

오픈소스 텍스트 분석 라이브러리

: TEANAPS

by FINGEREDMAN
admin@teanaps.com

Contents

Introduction

Background

What can you do with TEANAPS?

Use Cases

DEMO

TEXT MINING with TEANAPS

TEANAPS : Text Analysis APIs for Education



본 자료는 텍스트 마이닝(text mining)에 보다 쉽게 접근할 수 있도록 도와주는 **Python** 라이브러리입니다. 텍스트 마이닝을 위해서는 **Python** 언어를 배운 후에도 다양한 라이브러리를 활용할 줄 알아야합니다(e.g., KoNLPy, NLTK, Gensim). 하지만 배워야하는 외부 라이브러리가 적지않고 난이도도 높아 **Python** 언어에 충분히 익숙하지 않으신 분들은 접근하기가 쉽지 않습니다. TEANAPS 는 텍스트 마이닝과 관련된 외부 라이브러리들을 하나의 인터페이스(API) 형태로 통합하고 **Google Colabotory** 를 활용해 설치환경을 통일하여, 텍스트 마이닝을 위한 사전작업을 최소화하고 필요한 프로그래밍 코드를 최소화 할 수 있도록 도와줍니다. 본 패키지를 활용하기 전 **Python** 기초문법과 텍스트 마이닝에 필요한 필수 사전지식을 먼저 학습하시기를 추천드리며, **User Guide** 와 **Tutorial** 을 참조하시어 TEANAPS 활용법을 따라해보시길 권장드립니다.

- 본 자료는 텍스트 마이닝을 활용한 연구 및 강의를 위한 목적으로 제작되었습니다.
- 본 자료를 강의 또는 연구 목적으로 활용하고자 하시는 경우 꼭 아래 메일주소로 연락주세요.
- 본 자료에 대한 상업적 활용과 허가되지 않은 배포를 금지합니다.
- 강의, 저작권, 출판, 특허, 공동저자에 관련해서는 문의 바랍니다.
- **Contact : ADMIN(admin@teanaps.com)**

Background

What is Text Mining?

