ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)

Review

Business Understanding

START

Check 텍스트 마이닝을 통해 해결하고자 하는 문제가 명확한가?

Yes / No

Check 해결하고자 하는 비즈니스에 대한 충분한 이해를 가지고 있는가?

Yes / No

비즈니스 문제정의

Data Preparation

Check 문제해결을 위해 분석할 텍스트 데이터를 보유하고 있는가?

Yes / No

Check 해당 데이터에 접근할 권한을 가지고 있는가?

Yes / No

Check 문제해결을 위해 필요한 데이터가 온라인상에 존재하는가?

Yes / No

Check 해당 온라인 데이터 소스에서 원하는 정보를 가져오기 위한 API를 제공하는가?

Yes / No

- API 호출을 통한 데이터 수집
- 웹 스크래핑을 통한 데이터 수집
- 오프라인 데이터 수집

Data Pre-processing

Check 필요한 텍스트가 일정한 형태로 정규화되어 있는가?

Yes / No

데이터 정규화

Check 텍스트 데이터를 특성에 따라 분리하여 분석할 필요가 있는가?

Yes / No

데이터 분리

Check 분석할 텍스트가 영문이거나 영어단어가 다수 포함되어 있는가?

Yes / No

형태소분석

개체명인식

원형복원

불용어제거

단어빈도분석

Check 불용어 및 빈출어가 충분히 제거되었는가?

Yes / No

Check 필요한 단어들이 올바르게 추출되었는가?

Yes / No

- 동시출현분석
- 문서분류
- 문서요약
- 키워드 추출
- 클러스터링

- 임베딩
- 알고리즘 선정
- 학습
- 평가
- 배포

시각화

분석결과 비교/해석

분석결과 리포팅

COMPLETE

Data Analysis, Visualization, and Reporting

* references
** references
*** references

WEEK 11

Summarization

문서를 요약하는 방법

문서에서 핵심 문장 추출하기


세계일보

네이버 쇼핑라이브, 이젠 '검색'으로 만난다

기사입력 2020.11.15. 오후 6:43 기사원문 스크랩 본문듣기 • 설정

공감 댓글 요약봇 가

쇼핑 LIVE '온앤온 원피스' 관련 라이브 영상

LIVE 9,999명 시청중

온앤온 원피스 30% 싹쓰리 할인
방송 중에만 혜택!

10.3만

온앤온 · 현대백화점 판교점

쇼핑 라이브 더보기 →

네이버가 '쇼핑 라이브'를 통합 검색에 노출한다.

쇼핑라이브는 지난 3월 말 코로나19로 어려움을 겪고 있는 오프라인 판매자들이 비대면 환경에 적응하고 새롭게 온라인 판로를 확보할 수 있도록 제공된 '라이브 커머스 툴'로 시작했다. 이후 7월 말부터는 별도의 서비스로 개편, 판매자와 사용자들이 실시간을 소통하며 상품을 판매·구매할 수 있어 많은 관심을 받고 있다.

네이버는 그동안 서비스 내에서만 확인할 수 있었던 쇼핑라이브를 사용자들이 더욱 손쉽게 활용할 수 있도록, 쇼핑라이브 내 소개됐거나 소개될 예정인 제품 등 관련 키워드 검색시, 통합검색 결과를 통해 쇼핑라이브 콘텐츠를 제공한다.

사용자들은 통합검색에서 현재 진행중인 라이브 콘텐츠 뿐 아니라, 과거, 앞으로 방송 예정인 라이브 콘텐츠까지 모두 확인할 수 있다.

특히 구매하려는 상품이 쇼핑라이브로 소개될 예정이라면, 방송알림 설정을 통해, 잊지않고 라이브 방송을 시청하며 다양한 혜택을 받으며 쇼핑할 수 있다.

한편, 쇼핑라이브는 별도의 스튜디오나 전문 장비가 없어도 스마트폰 하나만으로 편리하게 라이브로 진행할 수 있어 새로운 판로가 필요한 SME (Small and Medium-sized Enterprise)들에게 각광받고 있다. 사용자들에게도 새로운 쇼핑의 즐거움을 제공하며, 누적 시청자수 3000만명을 돌파하는 등 지속적으로 성장하고 있다.

현화영 기자 hhy@segye.com
사진=네이버

* 현화영(세계일보), 네이버 쇼핑라이브, 이젠 '검색'으로 만난다, 2020.11.15., https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=001&oid=022&aid=0003523331/.
** references
*** references

문서를 요약하는 방법

문서를 핵심 문장 하나로 표현하기

매일경제

코로나로 승객 29% 감소…지하철 적자 1.7조

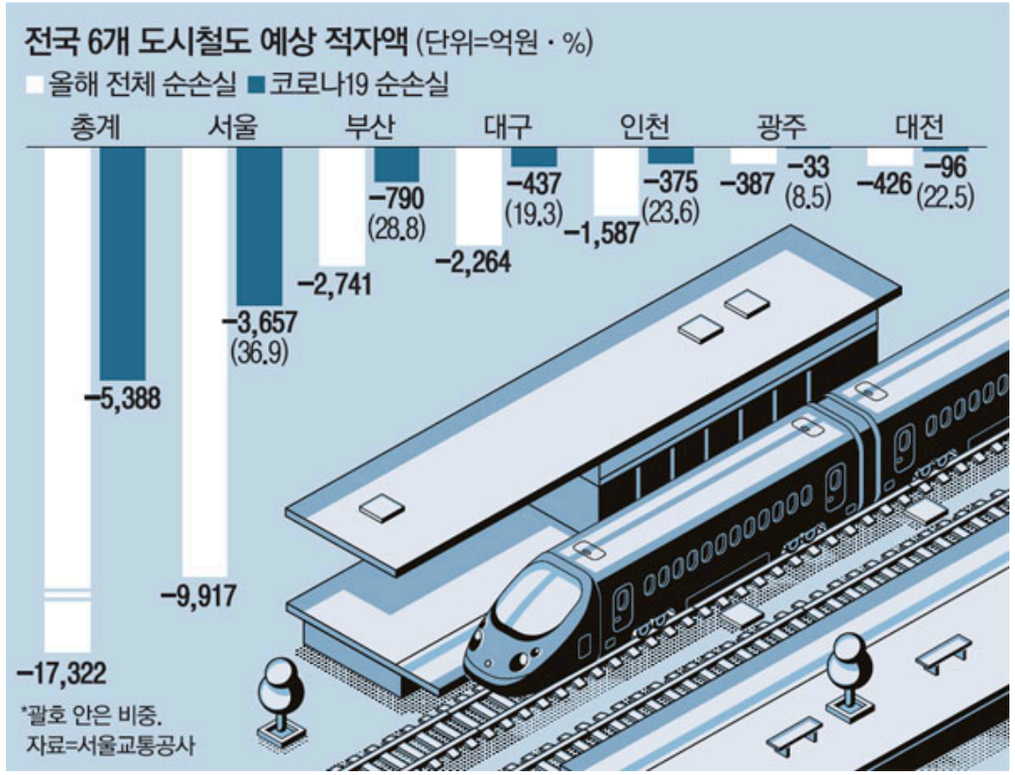
A29면 TOP | 기사입력 2020.11.15. 오후 5:49 기사원문 스크랩 본문듣기 • 설정

공감 댓글 요약봇 가

서울을 비롯한 전국 6개 도시철도의 올해 예상 적자액이 무려 1조7000억원에 달하는 것으로 나타났다. 도시철도의 만년 적자 원인인 '무임승차 손실' 보전 문제가 해결되지 않고 있는 가운데 코로나19 감염 우려에 따른 승객 감소가 겹친 탓이다. 코로나19발 도시철도 경영 위기가 현실화하면서 낮은 요금 체계 개선, 중앙정부의 법정 무임수송 손실분 보전 등 대책을 촉구하는 목소리가 높지만 정부와 지방자치단체 대응이 미온적이라는 비판이 나온다.

