DOI : 10.17314/jjlc.2024..69.005

LDA 토픽모델링 기법을 활용한 소셜미디어의 일본 관련 텍스트의 토픽 분석*

김 유 영**

<目 次>

Key Word : 텍스트 마이닝(Text Mining), 토픽 모델(Topic Model), LDA(Latent Dirichlet Allocation, 소셜 빅데이터(Social big data), 블로그(Blog), 네이버(Naver)

1. 들어가며

Melvin M. Vopson(2021)[1]에 따르면, 전 세계적으로 2020년 한 해에만 생성된 데이터가 약 59제타바이트(ZB)에 이른다고 하는데, 이는 무려 약 59조 기가바이트(GB)에 해당하는 정보량이 된다. 게다가 2025년에 이르러서는 정보량이

* 이 논문은 2023년도 동덕여자대학교 연구비 지원에 의하여 수행된 것임(연구번호: 20230266). This study was supported by the Dongduk Women's University grant(No. 20230266)

** 동덕여자대학교 일어일본학과 부교수, 일본어학

1) Fourth Industrial Revolution – The world's data explained: how much we're producing and where it's all stored(2021).

한해에만 2020년의 거의 세 배에 가까운 약 175ZB에 이를 것으로 예상된다. 단적으로 매일 전 세계적으로 2억 940억 개 이상의 이메일과 5억 개 이상의 트윗 그리고 400만GB의 페이스북 데이터가 생산되고 있다. 이처럼 IT기술과 디지털 미디어의 발전, 그리고 인공지능의 등장으로 인류가 생산해 내고 있는 정보량이 기하급수적으로 증가하고 있어, 필연적으로 정보의 효율적인 처리 및 분석 방법에 대한 관심 또한 증가하고 있다. 학술분야에서는 자연어와 음성 데이터와 같은 비정형 빅데이터 속에 잠재되어 있는 유의미한 정보를 추출할 수 있는 기술인 텍스트 마이닝(Text Mining)이 특히 주목을 끌게 되었다.

이에 본 연구에서는 텍스트 마이닝 기법, 그리고 그 중에서도 토픽 모델(Topic Model, 혹은 토픽 모델링(Topic Modeling)) 기법을 사용하여 텍스트 빅데이터를 분석하는 것을 통해 대규모 텍스트 속의 주요 키워드를 추출하고 이를 바탕으로 한 눈에 파악하기 어려운 주요 토픽을 추출하여 유의미한 관점을 추출하고자 한다. 그리고 일본학 분야에 있어서도 대규모 텍스트 데이터를 다룰 필요성이 높아만 가고 있는 지금, 효율적인 텍스트 분석 기술인 텍스트 마이닝 기법의 확산 또한 필요하다고 할 수 있겠다.

그리고 종래의 데이터 생성과 발신 기능이 출판물을 중심으로 한 오프라인 미디어와 일부 대중매체에 집중되어 있었다고 한다면, 웹 5.0시대가 도래함에 따라 일반 대중에 의한 블로그, 커뮤니티, SNS 등 개인의 경험과 생각 그리고 감정이 담겨 있는 소셜 미디어로 빅데이터의 중심이 옮겨가고 있다. 이에 본 연구에서도 텍스트 빅데이터 중에서도 수많은 개인들이 생성한 방대한 양의 소셜 빅데이터를 기반으로 텍스트 마이닝을 수행하여 장기간에 걸친 한국인의 일본에 대한 관심 양상을 밝혀내고자 한다.

2. 선행연구 및 문제제기

2.1 텍스트 마이닝(Text Mining)

데이터 마이닝이 정제된 구조적 데이터를 분석하는 데에 반해, 텍스트 마이

닝은 자연어와 같이 정제되지 않은 비구조적인 데이터(Unstructured data)인 텍스트의 패턴 및 상호 연관성 등을 추출하는 통계적 연구방법이다. 즉, 텍스트 마이닝은 숫자와 같은 정형 데이터가 아니라, 텍스트나 음성 언어 등과 같은 형태와 형식 그리고 구조가 일정치 않은 데이터를 대상으로 자연어 처리기술을 사용하여 의미 있는 정보를 추출하여 가공 및 분석하는 것을 목표로 한다. 따라서 텍스트 마이닝은 방대한 언어 데이터를 처리하기 위해서는 필수불가결하다.

이에 일본학 분야에서도 이와 같은 텍스트 마이닝을 활용한 연구가 주목을 받기 시작하여 이경숙 외(2018), 落合由治(2020) 등의 연구에서는 텍스트 마이닝의 연구사례 등을 소개하며 그 효용성에 대해 언급하고 있다. 또한 여타 연구분야와 마찬가지로, 이경숙(2021) 등과 같이 논문 DB에 대한 조사를 통해 연구동향을 분석한 연구도 눈에 띈다. 이어서 이유희(2021), 김유영(2020·2022), 金明哲 외(2020), 김소희(2021), 김혜연(2022) 등의 연구에서는 실제 텍스트의 분석을 수행했으며, 그 중에서도 기계적으로 텍스트를 구성하고 있는 요소의 빈도를 집계하기 위한 전용 프로그램을 개발한 金明哲 외(2020)과 구 일본어능력시험 문항에 대한 텍스트 빅데이터 분석을 수행한 김유영(2022) 등이 본 연구에 시사하는 바가 크다. 그러나 특정 단어의 빈도나 공기하는 단어와의 관련성을 도출하는 것을 목표로 하는 연구가 많아, 종래의 코퍼스 언어학과 콜로케이션 연구와의 차별화되기 어려운 점을 지적할 수 있겠다. 이에 본고에서는 텍스트 마이닝에 의한 빈도분석이나 단어 간 콜로케이션 분석에서 한 발 더 나아가 Latent Dirichlet Allocation(잠재 디리클레 할당, 이하 LDA) 통계기법을 사용하는 것을 통해, 직관적으로 파악하기 어려운 텍스트의 잠재적 토픽과 이들 토픽간 연관성을 도출하고자 한다.

2.2 토픽모델(Topic Model)과 소셜 빅데이터 분석

텍스트마이닝 중에서도 토픽모델은 구조화되어 있지 않은 방대한 텍스트 데이터 중에서 토픽을 찾아내기 위한 알고리즘으로, 문서가 생성되는 과정을 확률을 사용하여 모델화한 확률모델이라고도 할 수 있다. 간단히 말하자면 특정 토픽(주제)의 문서에는 그 토픽과 관련된 단어가 다수 나타나게 된다. 예를

들어 토픽이 '경제'일 경우 문서에는 '경기, 주식, 회사, 환율, 금리…' 등과 같은 단어가 함께 나타날 확률이 높은 것에 반해, '여행'이 토픽인 문서에는 '숙박, 교통, 미식, 풍경, 사진…' 등과 같은 단어가 함께 나타날 확률이 높아질 것이다. 즉, 토픽모델에서는 함께 출현할 확률이 높은 단어의 그룹을 잠정적 토픽이라고 정의할 수 있으나, 이와 같은 그룹을 구성하는 것이 토픽모델이 된다는 개념이다(김유영:2022).