텍스트 마이닝

“언어학, 통계학, 기계학습 등을 기반으로

자연언어 처리 기술을 활용하여,

반정형/비정형 텍스트 데이터를 정형화하고,

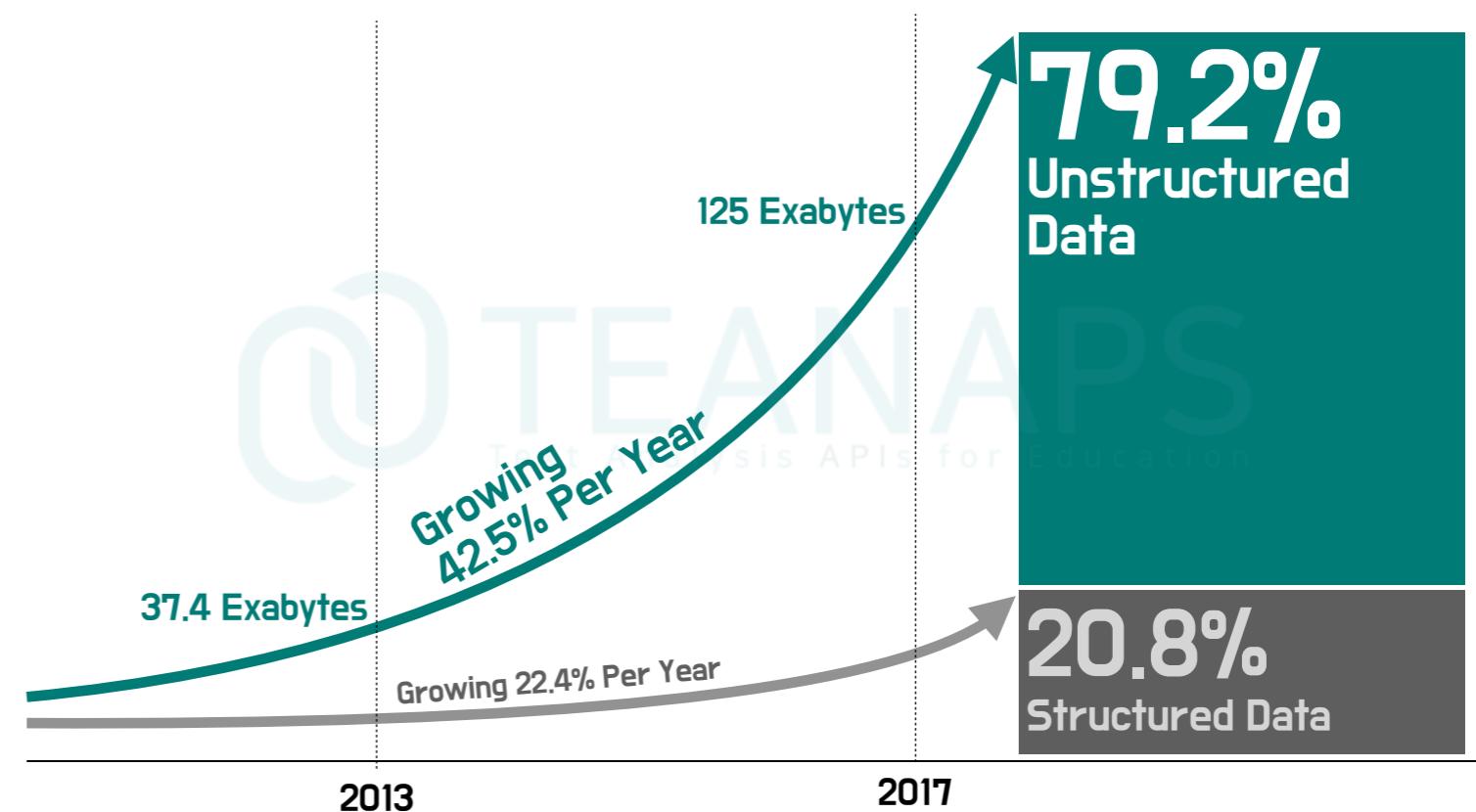
특징을 추출하기 위한 기술과 추출된 특징으로부터

유의미한 패턴 또는 지식을 추출하는 과정”

“Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.”

Why Text Mining?

01. 비정형 데이터의 폭발적 증가



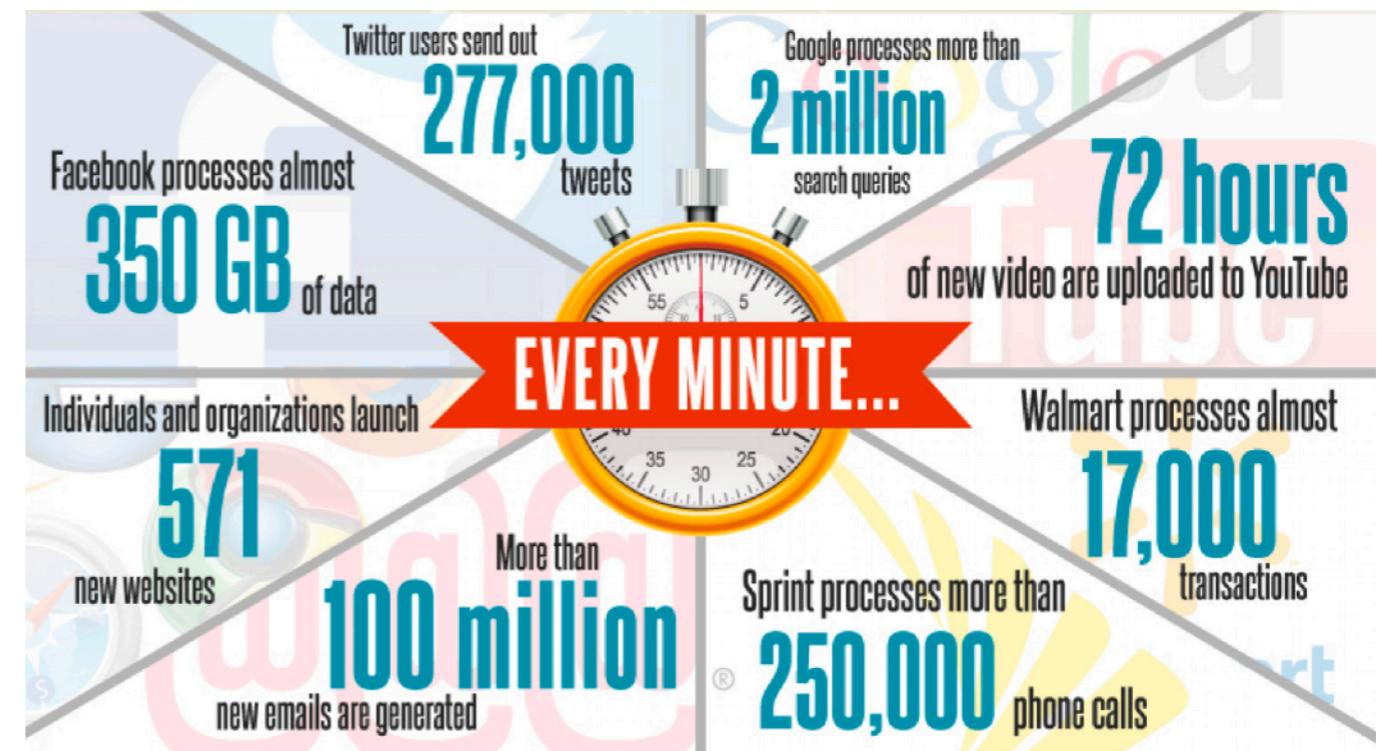
4차산업혁명과 사물인터넷 등

빅데이터 관련 기술의 급진적 발전

실제로 생산되는 데이터의 70~80%는 비정형 데이터

Why Text Mining?