15일 서울교통공사에 따르면 서울, 부산, 대구, 인천, 광주, 대전 등 6개 도시철도의 올해 예상 당기순손실은 1조7322억원으로 집계됐다. 보유한 노선이 가장 많은 서울이 9917억원으로 예상 적자 폭이 가장 컸으며, 부산(2741억원) 대구(2264억원) 인천(1587억원) 대전(426억원) 광주(387억원)가 뒤를 이었다. 전체 예상 적자 가운데 코로나19에 따른 승객 감소가 불러온 순손실액만 5388억원으로 30%가 넘는 비중을 차지했다.

올해 전국 도시철도의 예상 적자액이 이토록 심각한 이유는 무임승차 손실을 철도 운영기관이 온전히 떠안아왔고, 코로나19로 인한 큰 폭의 승객 감소가 더해졌기 때문이다. 예상 적자 폭이 가장 큰 서울교통공사는 2016년부터 지난해까지 연간 3000억원대 무임수송 손실이 발생해 적자가 지속적으로 쌓여 왔다.



* 현화영(세계일보), 네이버 쇼핑라이브, 이젠 '검색'으로 만난다, 2020.11.15., https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=001&oid=022&aid=0003523331/.
** references
*** references

문서를 요약하는 방법

문서에서 핵심 키워드(keyword, key-phrase) 추출하기

Text mining

From Wikipedia, the free encyclopedia

Text mining, also referred to as ***text data mining***, similar to **text analytics**, is the process of deriving high-quality information from text. It involves "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources."[1] Written resources may include websites, books, emails, reviews, and articles. High-quality information is typically obtained by devising patterns and trends by means such as statistical pattern learning. According to Hotho et al. (2005) we can differ three different perspectives of **text mining**: information extraction, data mining, and a KDD (Knowledge Discovery in Databases) process.[2] Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interest. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (*i.e.*, learning relations between named entities).

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP), different types of algorithms and analytical methods. An important phase of this process is the interpretation of the gathered information.

A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted. The document is the basic element while starting with text mining. Here, we define a document as a unit of textual data, which normally exists in many types of collections.[3]

* 위키피디아, https://en.wikipedia.org/wiki/Text_mining/.
** references
*** references

추출요약과 생성요약

문서 요약 (Document Summurization)

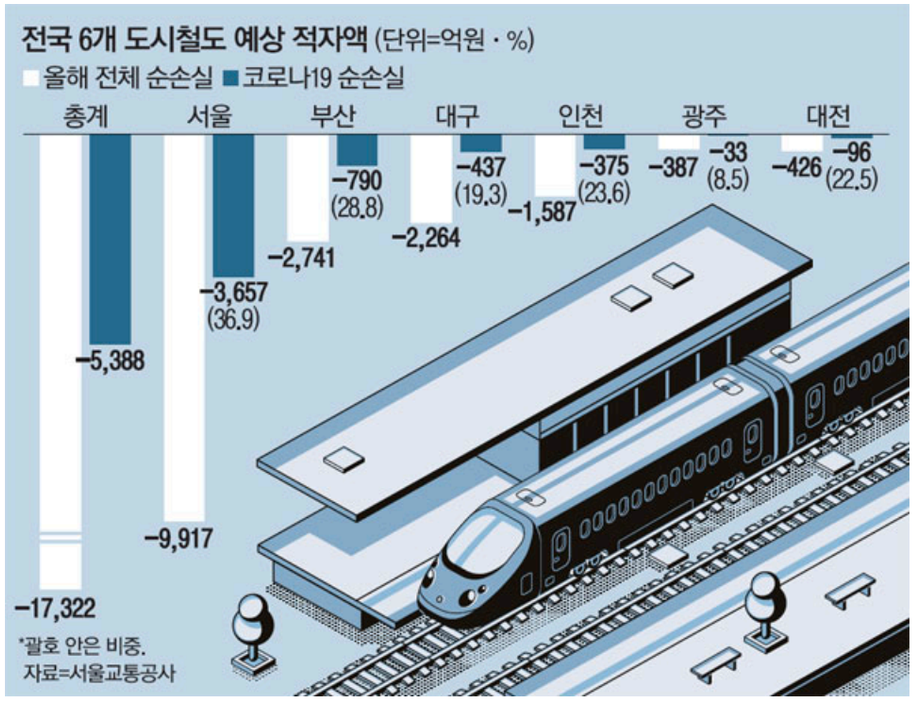
- 문서를 목적에 맞게 축약된 형태의 문서로 표현하는 방법
- 문서의 복잡도를 줄이면서도 필요한 정보는 유지하고 강조하여 표현하는 것이 중요함
- **추출요약** (Extractive Summarization) : 문서에 존재하는 단어나 구, 문장을 그대로 활용하여 요약하는 방법
- **생성요약** (Abstract Summarization) : 문서의 내용을 요약하여 표현한 새로운 문서를 작성하는 방법

매일경제

코로나로 승객 29% 감소…지하철 적자 1.7조

A29면 TOP | 기사입력 2020.11.15. 오후 5:49 기사원문 스크랩 본문듣기 · 설정

공감 댓글 요약봇 가



서울을 비롯한 전국 6개 도시철도의 올해 예상 적자액이 무려 1조7000억원에 달하는 것으로 나타났다. 도시철도의 만년 적자 원인인 '무임승차 손실' 보전 문제가 해결되지 않고 있는 가운데 코로나19 감염 우려에 따른 승객 감소가 겹친 탓이다. 코로나19발 도시철도 경영 위기가 현실화하면서 낮은 요금 체계 개선, 중앙정부의 법정 무임수송 손실분 보전 등 대책을 촉구하는 목소리가 높지만 정부와 지방자치단체 대응이 미온적이라는 비판이 나온다.

15일 서울교통공사에 따르면 서울, 부산, 대구, 인천, 광주, 대전 등 6개 도시철도의 올해 예상 당기순손실은 1조7322억원으로 집계됐다. 보유한 노선이 가장 많은 서울이 9917억원으로 예상 적자 폭이 가장 컸으며, 부산(2741억원) 대구(2264억원) 인천(1587억원) 대전(426억원) 광주(387억원)가 뒤를 이었다. 전체 예상 적자 가운데 코로나19에 따른 승객 감소가 불러온 순손실액만 5388억원으로 30%가 넘는 비중을 차지했다.

올해 전국 도시철도의 예상 적자액이 이토록 심각한 이유는 무임승차 손실을 철도 운영기관이 온전히 떠안아왔고, 코로나19로 인한 큰 폭의 승객 감소가 더해졌기 때문이다. 예상 적자 폭이 가장 큰 서울교통공사는 2016년부터 지난해까지 연간 3000억원대 무임수송 손실이 발생해 적자가 지속적으로 쌓여 왔다.

