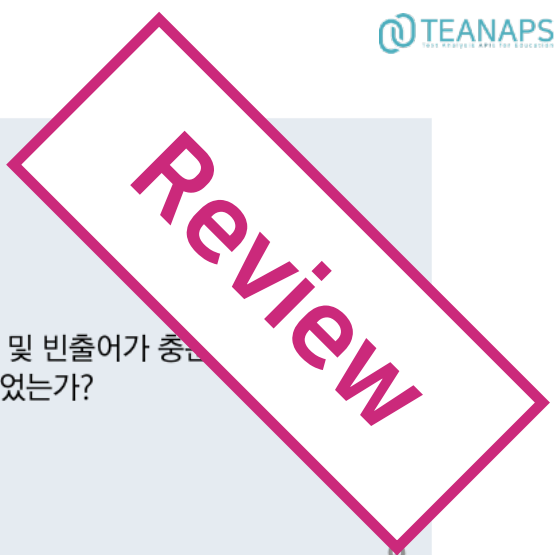


ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)



WEEK 09

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



N쇼핑 LIVE '온앤온 원피스' 관련 라이브 영상 ⓘ



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

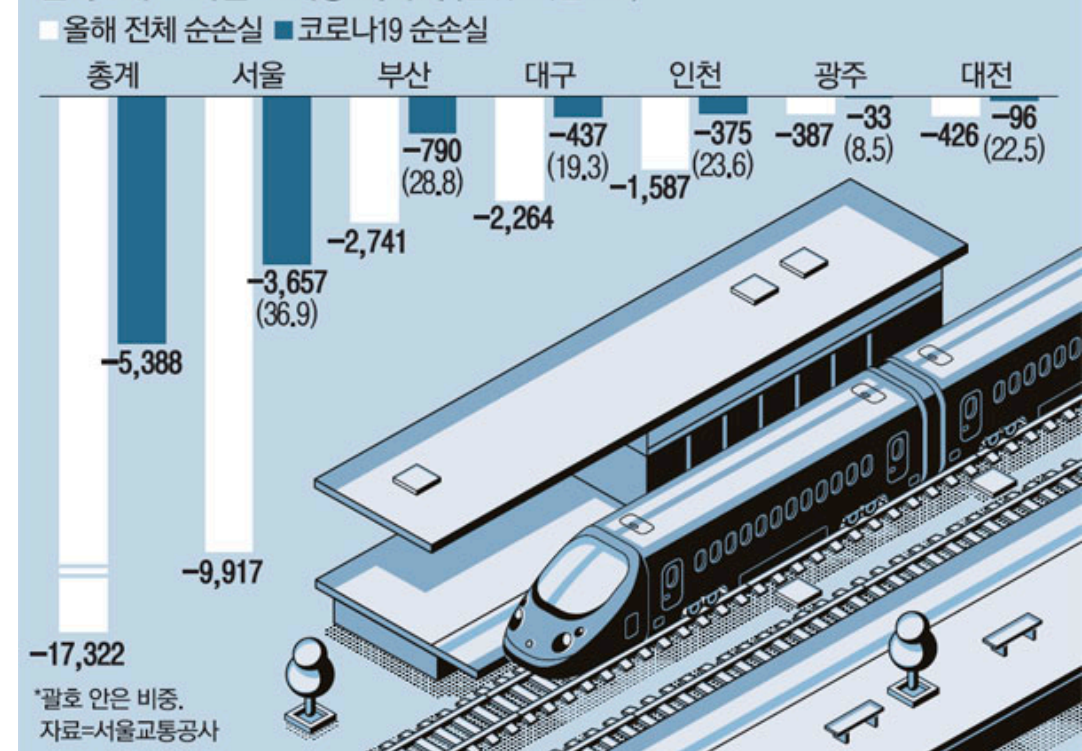


서울을 비롯한 전국 6개 도시철도의 올해 예상 적자액이 무려 1조7000억원에 달하는 것으로 나타났다. 도시철도의 만년 적자 원인인 '무임승차 손실' 보전 문제가 해결되지 않고 있는 가운데 코로나19 감염 우려에 따른 승객 감소가 겹친 탓이다. 코로나19발 도시철도 경영 위기가 현실화하면서 낮은 요금 체계 개선, 중앙정부의 법정 무임수송 손실분 보전 등 대책을 촉구하는 목소리가 높지만 정부와 지방자치단체 대응이 미온적이라는 비판이 나온다.