02. 가장 흔하고 찾기 쉬운 데이터



온라인 텍스트 데이터의 대부분이 **SNS**에서 발생

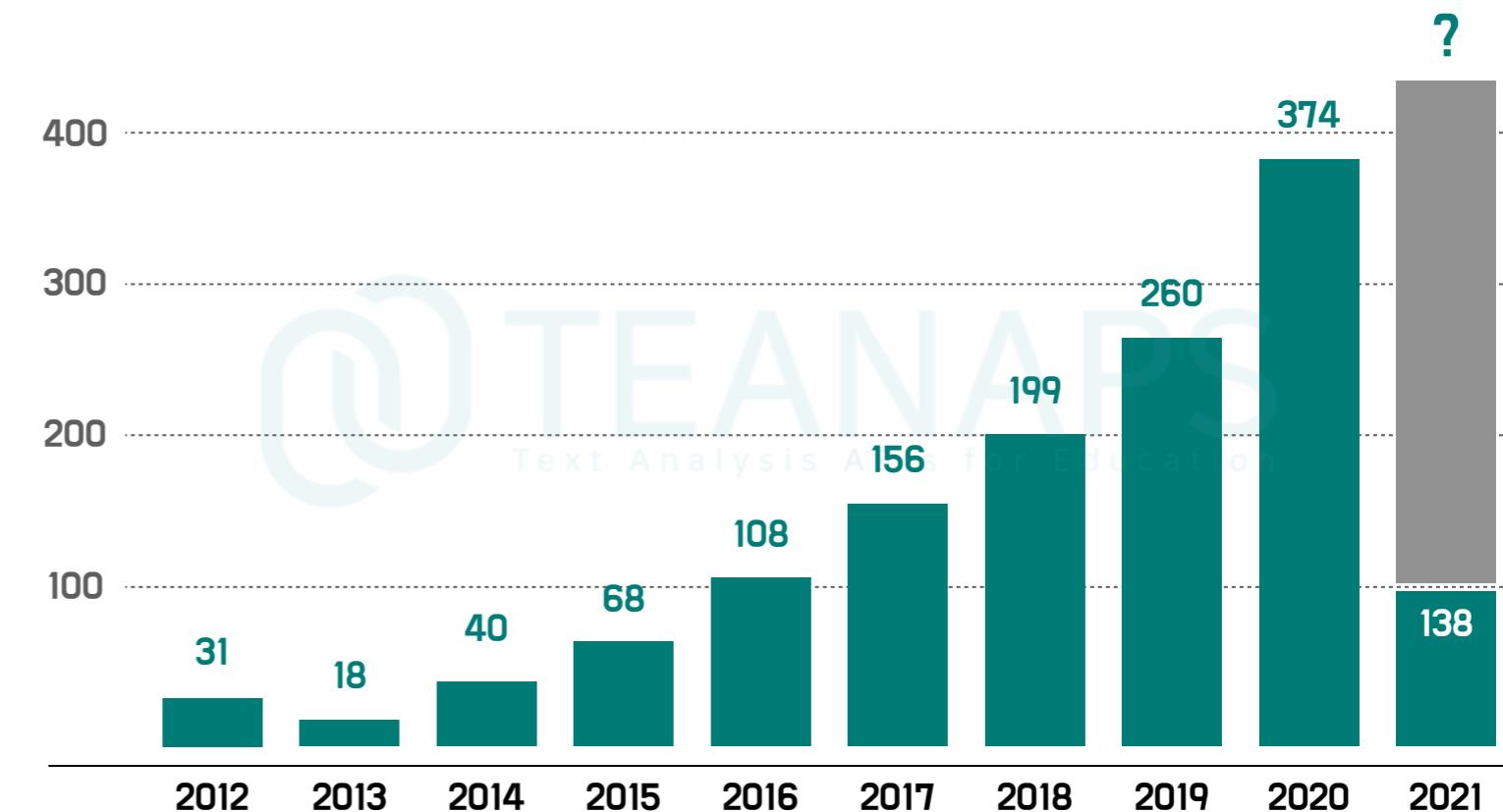
다양한 형태의 데이터가 텍스트 형태로 변형되어 활용

→ **Speech/Image/Video to TEXT**

웹크롤링, Open API를 활용한 데이터 수집

Text Mining based Research

01. “텍스트 마이닝” 관련 논문 수



국내 텍스트 마이닝 활용 연구사례는 최근 10년 동안
1,392건으로 2012년 대비 **12배** 증가

Text Mining based Research

02. “텍스트 마이닝” 연관 키워드

지능정보연구에서 발행한 학술지의 논문에 등록된 키워드 Cloud



키워드 「텍스트 마이닝」 연관 키워드



최근 학술지에 가장 많이 등록된 키워드는
“텍스트 마이닝”

연관 키워드에 가장 많이 출현하는 분석기법은
토픽모델링, 네트워크 분석, 감성분석, 군집분석

Challenges of Text Mining

01. 텍스트 마이닝이 어려운 이유

동의어, 동형(동음) 이의어가 포함되거나

단어의 의미가 도메인 별로 다를 수 있음

사람들이 작성한 문장은 **맞춤법**과 **철자**가 틀리고.

