

# 데이터사이언스와 R

토픽모델링을 활용한  
최근 5년 간 국내 문헌정보학의 연구동향 분석

데이터사이언스와 R 1조  
2017314643 김현우  
2018312990 조명재  
2021313897 허지원  
2021314373 안세연  
2021314642 신민서

# Index

---

- 1 연구배경 및 연구목표 소개
- 2 연구 설계
- 3 데이터 분석 및 결과 해석
- 4 연구의 의의 및 한계





# Part 1.

## 연구배경 및 연구목표 소개

# 연구 배경



한국도서관·정보학회  
Korea Library and Information Science Society

韓國情報管理學會

Korea Society for Information Management

한국비블리아학회


Korean Biblia Society for  
Library and Information Science

한국문헌정보학회

KOREAN SOCIETY FOR LIBRARY AND INFORMATION SCIENCE


1970년, 한국문헌정보학회를 시작으로  
문헌정보학 분야 주요 학회들이 설립되며 활발한 연구가 진행

# 연구 배경

 한국학술지인용색인  
Korea Citation Index

KCI 소개   논문검색   학술지검색   기관정보검색   인용정보검색   통계정보

**KCI 통합검색**   통합검색   검색어를 입력해 주세요

 KCI 등재

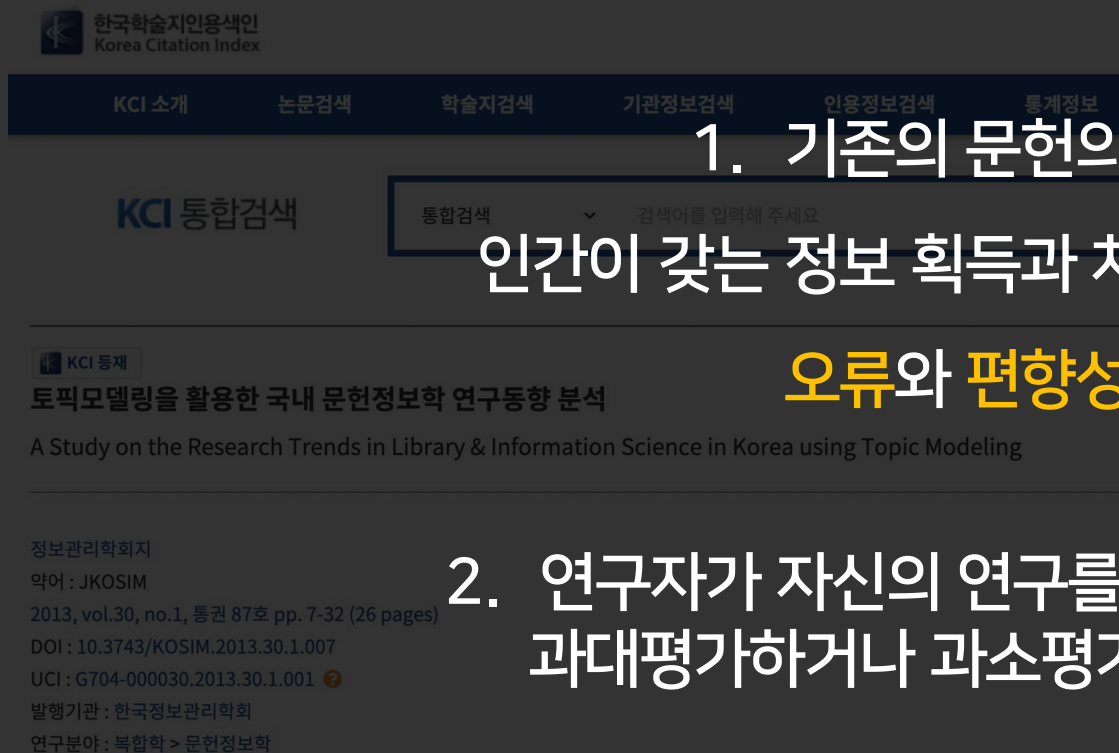
**토픽모델링을 활용한 국내 문헌정보학 연구동향 분석**  
A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling

정보관리학회지  
약어 : JKOSIM  
2013, vol.30, no.1, 통권 87호 pp. 7-32 (26 pages)  
DOI : 10.3743/KOSIM.2013.30.1.007  
UCI : G704-000030.2013.30.1.001  
발행기관 : 한국정보관리학회  
연구분야 : 복합학 > 문헌정보학



연구자들은 문헌정보의 관심분야가 어떻게 변화하는지  
다각적인 측면에서의 연구동향 분석을 시도

# 연구 배경



1. 기존의 문헌의 연구 방식은

인간이 갖는 정보 획득과 처리의 인지적 한계로 인해

오류와 편향성이 발생 가능

2. 연구자가 자신의 연구를 관련 연구로 추가할 때  
과대평가하거나 과소평가하는 편향성 또한 존재

3. 영문 초록을 활용하며 토픽 모델링을 한 연구 사례만이 존재

연구자들은 문헌정보의 관심분야가 어떻게 변화하는지

한글 초록에 대한 토픽 모델링 연구 미비

다각적인 측면에서의 연구동향 분석을 시도



# 선행연구

## 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석\*

A Study on the Research Trends in Library & Information Science  
in Korea using Topic Modeling

박자현 (Ja-Hyun Park)\*\*  
송민 (Min Song)\*\*\*

### 초록

본 연구는 국내 문헌정보학 분야의 연구동향을 규명하기 위하여 문헌정보학 주요 학술지인, 정보관리학회지, 한국문헌정보학회지, 한국도서관·정보학회지, 한국비블리아학회지의 1970년도부터 2012년도까지 발표 논문 초록을 수집하여 LDA(Latent Dirichlet Allocation)기반의 토픽 모델링 실험을 수행하였다. 그 결과를 종합하면 다음과 같다. 첫째, 토픽모델링 실험에서 도출된 연구주제를 문헌정보학 주제분류표와 비교·분석한 결과, '정보학'영역의 디지털도서관, 이용연구, 인터넷, 전문가시스템, 계량정보학, 자동화, 정보검색, 정보시스템, '도서관 서비스'영역의 정보서비스, 도서관 유형별 서비스, 이용자 교육/정보리터러시, 서비스 평가, '문헌정보학 기초'영역의 도서관과 사회, 전문성, '자료조직'영역의 분류, 편목, 메타데이터, '도서관 경영'영역의 도서관 평가, 장서개발/관리, '서지학'영역의 고서지, '도서관 체제'영역의 도서관 및 정보정책, '출판'영역의 도서/출판, '기록관리학'영역의 하위주제 등과 연결할 수 있었다. 또한 가장 많은 연구주제가 발견된 학문영역은 정보학과 도서관서비스로 나타났다. 둘째, 문헌정보학의 주요 연구주제에서 도서관 유형별 서비스 및 평가, 인터넷, 메타데이터의 연구주제는 상승세를 보였으나, 도서, 분류, 편목, 고서지에 관한 연구주제는 하강세를 보였다. 셋째, 학술지를 구분하여 비교·분석한 결과, 정보관리학회지는 도서관에 관한 연구주제보다 정보학에 관한 연구주제가 많이 출현하였고, 한국문헌정보학회지와 한국도서관·정보학회지, 한국비블리아학회지는 도서관에 관한 연구주제가 정보학에 관한 주제보다 많이 나타났다.

## 토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구\*

Topic Modeling based Interdisciplinarity Measurement  
in the Informatics Related Journals

진설아 (Seol A Jin)\*\*  
송민 (Min Song)\*\*\*

### 초록

본 연구는 인용 정보와 주제범주 분류체계를 기반으로 한 기존 하향식 접근법과 달리 문헌에 출현한 단어정보를 기반으로 세부주제를 자동 추출하는 토픽 모델링을 사용하여 학제성을 측정하였다. JCR 2013의 Information & Library Science 주제범주에서 5년 영향력 지수 상위 20개 학술지의 최근 5년 동안의 논문 제목과 초록 텍스트를 분석대상으로 사용하였다. 학제성을 측정하기 위한 지수로 '분야적 다양성'을 나타내는 Shannon 엔트로피 지수와 Stirling 다양성 지수, '네트워크 응집성'을 나타내는 지수로는 토픽 네트워크의 평균 경로길이를 사용하였다. 계산된 다양성과 응집성 지수를 통해 학제성의 유형을 분류한 후 각 유형을 대표하는 학술지들의 토픽 네트워크를 비교하였다. 이를 통해 본 연구의 텍스트 기반 다양성 지수는 기존의 인용정보 기반 다양성 지수와 다른 양상을 보이고 있어 상호보완적으로 활용될 수 있으며, 다양성과 응집성을 모두 고려하여 분류된 각 학술지의 토픽 네트워크를 통해 개별 학술지가 다루는 세부주제의 특성과 연결 정도를 직관적으로 파악할 수 있었다. 이를 통해 토픽 모델링을 통한 텍스트 기반의 학제성 측정이 학술지의 학제성을 나타내는 데에 다양한 역할이 가능함을 확인하였다.