이와 같은 토픽모델은 대규모 텍스트를 다루는 다양한 분야의 연구에 있어서 효과적인 분석 도구로 사용되고 있다. 일본학 분야에도 黒田絢香(2021)은 토픽모델의 하나인 LDA를 사용하여 효과적인 시각화 툴의 개발을 수행한 연구를 찾아볼 수 있으며, 黄晨雯(2021)은 토픽모델 Top2Vec을 사용하여 토픽이라는 관점에서 소설을 읽어내고자 시도했다. 또한 김유영(2022)는 과거 20년간의 구일본어능력시험의 독해지문에 대한 토픽 모델링을 통해 주요 토픽을 도출하여 문제출제의 방향성을 분석했으며, 李広微 외(2020)은 토픽모델을 사용하여 현대소설에 나타난 접속표현의 사용 양상의 변화과정을 분석했다. 이와 같이 일본학 분야에서도 토픽모델에 대한 관심이 높아지고 있으며, 분석 툴의 개발부터 응용에 이르기까지 관련 연구가 확대되고 있다고 할 수 있으나, 일본학 분야는 토픽모델 관련 연구에 있어서 타 연구분야와 비교해 특히 양적차원에서 뒤쳐져 있어, 앞으로 더욱 더 많은 관련 연구가 필요한 실정이라고 볼 수 있겠다.

한편, 텍스트 마이닝의 조사 대상이라는 측면에서 보자면, 대부분의 선행연구는 뉴스와 출판서적 그리고 학습서 등 전통적 미디어를 주된 대상으로 하고 있어, 블로그, 커뮤니티, SNS 등과 같은 소셜 빅데이터를 조사대상으로 하여 텍스트 마이닝을 수행한 연구는 일본 애니메이션 영화의 흥행 요인을 분석한 김다흰 외(2021)를 제외하고는 찾아보기 어렵다. 이에 본 연구에서는 실제적이고 생생한 한국인들의 일본에 대한 관심의 양상을 함께 밝혀내고자 소셜 미디어 중에서 네이버 블로그의 게시물을 10년간에 걸쳐 수집·정제하여 분석을 수행했다.

3. 조사 대상 및 연구방법

본고에서는 일본에 대한 한국인의 관심 토픽 및 그 추이를 조사하기 위해 소셜 미디어 중 블로그의 게시물을 조사 대상으로 선정했는데, 국내 블로그와 검색엔진 점유율을 기반으로 네이버 블로그[2] 플랫폼의 게시물을 조사 대상으로 자료를 수집했다. 그 구체적인 조사대상 및 자료 수집 방법 그리고 데이터 처리 과정 단계는 다음과 같다.

3.1 조사 대상 및 자료 수집 툴

본고에서는 네이버 블로그의 게시물 중 '일본'을 언급한 게시물 중 2014년 3월 31일부터 2023년 12월 31일까지 약 10년간의 게시물을 웹크롤링 방식으로 수집하여 코퍼스를 구축했다. 그리고 이를 텍스트 마이닝 기법을 사용하여 분석하는 것을 통해 일본에 대한 한국인의 관심 토픽 및 그 추이를 고찰했다. 단, 전체 연도의 게시물을 모두 크롤링 하는 데에는 서버의 크롤링 제한 및 데이터 량 등의 물리적인 한계를 고려하여 각 연도의 각각 3월 31일, 6월 30일, 9월 30일, 12월 31일에 작성된 게시물에 한정하여 데이터를 수집하고, 그 중 우선적으로 검색 되는 4,000 건 내외의 데이터를 수집 및 정제했다. 그리고 '일본'이라는 키워드 전후로 170자 내외의 문장으로 데이터 수집을 한정하는 것을 통해 토픽의 오염을 막되, 게시글의 토픽이 가장 선명히 드러나는 제목 필드는 예외로 두었다. 이를 위해 개발한 오리지널 파이선 코드의 일부는 <표1>과 같으며, 자료 수집 결과인 Raw 데이터의 상세 사양은 아래 (1)과 같다.

---

2) a. 블로그 국내 점유율 : 네이버 1위, 88.8%(2024년 8월 29일 기준)
출처: 인터넷트렌드 데이터
b. 검색엔진 국내 점유율 : 네이버 1위, 60.70%(2023년 12월 31일 기준)
출처: 블로그차트

<표1> 네이버 블로그용 크롤링 파이선 코드 일부

```
import requests
from bs4 import BeautifulSoup
from datetime import datetime
from tqdm.notebook import tqdm

# 사용자에게 검색할 키워드와 기간 입력받기
keyword = input("검색할 키워드를 입력하세요: ")
start_date = input("시작 날짜를 입력하세요 (예: 2024-08-01): ")
end_date = input("종료 날짜를 입력하세요 (예: 2024-08-02): ")

# 기본 URL 설정
base_url =
f"https://section.blog.naver.com/Search/Post.naver?rangeType=PERIOD&order
By=sim&startDate={start_date}&endDate={end_date}&keyword={keyword}"

# 네이버 블로그의 특정 페이지 URL을 생성하는 함수
def get_page_url(base_url, page_no):
    return f"{base_url}&pageNo={page_no}"

def get_post_info(post_url):
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.3',
    }

    try:
        response = requests.get(post_url, headers=headers)
        response.raise_for_status()  # HTTP 에러 발생시 예외 발생시킴
    except requests.exceptions.RequestException as e:
        print(f"Error accessing {post_url}: {e}")
        return None

    soup = BeautifulSoup(response.text, 'html.parser')

    # 게시물 제목
    title_element = soup.find("h3", class_="se_textarea")
    title = title_element.get_text(strip=True) if title_element else "No title"
```

```
    # 게시물 내용
    content_element = soup.find("div", class_="se-main-container")
    content_text = content_element.get_text(strip=True) if content_element
else "No content"
                                … 이하 생략 …
```