단어를 변형하여 활용하는 등 규칙을 지키지 않음

낯선 **자연어처리** 기술에 대한 이해와

정형데이터에 비해 **복잡한 전처리** 과정이 필요함

Challenges of Text Mining

02. 더 복잡한 한국어 텍스트 분석

한국어는 흔치 않은 **교착어**

- 초성, 중성, 종성의 조합이 하나의 음절을 형성하며,
조사와 접사가 붙어 문법적 관계를 형성함

복잡한 **용언의 변형**

- 용언이 변하는 경우의 수가 매우 많고 인지하기 어려움

복잡한 한국어의 특성에 비해 **부족한 데이터**

- 형태소 분석기의 한계 및 미비한 어휘사전
신조어, 미등록어, 새로운 용어의 조합을 반영하기 어려움

Challenges of Text Mining

03. 부족한 텍스트 분석 도구

한국어는 **KoNLPy**. 영어는 **NLTK**

KoNLPy의 마지막 업데이트는 2015년…

형태소분석기 외 오픈소스 **라이브러리 부재**

- 띄어쓰기 보정, 문장분리 등 산재된 라이브러리
- 개체명인식기, 감성분석 등 널리 활용되는 학습모델을 필요시마다 학습해야함

라이브러리 별 **서로다른 지원환경**

- Mac/Windows 운영체제, Python 버전
- 라이브러리에 따라 JAVA, C++ 설치까지 요구

**What can you do
with TEANAPS?**

Why TEANAPS?

01. 오픈소스 대비 높은 퍼포먼스

형태소분석기 및 개체명인식기 제공

오픈소스 형태소분석기 대비 향상된

고유/복합 명사, 인물, 지명 등 인식성능

"**손흥민**(28)이 4경기 연속 골이자 자신의 시즌 14호 골을
작렬하며 **토트넘** 홉스퍼를 **잉글랜드축구협회(FA)**컵 16강에 올려놨다."

	손흥민	토트넘 홉스퍼	잉글랜드축구협회
TEANAPS	손흥민	토트넘 홉스퍼	잉글랜드축구협회
K	손+흥+민	토트넘+홉스퍼	잉+글+랜드+축구+협회
O	손흥민	토트넘+홉스퍼	잉글랜드+축구+협회

Why TEANAPS?

02. 다양한 텍스트 분석기법 제공

가장 많이 활용되는 텍스트 분석 기법 제공

단어빈도/TF-IDF, 분류기, 토픽모델링,

네트워크 분석, 군집분석, 문서요약 등

학습없이 바로 활용가능한 머신러닝/딥러닝 모델 제공

개체명인식기, 감성분석, 키워드추출

지저분한 텍스트 데이터 처리를 위한 전처리 기능 제공

불용어처리, 복합명사처리, 띠어쓰기보정,

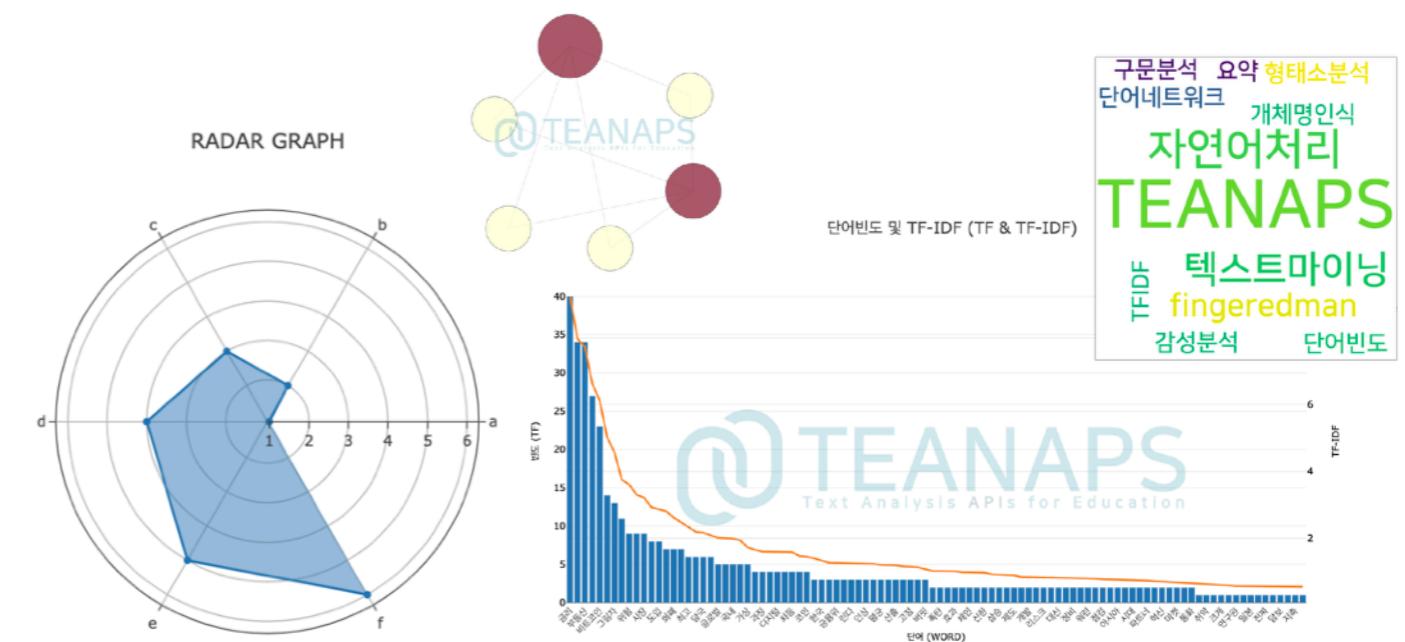
문장분리, 언어인식, 원형복원, 반복제거,

개인정보 마스킹 등

Why TEANAPS?

