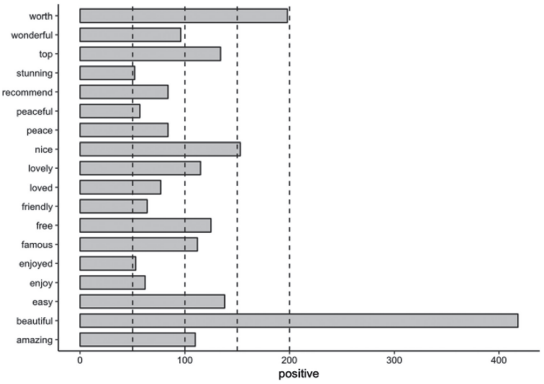



図 5　多頻度形態素

| | word | positive |
|---|---|---|
| 1 | beautiful | 418 |
| 2 | worth | 198 |
| 3 | nice | 153 |
| 4 | easy | 138 |
| 5 | top | 134 |
| 6 | free | 125 |
| 7 | lovely | 115 |
| 8 | famous | 112 |
| 9 | amazing | 110 |
| 10 | wonderful | 96 |
| 11 | peace | 84 |
| 12 | recommend | 84 |
| 13 | loved | 77 |
| 14 | friendly | 64 |
| 15 | enjoy | 62 |
| 16 | peaceful | 57 |
| 17 | enjoyed | 53 |
| 18 | stunning | 52 |

この分析から宮島について positive なイメージをもっている外国人観光者が多いことが確認できた。また，自然景観との関連性が強い形態素の出現回数が多く，宮島の自然が魅力要因になっていることが確認できる。その他，本研究で用いた分析だけで断定できないが，島であることが影響していると考えられる形態素や，アクセスのしやすさなど，都市観光地で魅力要因となる形態素が多数観測された。

以上を踏まえ，宮島は都市観光地広島の一観光スポットとして認識されており，都市内の観光スポットとしては珍しい島であるといえること，また，自然豊かな点が positive なイメージを与えていることが確認できた。

・negative 要因

ここではレビューの中から negative な形態素を抽出し，さらに出現頻度50回超の形態素をグラフ（図 6 参照）に可視化した。

分析では negative なイメージをもつ形態素が234種検出されたが，これは positive な形態素に比べ少なく，出現頻度 1 回の形態素が148種を占めていた。次に，形態素からイメージを分析するため，出現頻度が10回超の形態素を抽出し，可視化を行った。結果は，図 6 及び図 7 に示したとおりであり，出現頻度が多い順に rail（95回），wild（62回），crowded（51回）となった。

```
# negative 表現の抽出・グラフ表示
nega <- trip_freq_sentiment %>% filter(negative > 10)
ggplot(data = nega, aes(word,negative)) +
  theme_classic() +
  geom_col(width = .7, fill = "gray80", colour = "black") +
  xlab(NULL) +
  # グラフを90度回転させる
  coord_flip() +
  geom_hline (yintercept = 10, lty = 2, colour = "red") +
  geom_hline (yintercept = 20, lty = 2, colour = "red") +
  geom_hline (yintercept = 30, lty = 2, colour = "red") +
  geom_hline (yintercept = 50, lty = 2, colour = "red")
```

前節の positive 要因の分析により，当地の魅力要因の抽出ができるのに対し，本節の negative 要因の分析では，改善が望ましい要因が抽出できる。本研究の分析から明らかになったことは，騒がしさや値段が高いことなど，社寺観光地や都市観光地特有の negative イメージである。その他，本研究で用いた分析のみでは断定できない点が多く観測された。例えば，図 6 及び図 7 で確認できるように rail（95回），wild（62回），bomb（24回）など，negative と positive の両方に捉えられる形態素が多いことが分かる。この点は感情分析がもつ限界といえ，状況に合わせた解析が必要であり，本研究ではどちらの要因としても採択しないこととした。

図 6　ネガティブ表現（頻度10回以上）

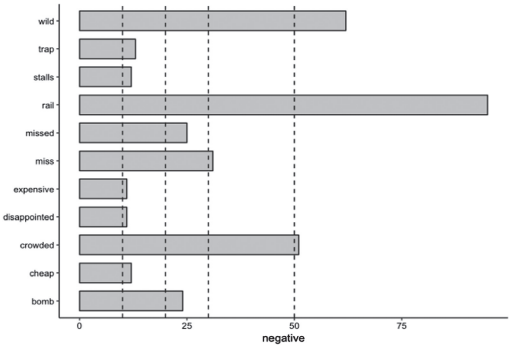



図 7　多頻度形態素

| | word | negative |
|---|---|---|
| 1 | rail | 95 |
| 2 | wild | 62 |
| 3 | crowded | 51 |
| 4 | miss | 31 |
| 5 | missed | 25 |
| 6 | bomb | 24 |
| 7 | trap | 13 |
| 8 | cheap | 12 |
| 9 | stalls | 12 |
| 10 | disappointed | 11 |
| 11 | expensive | 11 |