15일 서울교통공사에 따르면 서울, 부산, 대구, 인천, 광주, 대전 등 6개 도시철도의 올해 예상 당기순 손실은 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.

전국 6개 도시철도 예상 적자액 (단위=억원 · %)



* 현화영(세계일보), 네이버 쇼핑라이브, 이젠 '검색'으로 만나다, 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

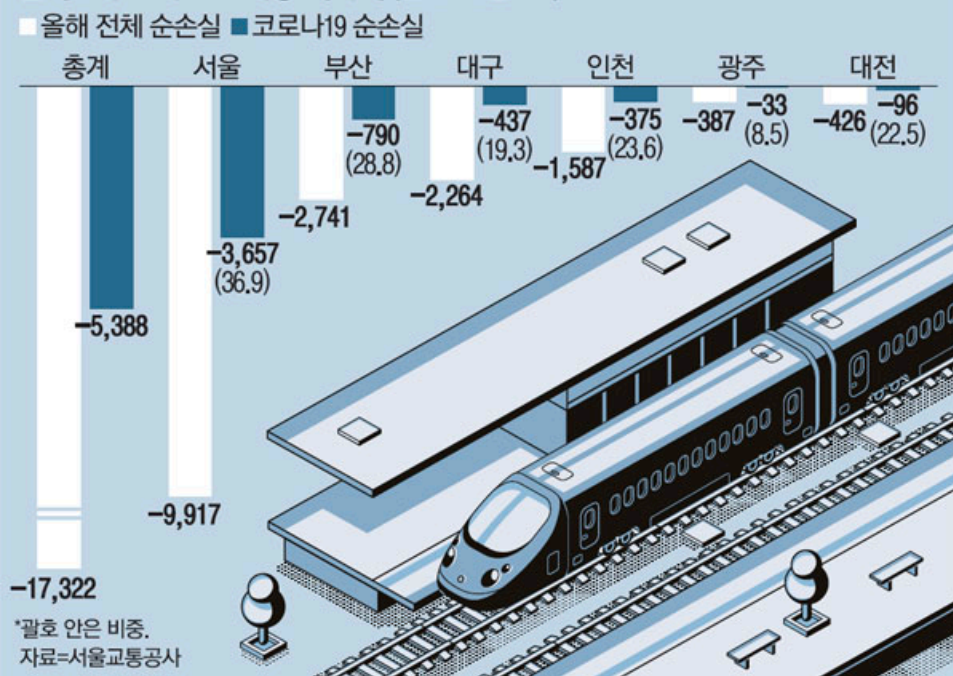
코로나로 승객 29% 감소...지하철 적자 1.7조

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

공감 댓글

요약본 가

전국 6개 도시철도 예상 적자액 (단위=억원 · %)



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 0는 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 대세를 이루고 있음