03. 텍스트 시각화 기능 제공

히스토그램, 라인그래프, 네트워크, 레이다,
워드클라우드, 하이라이팅,
클러스터링, 토픽모델링 분석결과 시각화 제공



늘 배우고 배우는 자세가 필요합니다 .

@TEANAPS
Text Analysis APIs for Education

욕심쟁이에게 스트레스 받으며 살다가 떠나고 나니 너무 행복해요 !

@TEANAPS
Text Analysis APIs for Education

Why TEANAPS?

04. OPEN API 지원

자연어처리 기술을 OPEN API 형태로 제공하여

텍스트 및 자연어처리 기반 **서비스 구축 지원**

05. 텍스트 수집-분석을 위한 자료제공

데이터 수집을 위한 **웹스크래핑 도구** 제공

- NAVER 뉴스기사 및 댓글 수집기
- 온라인 카페 게시글 및 댓글 수집기
- Google PlayStore, Apple AppStore 리뷰 수집기

TEANAPS 활용을 위한 **Documentation** 및

단계별 **이론/실습 교육자료** 제공

- Text Mining for Beginner
- Text Mining for Practice
- Advanced Text Mining



Use Cases

Journal

KBR 제24권 제2호 2020년 5월

<http://dx.doi.org/10.17287/kbr.2020.24.2.121>

Joint Sentiment 토픽모델링 기반 국내 여행 불만족 요인 연구*

최 윤 진**
이 소 현***
윤 상 혁****
김 희 웅*****

최근 개인의 삶의 질, 일과 삶의 균형을 중요하게 생각하게 되면서 여행 활동이 증가하고 있다. 이에 따라 국내 여행 산업도 성장하고 있으나 동시에 여행지에 따른 불만족에 대한 개선은 과제로 남아있다. 본 연구는 국내 여행에 대한 실제 여행자들의 주요 여행지별 불만족 요인을 도출하고 이에 대한 개선방안을 제시하는 것을 목적으로 한다. 소비자의 만족과 불만족의 개념을 설명하는 기대 불일치 이론을 기반으로 텍스트마이닝 기법과 Joint Sentiment 토픽 모델링을 활용하여 실제 여행자들이 사용하는 여행 애플리케이션에서 국내 주요 여행지별 불만족 사항을 도출 및 분석하고, 이를 기반으로 여행지별 불만족 사항의 해결 및 개선방안을 제시하였다.

연구 결과, 첫째, 국내 요인 전반에 대한 불만족 요인은 '물가, 교통, 바가지요금, 위생, 상업화'에 대한 토픽이었다. 둘째, 주요 여행지별 토픽으로 '제주도'는 '교통, 외국인 관광객, 물가, 특색 없는 명소', '전주'는 '먹거리 중심, 바가지요금, 상업화', '부산'은 '호객행위, 위생, 바가지요금, 교통', '경주'는 '주차, 바가지요금, 관광객, 교통'이 각각 도출되었다.

텍스트마이닝 기법의 토픽모델링에 감성 차원을 더한 Joint Sentiment 토픽모델링을 사용하여 국내 주요 여행지별 불만족 요인을 도출하였고 토픽과 키워드 간의 관계를 네트워크 다이어그램으로 시각화하여 직관적인 이해와 차별적인 개선방안을 도출한 것에 학술적 의의가 있다. 실무적으로는 여행지별 불만족 사항에 대한 구체적인 개선방안을 제안함으로써 국내 여행의 활성화를 통한 관광 산업의 수익 및 경쟁력 향상에 이바지할 수 있다.

주제어: 국내 여행, 텍스트마이닝, Joint Sentiment 토픽모델링, 기대 불일치, 온라인 고객 리뷰

I. 서 론

최근 개인의 삶의 질 향상 및 일과 삶의 균형(Work-Life Balance) 요구가 증대되면서 여가활동과 소비 행동에도 변화가 일어나고 있다. 특히, 현재의 만족과 행복을 중요시하는 가치 변화를 통해 소소하지만 확실한 행복을 중요시하는 현상이 확산되고 있다. 이는 다양한 여가 및 취미 활동에 참여뿐