(1) [네이버 블로그 수집 Raw 데이터 상세 사양]

a. 수집기간 : 2013년 3월 31일 ~ 2023년 12월 31일 기간 중, 각각의 연도별 3월 31일/6월 30일/9월 30일/12월 31일의 게시물 데이터 수집
b. 수집 사이트 : 네이버 블로그 / https://blog.naver.com
c. 정보 필드 : 제목, 제목링크_URL, 내용, 작성자, 블로그링크_URL, 출처블로그, 게시시간, 이미지_URL, 추출시간
d. 정보량 : 글 154,984건 / 81,209,056자(공백포함) / 40개 파일, 약 119MB
e. 문자코드(인코딩) : UTF-8(유니코드)
f. 파일명 예시 : NBlog_Japan_20140331_.csv(2014년 1분기), NBlog_Japan_20140930_.csv(2014년 3분기)
g. 코퍼스 배포 : 네이버 블로그 일본 코퍼스(NBlog_Japan_Corpus) http://japanese.or.kr/JapaneseStudy_corpus.aspx (단, 배포 코퍼스는 중복 및 오류 데이터를 삭제한 정제 데이터임)

<표2> NBlog_Japan_Corpus의 10개 데이터 예시

연번	게시시간	제목
1	2014. 3. 31.	아이패드 에어 케이스 추천
2	2014. 3. 31.	절식남의 사랑
3	2014. 3. 31.	구제 의류 쇼핑몰 라뽕 3월 31일 업데이트(adidas,UNIQLO,LACOSTE,levi's,abercrombie...
4	2014. 3. 31.	일본일주여행 디카사진 #38 no2
5	2014. 3. 31.	주권의 너머에서를 읽고
…	…	…
154980	2023. 12. 31.	싱가포르 럭셔리몰 | 마리나베이샌즈 라사푸라 마스터즈(Rasapura Masters) 푸드코트...
154981	2023. 12. 31.	삼평동맛집 규카츠정 판교점에 다녀왔어요
154982	2023. 12. 31.	: 수원 동탄, 맛있고 분위기 좋은 일식당 [하카타식당]
154983	2023. 12. 31.	원하고 바라옵건대
154984	2023. 12. 31.	Snow Planner 개봉기

참고로 본 연구의 데이터 분석 툴의 기본적인 사양과 패키지 등 상세 내용은 다음의 (2)와 같다.

(2) **[데이터 분석 툴 및 환경]**

a. Language : Python 3
b. 형태소분석 툴 : Kiwi(Korean Intelligent Word Identifier)
c. Python 통계 패키지 : Gensim / pyLDAvis / Wordcloud / Pandas 등
d. OS 등 : Colab - Python 3 Google Compute Engine / 가속 - TPU v2

그리고 본 연구의 검증과 후속 연구를 위해 본고에서 구축한 전체 코퍼스 데이터는 (1)g와 같이 웹페이지를 통해 공개해 두었으며, 이어지는 데이터 처리와 분석에 있어서 사용된 Python 코드 또한 모두 기술한다.

3.2 데이터 전처리

텍스트 마이닝은 크게 두 과정으로 나눌 수 있는데 첫 번째가 자료 처리과정(data processing) 그리고 두 번째가 자료 분석(data analysis)단계라고 할 수 있다. 우선, 자료처리과정은 비구조화 데이터를 분석 가능한 형태로 가공 및 정제하는 단계이고, 자료 분석은 데이터마이닝이나 기계학습(machine learning) 그리고 통계학(statistics) 등을 활용하여 텍스트로부터 유의미한 정보를 추출하는 단계이다(김유영:2020). 본고에서는 3.1절에서 수집한「*.csv」형식의 텍스트 raw data(n=154,984)를 형태소분석·통계처리 등의 데이터 분석에 적합한 형식으로 정제하는 전처리를 수행했는데, 우선 (1)c와 같은 분석 대상이 된 텍스트 필드에서 '작성자', 'URL', '추출시간' 등 토픽모델링에 불필요한 문자열 제거하고 제목과 내용을 통합하여 '텍스트(본문+내용)'의 정보 필드를 가진 분석용 데이터 세트를 구축했다.

<표3> 데이터 전처리 = 정제

```
import pandas as pd
#아래 NBlog_Japan_Corpus_data_OnlyTxt.csv 파일은 분석하고자 하는 파일
df = pd.read_csv('/content/drive/MyDrive/NBlog_Japan_Corpus/NBlog_Japan_Corpus_
data_OnlyTxt.csv', encoding='utf-8')
df = df[['제목', '제목링크_URL', '내용', '작성자', '블로그링크_URL',
'출처블로그', '게시시간', '이미지_URL', '추출시간']]
df['doc_NBlog_Jpn'] = df['제목'] + df['내용']
df['doc_NBlog_Jpn'] = df['doc_NBlog_Jpn'].apply(lambda x: str(x))
```

3.3 데이터 분석

3.3.1 형태소 분석 및 불용어 배제

전처리를 끝낸 데이터 세트를 한국어의 토크나이저 Kiwi(Korean Intelligent Word Identifier)를 사용하여 <표4>와 같이 형태소 분석을 수행했다. 형태소 분석 후에는 '명사'에 한정하여 키워드를 추출한 후, 한 글자의 단어를 제외하고 df['doc_NBlog_Jpn_clean'] 변수에 결과를 저장했다.

<표4> 형태소 분석 - 토크나이저 Kiwi

```
from kiwipiepy import Kiwi
kiwi = Kiwi()
df['doc_NBlog_Jpn_token'] = None
for idx, row in df.iterrows():
    doc_NBlog_Jpn = row['doc_NBlog_Jpn']
    nouns = []
    for sentence in kiwi.analyze(doc_NBlog_Jpn):
        for token in sentence[0]:
            if token.tag.startswith('NN'):
                nouns.append(token.form)
    df.at[idx, 'doc_NBlog_Jpn_token'] = nouns
df['doc_NBlog_Jpn_clean'] = df['doc_NBlog_Jpn_token'].apply(lambda x: [word
for word in x if len(word) > 1])
```

단, 형태소 분석 결과에서 우선 '일본'과 '한국'은 순서대로 본고에서 설정한 검색 키워드이며 한국어로 작성된 블로그인 만큼 일본과 연관된 키워드로서 빈도는 높으나 토픽으로 부적절 하다. 따라서 이와 같은 단어들을 아래 <표5>와 같이 제거했으며, 앞선 절에서 한 글자의 단어는 이미 배제했으나, 최대한의 무결성을 위해 '이', '그', '저', '년', '월', '일' 등과 같은 일부 단어를 재차 배제했다.