* 나의 큰 O는 log x야(티스토리), 자동 요약 기법의 연구 동향 정리, https://bab2min.tistory.com/625?category=673750/.
** 최현재(매일경제), 코로나로 승객 29% 감소…지하철 적자 1.7조, 2020.11.15., https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=103&oid=009&aid=0004695804/.
*** references

추출요약과 생성요약

추출요약 (Extractive Summarization)

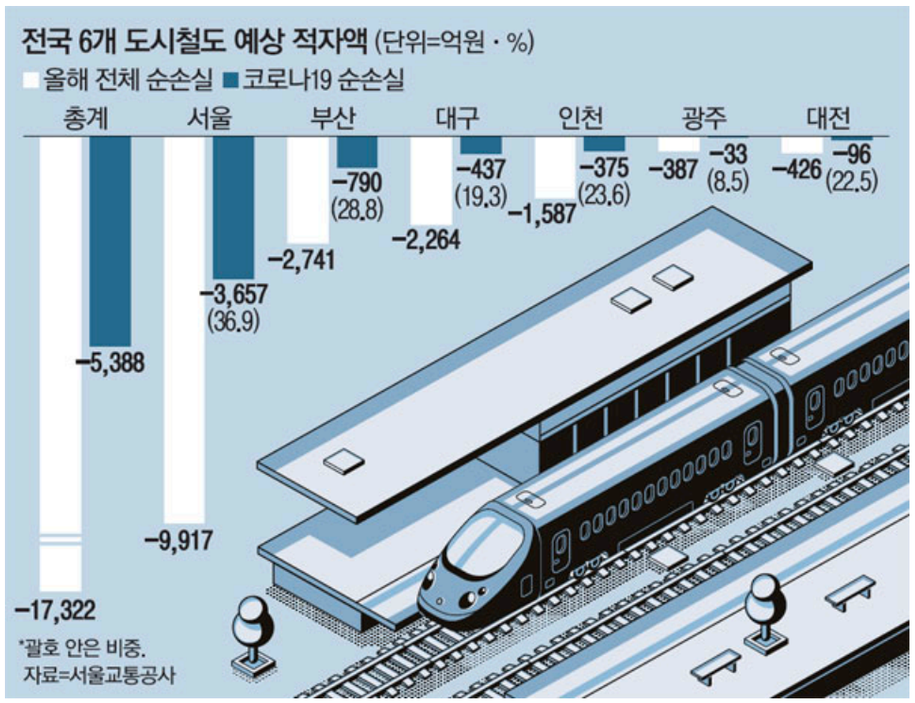
- 문서 내에서 중요한 핵심 문장(또는 단어, 구)들을 뽑아 이를 조합하여 요약문을 만드는 방법 (보통 2~3개 문장으로 요약)
- 어떤 문장이 핵심인지만을 파악하여 그걸 골라내고 순서에 맞게 이어 붙여 요약문을 생성함
- 딥러닝(deep learning)의 출현 전까지는 주로 추출요약에 대한 연구가 자동 요약 분야의 대세를이루고 있음

매일경제

코로나로 승객 29% 감소…지하철 적자 1.7조

A29면 TOP | 기사입력 2020.11.15. 오후 5:49 기사원문 스크랩 본문듣기 · 설정

공감 댓글 요약봇 가



서울을 비롯한 전국 6개 도시철도의 올해 예상 적자액이 무려 1조7000억원에 달하는 것으로 나타났다. 도시철도의 만년 적자 원인인 '무임승차 손실' 보전 문제가 해결되지 않고 있는 가운데 코로나19 감염 우려에 따른 승객 감소가 겹친 탓이다. 코로나19발 도시철도 경영 위기가 현실화하면서 낮은 요금 체계 개선, 중앙정부의 법정 무임수송 손실분 보전 등 대책을 촉구하는 목소리가 높지만 정부와 지방자치단체 대응이 미온적이라는 비판이 나온다.

15일 서울교통공사에 따르면 서울, 부산, 대구, 인천, 광주, 대전 등 6개 도시철도의 올해 예상 당기순손실은 1조7322억원으로 집계됐다. 보유한 노선이 가장 많은 서울이 9917억원으로 예상 적자 폭이 가장 컸으며, 부산(2741억원) 대구(2264억원) 인천(1587억원) 대전(426억원) 광주(387억원)가 뒤를 이었다. 전체 예상 적자 가운데 코로나19에 따른 승객 감소가 불러온 순손실액만 5388억원으로 30%가 넘는 비중을 차지했다.

올해 전국 도시철도의 예상 적자액이 이토록 심각한 이유는 무임승차 손실을 철도 운영기관이 온전히 떠안아왔고, 코로나19로 인한 큰 폭의 승객 감소가 더해졌기 때문이다. 예상 적자 폭이 가장 큰 서울교통공사는 2016년부터 지난해까지 연간 3000억원대 무임수송 손실이 발생해 적자가 지속적으로 쌓여 왔다.

* 나의 큰 O는 log x야(티스토리), 자동 요약 기법의 연구 동향 정리, https://bab2min.tistory.com/625?category=673750/.
** 최현재(매일경제), 코로나로 승객 29% 감소…지하철 적자 1.7조, 2020.11.15., https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=103&oid=009&aid=0004695804/.
*** references

추출요약과 생성요약

생성요약 (Abstractive Summarization, 추상요약)

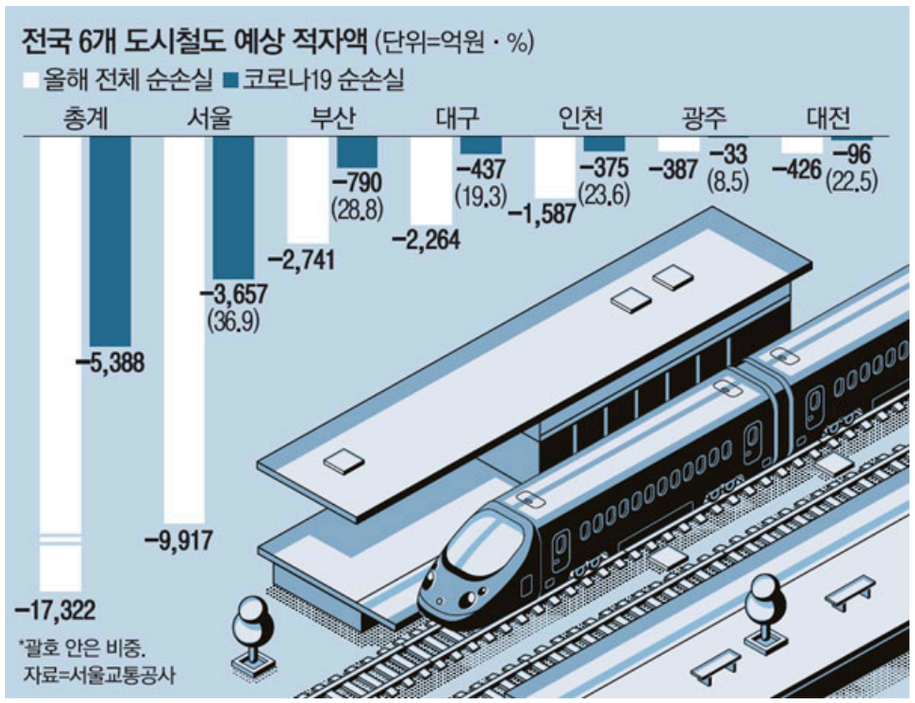
- 문서의 내용을 잘 표현할 수 있는 새로운 문장을 직접 생성함으로써 요약문을 만드는 방법
- 문서 전체 내용을 이해하고 그 내용을 잘 표현할 수 있는 간결한 문장을 직접 작성해야하므로 상당히 고난도의 자연어생성(natural language generation, NLG) 기술이 필요함
- 최근 딥 러닝 기반의 자연언어처리 기술이 발전하면서 추상요약에 대한 많은 연구가 이루어지고 있음