아니라 여행의 증가로도 나타나고 있다. 실제 여행 전문 연구 기관인 컨슈머인사이트가 수행한 '여름휴가 여행 조사'에 따르면, 설문대상자 중 78.8%가 여름휴가 여행을 다녀온 것으로 이는 2018년 76.7% 보다 2.1% 증가하였다(김태형, 2019). 특히, 2019년 기준으로 3년 만에 국내 여행 비율은 70.5%로 증가하고, 해외여행 비율은 24.8%로 감소하면서 여행 트렌드도 변화하였다. 그뿐만 아니라, 주로 장기간 휴가를 이용하여 여행하던 것에서 짧은 연휴를

Journal

『정보시스템연구』 제29권 제4호
한국정보시스템학회
2020년 12월, pp. 137~151

<http://dx.doi.org/10.5859/KAIS.2020.29.4.137>

텍스트 마이닝 기반의 자산관리 핀테크 기업 핵심 요소 분석: 사용자 리뷰를 바탕으로*

손애린** · 신왕수*** · 이준기****

〈목 차〉

I. 서론	3.3 데이터 분석
II. 이론적 배경	IV. 분석결과
2.1 자산관리 핀테크 기업 - 뱅크샐러드, 토스	4.1 토픽모델링 결과
2.2 텍스트 마이닝	4.2 속성별 감성 분석 결과
2.2.1 토픽모델링	V. 결론 및 시사점
2.2.2 감성 분석	5.1 연구 결과 토의
2.2.3 속성별 감성 분석	5.2 연구의 한계 및 향후 연구 방향
III. 연구 방법	5.3 연구의 시사점
3.1 데이터 수집	참고문헌
3.2 데이터 전처리	<Abstract>

I. 서 론

자산관리 애플리케이션은 기본적으로 계좌 통합조회 서비스를 제공하며, 이를 통해 가계부를 자동으로 작성하고 소비패턴을 분석하여 맞춤형 금융상품을 추천하는 등 개인 금융 데이

하는 사람들이 늘고 있으며, 이에 따라 가계부, 카드 사용내역 관리, 세금 계산 등이 가능한 애플리케이션이 인기를 끌며 이른바 ‘애텐트’(애플리케이션과 재테크의 합성어)라는 신조어가 등장하기도 했다(조선비즈, 2017). 게다가 지난 8월 데이터 3법(개인정보보호법·신용정보법·정

Journal

韓國IT서비스學會誌
第19卷 第3號
2020年 6月, pp.117-137

Journal of Information Technology Services
<https://doi.org/10.9716/KITS.2020.19.3.117>

텍스트마이닝 기법을 이용한 모바일 피트니스 애플리케이션 주요 요인 분석 : 사용자 경험 관점*

이소현** · 김진솔*** · 윤상혁**** · 김희웅*****

An Analysis on Key Factors of Mobile Fitness Application by
Using Text Mining Techniques : User Experience Perspective*

So-Hyun Lee** · Jinsol Kim*** · Sang-Hyeak Yoon**** · Hee-Woong Kim*****

Abstract

The development of information technology leads to changes in various industries. In particular, the health care industry is more influenced so that it is focused on. With the widening of the health care market, the market of smart device based personal health care also draws attention. Since a variety of fitness applications for smartphone based exercise were introduced, more interest has been in the health care industry. But although an amount of use of mobile fitness applications increase, it fails to lead to a sustained use. It is necessary to find and understand what matters for mobile fitness application users. Therefore, this study analyze the reviews of mobile fitness application users, to draw key factors, and thereby to propose detailed strategies for promoting mobile fitness applications. We utilize text mining techniques - LDA topic modeling, term frequency analysis, and keyword extraction - to draw and analyze the issues related to mobile fitness applications. In particular, the key factors drawn by text mining techniques are explained through the concept of user experience. This study is academically meaningful in the point that the key factors of mobile fitness applications are drawn by the user experience based text mining techniques, and practically this study proposes detailed strategies for promoting mobile fitness applications in the health care area.

Keyword : Mobile Fitness Application, User Experience, Text Mining Technique, LDA Topic Modeling, Term Frequency Analysis, Keyword Extraction Analysis, Honeycomb

Journal

『인터넷전자상거래연구』 제20권 제2호
2020년 4월, pp. 107~125
한국인터넷전자상거래학회

<https://doi.org/10.37272/JIECR.2020.04.20.2.107>

이커머스 유료회원제 가입자/비가입자 주요 이슈 비교: 텍스트마이닝 기법 활용

A Comparative Analysis of E-commerce Issues between Customers With and Without Paid Membership: Using Text mining technique

우 유 란* · 이 중 정** · 이 소 현***

Youran Woo · Choong C. Lee · So-Hyun Lee

... Abstract ...