<표5> 불용어 배제 - stopwords

```
stop_words_NBlog_Jpn =
['일본','한국','년','수','월','일','달','것','때','나','중','거','곳','때','논문','중국','등','번','후',
'시','분','전','앞','뒤','좌','우','하나','둘','셋','넷','다섯','여섯','일곱','여덟','아홉','열','날',
'만','뿐','개','내','너','나','우리','정도','저','데','이','그','저','어느','위','아래','옆','인','점',
'오','미','돈','속','간','회','안','밖','쪽','기타','여타','타','피','저번','지난','오늘','어제','그
제','내일','모레','글피','올해','작년','금년','내년','재작년', '']
def remove_stopwords(word_list, stop_words_NBlog_Jpn):
  filtered_words = []
  for word in word_list:
    if word not in stop_words_NBlog_Jpn:
      filtered_words.append(word)
  return filtered_words
df['doc_NBlog_Jpn_clean'] = df['doc_NBlog_Jpn_clean'].apply(lambda x:
remove_stopwords(x, stop_words_NBlog_Jpn))
```

3.3.2 Dictionary·BoW·Corpus 생성

형태소 분석을 마친 텍스트는 아래 <표6>의 코드를 통해 본 연구를 위한 학습 사전(Dictionary)을 <표7>과 같이 구축하고 이를 Bow형식[3]으로 변환했다. 참고로 학습 사전은 계 154,085어 규모가 된다. 그리고 Bow형식 데이터를 사용하여 최종적으로 LDA 분석용 코퍼스(Corpus)를 생성했다.

3) BoW(Bag of Words)형식이란 문서군을 문서(Document)와 단어(Word)의 행렬로 표현하는 형식으로, 각각의 요소에 출현빈도를 카운트하여 값을 부여한 데이터 형식.

<표6> LDA 분석용 '사전' 및 '코퍼스' 생성

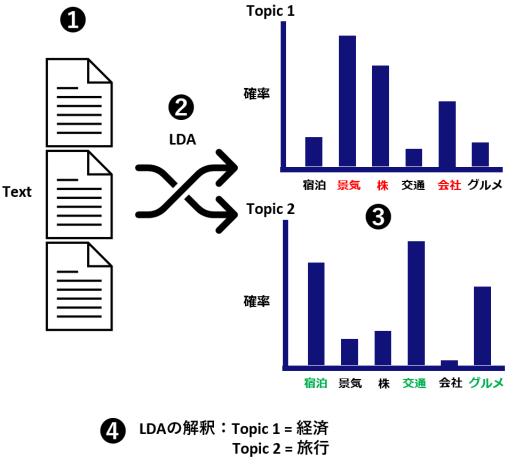
```
import gensim
all_tokens = df['doc_NBlog_Jpn_clean'].tolist()
#파일을 Gensim 의 corpora.Dictionary로 변환
dictionary_NBlog_Jpn = gensim.corpora.Dictionary(all_tokens)
#doc2bow를 이용해 corpus에 단어를 저장
corpus_NBlog_Jpn = [dictionary_NBlog_Jpn.doc2bow(text) for text in
all_tokens]
```

<표7> Dictionary 전체 154,085 중 50개 단어 예시

각지, 러시아, 말씀, 무대, 미국, 베루스, 선배, 세계, 수출, 아시아, 아이패드, 에어, 오디오, 추천, 케이스, 크로커다, 크로커다일, 한때, 홍콩, 공통, 라인, 분모, 사랑, 스토리, 시민, 양면, 영화, 절식, 주인공, 1, 최근, 특별, 프레임, 가디건, 구제, 데코, 라뽕, 메인, 명품, 블루, 셔츠, 쇼핑몰, 수입, 스몰, 업데이트, 의류, 자수, 체크, 카라, 프린팅, 혼방 ……

4. LDA 분석 및 결과

4절에서는 [그림1]과 같이, LDA(Latent Dirichlet Allocation) 분석을 수행했는데, 우선 대량의 문서 데이터([그림1-❶])를 단어 단위로 '그룹핑'을 수행할 때 이 그룹 하나하나가 '토픽'이 된다([그림 2-❸]). 단, LDA 분석을 실행하는 데에는 전제조건이 두 가지 있는데, 분석을 수행하기 전에 먼저 '토픽의 수를 미리 정해야만 하며', '어떠한 그룹이 될 것인가는 산출된 확률에 기반 한 인간의 해석([그림 2-❹], '경제'와 '여행')이 필요하다'.



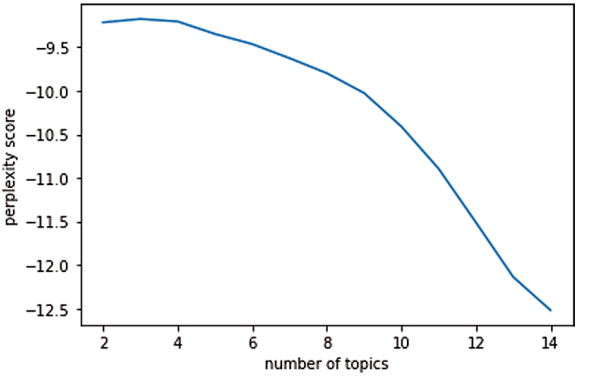
[그림1] LDA 개념 : '경제'와 '여행' 토픽의 텍스트(김유영:2022)

4.1 LDA에 의한 토픽 개수 설정

본고에서는 몇 개의 토픽으로 정하고 LDA 모델을 학습시켜야 할 것인가를 결정하여 최적의 결과를 얻어낼 수 있도록 <표8·9>와 같이 「Perplexity」와 「Coherence」를 동시에 고려한 결과로부터 토픽 수 최적화를 수행했으며, 그 결과는 이어지는 [그림 2·3]과 같다. 우선 [그림2]와 같이 토픽의 증가에 따른 스코어가 마이너스를 가리키고 있다. 따라서 보다 유의미한 분석을 위해 [그림3]과 같이 Coherence Model을 주된 근거로 삼아 가장 높은 스코어를 기록한 토픽 수 「6」을 최적 토픽 수로 산정했다. 참고로 Perplexity는 확률 모델의 예측 정확도를 판단하는 평가지표로, 선정된 토픽 개수마다 학습을 수행하여 가장 작은 값을 보이는 구간을 찾아 토픽의 개수 선정에 도움을 주며 낮을수록 정확하게 예측할 수 있다. 그리고 Coherence는 토픽의 품질을 평가하는 지표로, 추출된 토픽을 인간이 얼마나 해석하기 쉬운가를 나타낸다. 값이 클수록 좋은 확률 모델이 된다.