매일경제

코로나로 승객 29% 감소…지하철 적자 1.7조

A29면 TOP | 기사입력 2020.11.15. 오후 5:49 기사원문 스크랩 본문듣기 · 설정

공감 댓글 요약봇 가



서울을 비롯한 전국 6개 도시철도의 올해 예상 적자액이 무려 1조7000억원에 달하는 것으로 나타났다. 도시철도의 만년 적자 원인인 '무임승차 손실' 보전 문제가 해결되지 않고 있는 가운데 코로나19 감염 우려에 따른 승객 감소가 겹친 탓이다. 코로나19발 도시철도 경영 위기가 현실화하면서 낮은 요금 체계 개선, 중앙정부의 법정 무임수송 손실분 보전 등 대책을 촉구하는 목소리가 높지만 정부와 지방자치단체 대응이 미온적이라는 비판이 나온다.

15일 서울교통공사에 따르면 서울, 부산, 대구, 인천, 광주, 대전 등 6개 도시철도의 올해 예상 당기순손실은 1조7322억원으로 집계됐다. 보유한 노선이 가장 많은 서울이 9917억원으로 예상 적자 폭이 가장 컸으며, 부산(2741억원) 대구(2264억원) 인천(1587억원) 대전(426억원) 광주(387억원)가 뒤를 이었다. 전체 예상 적자 가운데 코로나19에 따른 승객 감소가 불러온 순손실액만 5388억원으로 30%가 넘는 비중을 차지했다.

올해 전국 도시철도의 예상 적자액이 이토록 심각한 이유는 무임승차 손실을 철도 운영기관이 온전히 떠안아왔고, 코로나19로 인한 큰 폭의 승객 감소가 더해졌기 때문이다. 예상 적자 폭이 가장 큰 서울교통공사는 2016년부터 지난해까지 연간 3000억원대 무임수송 손실이 발생해 적자가 지속적으로 쌓여 왔다.

* 나의 큰 O는 log x야(티스토리), 자동 요약 기법의 연구 동향 정리, https://bab2min.tistory.com/625?category=673750/.
** 최현재(매일경제), 코로나로 승객 29% 감소…지하철 적자 1.7조, 2020.11.15., https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=103&oid=009&aid=0004695804/.
*** references

키워드 추출 알고리즘

키워드 추출 (Keyword Extraction)

- 문서의 주제를 가장 잘 설명하는 키워드를 자동으로 식별하는 작업
- 키워드란 하나의 단어 뿐만 아니라 구(phrase)단위까지를 의미하며, 영문 표현으로 Keyword, Key-phrase, Key-term, Key-segment 등이 모두 같은 의미를 가짐 (주로 명사구 단위를 취급함)
- 키워드 추출은 텍스트 마이닝, 정보검색(information retrieval, IR) 및 자연어처리(natural language processing, NLP) 분야에서 오랫동안 중요한 문제로 인식되어 왔으며 다양한 키워드 추출 알고리즘이 제안됨 (TF-IDF, TextRank, PKEA, RKEA, …)

구분	내용
원문	손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홋 스 퍼 의 승 리 를 이 끌 었 다 .
토큰화	손흥민 | 이 | 골 | 을 | 작렬 | 하 | 며 | 토트넘 홋스퍼 (?) | 의 | 승리 | 를 | 이끌 | 었 | 다 | .
키워드 추출	1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0

* references
** references
*** references

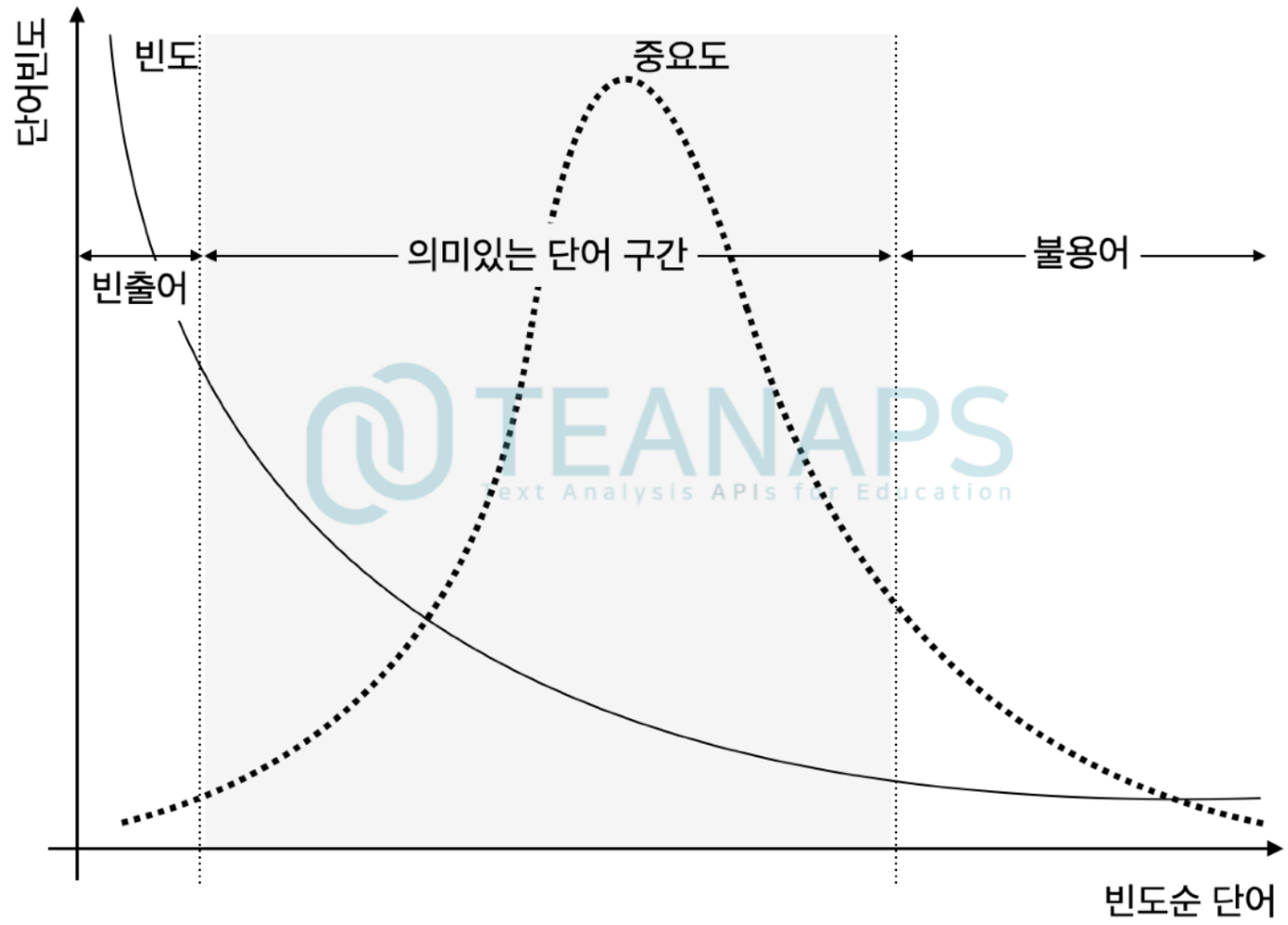
Review

단어 가중치: 단어빈도

단어빈도 (Term Frequency, TF)