매일경제

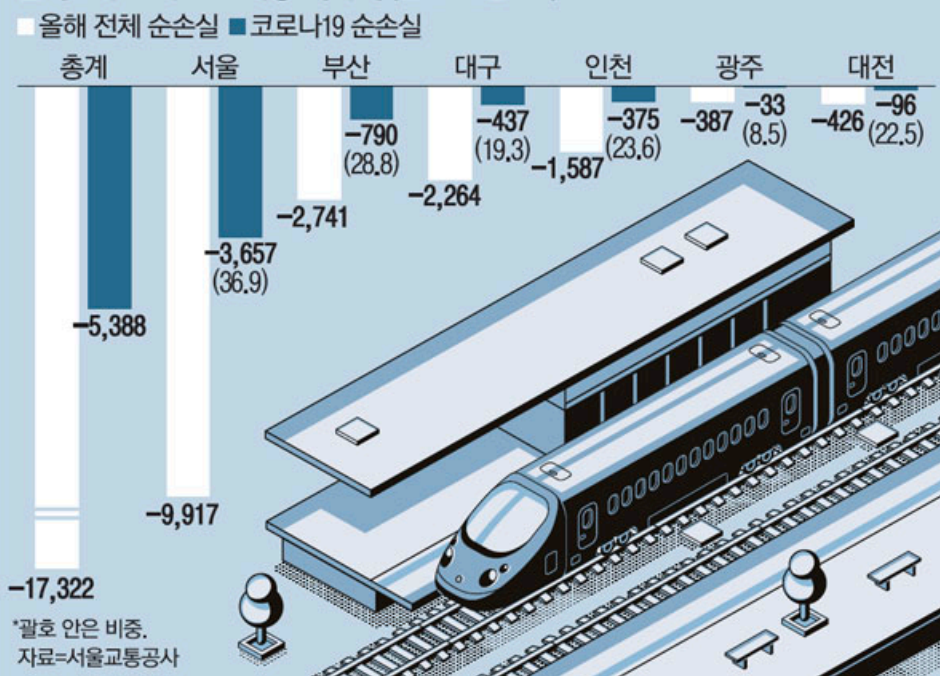
코로나로 승객 29% 감소...지하철 적자 1.7조

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

공감 댓글

요약봇 가

전국 6개 도시철도 예상 적자액 (단위=억원 · %)



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 당기순 손실은 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 0는 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 대세를 이루고 있음

매일경제

코로나로 승객 29% 감소...지하철 적자 1.7조

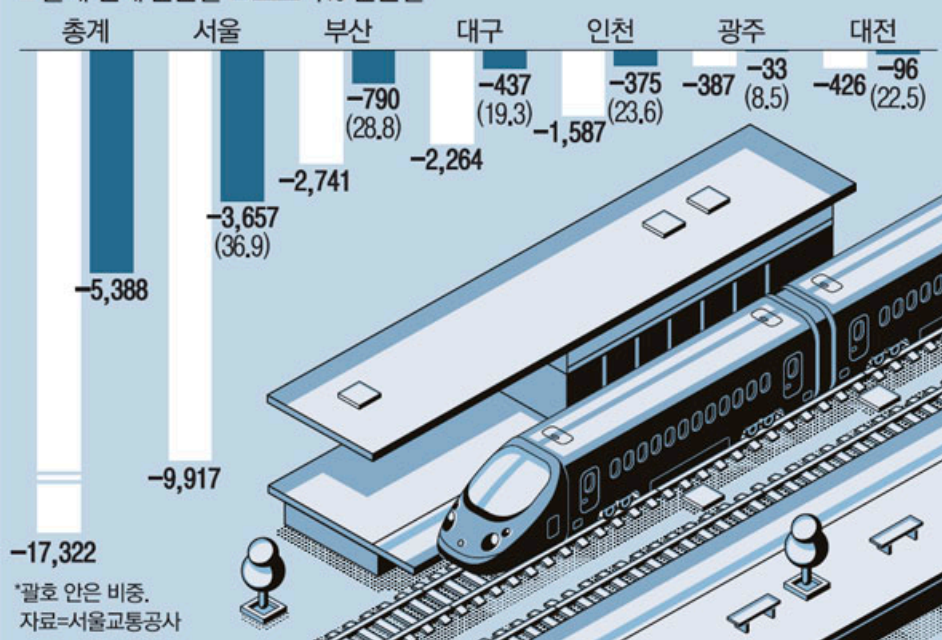
A29면 TOP | 기사입력 2020.11.15. 오후 5:49 | 기사원문 | 스크랩 | 본문듣기 · 설정

공감 댓글

요약봇 가

전국 6개 도시철도 예상 적자액 (단위=억원 · %)

■ 올해 전체 순손실 ■ 코로나19 순손실



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 스개 가스가 겨치 타이다. 크리 10반 도시철도 겨여 의기가 형식하하면서 낮은 요금 체계가 방자치단

15일 서울
손실은 1조
가장 컸으
이었다. 전
30%가 남

본문 요약봇

자동 추출 기술로 요약된 내용입니다. 요약 기술의 특성상 본문의 주요 내용이 제외될 수 있어, 전체 맥락을 이해하기 위해서는 기사 본문 전체보기를 권장합니다.