## 텍스트마이닝 기법을 활용한

문헌정보학의 연구동향분석이 이미 많이 시행되고 있음

# 선행연구

## 토픽 모델링을 이용한 핀테크 기술 동향 분석

김태경, 최희련, 이홍철\*  
고려대학교 산업경영공학과

### A Study on the Research Trends in Fintech using Topic Modeling

TaeKyung Kim, HoeRyeon Choi, HongChul Lee\*  
Department of Industrial and Management Engineering, Korea University

**요 약** 최근 인터넷과 모바일 환경을 기반으로 금융과 IT가 융합된 핀테크(Fintech) 산업이 급속히 성장하고 있으며 간편성, 편리성 등으로 무장한 핀테크 서비스는 모든 금융서비스의 온라인·모바일 화를 주도하고 있다. 그러나 핀테크 산업의 급격한 성장에도 불구하고, 핀테크 기술에 대한 세부기술 분류와 주요 시장국의 기술개발 동향을 분석하고 기술기획을 지원하기 위한 연구는 매우 미흡한 실정이다. 이에 본 연구는 핀테크 기술의 비정형 데이터 형태의 특허 데이터를 이용하여 토픽모델링 기법을 통해, 핀테크 세부 기술을 추출하고 정의한다. 도출된 핀테크 세부 기술에 대해 Hot&Cold topic 을 파악하여 핀테크 기술의 트렌드를 파악한다. 또한 핀테크 산업의 주요 기술에 대한 주요 시장국인 미국, 한국, 중국의 기술개발 동향을 각각 분석한다. 마지막으로 핀테크 세부 기술 간 네트워크 분석을 통해 기술 간의 연계 관계를 살펴본다. 본 연구를 통해 파악된 핀테크 산업 기술 동향은 핀테크 산업분야의 정책 수립과 핀테크 관련 기업의 기술 전략 수립에 효과적으로 활용될 수 있을 것으로 기대된다.

## LDA 토픽모델링을 활용한 인공지능 관련 국가R&D 연구동향 분석

### A Study on Analysis of national R&D research trends for Artificial Intelligence using LDA topic modeling

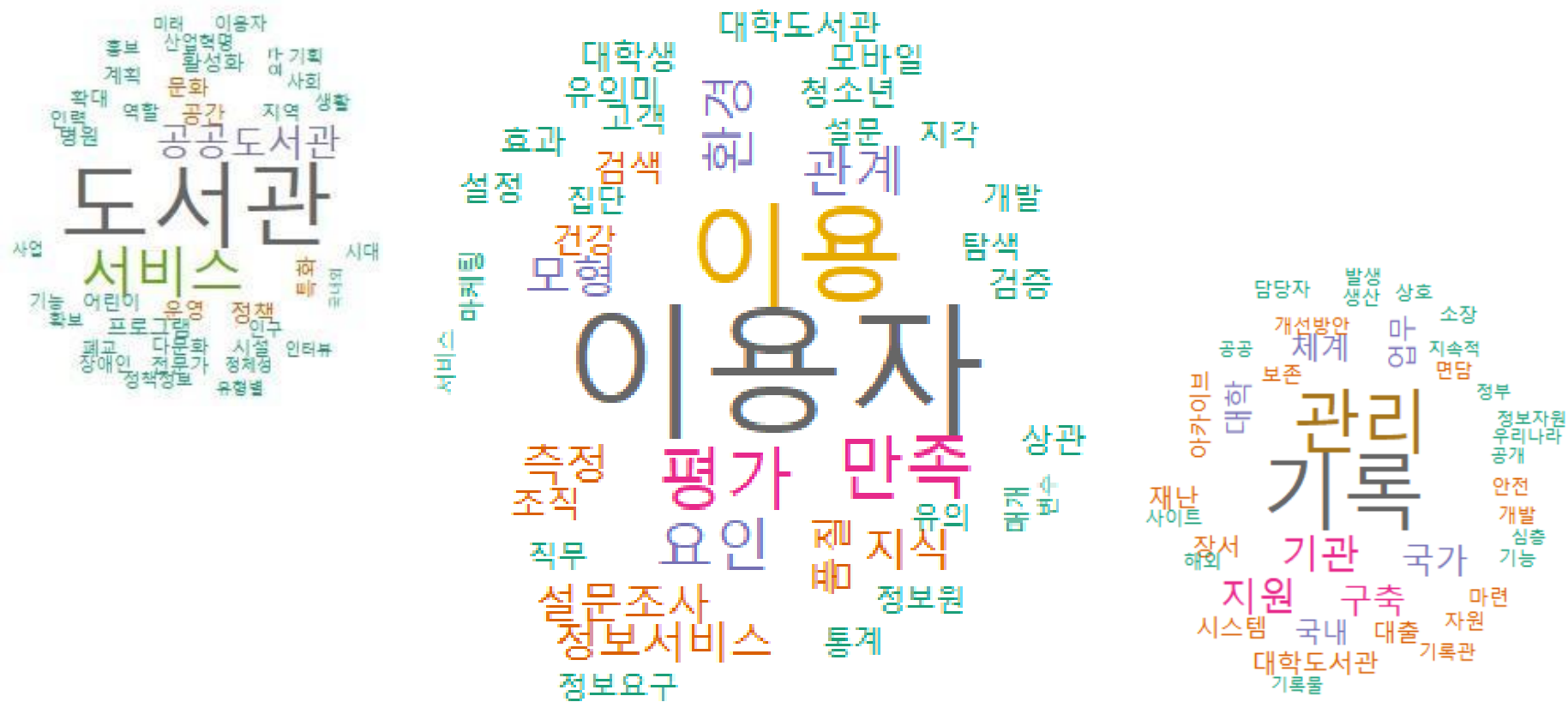
양 명 석<sup>1</sup> 이 성 회<sup>2</sup> 박 근 회<sup>2</sup> 최 광 남<sup>2</sup> 김 태 현<sup>2</sup>  
MyungSeok yang SungHee Lee KeunHee Park KwangNam Choi TaeHyun Kim

#### 요 약

특정 주제분야에 대한 연구동향 분석은 대부분 논문, 특허 등 문헌정보를 대상으로 한 키워드 추출을 통해 토픽모델링 기법을 적용하여 주요 연구주제와 연도별 추이 등을 살펴보는 방식을 활용하고 있다. 본 논문에서는 국가과학기술지식정보서비스(NIS)에서 제공하는 인공지능 관련 국가연구개발사업 과제정보를 대상으로 LDA(Latent Dirichlet Allocation) 토픽모델링 기법을 활용하여 연구주제와 관련된 토픽들을 추출·분석하여 국가연구개발사업에 대한 연구주제와 투자방향에 대하여 분석하고자 한다. NIS는 국가연구개발사업·과제정보를 비롯하여, 논문, 특허, 보고서 등 연구를 통해 생성된 주요 연구개발성과에 이르기까지 방대한 양의 국가R&D 정보를 제공하고 있다. 본 논문에서는 NIS 통합검색에서 인공지능 키워드와 관련된 분류 검색을 수행하여 검색결과를 확인하고, 최근 3개년 과제정보를 다운로드 받아 기초데이터를 구축하였다. 파이썬에서 제공하는 LDA 토픽모델링 라이브러리를 활용하여 기초데이터 (연구목표, 연구내용, 기대효과, 키워드 등)를 대상으로 관련 토픽과 주제어를 추출하고 분석하여 연구투자방향에 대한 인사이트를 도출하였다.

다른 분야에서도 LDA를 통해  
토픽모델링을 진행하여 유의미한 결론을 도출해 낸 것을 확인함





## -> 5년 간의 문헌정보학 분야의 연구 현황 분석

# 연구목표

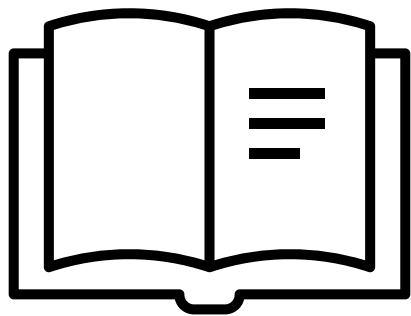
이때, 한글 초록데이터를 수집하여  
토픽 모델링을 진행함

향후 문헌정보학 연구 설계의 기초 자료를 생성하는 것을  
목표로 하는 “연구 동향 분석”

텍스트 마이닝 기법 LDA를 활용한 토픽 모델링

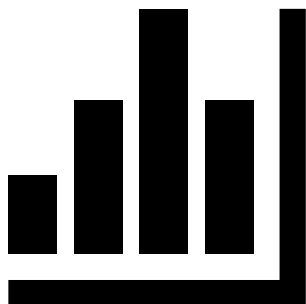
-> 5년 간의 문헌정보학 분야의 연구 현황 분석

# 연구목표



## 연구문제1

최근 5개년 간 문헌정보학 분야에서 활발하게 연구되는  
최적의 연구 주제 개수와 그 연구 주제가 무엇인가?