<표8> Perplexity

```
import matplotlib.pyplot as plt
perplexity_values=[]
for i in range(2,15):
    ldamodel=gensim.models.ldamodel.LdaModel(corpus_NBlog_Jpn,
num_topics=i, id2word=dictionary_NBlog_Jpn)
    perplexity_values.append(ldamodel.log_perplexity(corpus_NBlog_Jpn))
x=range(2,15)
plt.plot(x, perplexity_values)
plt.xlabel("number of topics")
plt.ylabel("perplexity score")
plt.show()
```

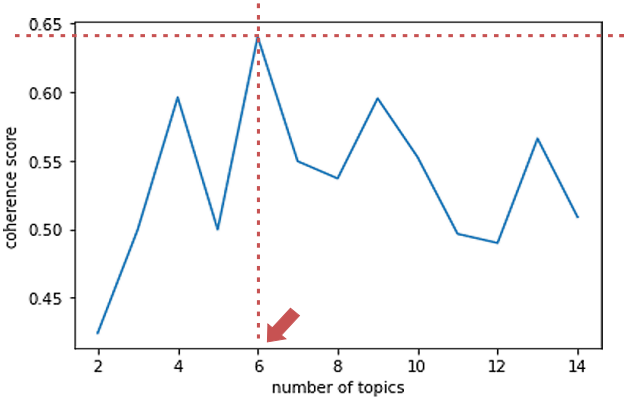


[그림2] 토픽 수 대비 Perplexity Score

<표9> Coherence

```
import matplotlib.pyplot as plt
from gensim.models import CoherenceModel
high_score_reviews = all_tokens
high_scroe_reviews = [[y for y in x if not len(y)==1]
                      for x in high_score_reviews]
coherence_values=[]
for i in range(2,15):
```

```
    ldamodel=gensim.models.ldamodel.LdaModel(corpus_NBlog_Jpn,
num_topics=i, id2word=dictionary_NBlog_Jpn)
    coherence_model_lda=CoherenceModel(model=ldamodel,
texts=high_score_reviews, dictionary=dictionary_NBlog_Jpn,topn=10)
    coherence_lda=coherence_model_lda.get_coherence()
    coherence_values.append(coherence_lda)
x=range(2,15)
plt.plot(x, coherence_values)
plt.xlabel("number of topics")
plt.ylabel("coherence score")
plt.show()
```



[그림3] 토픽 수 대비 Coherence Score

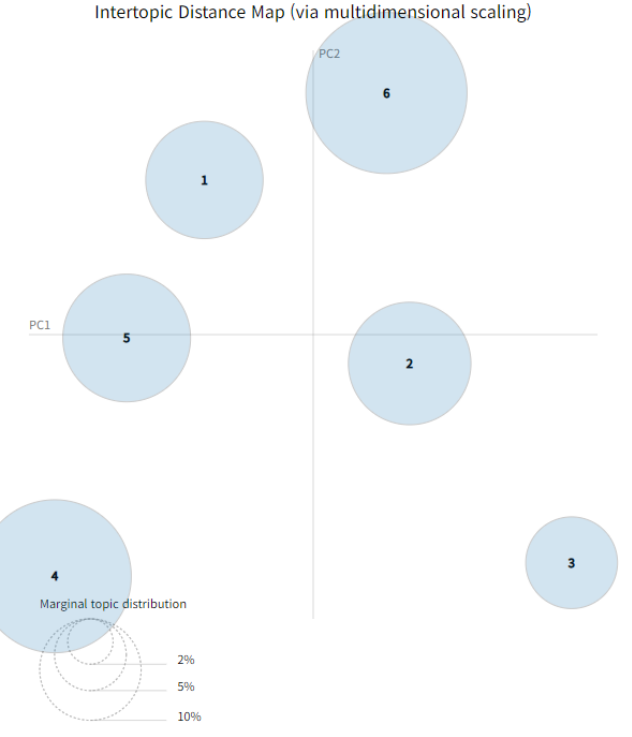
4.2 LDA에 의한 토픽 모델링

위 [그림3]과 같이 텍스트의 토픽 수를 「6」으로 설정하여 LDA 분석을 수행하고, 이를 <표10>과 같이 pyLDAvis 패키지를 사용하여 토픽 별 거리 및 단어 분포 결과를 시각화 하면 다음 [그림4·5]와 같다.

<표10> LDA 분석 및 pyLDAvis 시각화

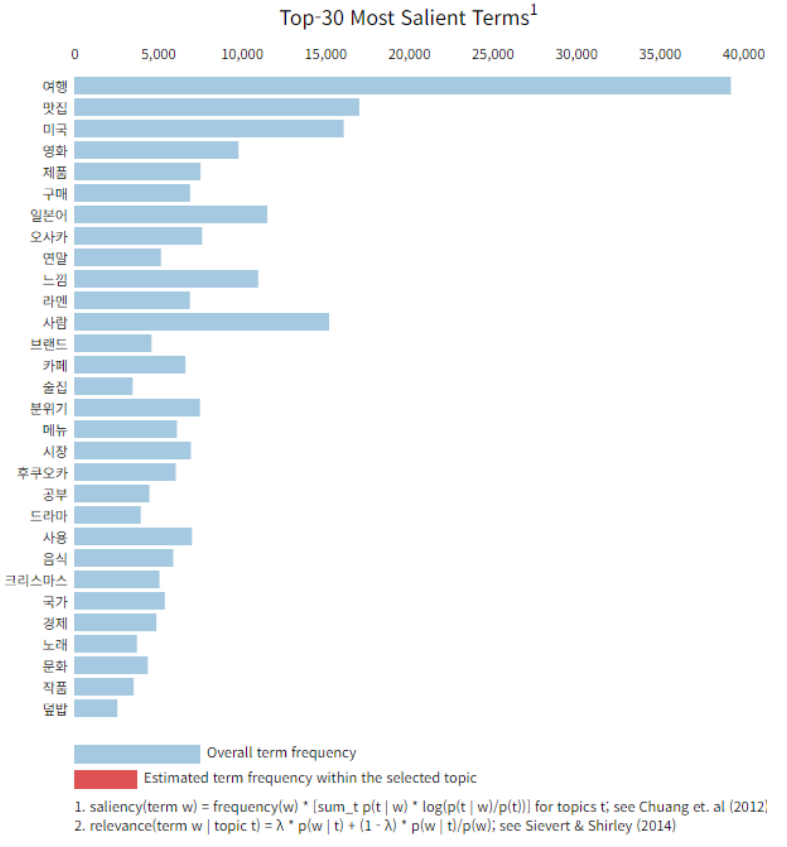
```
import pyLDAvis
import pyLDAvis.gensim

pyLDAvis.enable_notebook() # added
ldamodel=gensim.models.ldamodel.LdaModel(corpus_NBlog_Jpn, num_topics=6,
id2word=dictionary_NBlog_Jpn)
ldamodel.print_topics(num_words=6)
vis = pyLDAvis.gensim.prepare(ldamodel, corpus_NBlog_Jpn,
dictionary_NBlog_Jpn, sort_topics=False)
pyLDAvis.save_html(vis, 'lda_NBlog_Japan.html')
display(vis)
```



[그림4] Intertopic Distance Map(via multidimensional scaling)