서울을 비롯한 전국 6개 도시철도의 올해 예상 적자액이 무려 1조7000억원에 달하는 것으로 나타났다.

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.

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.

상 당기순
적자 폭이
원)가 뒤를
으로

* 나의 큰 0는 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 딥러닝 기반의 자연언어처리 기술이 발전하면서 추상요약에 대한 많은 연구가 이루어지고 있음

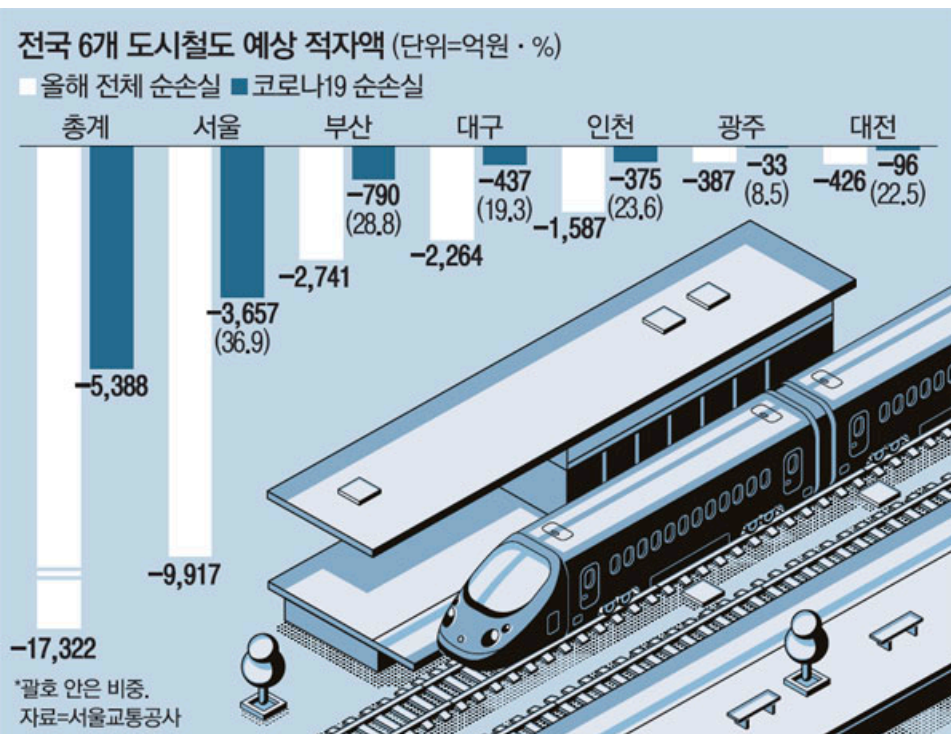
매일경제

코로나로 승객 29% 감소...지하철 적자 1.7조

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

공감 댓글

요약본 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 당기순 손실은 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.

* 나익의 큰 0는 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)단위**까지를 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 려 하 며 **토 트 념 핫 스 퍼** 의 승 리 를 이 끌 었 다 .

토큰화

손흥민

이

골

을

작렬

하

며

토트넘 핫스퍼 (?)

의

승리

를

이끌

었

다

.