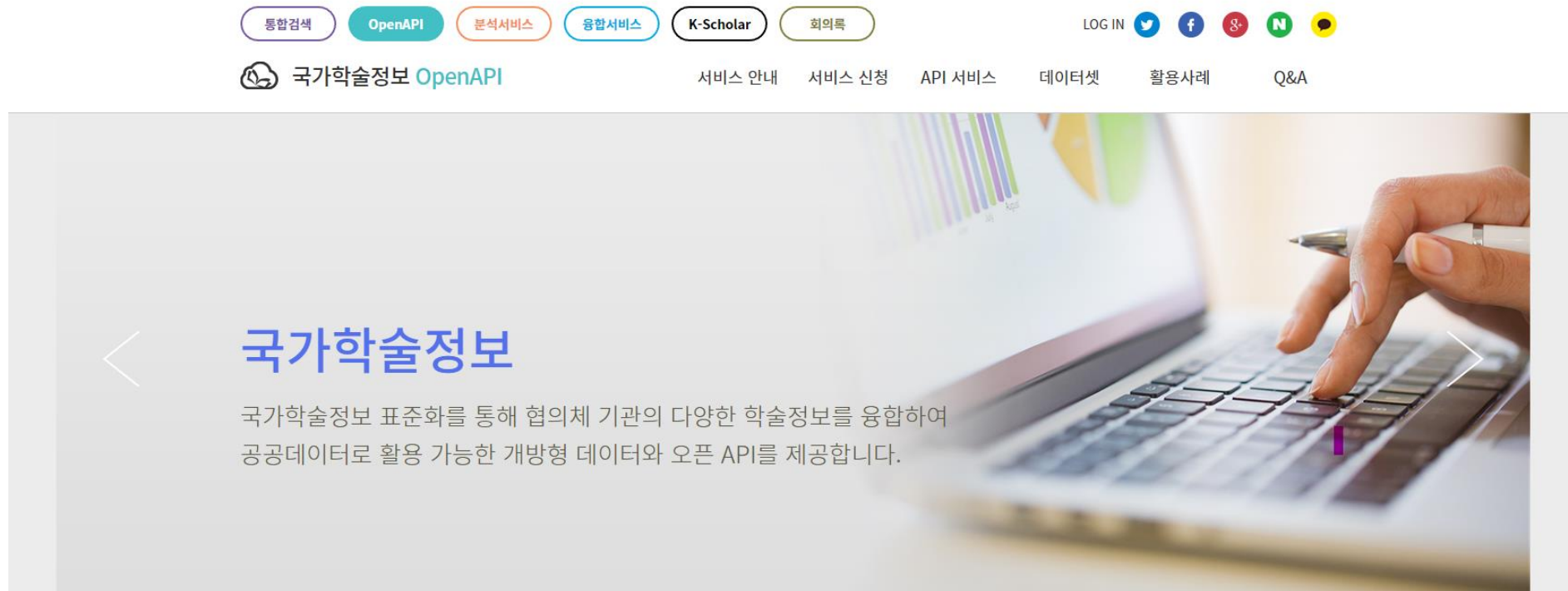
## 연구문제2

2017~2020년 동안 문헌정보학 분야에서  
연구되고 있는 주제들의 비율은 어떻게 변화하였는가?

## Part 2.

### 연구설계

# 논문 초록데이터 수집



국회도서관에서 제공하는  
국가학술정보 OpenAPI 활용

# 논문 초록데이터 수집

## | 검색

[홈](#) > [API서비스](#) > 검색

- 오픈 API 활용 신청을 하시면 현재 로그인한 아이디에 사용 권한이 부여됩니다.
- 서비스 제공 기간은 오픈 API 활용 신청이 승인된 날로부터 1년이며, 1년에 한해 자동 갱신됩니다. 자세한 사항은 이용 약관을 참조하시기 바랍니다.
- 오픈 API 활용 신청을 위해서는 먼저 인증키를 발급받으셔야 합니다. [서비스 신청] → [인증키 발급] 메뉴에서 인증키를 발급받으실 수 있습니다.

전체 4 건 | 페이지 1 / 1

API명	등록일	활용 현황	제공 방식
연구자 정보 보기	2017-11-29	10	JSON
주제어 정보 보기	2017-11-29	16	JSON
상세보기	2017-09-04	20	JSON
통합검색	2017-09-04	40	JSON

연구자 정보, 주제어 정보, 논문 통합검색, 논문 상세정보  
POST 방식으로 데이터 호출, JSON 형식으로 반환



# 논문 초록데이터 수집

## 통합검색 API

문헌정보학 관련 학술지 이름을  
검색 쿼리로 설정  
작성기간이 2017~2021년인  
[LOD 아이디, 논문 제목, 발행년도] 수집

## 상세보기 API

통합 검색을 통해 수집한  
학술지 별 논문 데이터를 바탕으로 실행  
[논문 제목, 발행년도, 논문 초록정보] 수집

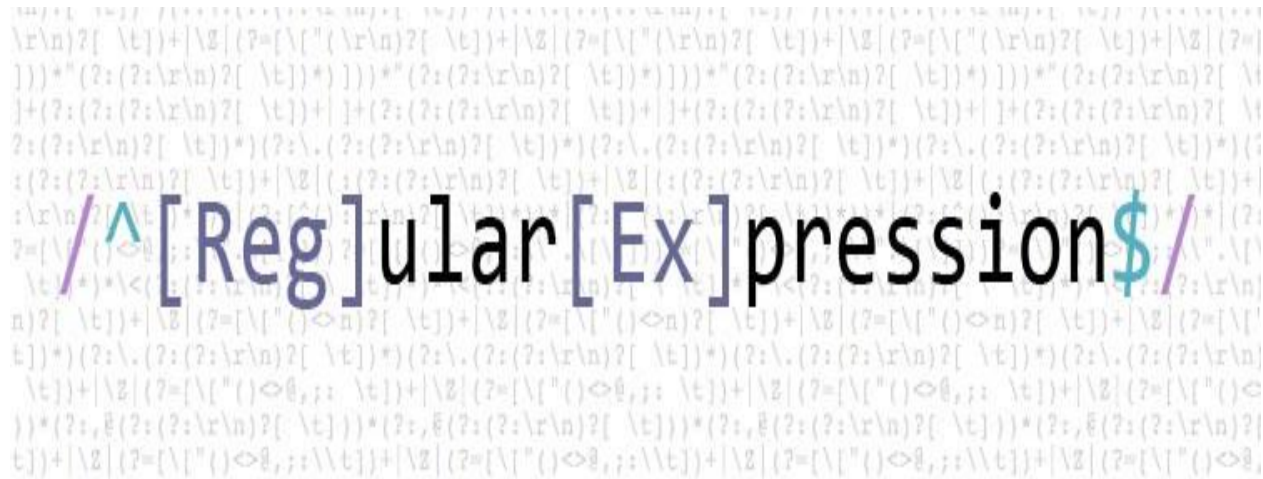
국가학술정보 OpenAPI의 통합검색 API와 상세보기 API를 활용하여  
각 학술지별 논문 제목과 발행년도, 초록정보를 수집

# 논문 초록데이터 수집

학술지명	수집한 논문수	초록 결측치 처리 후 논문수	비고
한국문헌정보학회지	320	273	총 47건 제외 (API 오류로 인한 누락 1건, 초록 미등재 46건) 2021년 발행된 논문의 초록데이터가 누락됨
정보관리학회지	225	224	총 1건 제외 (API 오류로 인한 누락 1건)
한국비블리아학회지	293	242	총 51건 제외 (초록 미등재 51건)
한국도서관·정보학회지	-	-	API에서 검색이 되지 않아 분석 대상에서 제외

초록이 정상적으로 수집 안됐거나, API상에 초록 미등재된 경우  
분석 대상에서 제외하여 총 838건 중 739건을 분석대상으로 선정함

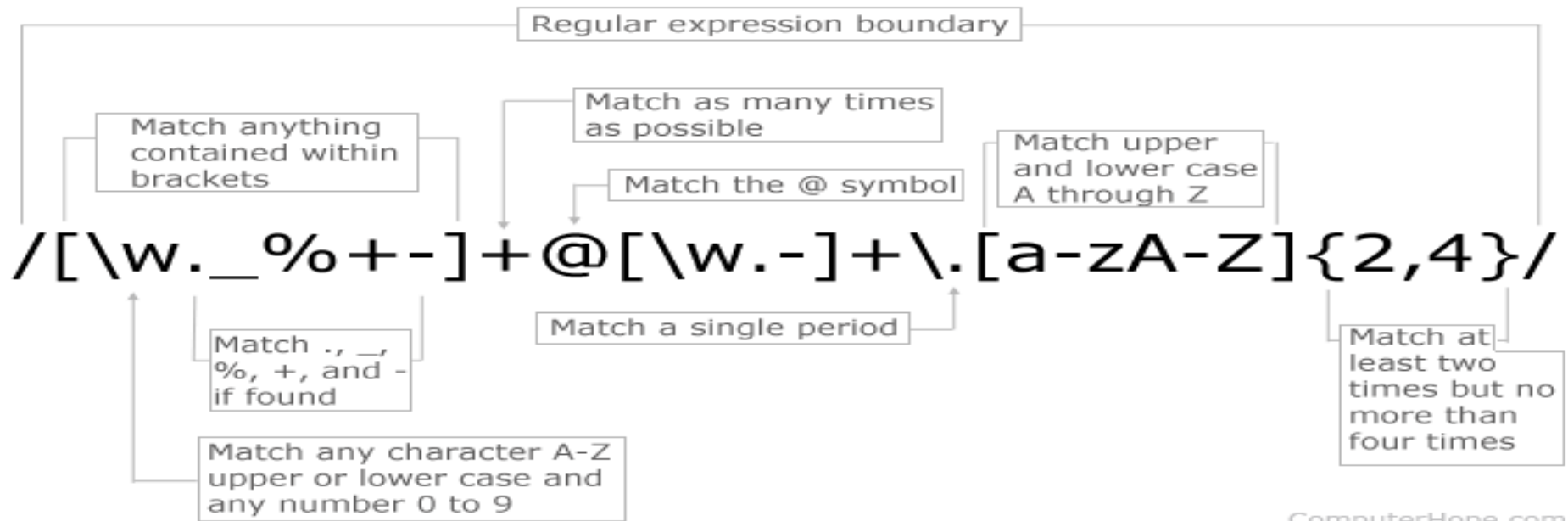
# 논문 초록데이터 전처리

  
/<sup>^</sup>[Reg]ular[Ex]pression\$/

수집한 논문 초록 데이터에서  
한글 초록만을 추출하기 위하여 정규표현식을 사용

# 논문 초록데이터 전처리

## Regular Expression E-mail Matching Example



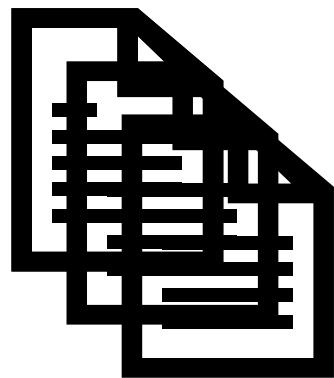
정규 표현식을 통해 텍스트에서 특정 패턴을 가진 문자들을 추출할 수 있고,  
이를 통해 초록 데이터에서 한글 초록데이터만을 추출할 수 있음.