위 [그림4]와 아래 [그림5]의 시각화 결과는 웹페이지를 통해 공개해, 본 분석 내용을 검증할 수 있도록 했으며, 결과 페이지에서 토픽을 변경하는 것을 통해 상세한 토픽 별 워드 리스트를 확인할 수 있다. 접속 링크는 아래와 같으며, 각 0~5로 표기된 토픽은 본고의 1~6에 해당한다는 점에 주의할 것.



pyLDAvis : http://japanese.or.kr/Lda_NBlog_Japan_V.240831-1404.html

[그림5] Top-30 Most Salient Terms

4.3. 토픽 분석

위 [그림4]와 같이 pyLDAvis 시각화 분석 결과는 서로 겹치지 않으며 토픽 간 거리 및 분포가 적절히 분산되어 있어, 토픽 분석을 통한 그룹핑이 효과적으로 수행된 것을 확인할 수 있다. 이에 본 절에서는 각각의 토픽을 빈출 순위에 따라 차례로 토픽 해석을 수행하여 실제 현상과 비교·대조 분석하는 것을 통해, 한국인의 일본에 대한 관심 양상과 토픽모델의 효용성을 명확히 하고자 한다. 본격적인 분석에 앞서 각각 6개의 토픽으로 그룹핑 된 빈출 어휘 중 상위 30 단어를 나열하면 아래 <표11>과 같다.

<표11> 토픽 별 순위 및 상위 빈출 20 단어

순위	토픽	상위 빈출 30 단어	해석
1	**토픽 6**	여행, 맛집, 느낌, 도쿄, 사진, 오사카, 분위기, 라멘, 시간, 친구, 카페, 메뉴, 후쿠오카, 음식, 사람, 추천, 생각, 마지막, 호텔, 식당	여행
2	**토픽 4**	미국, 세계, 시장, 국가, 크리스마스, 경제, 기업, 유럽, 해외, 정부, 영국, 나라, 국내, 독일, 지역, 대만, 투자, 기술, 수출, 달러	경제
3	**토픽 5**	일본어, 사람, 나라, 공부, 문화, 학교, 생각, 역사, 시작, 결산, 일본인, 운동, 대학, 조선, 이야기, 유학, 시험, 학생, 영어, 수업	학습
4	**토픽 2**	제품, 구매, 사용, 브랜드, 판매, 과자, 코로나, 가능, 가격, 대행, 상품, 후기, 골프, 직구, 디자인, 선수, 선물, 게임, 사이즈, 카드	쇼핑
5	**토픽 1**	영화, 연말, 드라마, 노래, 생각, 작품, 사랑, 사람, 이야기, 마지막, 게임, 소설, 시작, 작가, 만화, 애니메이션, 새해, 방송, 추천, 영상	대중 문화
6	**토픽 3**	술집, 덮밥, 푸딩, 고양이, 꼬치, 가성비, 금리, 사용, 사케, 건강, 딸기, 나무, 크림, 퇴근, 연어, 하이, 화장실, 대기, 스토어, 쿠키	일식

우선 1순위로 LAD 분석 결과에서 가장 큰 비율을 점하고 있는 [토픽 6]의 경우, '여행', '맛집', '사진', '음식', '호텔', '방문' 등의 단어(키워드)와 여러 일본의 지명이 주요 단어가 상위권 키워드로 올라 있는 것을 확인할 수 있다. 이는

직관적으로「여행」토픽으로 해석할 수 있을 텐데, 조사 대상 텍스트 속에 '여행지 혹은 관광지로서의 일본'을 다룬 텍스트가 많았고 이것이 자연스럽게「여행」토픽을 최상위에 위치하게 한 것이라고도 해석할 수 있겠다. 토픽 6의 빈출 어휘를 통해 워드 클라우드를 작성하면 아래 [그림6]과 같은데 시각적으로도 명확하여 여행 토픽으로 어렵지 않게 해석할 수 있다는 것을 알 수 있다.

토픽 6-여행 토픽의 전체 토픽 안에서의 순위를 고려하면, 한국인들에게 있어서 일본은 가장 먼저 '여행지/관광지'의 대상으로서 소비되고 있다는 것을 의미한다. 그리고 일본 지명 중 도쿄, 오사카, 후쿠오카, 교토 순으로 빈출했다는 것을 통해 한국인이 가장 많이 관심을 갖고 있고 방문했거나 방문을 희망하는 지역이 어디인지도 읽어낼 수 있는데, 이는 실제 통계 수치[4]와도 정확히 일치 한다. 이는 본고의 토픽 모델링 결과의 유효성을 뒷받침하는 근거라고도 볼 수 있겠다.

[그림6]「토픽 6」의 워드 클라우드 :「여행」

이어 2순위 [토픽 4]의 경우 '세계', '시장', '국가', '경제', '기업', '투자', '기술' 등의 키워드가 그룹핑 되어 있어, '국제 관계 속 경제 주체로서의 국가인 일본'을 다룬「경제」토픽이라고 해석될 수 있다. 예전보다 경제대국으로서의 그 위상은 약해지기는 했으나, 한국인에게 있어서 일본은 여전히 세계경제의 흐름

4) e-Stat의 2023년 통계에 따르면 공항별 한국인 입국자는 도쿄, 오사카, 후쿠오카 순으로 많았다.

을 논하는 데에 있어 필수적인 대상이자 관심 혹은 평가의 척도로서 자리 잡고 있음을 확인할 수 있다.

3순위 [토픽 5]의 경우, '일본어', '공부', '학교', '대학', '유학' 등의 빈출 키워드를 통해 「학습」토픽으로 해석했다. 이와 같은 키워드가 상위권에 위치한다는 것은 한국인들이 '학습의 대상 혹은 유학의 대상으로서의 일본어와 일본'도 높은 확률로 상정하고 있다는 것을 나타내는 것으로, 실제로도 일본국제교류기금(Japan Foundation)의 2023년 통계에 따르면 한국은 일본어능력시험(JLPT) 응시자 수로 절대 수치 2위, 인구대비 응시자 수 1위를 기록하고 있는 국가이며, 일본학생지원기구(JASSO)의 2023년 통계에서도 한국은 일본 내 외국인 유학생 수에 있어서도 4위를 기록했다.

4순위 [토픽 2]는 '제품', '브랜드', '과자', '가격'. '대행', '후기', '배송', '출시' 등의 키워드를 통해 이 그룹은 「쇼핑」 토픽에 해당한다고 해석할 수 있다. 한국인의 일본 관련 블로그 게시물에 '상품 구매처로서의 일본'을 다룬 텍스트가 네 번째로 많았고 이를 통해 일본 상품에 대한 한국인들의 관심과 선호도를 읽어낼 수 있다. 특히 '과자', '골프', '게임', '안경' 순으로 자주 언급되는 제품 키워드를 통해 거시적인 통계로 확인하기 어려운 일반 대중들이 선호하는 구체적인 일본산 제품이 무엇인지를 확인할 수 있었다.