- 특정 단어가 문서에 출현한 횟수로 단어의 특징을 표현하는 가장 간단한 방법
- 간단하지만 가장 빠르게 문서를 표현하고 파악할 수 있으며 기초통계와 같이 분석 전 반드시 거쳐야 하는 과정
- 단어가 너무 희귀한 경우 큰 의미를 부여하기 어려우며, 너무 흔한 경우 의미가 과도하게 부여될 가능성이 있음

$$TermFrequency = count(word|document)$$



* 전병진, (2019). 소셜미디어 텍스트 마이닝을 위한 통합 애플리케이션 개발 : KoALA. 연세대학교 정보대학원.
** references
*** references

Review

단어 가중치: 단어빈도

TF-IDF (TF-Inverse Document Frequency)

- **역문서빈도** (Inverse Document Frequency, IDF) : 단어가 출현한 문서가 적을수록 단어의 가중치를 낮게 표현하는 방법 (희박성)

  $IDF = 1 + count(word/document)/count(document)$

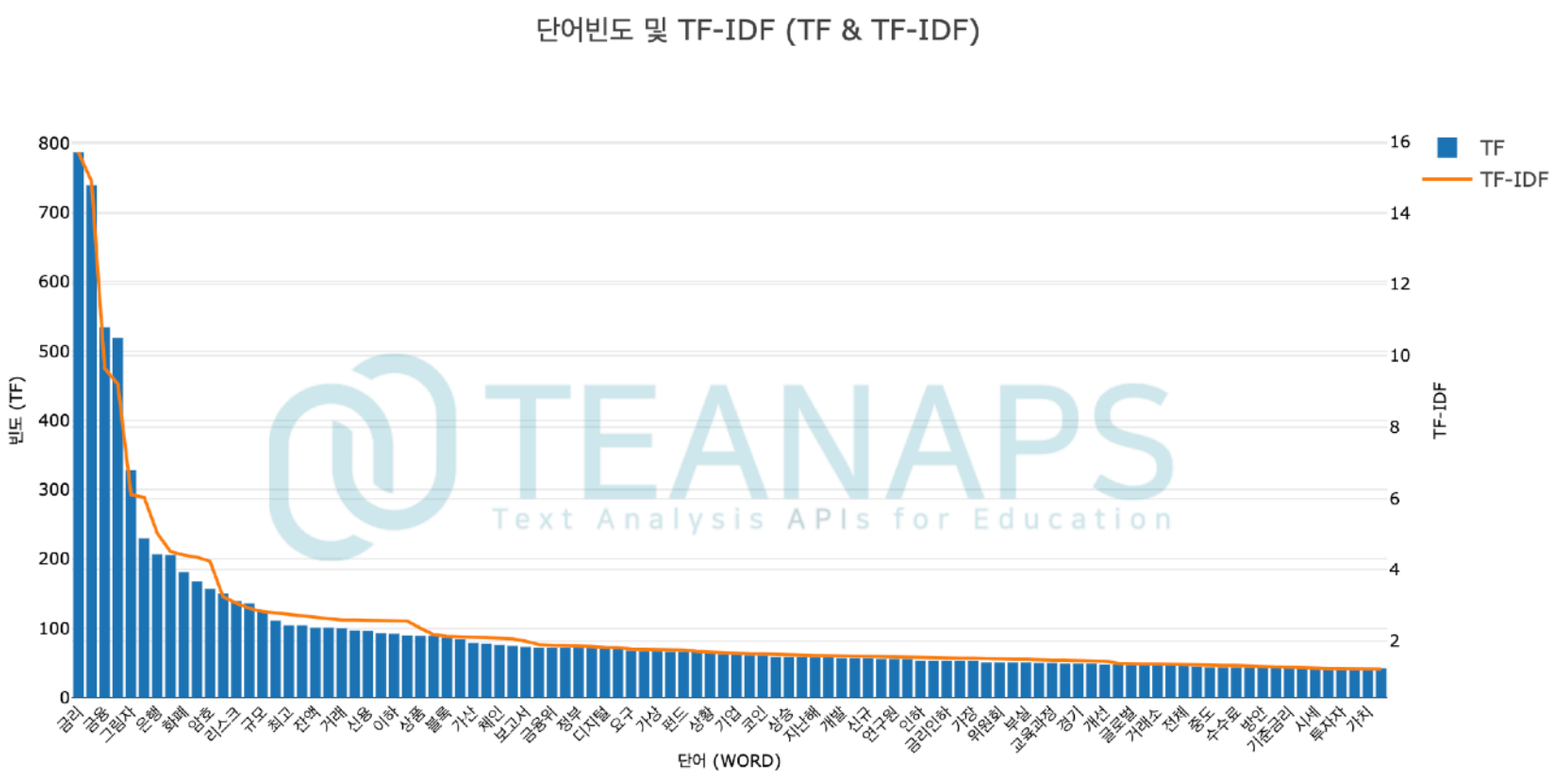
- TF와 IDF 개념을 통합하여, 단어가 문서에 출현한 횟수와 희박성을 동시에 활용해 가중치를 표현하는 방법

  $TF - IDF = Frequency * IDF$

- **지프의 법칙** (Zipf's law)

  1) 자연어에 나타나는 단어들을 출현 횟수가 높은 순으로 정렬하면, 단어의 출현 횟수는 순위에 반비례함

  2) 가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높음



* 김수인, 김재원, and 배휘동, 왜 프로그래밍에는 창의성이 필요하다고 할까요, 2017.5.25., https://medium.com/elice/.