# 논문 초록데이터 수집

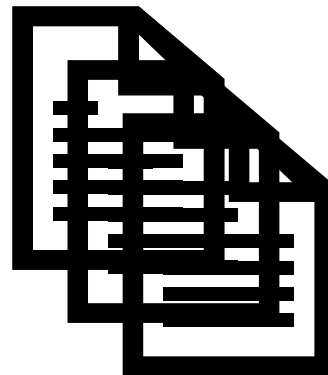
정규표현식	의미	사용처
[^[가-힣]]+\$	텍스트 맨 뒤에서 한글이 아닌 모든 단어들의 모음을 찾기	한글 초록 뒤에 있는 영문 초록 제거
^[^[가-힣]]+	텍스트 맨 앞에서 한글이 아닌 모든 단어들의 모음을 찾기	한글 초록 앞에 있는 영문 초록 제거
[一-龠]+	한자로 되어있는 모든 단어를 텍스트 내에서 찾기	초록 데이터에 존재하는 한문단어 제외

표와 같이 정규표현식을 사용하여 초록 데이터에서 영문 초록 데이터를 제거하고,  
유의하지 않을 것으로 판단되는 한문데이터를 초록 데이터에서 제거

# 논문 초록데이터 전처리



총 논문 초록 데이터  
739건

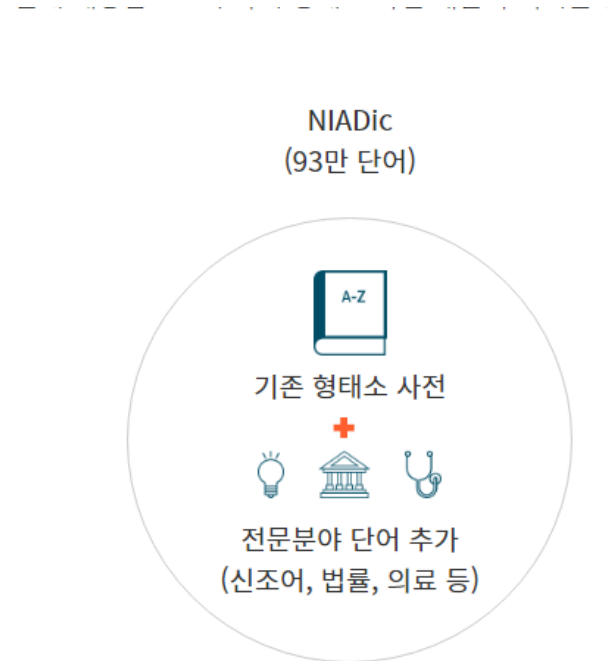
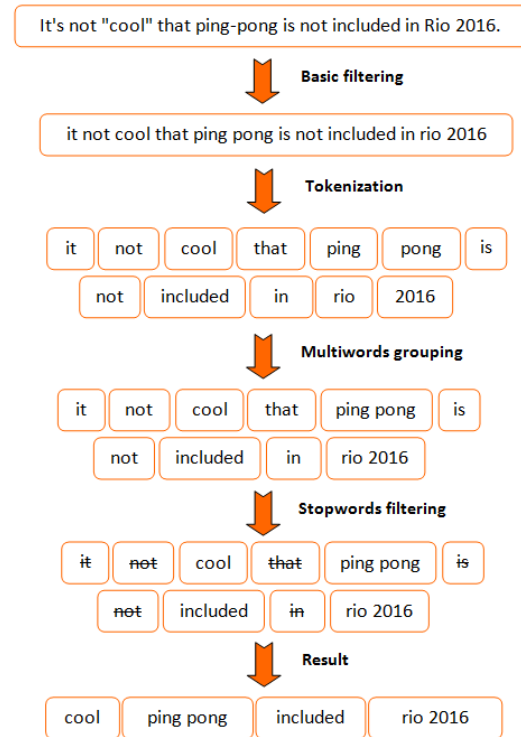


한글 초록 데이터  
611건

이 과정을 통해 분석 대상이 총 논문 초록데이터 739건에서  
한글 초록만 가진 데이터 611건으로 변경함



# 논문 초록데이터 전처리



한글 초록 데이터에 대한 토큰화를 진행하여 초록을 형태소 단위로 분해  
이때, 형태소를 분류하는 기준으로 NIADic을 사용

# 논문 초록데이터 전처리

## 분해 전 초록 데이터

실용성을 강조하는 여타  
학문 분야처럼 문헌정보학  
은 실무자인 사서의 ...

이 연구는 국립중앙도서관  
의 국가선거공동활용시스  
템에 참여하여 ...

본 연구는 청소년 메이크  
업에 관한 유튜브 동영상  
의 주요 시청자가 ...

...

extractNoun()



## 분해 후 초록 데이터

실용성 강조 하 여타 학문  
분야 문헌정보학 실무자  
사 연구 활동 학술논문 ...

연구 국립중앙도서관 국가  
선거 공동 활용 시스템 참  
여 로컬 도서관 선거 ...

국가 운영 정책 기획 실행  
평가 하기 영역 의사 결정  
단계 정책연구 정보 ...

...

KoNLP의 extractNoun()을 통해 초록 데이터에서 명사만을 추출한 후,  
추후 문서-단어 행렬 형성을 위해서 띄어쓰기로 연결하여 저장

# 논문 초록데이터 전처리

이때, 각 논문 별 키워드를 NIADic에 추가하여  
KoNLP의 형태소 분해의 성능을 높이하고자 하였으나  
Java 문제로 인해 키워드를 사전에 추가할 수 없었음



추후 말뭉치에서 제대로 분리되지 않은 단어들을  
직접 제거해주는 것으로 이를 보완함

KoNLP의 extractNoun()을 통해 초록 데이터에서 명사만을 추출한 후,  
추후 문서-단어 행렬 형성을 위해서 띄어쓰기로 연결하여 저장

# 논문 초록데이터 전처리

## 분해 후 초록 데이터

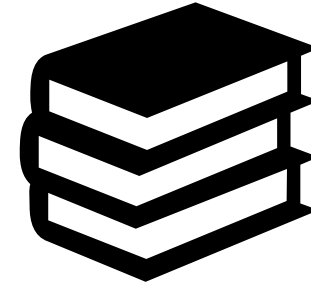
실용성 강조 하 여타 학문  
분야 문헌정보학 실무자  
사 연구 활동 학술논문 ...

연구 국립중앙도서관 국가  
전거 공동 활용 시스템 참  
여 로컬 도서관 전거 ...

국가 운영 정책 기획 실행  
평가 하기 영역 의사 결정  
단계 정책연구 정보 ...

...

Vcorpus()



말뭉치  
(Corpus)

형태소 분해가 완료된 초록 데이터를 Vcorpus()를 통해  
말뭉치(Corpus)의 형태로 변환함

# 논문 초록데이터 전처리

함수명	기능	비고
removePunctuation	특수문자 제거	!, " 등의 특수문자 제거
removeNumbers	숫자 제거	-
tolower	영문 단어 소문자 변환	영문 단어를 통일시키는 것으로 불용어인 영단어를 삭제하는데 도움을 줌
removeWords	불용어 사전에 저장된 단어들을 말뭉치 내에서 제거	불용어 사전에 저장된 단어들만 제거