키워드
추출

ㄱ

ㅇ

ㄱ

ㅇ

ㅇ

ㅇ

ㅇ

ㄱ

ㅇ

ㄱ

ㅇ

ㅇ

ㅇ

ㅇ

ㅇ

키워드 추출 알고리즘: PKEA

```

-----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-q
text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction,
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
**** 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-qual
text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction,
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
-- 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
natural 0.0561
natural language 0.0561
language 0.0561
set 0.0561
set of documents 0.0561
patterns 0.0238
concept/entity 0.0008

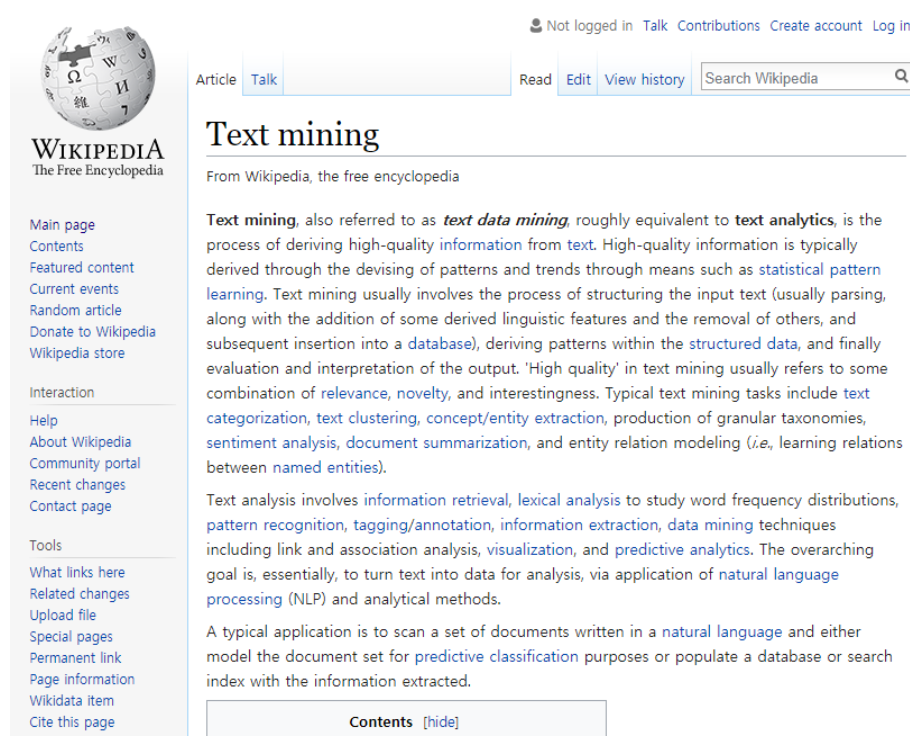
```

TEANAPS
Text Analysis APIs for Education

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

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 거리 (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

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	-- Extracting Keyphrases...
2	mining	1.1000	1	0	-- Reading instance
3	text mining	1.4555	1	1	-- Converting instance
4	refer	0.7500	3	0	mining 0.5171
5	data	0.7500	7	0	text mining 0.5171
6	text data	1.4555	7	1	data 0.5171
7	data mining	1.4666	8	1	text data 0.5171
8	text data mining	data mining 0.5171
9	equivalent	text 0.5171
10	Analytics	information 0.262
11	analytics 0.262

```

-- 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

```


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

anonymity	<i>atomicity</i>
atomicity	<i>auction</i>
auction	customer
electronic	<i>electronic</i>
commerce	commerce
privacy	intruder
real-time	merchant
security	protocol
transaction	<i>security</i>
	third party
	<i>transaction</i>

Neural multigrid for gauge theories and other disordered systems

disordered	disordered
systems	gauge
<i>gauge fields</i>	<i>gauge fields</i>
<i>multigrid</i>	interpolation kernels
neural multigrid	length scale
neural networks	<i>multigrid</i>
	smooth

Proof nets, garbage, and computations

<i>cut-elimination</i>	cut
linear logic	<i>cut elimination</i>
<i>proof nets</i>	garbage
sharing graphs	<i>proof net</i>
typed lambda-calculus	weakening

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