** references

*** references

단어 가중치: 네트워크 중심성

Review

동시출현 분석 (Co-word Analysis)

- 문서에 서로다른 두 단어의 동시출현 횟수와 네트워크 중심성을 통해 단어의 특징을 표현하는 방법
- 두 단어 사이의 동시출현을 연관성의 척도로 취급하고, 그 관계를 네트워크 중심성으로 표현하여 가중치를 계산함
- **연관어** (공기어, Co-word) : 하나의 문서에서 함께 출현하여 서로 밀접한 의미관계를 가지는 단어

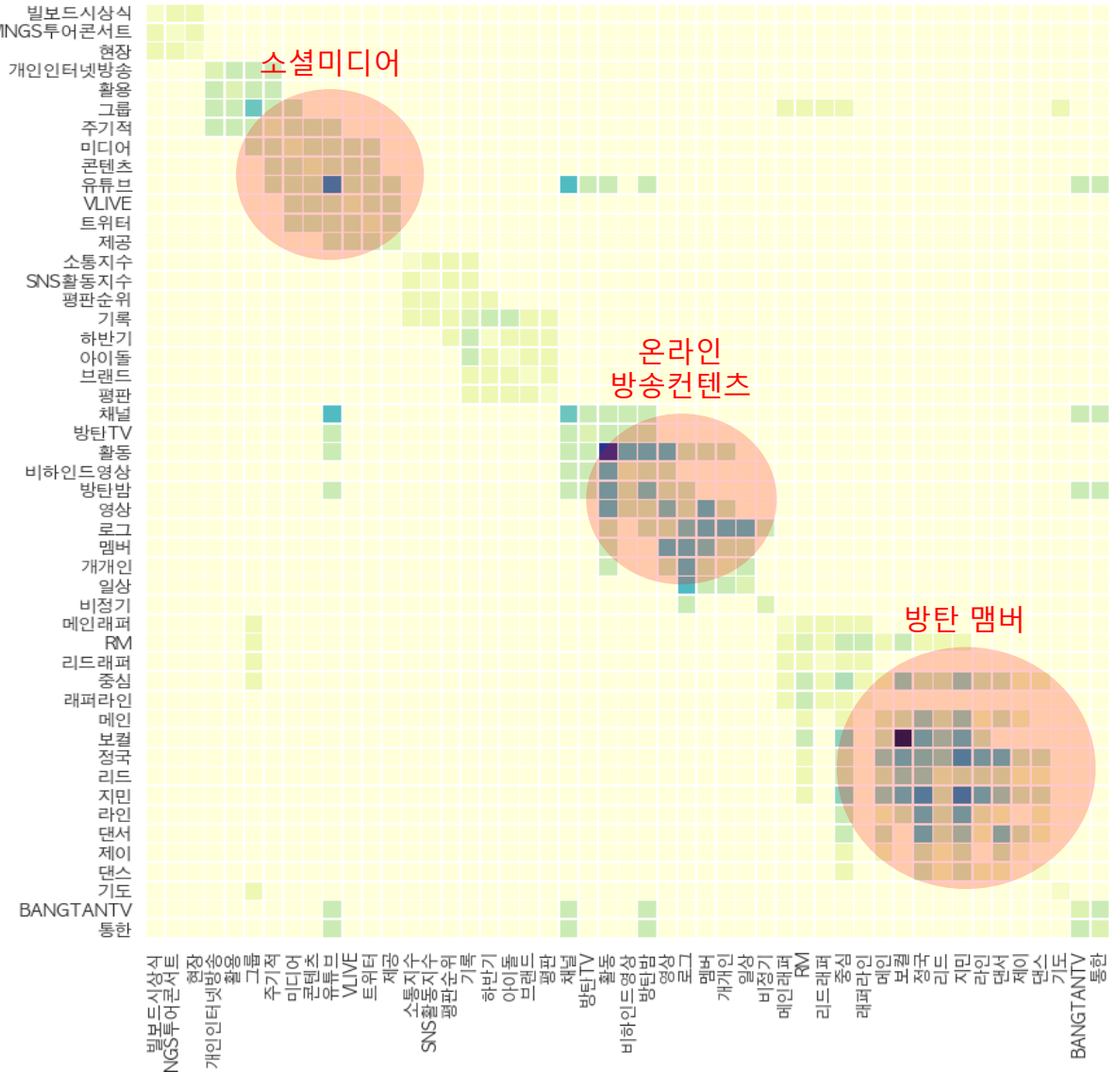


표1 '야쿠르트 아줌마' 연관어 변화

야쿠르트 아줌마는 여전히 '야쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 야쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.

2013년			2014년			2015년			2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	야쿠르트	21.3%	1	야쿠르트	26.3%	1	야쿠르트	26.6%	1	야쿠르트	13.1%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%	2	콜드브루	8.2%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%	3	커피	7.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%	4	맛	6.6%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%	5	끼리	5.7%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%	6	치즈	5.3%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%	7	과자	5.0%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%	8	아메리카노	4.1%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%	9	먹다	3.3%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%	10	크림치즈	3.1%
11	야쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%	11	라떼	2.8%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%	12	만나다	2.7%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%	13	가격	2.4%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%	14	찾다	1.9%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%	15	아침	1.8%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%	16	10일	1.6%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%	17	엄마	1.5%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%	18	우유	1.4%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%	19	팔다	1.3%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%	20	발견하다	1.3%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%	21	사다	1.2%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%	22	인기	1.2%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%	23	편의점	1.2%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%	24	끼리딥앤크런치	1.1%
	...			...			...			...	
29	팔다	1.4%	29	묻다	1.3%	29	배달하다	1.3%	29	구입하다	1.0%

상승 키워드 하락 키워드 신규 키워드

* 전병진, 신한은행 파이썬으로 시작하는 데이터분석: 텍스트 마이닝 기초, 2018.12.12.
** 백경혜(DBR), "매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다", 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.
*** references

키워드 추출 알고리즘: PKEA

Algorithm: Step 1. 키워드 후보군 선정

- (1) 텍스트 전처리 : 문서를 문장 단위로 분리하고, 토큰화를 통해 단어주머니 생성


```
Text mining, also referred to as text data mining, roughly equivalent to text analytics, is
the process of deriving high-quality information from text. High-quality information is
typically derived through the devising of patterns and trends through means such as
statistical pattern learning. Text mining usually involves the process of structuring the
input text (usually parsing, along with the addition of some derived linguistic features and
the removal of others, and subsequent insertion into a database), deriving patterns within
the structured data, and finally evaluation and interpretation of the output. 'High quality'
in text mining usually refers to some combination of relevance, novelty, and
interestingness. Typical text mining tasks include text categorization, text clustering,
concept/entity extraction, production of granular taxonomies, sentiment analysis, document
summarization, and entity relation modeling (i.e., learning relations between named
entities).
Text analysis involves information retrieval, lexical analysis to study word frequency
distributions, pattern recognition, tagging/annotation, information extraction, data mining
techniques including link and association analysis, visualization, and predictive analytics.
The overarching goal is, essentially, to turn text into data for analysis, via application
of natural language processing (NLP) and analytical methods.
A typical application is to scan a set of documents written in a natural language and either
model the document set for predictive classification purposes or populate a database or
search index with the information extracted.
The term text analytics describes a set of linguistic, statistical, and machine learning
techniques that model and structure the information content of textual sources for business
intelligence, exploratory data analysis, research, or investigation.[1] The term is roughly
synonymous with text mining; indeed, Ronen Feldman modified a 2000 description of "text
mining"[2] in 2004 to describe "text analytics".[3] The latter term is now used more
frequently in business settings while "text mining" is used in some of the earliest
application areas, dating to the 1980s,[4] notably life-sciences research and government
intelligence.
The term text analytics also describes that application of text analytics to respond to
business problems, whether independently or in conjunction with query and analysis of
fielded, numerical data. It is a truism that 80 percent of business-relevant information
originates in unstructured form, primarily text.[5] These techniques and processes discover
and present knowledge - facts, business rules, and relationships - that is otherwise locked
in textual form, impenetrable to automated processing.
```