5순위 [토픽 1]은 '영화', '드라마', '노래', '작품', '게임', '애니메이션' 등의 키워드 단어를 통해 알 수 있듯이 「대중문화」 토픽으로 해석하는 것이 타당할 것이다. 즉 매력적인 '대중문화 발신지로서의 일본'을 소비하고 있다고 할 수 있는데, 여기에서 특기할만한 점은 '애니메이션' 혹은 '만화' 등 일본의 서브컬쳐에 그 관심이 치중되어 있을 것이라 예상되었던 것과 달리, 실제로 한국인들이 '영화', '드라마', '게임', '소설', '만화', '애니메이션(애니)', '음악' 등 일본 대중문화 전반에 대해 높은 관심을 갖는 소비 패턴을 보인다는 것을 확인할 수 있다.

마지막으로 6순위 [토픽 3]은 '술집', '덮밥', '꼬치', '사케', '대기' 등의 키워드로 구성되어 있어 「일식」 토픽으로 해석할 수 있는데, 이는 '식문화와 요식산업이라는 측면에서 일본'을 한국 대중들이 폭 넓게 소비하고 있음을 의미한다. 특히 이는 폭발적으로 증가하고 있는 한국 내 일본식 요식업체의 수를 고려할

때, 이와 같은 토픽은 한국 내 일본 요식문화의 대중화라는 현실을 정확히 반영하고 있다고 할 수 있겠다. 실제로 2021년 통계청 전국 사업체 조사에 따르면 전국의 일식 전문점은 2016년에 비해 약 5년 사이에 69.2% 증가한 수치를 보이고 있는데, 그 다음 해 2022년에는 더욱 증가하여 21,553개 업체에 이르게 된다(동대신문:2024.03.24.).

5. 나가며

본고에서는 방대해지고 있는 텍스트 데이터의 효율적인 활용을 위한 시도의 일환으로 토픽모델이라는 텍스트 마이닝 기법을 일본학 분야에 도입했다. 또한 보다 생생한 한국인의 일본에 대한 관심 양상을 밝혀내기 위해 그 연구대상으로 기존의 매스 미디어에 의해 생성된 빅데이터가 아닌 일반 대중에 의해 생성된 소셜 빅데이터를 기반으로 텍스트 마이닝을 수행했다. 구체적으로는 10년간의 네이버 블로그의 '일본' 관련 게시물 텍스트를 수집하여 코퍼스를 구축한 후, 이에 대한 토픽모델 분석을 수행했다.

그 결과 본고에서는 일본과 관련된 블로그 게시물을 크게 6개의 토픽으로 분류할 수 있었으며 각각의 토픽으로 분류된 빈출 키워드 단어들을 바탕으로 각각의 토픽을 「여행」,「경제」,「학습」,「쇼핑」,「대중문화」,「일식」순으로 해석했다. 이를 통해 한국인의 일본에 대한 가장 큰 관심은 「여행」토픽, 즉 '1. 여행지/관광지의 대상으로서의 일본'이었으며, 이어서 순서대로 '2. 국제 관계 속 경제 주체로서의 국가인 일본', '3. 학습의 대상 혹은 유학의 대상으로서의 일본어와 일본', '4. 상품 구매처로서의 일본', '5. 대중문화 발신지로서의 일본', '6. 식문화와 요식산업이라는 측면에서 일본'을 소비하고 있다는 것을 확인할 수 있었다. 게다가 이와 같은 한국인의 '일본'에 대한 관심과 소비 양상 해석은 실제 사회현상을 정확히 반영하고 있다는 점에 토픽 모델링의 가능성을 재확인할 수 있었다. 이에 일본학 분야에서도 데이터 마이닝, 그 중에서도 토픽 모델 분석을 통해 대규모 빅데이터를 처리할 수 있다는 점과 유의미한 분석 결과 도출 가능

성을 고려할 때 앞으로도 더욱 널리 응용 가능할 것이라는 차원에서도 본 연구의 의의가 있다고 하겠다.

그러나 본고에서 10년간의 데이터를 조사했다고는 하지만 모든 데이터를 하나로 묶어 처리하고 분석을 수행했다는 점과 해당 기간 전체 데이터가 아닌 일 년을 4분기로 나누고, 그 중에서도 특정기간에 한정하여 자료를 수집했다는 점에서 아쉬운 점이 많다고 하겠다. 이에 금후의 연구에서는 자료 조사 기간뿐만 아니라 타 소셜 미디어로 자료 조사를 확대하는 것을 통해 데이터양을 늘리고, 이를 바탕으로 시간의 흐름에 따라 한국인의 일본에 대한 관심 토픽의 변화 양상을 중점적으로 파악하고자 한다.

【参考文献】

김소희(2021)「テキストマイニングを活用した「X{まで}」構文の語彙分析―「X{さえ}」「X{も}」との比較·対照を中心に―」『일본학보』126, 한국일본학회, pp. 121-143

김유영(2020)「テキストマイニングを用いた日本のメディアの韓国ニュースにおける感情の推移に対する分析―Pythonを用いた「単語感情極性対応表」の分析を活用して―」『일본어학연구』65, 한국일본어학회, pp.25-43

김혜연(2022)「텍스트마이닝을 활용한 한일 어휘·문화 교육의 가능성-'トイレ', '화장실' 관련 어휘 및 문화를 중심으로-」『일본어문학』92, 한국일본어문학회, pp.139-159

이경숙·곽내정(2018)「ICT時代の日本語研究と教育におけるテキストマイニングの有效性」『한국일본어교육학회 학술발표논문집』 2018, 한국일본어교육학회, pp.65-69

이경숙(2021)「텍스트마이닝 기법을 활용한 일본어 학습자의 음성에 관한 연구 동향 분석」『일본어학연구69, 한국일본어학회, pp.93-106

이유희(2021)「텍스트 마이닝 기법을 통한 일본어능력시험 분석과 학습 교육-JLPT 4급 문자·어휘를 중심으로-」『일본문화학보』91, 한국일본문화학회, pp.283-304