말뭉치에 다음과 같은 함수들을 사용하여  
각 초록 데이터에서 유의하지 않은 단어들이나 표현들을 제거함

# 논문 초록데이터 전처리

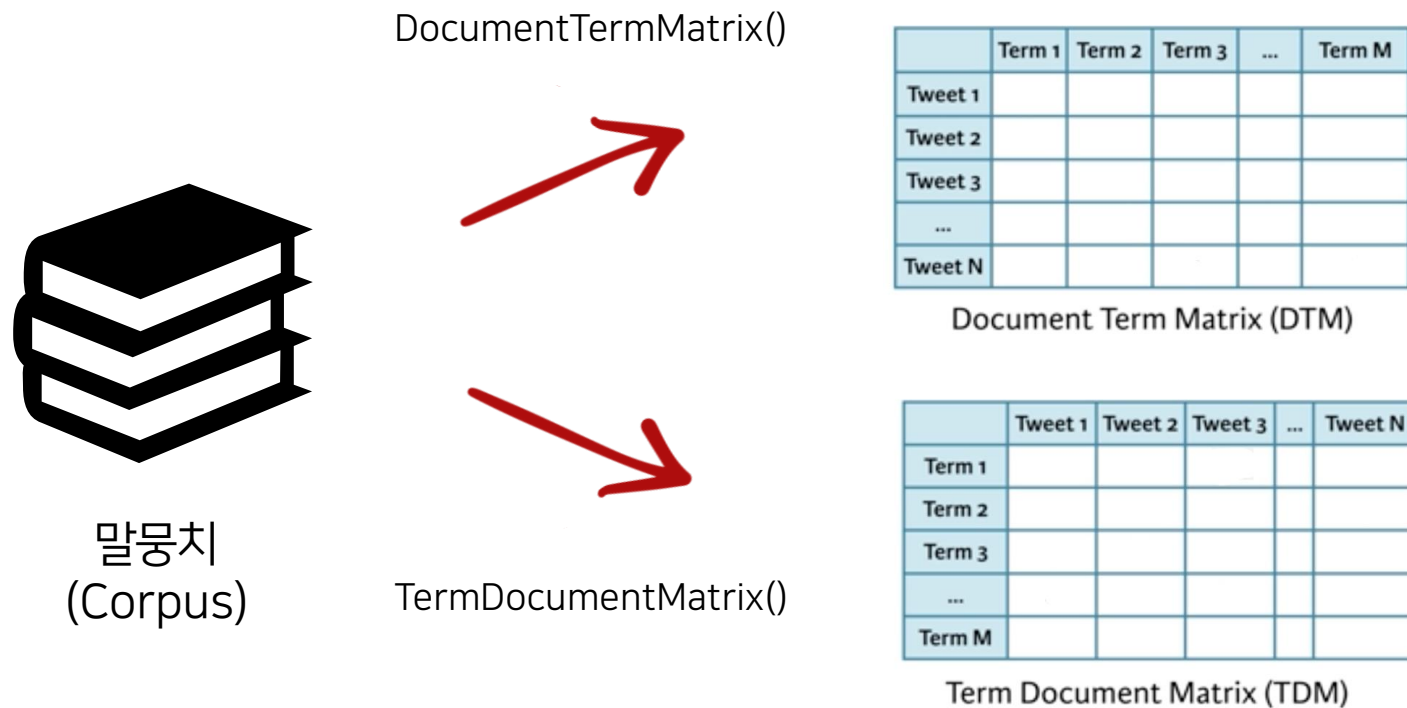
기능명	기능	비고
	특수문자 제거	!, " 등의 특수문자 제거
	숫자 제거	-
	영문 단어 소문자 변환	영문 단어를 통일시키는 것으로 불용어인 영단어를 삭제하는데 도움을 줌
removeWords	불용어 사전에 저장된 단어들을 말뭉치 내에서 제거	불용어 사전에 저장된 단어들만 제거

유형	불용어 예시
많이 나타나지만 토픽에서 주요한 의미를 가지지 않는 단어	연구, 대표, 주요, 현황 등
특수문자가 완전히 제거되지 않은 단어	정보', '진로, '범죄데이터' 등
영어단어	Lda, academy, library
형태소 분해가 제대로 이루어지지 않은 단어	제안하였, 강조하였, 파악하였

이때 제거한 불용어를 유형화하면 다음과 같고,  
이 중 영어표현은 분석 기법 내지는 한글과 병기한 표현이므로 제거해주었음



# 논문 초록데이터 전처리



전처리가 완료된 말뭉치를 문서-단어 행렬(Document-Term Matrix)와  
단어-문서 행렬(Term-Document Matrix)로 변환하여 분석을 진행

# 분석 방법 소개 – LDA(잠재 디렉클레 할당)

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

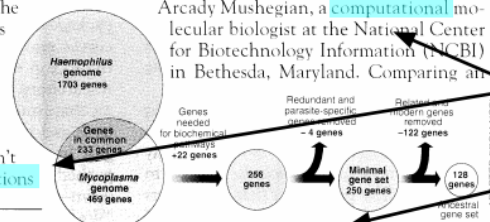
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

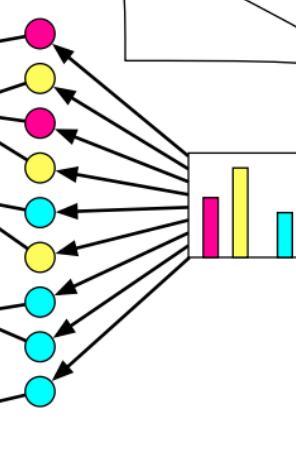


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

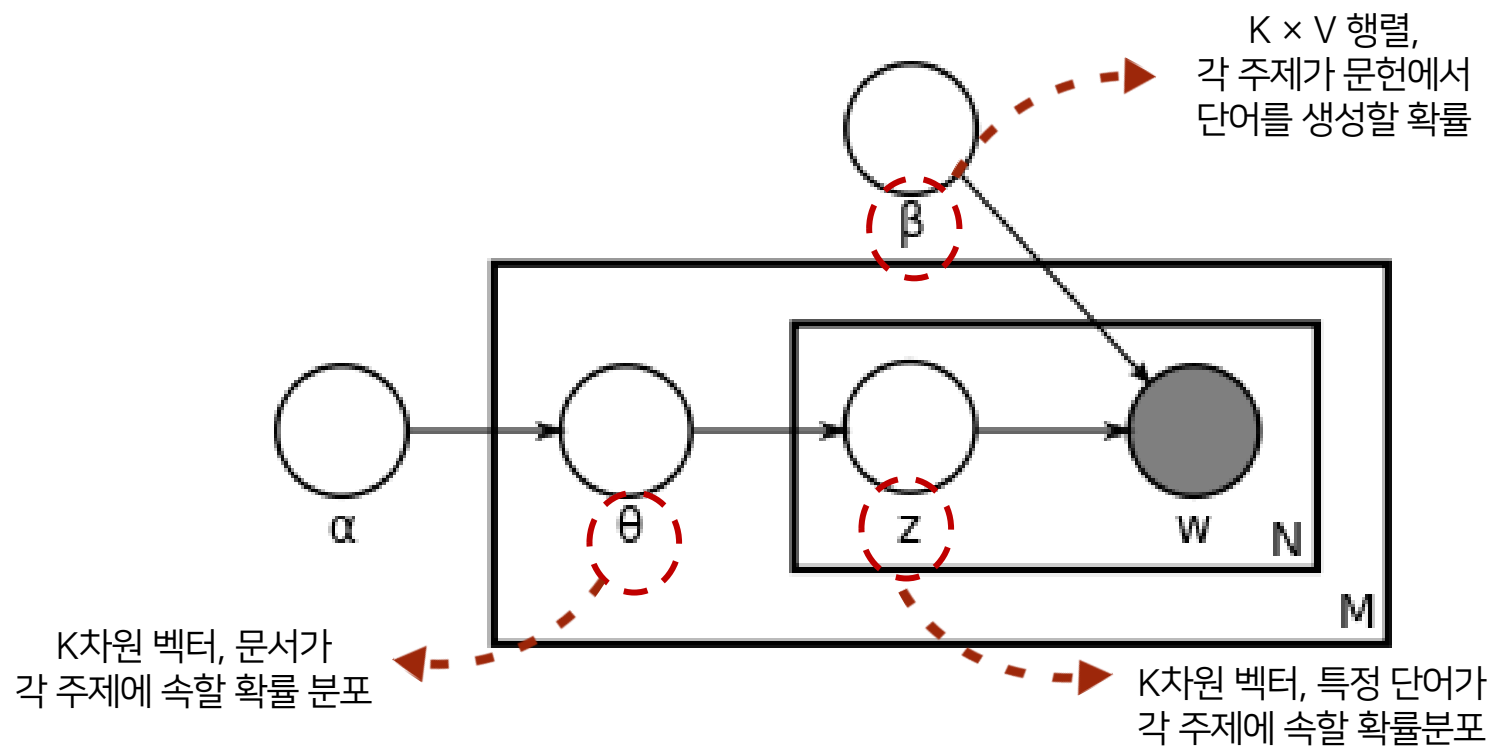
## Topic proportions & assignments



LDA는 베이지 기반의 머신러닝 모델이며,

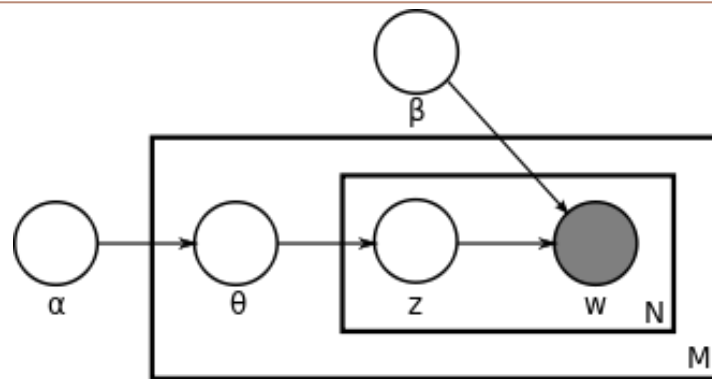
주제별 단어 분포를 활용해 각 문서별 단어 분포 분석 및 주제를 예측함

# 분석 방법 소개 - LDA(잠재 디렉클레 할당)



LDA를 통한 문서의 생성 과정을 그래프로 도식화 하면 다음과 같음

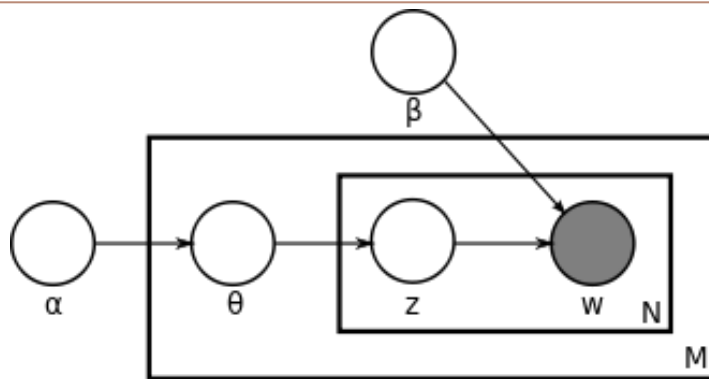
# 분석 방법 소개 – LDA(잠재 디레클레 할당)