## Ⅳ　ま　と　め

本研究では，宮島についてのレビューを題材にテキストマイニング手法を用いて分析を行い，外国人観光者が抱く宮島に対するイメージ，及びその影響要因を明らかにした。

形態素の頻度分析では，宮島は都市観光地広島の一観光スポットであることが確認される結果となった。その理由に，景観がいいこと，水上に浮かぶ厳島神社，JR で簡単にアクセスできる（日帰りできる）こと，日本（のイメージ）を表していることなどの表現が多種かつ，多頻度で用いられていることも確認できた。

感情分析では，外国人観光者は宮島を positive なイメージで捉えていることが確認された一方，出現頻度は少ないが，‘crowded’ や ‘expensive’ など negative な表現が抽出され，宮島観光の課題を明らかにすることができた。

本研究では，レビューの分析をとおして，観光者による宮島のイメージの検出と観光地としての課題を外国人観光者の目線から検出する成果をあげることができた。また，近年 SNS から大量の文字や画像を収集，分析することから有益な情報抽出を研究に応用することが進む中，本研究に用い

たデータの収集や分析の手法が観光研究の有効なツールになることを実証でき，観光研究において新たな手法を提示することができた。

## 参 考 文 献

金　徳謙（2013）：香川県直島にみる SNS 書込内容の分析に基づく観光者の類型化，日本観光研究学会全国大会論文集，第28回，pp. 313–316.

金　徳謙（2015）：観光資源の利用実態の解明に向けた画像ビッグデータの空間分析，中四国商経学会第56回研究発表大会，2015年12月.

抜井ゆかり（2012）：テキストマイニンクを用いたトラヘルライティンク分析による観光シソーラスの構築，観光科学研究，Vol. 5，pp. 177–184.

馬場　武・萩野　誠（2017）：SNS の話題性分析から見る離島観光，奄美ニュースレター，No. 41，pp. 1–6.

若山公威（2016）：ツイートからの観光ルート抽出，名古屋外国語大学外国語学部紀要，第50号，pp. 167–177.

Amaro, S.; P. Duarte; and C. Henriques. 2016. Travelers' use of social media: A clustering approach. *Annals of Tourism Research* 59: 1–15.

Chua, A.; L. Servillo; E. Marcheggiani; and A.V. Moere. 2016. Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management* 57: 295–310.

Coletto, M.; A. Esuli; C. Lucchese; C.I. Muntean; F.M. Nardini; R. Perego; and C. Renso. 2017. Perception of social phenomena through the multidimensional analysis of online social networks. *Online Social Networks and Media* 1: 14–32.

Fang, B.; Q. Ye; D. Kucukusta; and R. Law. 2016. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management* 52: 498–506.

Kim, S.-E.; K.Y. Lee; S.I. Shin; and S.-B. Yang. 2017. Effects of tourism information quality in social media on destination image formation: The case of Sina Weibo. *Information & Management* 54: 687–702.

Lansley, G. and P.A. Longley. 2016. The geography of Twitter topics in London. *Computers, Environment and Urban Systems* 58: 85–96.

Liu, Z. and S. Park. 2015. What makes a useful online review? Implication for travel product websites. *Tourism Management* 47: 140–151.

Lloyd, A. and J. Cheshire. 2017. Deriving retail centre locations and catchments from geo-tagged Twitter data. *Computers, Environment and Urban Systems* 61: 108–118.

Mariani, M.M.; M. Di Felice; and M. Mura. 2016. Facebook as a destination marketing tool: Evidence from Italian regional Destination Management Organizations. *Tourism Management* 54: 321–343.

Marine-Roig, E. and S. Anton Clavé. 2015. Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management* 4: 162–172.

Oku, K.; F. Hattori; and K. Kawagoe. 2015. Tweet-mapping Method for Tourist Spots Based on Now-Tweets and Spot-photos. *Procedia Computer Science* 60: 1318–1327.

Vecchio, P.D.; G. Mele; V. Ndou; and G. Secundo. 2017. Creating value from Social Big Data: Implications for Smart Tourism Destinations. *Information Processing & Management*.

Xiang, Z. and U. Gretzel. 2010. Role of social media in online travel information search. *Tourism Management* 31: 179–188.

Xiang, Z.; Z. Schwartz; J.H. Gerdes; and M. Uysal. 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management* 44: 120–130.

안　진현；김　응희；and 김　홍기. 2017. Big Data based Tourist Attractions Recommendation－Focus on Korean Tourism Organization Linked Open Data－. *Management & Information Systems Review* 36: 129–148.【韓国語】