류은주·오상훈·박시사(2014)「日本の高齢者の「観光」意識に関するテキストマイニング分析」『일본근대학연구』46, 한국일본근대학회, pp.371-386

落合由治(2020)「AI技術からみた日本語学, 日本語教育研究の展望と課題―日本語教

育の繋がりと協働の新領域をめざして―」『日本語教育研究』50, 韓国日本語教育学会, pp.23-34
金明哲・鄭弯弯(2020)「テキストコーパスマイニングツールMTMineR」『計量国語学』32(5), 計量国語学会, pp.265-276
黒田絢香(2021)「トピックモデル可視化ツールの開発に向けて」『言語文化共同研究プロジェクト』2021, 大阪大学大学院言語文化研究科, pp.5-13
日本学生支援機構(2024)「JASSO概要 2024 令和6年」 日本学生支援機構, p.12
黄晨雯(2021)「Top2Vec による小説の探索的研究: 程小青の作品解読を中心に」『言語文化共同研究プロジェクト』2020, 大阪大学大学院言語文化研究科, pp.43-53
藤田郁(2022)「LDAトピックモデルによるIPAテキスト分析の試み: アルフレッド・テニスンの韻文を用いて」『言語文化共同研究プロジェクト』2021, 大阪大学大学院言語文化研究科, pp.15-38
李広微・金明哲(2020)「トピックモデルに基づいた現代小説の接続表現の分析」『日本行動計量学会大会抄録集』48, 日本行動計量学会, pp.152-153
동대신문「[문화] 문화의 장, 한국에서 만나보는 일본」2024.03.24. https://www.donggukmedia.com/news/articleView.html?idxno=81671(검색일: 2024.08.01.)
AJ-All about Japanese Study「2. 일본어 코퍼스 - 2-2. 네이버 블로그 일본 관련 게시물 코퍼스(NBlog_Japan_Corpus)」 http://japanese.or.kr/JapaneseStudy_corpus.aspx(검색일: 2024.08.01.)
bab2min Minchul Lee「지능형 한국어 형태소 분석기(Korean Intelligent Word Identifier)-KIWI」 https://github.com/bab2min/Kiwi(검색일: 2024.08.01.)
BLOGchart「블로그 점유율」 https://www.blogchart.co.kr(검색일:2024.08.01.)
e-Stat(政府統計の総合窓口)「出入国管理統計」 https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00250011&tstat=000001012480&cycle=1&year=20230&month=11010303&tclass1=000001012481&result_back=1&tclass2val=0(검색일: 2024.08.01.)
Japan Foundation「JLPT」 https://www.jlpt.jp/statistics(검색일: 2024.08.01.)
KOTRA「일본 경제통상 리포트23-7-2023년 상반기 대일본 수출입 동향 분석」 https://dream.kotra.or.kr/kotranews/cms/com/index.do?MENU_ID=290(검색일: 2024.08.01.)
Google「Colab」 https://colab.research.google.com(검색일: 2024.08.01.)
Melvin M. Vopson, 2021, Fourth Industrial Revolution - The world's data explained:

how much we're producing and where it's all stored
https://www.weforum.org/agenda/2021/05/world-data-produced-stored-global-gb-tb-zb(검색일:2024.08.01.)
INTERNET TREND「검색엔진」
http://www.internettrend.co.kr/trendForward.tsp(검색일: 2024.08.01.)

◇논문마감: 2024. 11. 24.
◇심사개시: 2024. 11. 26.
◇게재확정: 2024. 12. 03.

〈Abstract〉

Topic Analysis of Japan-Related Texts on Social Media Using LDA Topic Modeling Technique

Kim, YuYoung

This study applies the topic modeling technique, a prominent method in text mining, to the field of Japanese studies, aiming to address the challenges posed by the growing volume of textual data. To provide a deeper understanding of Korean interest in Japan, the research focuses on social media data generated by the ordinary users, rather than traditional mass media sources. Specifically, the study collected and constructed a corpus from ten years of Japan-related blog posts on the Korean platform Naver, which was then analyzed using topic modeling.

The analysis revealed that these Japan-related blog posts could be broadly categorized into six topics. Based on the frequent keywords associated with each topic, these were identified as "Travel," "Economy," "Learning," "Shopping," "Popular Culture," and "Japanese Cuisine." The findings indicate that Koreans primarily view Japan as a travel destination, followed by its significance as an economic entity in international relations, a subject of study or destination for learning the Japanese language and culture, a shopping destination, a source of popular culture, and a provider of culinary experiences. These patterns demonstrate that topic modeling effectively identifies themes and consumption patterns, reflecting real-world social phenomena.

This study underscores the potential of data mining, particularly topic modeling, in Japanese studies by enabling the analysis of large-scale textual data and deriving meaningful insights. However, limitations include the consolidation of data into a single set for analysis and the restriction of data collection to specific periods within each year. Future research should extend the data collection timeframe and incorporate data from other social media platforms to capture evolving trends in Korean interest in Japan over time.

〈要旨〉

LDAトピックモデリング手法を用いたソーシャルメディアにおける日本関連テキストのトピック分析

金嚅泳

本稿では、増大するテキストデータの効率的な活用を目指し、テキストマイニング手法の一つであるトピックモデルを日本学分野に導入した。また、韓国人の日本に対する関心の様相をより鮮明に解明するため、従来のマスメディアによって生成されたビッグデータではなく、一般大衆によって生成されたソーシャル・ビッグデータを対象にテキストマイニングを行った。具体的には、10年間にわたる韓国・Naver社のブログの「日本」に関連する投稿テキストを収集し、コーパスを構築した後、これに対してトピックモデル分析を実施した。

その結果、本研究では、日本に関連するブログ投稿を大きく6つのトピックに分類することができた。各トピックにグループ化された頻出キーワードに基づき、それぞれのトピックを「旅行」、「経済」、「学習」、「買い物」、「大衆文化」、「和食」と解釈した。これにより、韓国人の日本に対する最大の関心は「旅行」トピック、すなわち「旅行地或いは観光地としての日本」であり、その次に「国際関係における経済主体としての日本」、「学習の対象または留学先としての日本語および日本」、「商品の購入先としての日本」、「大衆文化の発信地としての日本」、「食文化および飲食産業の側面からの日本」への関心が続いていることが確認された。また、このような韓国人の「日本」に対する関心と消費の様相は、実際の社会現象を正確に反映していることから、トピックモデルの可能性が再確認されたといえる。したがって、日本学分野においても、データマイニング、特にトピックモデル分析を用いることで、大規模なビッグデータを処理し、有意義な分析結果を導き出すことが可能であり、今後さらに広く応用される可能性がある点でも本研究には意義があると考えられる。

しかしながら、本研究では10年間のデータを調査したものの、全てのデータを一つにまとめて処理し分析を行った点、調査対象期間全体のデータではなく、1年を四半期に分け、特定の期間に限定して資料を収集した点において、限界があるといえる。今後の研究では、調査期間を拡大するとともに、他のソーシャルメディアにまで調査対象を広げることで、データ量を増加させ、時間の経過による韓国人の日本に対する関心トピックの変化をより深く理解することを目指したい。