## LDA를 통한 문서의 생성 과정

- 1  $N \sim \text{Poisson}(\xi)$ 을 선택한다.
- 2  $\theta \sim \text{Dir}(\alpha)$ 를 선택한다.  
 $\alpha$  = 디레클레 분포의 매개변수
- 3 문서 내의 단어  $w_n$ 에 대해서
  - $Z_n \sim \text{Multinomial}(\theta)$ 를 선택한다.
  - $Z_n$ 이 주어졌을 때  $w_n$ 은  $P(w_n|Z_n, \beta)$ 로부터 선택한다.

# 분석 방법 소개 - LDA(잠재 디렉클레 할당)



## LDA를 통한 문서의 생성 과정



$N \sim \text{Poisson}(\xi)$ 을 선택한다 실제 관측 가능한  $w_n$ 을 통해



$\theta \sim \text{Dir}(\alpha)$ 를 선택한다.

$\alpha$  = 디렉클레 분포의 매개변수



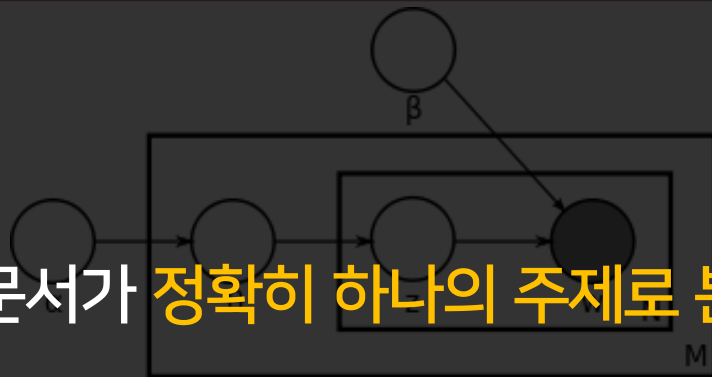
문서 내의 단어  $w_n$ 에 대해서

활발히 연구중인 주제를 확인하고, 각 문서의 주제를 확인할 수 있음

-  $z_n \sim \text{Multinomial}(\theta)$ 을 선택한다.

-  $z_n$ 이 주어졌을 때  $w_n$ 은  $P(w_n|z_n, \beta)$ 로부터 선택한다.

# 분석 방법 소개 - LDA(잠재 디렉클레 할당)



LDA의 특성상, 문서가 **정확히 하나의 주제로 분류되는 것이 아닌,**  
**여러 주제들에 대한 가중치의 형태로 혼합되어서 나타남**

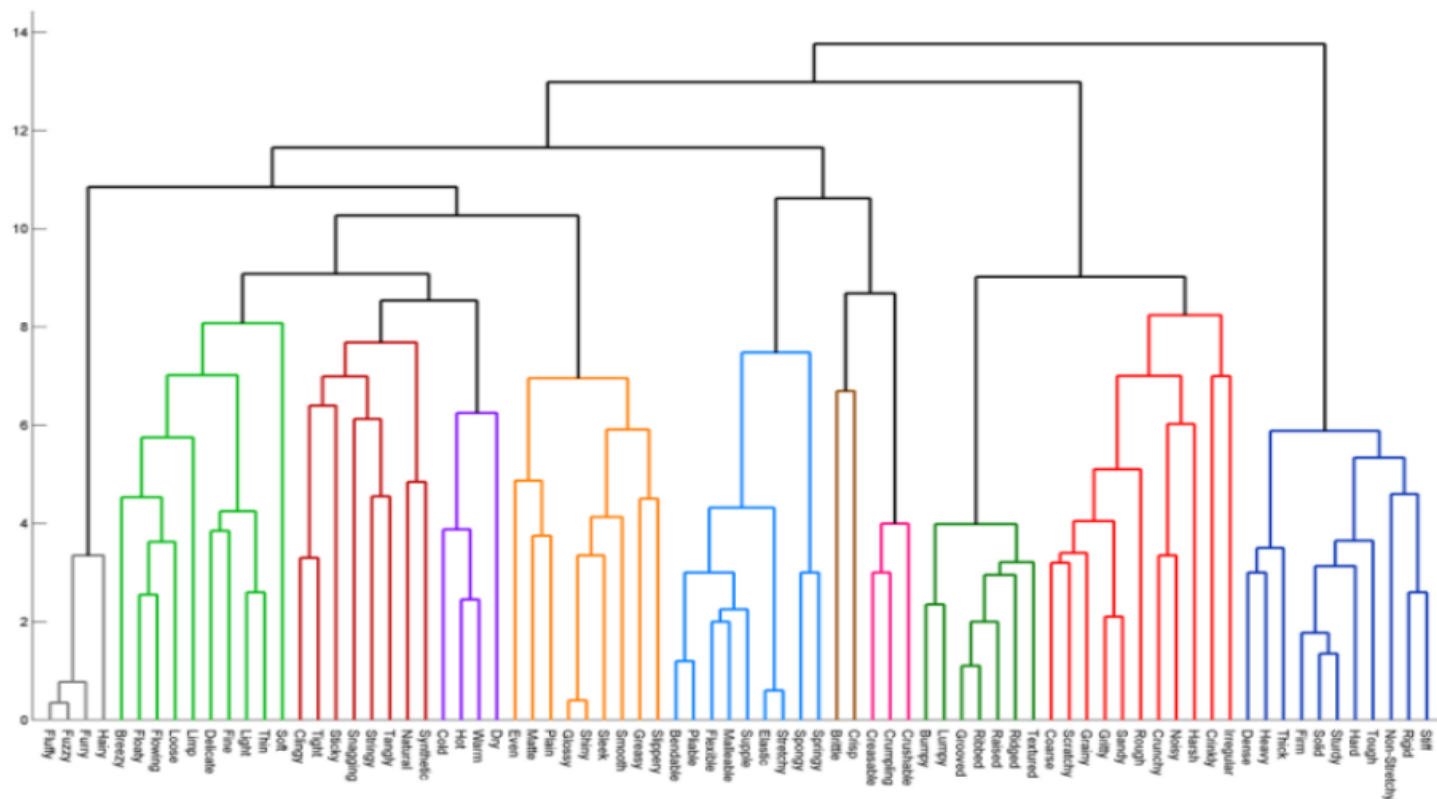


**여러 주제들이 복합적으로 나타날 것으로 예상되는**  
문헌정보학의 논문에서 나타나는 주제에 대한  
**토픽 모델링에 매우 효과적일 것으로 판단하여 선정**

활발히 연구중인 주제를 확인하고, 각 문서의 주제를 확인할 수 있음  
 $Z_n \sim \text{Multinomial}(\theta)$ 을 선택한다.  
 $Z_n$ 이 주어졌을 때  $w_n$ 은  $P(w_n|Z_n, \beta)$ 로부터 선택한다.