구분	문장	유니그램 (불용어 제거 후)
1	Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving …	Text, mining, *, referred, *, text, data, mining, *, equivalent, *, text, analytics, *, process, *, deriving, …
2	Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of …	Text, mining, *, involves, *, process, *, structuring, *, input, text, *, parsing, along, *, addition, *, …
3	High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.	High, quality, *, text, mining, *, refers, *, combination, *, relevance, *, novelty, interestingness, …

* Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2005). Kea: Practical automated keyphrase extraction. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific (pp. 129-152). IGI Global.
** references
*** references

키워드 추출 알고리즘: PKEA

Algorithm: Step 1. 키워드 후보군 선정

- (2) 키워드 식별 : 불용어를 제외한 문장내 모든 단어들의 연속된 조합(N-gram) 식별 (불용어를 기준으로 키워드 구분)
- (3) 원형복원 및 대/소문자 통일 : 어간추출(stemming) 또는 표제어추출(lemmatization)을 활용하여 단어를 원형을 복원

구분	문장	유니그램 (불용어 제거 후)
1	Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving ···	Text, mining, *, referred, *, text, data, mining, *, equivalent, *, text, analytics, *, process, *, deriving, ···
2	Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of ···	Text, mining, *, involves, *, process, *, structuring, *, input, text, *, parsing, along, *, addition, *, ···
3	High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.	High, quality, *, text, mining, *, refers, *, combination, *, relevance, *, novelty, interestingness, ···

구분	키워드 후보군
1	Text
2	mining
3	Text mining
4	referred
5	text
6	data
7	mining
8	Text data
9	data mining
10	text data mining

구분	키워드 후보군
1	text
2	mining
3	text mining
4	refer
5	data
6	text data
7	data mining
8	text data mining
9	equivalent
10	analytics

* Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2005). Kea: Practical automated keyphrase extraction. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific (pp. 129-152). IGI Global.
** references
*** references

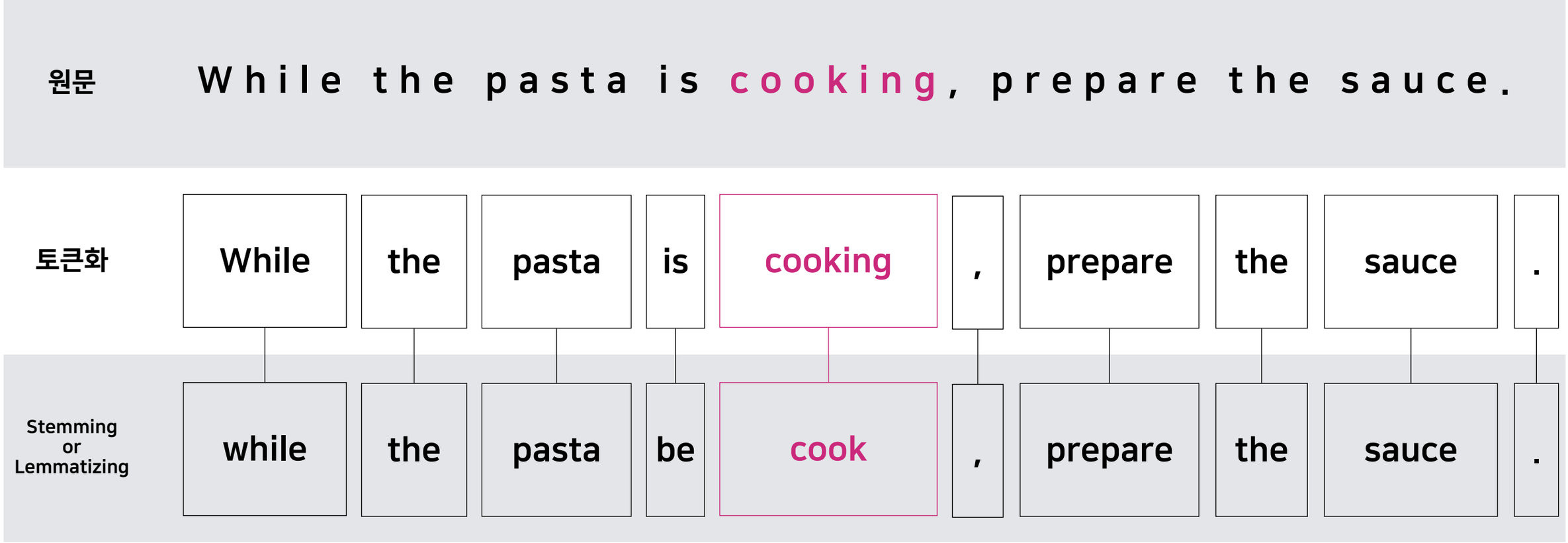
영어 형태소 분석

Word	Stemming	Lemmatizing
cooking(v)	cook	
cooking(n)	cook	c
cookbooks	cookbook	cook
believes	believ	believe
using	us	use

Review

원형복원: 어간추출과 표제어추출 (Stemming & Lemmatization)

- 원형에서 다른 형태로 변형되어 활용된 단어를 원형으로 복원하는 과정
- **어간추출** (Stemming) : 규칙 기반으로 단어의 변형된 형태를 제거하거나 치환하여 원형으로 복원하는 방법
  cooking (v) →cook, cooking (n) →cook, cookbooks → cookbook, believes → believ, using → us
- **표제어추출** (Lemmatizing) : 단어의 형변환 사전을 기반으로 대상 단어의 품사에 맞는 단어의 원형으로 복원하는 방법
  cooking (v) →cook, **cooking (n) →cooking**, cookbooks → cookbook, **believes → believe**, **using → use**



* references
** references
*** references

키워드 추출 알고리즘: PKEA

Algorithm: Step 2. 자질 추출 (Feature Extraction)

- TF-IDF : 단어가 문서에 출현한 횟수와 희박성을 동시에 활용해 가중치를 표현하는 방법
- First Occurrence : 단어가 문장의 첫 단어로부터 등장하는 평균거리 (가까울 수록 중요하다고 가정함)

구분	키워드 후보군
1	text
2	mining
3	text mining
4	refer
5	data
6	text data
7	data mining
8	text data mining
9	equivalent
10	analytics
11	…

구분	키워드 후보군	TF-IDF	First Occurrence
1	text	1.0833	0
2	mining	1.1000	1
3	text mining	1.4555	1
4	refer	0.7500	3
5	data	0.7500	7
6	text data	1.4555	7
7	data mining	1.4666	8
8	text data mining	…	…
9	equivalent	…	…
10	Analytics	…	…
11	…	…	…

* Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2005). Kea: Practical automated keyphrase extraction. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific (pp. 129-152). IGI Global.
** references
*** references

Review

단어 가중치: 단어빈도

TF-IDF (TF-Inverse Document Frequency)

- **역문서빈도** (Inverse Document Frequency, IDF) : 단어가 출현한 문서가 적을수록 단어의 가중치를 낮게 표현하는 방법 (희박성)