While the fierce competition in e-commerce market continues to be escalated, more and more e-commerce firms is trying to secure loyal customers through membership marketing. However, as many marketers see the importance of membership marketing, many firms still don't clearly understand the unique characteristics of their paid membership customers so that the firms can transform them into loyal customers. Surprisingly, there are very few researches to study the unique characteristics of paid membership customers, that is, investigating how they are different from no-membership customers in terms of the type of service, their interest/ concern, and their expectation in purchasing experience. Therefore, this study explore and compare the key issues between customers with and without paid membership in e-commerce business. The differences in the key issues were derived from text-mining technique(i.e., LDA topic modeling and word frequency analysis) with the customer review data. For the results of the LDA topic modeling, we derived topics of membership, purchased products, events, and payment methods from the customers with paid membership, and membership and purchased products from the customers without paid membership. In addition, for the result of the word frequency analysis, we derived issues of price discount from the customers with paid membership, and purchase/use experience and product information from the customers without paid membership. The result of this study has practical implication that contribute to provide the detailed marketing strategies for each customer group based on the better comparative understanding of paid membership customers over no membership customer group.

Journal



디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 2, pp. 291-299, Feb. 2021

Check for updates

텍스트마이닝 기법과 ARIMA 모형을 활용한 배달의 민족 앱 리뷰 분석

정지훈¹ · 정혜인² · 이준기^{3*}

^{1,2}연세대학교 정보대학원 석사과정

³*연세대학교 정보대학원 교수

An Analysis of Mobile Food Delivery App 'Baemin' by Using Text Mining and ARIMA Model

Ji-Hoon Jung¹ · Hye-In Chung² · Zoon-Ky Lee^{3*}

^{1,2}Master's Course, Graduate School of Information, Yonsei University, Seoul 03722, Korea

³*Professor, Graduate School of Information, Yonsei University, Seoul 03722, Korea

[요약]

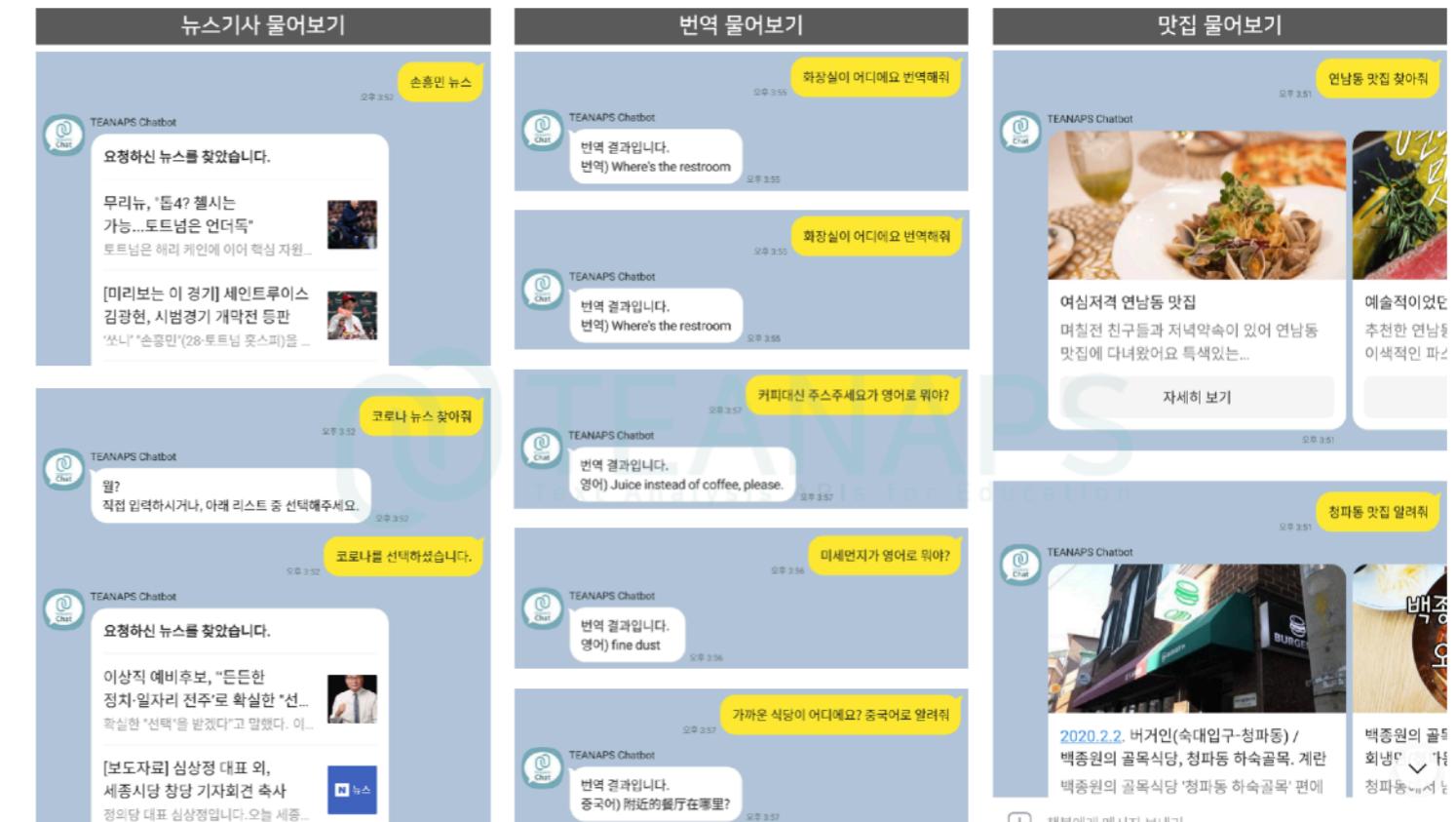
모바일 애플리케이션 시장의 빠른 성장과 함께 다운로드 수와 리뷰의 수도 급증하고 있다. 개발자는 사용자 리뷰 분석을 통해 서비스 이슈를 감지하고 사용자의 불만족을 개선하고자 하지만, 매일 다량으로 생성되는 리뷰를 처리하기에는 어려움이 있다. 본 연구에서는 수많은 리뷰 중에서 유의미한 서비스 이슈를 탐지하기 위해 텍스트마이닝 기법과 시계열 이상치 탐지 모형 Autoregressive Integrated Moving Average(ARIMA)을 이용하여 최근 빠르게 성장하고 있는 배달 어플리케이션 '배달의 민족' 리뷰를 분석하고자 한다. 이를 위하여 '배달의 민족' 앱 리뷰 71,435건을 크롤링하여 수집한 후, 토픽모델링 기법을 적용하여 5개의 토픽을 도출하였다. 또한, 시계열 기반 이상치 탐지 ARIMA 모델을 활용하여 리뷰수가 급증한 12개의 이상치 일자를 탐지하였고, 해당 일자를 토픽 별 감성분석 결과를 바탕으로 이슈 요인을 분석하였다.