# 분석 방법 소개 - 계층적 클러스터링

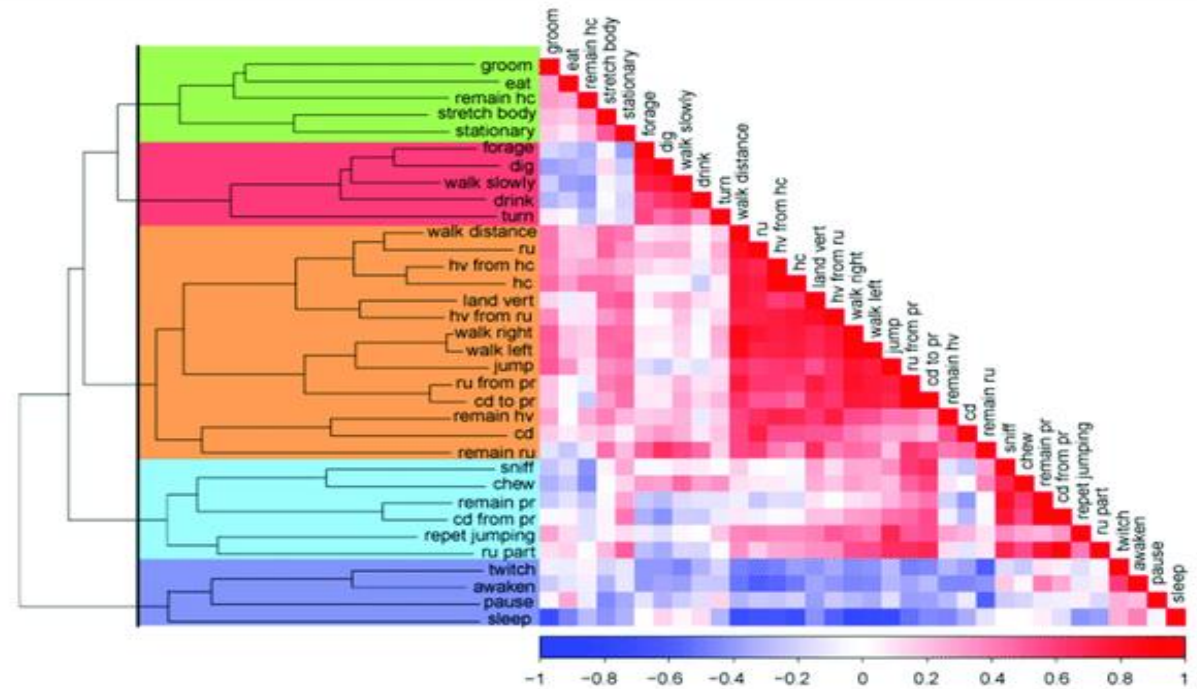


계층적 클러스터링은 계층적인 트리 모델을 사용하며,  
개별 개체들을 순차적, 계층적으로 유사한 개체 또는 그룹과 통합하는 알고리즘

# 분석 방법 소개 - 계층적 클러스터링

$$\text{Euclidean distance} = ||P - Q|| = \sqrt{(P - Q) \cdot (P - Q)}$$

$$\text{where } P = (p_1, \dots, p_n), \quad Q = (q_1, \dots, q_n)$$



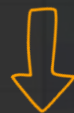
계층적 클러스터링을 위해서는 개체간 거리 혹은 유사도가 구해져야 하며,  
거리 혹은 유사도를 통해 거리 행렬을 만든 후 Ward 연결법으로 묶어 덴드로그램을 형성함.

# 분석 방법 소개 - 계층적 클러스터링

단순히 문서 전체에서 나타난

단어의 출현 빈도를 분석하는 것이 아닌,

단어의 문서 별 출현 빈도를 계층적인 구조로 파악



LDA 토픽 모델링 결과와 비교하는 것으로

단어의 구조와 주제들의 잠재적인 관계를 분석할 예정

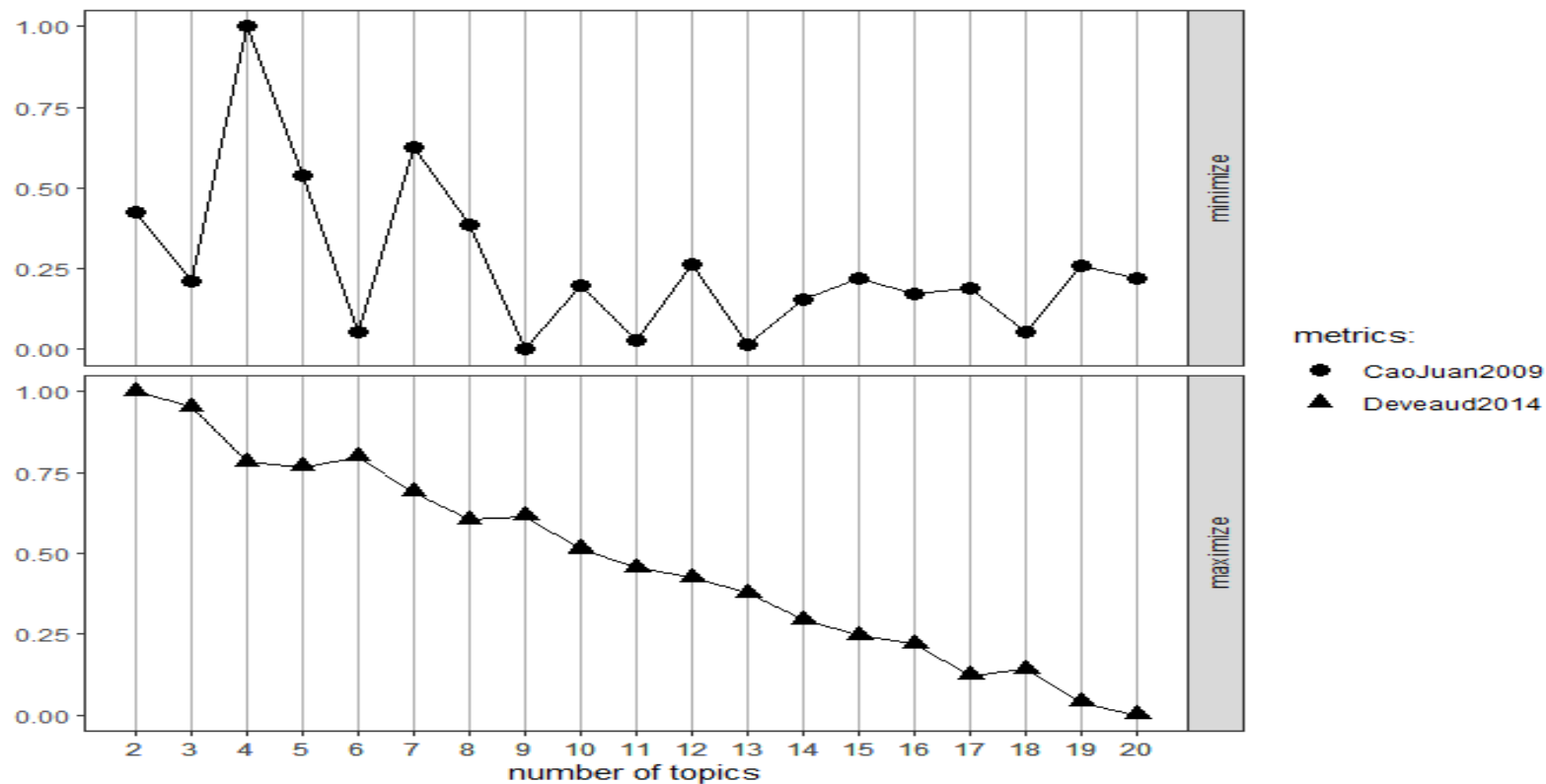
계층적 클러스터링을 위해서는 개체간 거리 혹은 유사도가 구해져야 하며,  
거리 혹은 유사도를 통해 거리 행렬을 만든 후 Ward 연결법으로 묶어 덴드로그램을 형성함.