  $IDF = 1 + count(word/document)/count(document)$

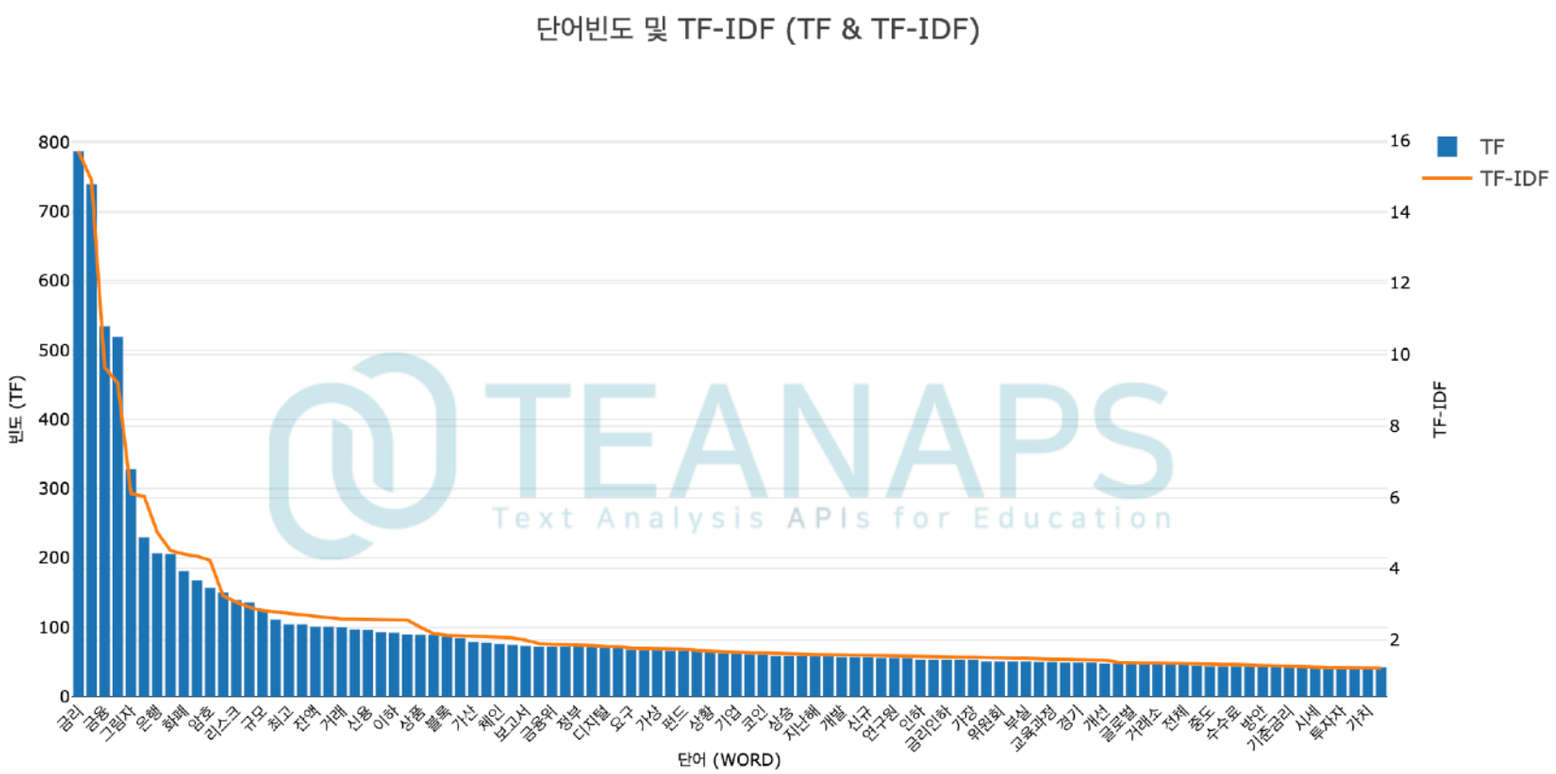
- TF와 IDF 개념을 통합하여, 단어가 문서에 출현한 횟수와 희박성을 동시에 활용해 가중치를 표현하는 방법

  $TF - IDF = Frequency * IDF$

- **지프의 법칙** (Zipf's law)

  1) 자연어에 나타나는 단어들을 출현 횟수가 높은 순으로 정렬하면, 단어의 출현 횟수는 순위에 반비례함

  2) 가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높음



* 김수인, 김재원, and 배휘동, 왜 프로그래밍에는 창의성이 필요하다고 할까요, 2017.5.25., https://medium.com/elice/.
** references
*** references

키워드 추출 알고리즘: PKEA

Algorithm: Step 3. 학습 (Traning)

- 학습될 문서와 그 문서의 키워드 후보군, 그리고 각 후보군의 특성값이 학습데이터로 활용됨
- 머신러닝 기반의 분류 알고리즘에 의해 모델을 학습함 (논문에서는 나이브 베이즈 알고리즘 활용)

구분	키워드 후보군	TF-IDF	First Occurrence	Label (키워드 여부)
1	text	1.0833	0	0
2	mining	1.1000	1	0
3	text mining	1.4555	1	1
4	refer	0.7500	3	0
5	data	0.7500	7	0
6	text data	1.4555	7	1
7	data mining	1.4666	8	1
8	text data mining	…	…	…
9	equivalent	…	…	…
10	Analytics	…	…	…
11	…	…	…	…

```
-- Extracting Keyphrases...
-- Reading instance
-- Converting instance
mining  0.5171
text mining     0.5171
data    0.5171
text data       0.5171
data mining     0.5171
text    0.5171
information     0.262
analytics       0.262
process 0.262
deriving        0.262
high-quality    0.262
high-quality information        0.262
derived 0.1455
involves        0.1455
analysis        0.0561
document        0.0561
database        0.0561
Typical 0.0561
text mining tasks include       0.0561
extraction      0.0561
predictive      0.0561
application     0.0561
```

* Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2005). Kea: Practical automated keyphrase extraction. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific (pp. 129-152). IGI Global.
** references
*** references

키워드 추출 알고리즘: PKEA

Algorithm: Step 4. 검증 (Validation)

- 학습된 키워드 추출 모델을 바탕으로 새로운 문서에 대해 키워드를 추출하고 결과를 비교함
- 모델이 추출한 키워드와 문서의 작성자가 선정한 키워드와 일치횟수를 계산하여 모델의 정확도를 측정함
  → 테스트 결과 저자가 선정한 키워드와 20~30% 일치하는 결과를 보임
  → 20개의 적은 학습데이터 셋으로도 좋은 성능의 모델을 만들 수 있음

Protocols for secure, atomic transaction execution in electronic commerce		Neural multigrid for gauge theories and other disordered systems		Proof nets, garbage, and computations	
anonymity	*atomicity*	disordered systems	disordered	*cut-elimination*	cut
atomicity	*auction*	*gauge fields*	gauge	linear logic	*cut elimination*
auction	customer	*multigrid*	*gauge fields*	*proof nets*	garbage
electronic commerce	*electronic commerce*	neural multigrid	interpolation kernels	sharing graphs	*proof net*
privacy	intruder	neural networks	length scale	typed lambda-calculus	weakening
real-time	merchant		*multigrid*		
security	protocol		smooth		
transaction	*security*				
	third party				
	transaction				

Figure 1 Examples of author- and Kea-assigned keyphrases

* Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2005). Kea: Practical automated keyphrase extraction. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific (pp. 129-152). IGI Global.
** references
*** references

E.O.D

Contact
http://www.teanaps.com
fingeredman@gmail.com