[Abstract]

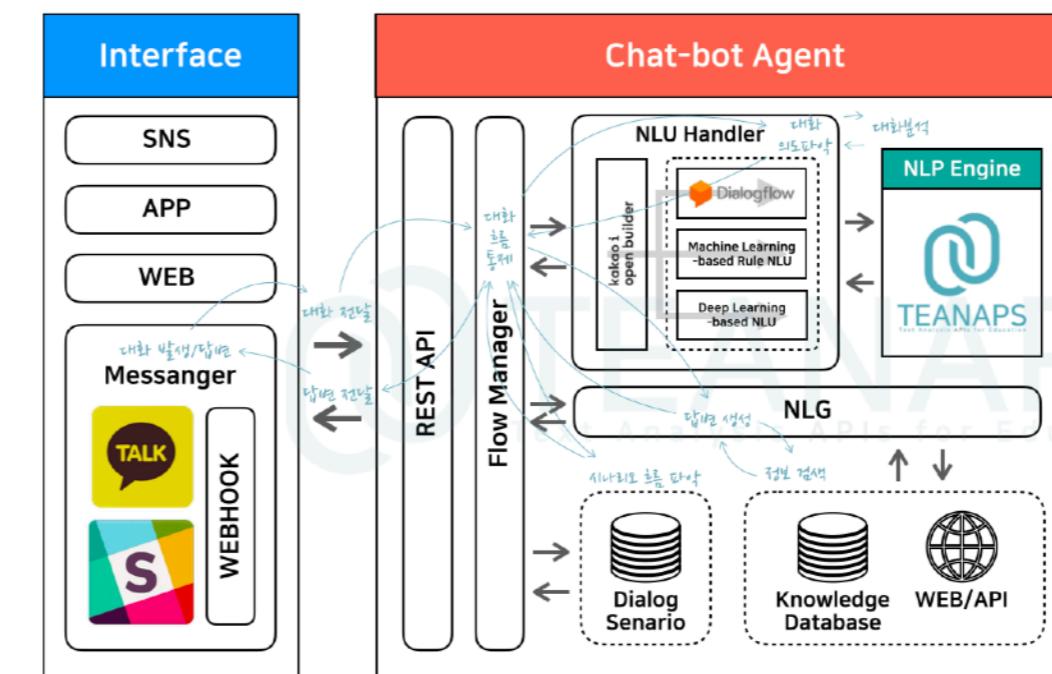
As the mobile application market is growing rapidly, the number of application downloads and reviews are increasing as well. Developers want to detect and improve service issues and user dissatisfaction through application review analysis, but it is difficult to handle large quantities of reviews that are generated every day. In this paper we analyze review of the delivery application

Project

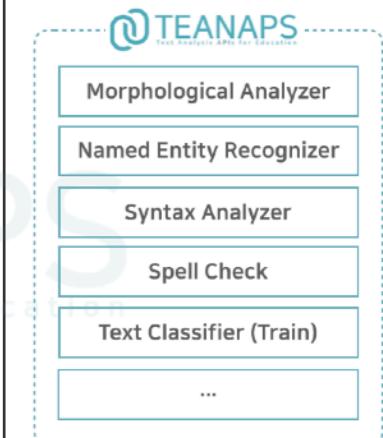
Chat-bot with TEANAPS



@TEANAPS



@TEANAPS



© 2020. FINGEREDMAN all rights reserved.

DEMO

**Let's try
TEANAPS**

TEANAPS 공식 Github
<http://www.teanaps.com>

Contact

E-mail

admin@teanaps.com

Homepage

<http://teanaps.com>

E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com