# 데이터 분석 및 결과 해석

# 데이터 분석 및 결과 해석

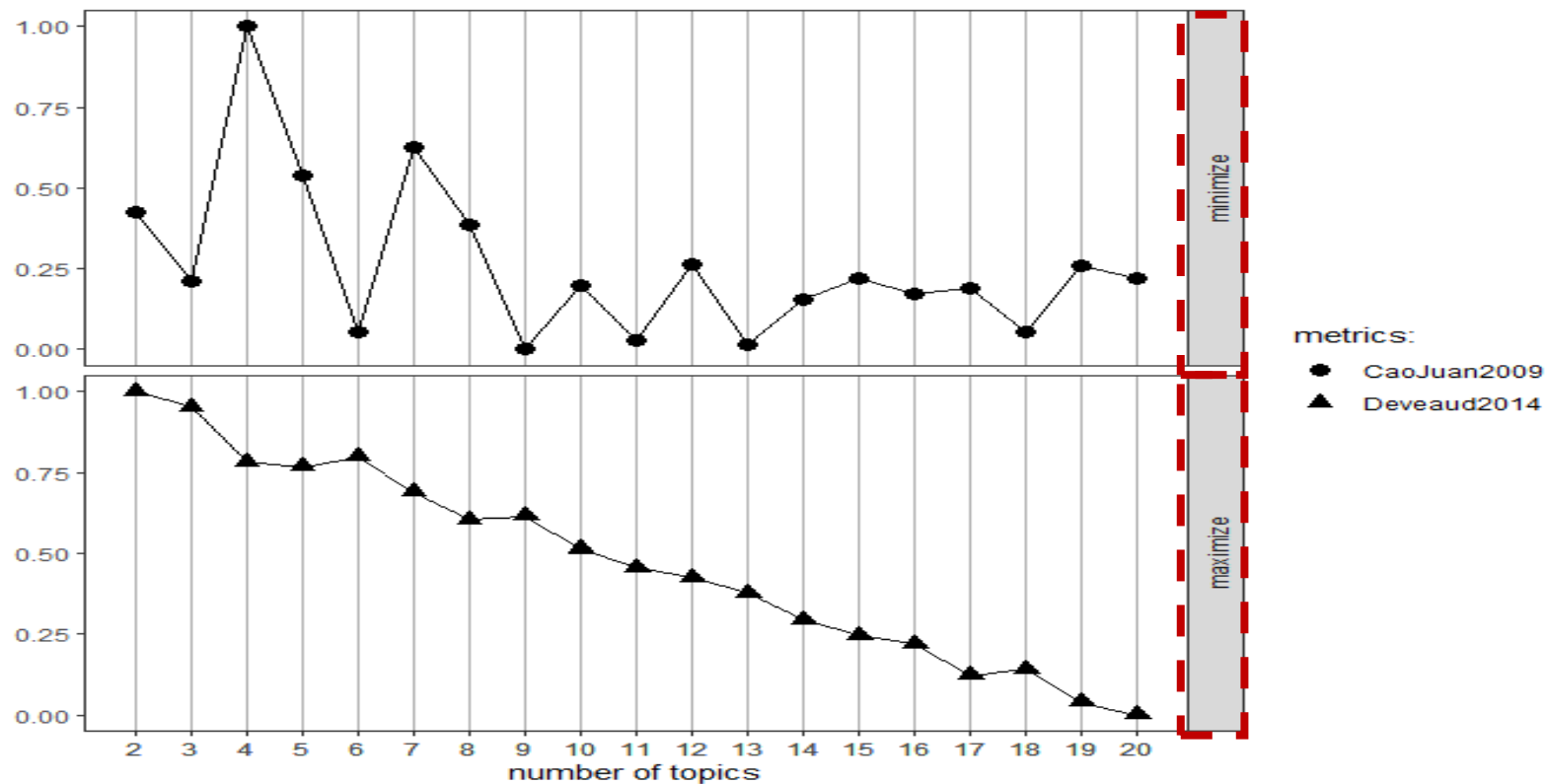


# 최적의 주제 수 K 파라미터 튜닝



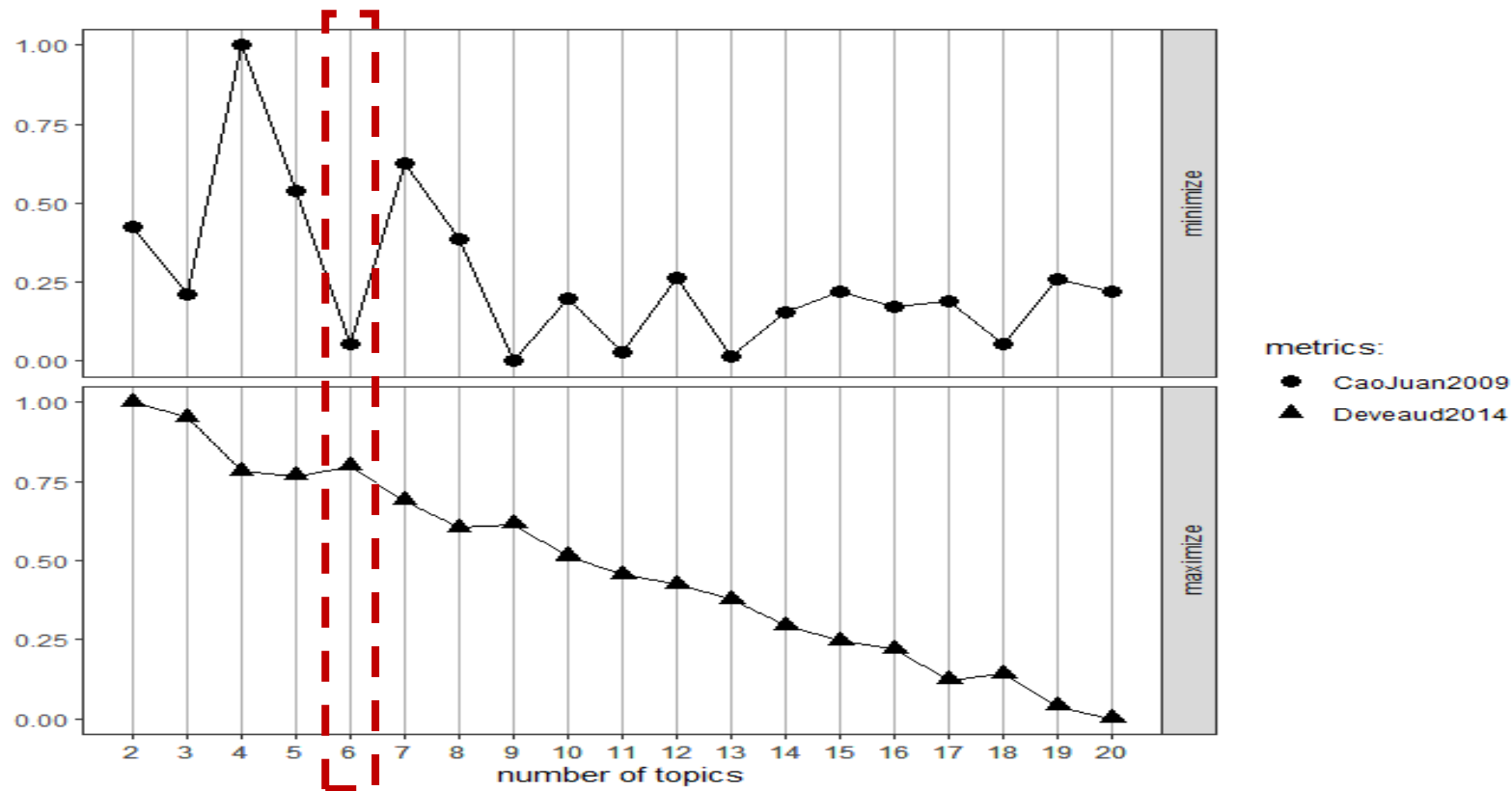
LDA를 진행하기 전 최적의 주제 수 K를 구하기 위하여  
ldatuning 패키지를 활용하여 파라미터 튜닝을 진행함

# 최적의 주제 수 K 파라미터 튜닝



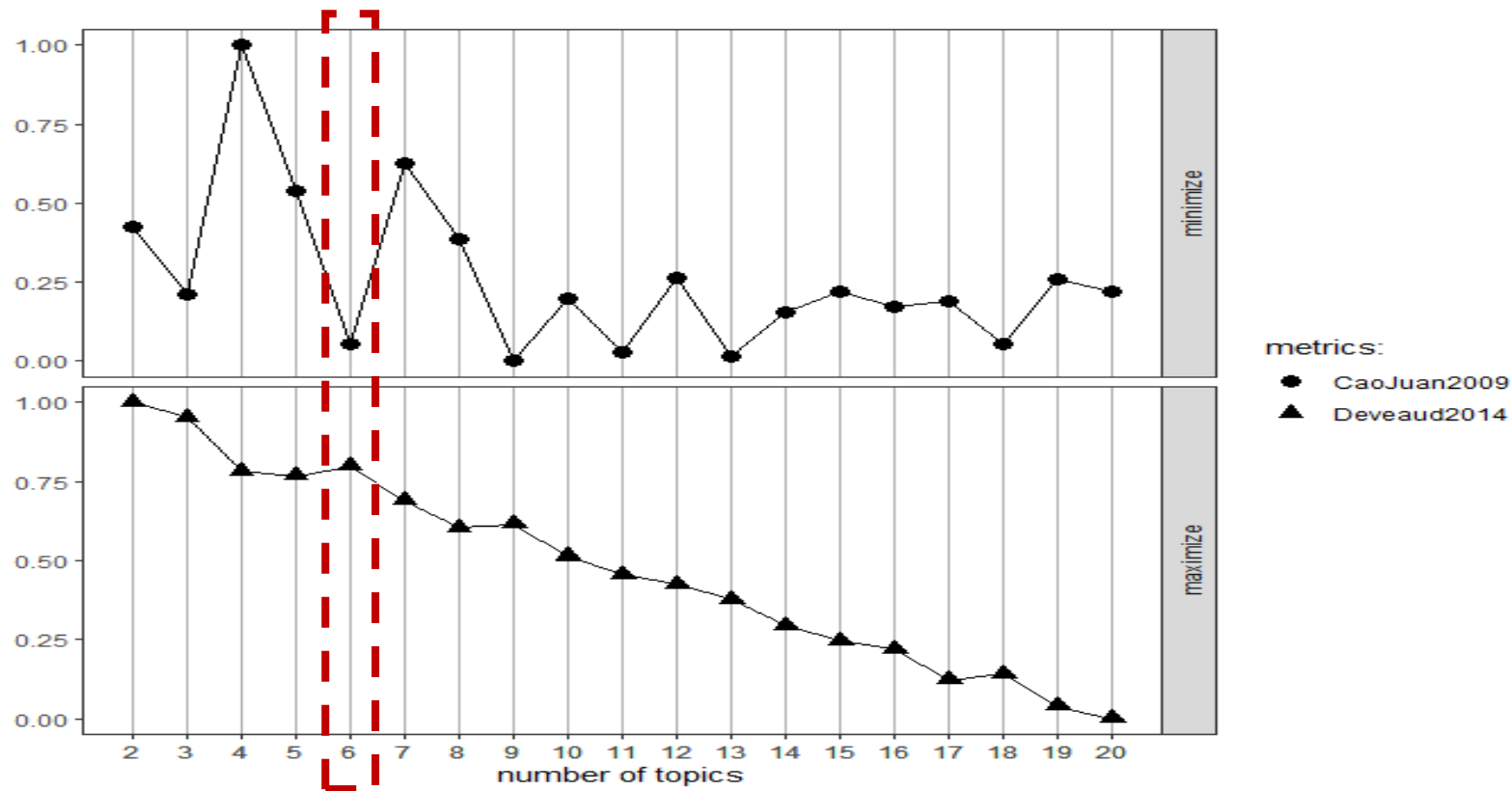
모델의 성능 지표로 CaoJuan2009와 Deveaud2014가 있고,  
해당 지표는 각각 최소화, 최대화되어야 하는 값임

# 최적의 주제 수 K 파라미터 튜닝



Plot을 통해 확인해 본 결과, K=6일때 각 지표들이  
최적화되어 나타나는 모습을 확인 할 수 있었음.

# 최적의 주제 수 K 파라미터 튜닝



따라서, 최적의 주제 수  $K=6$ 임을 확인하여  
이후 LDA를 진행 할 때 주제 수  $K$ 를 6으로 설정하여 분석을 진행



# LDA 토픽 모델링 및 결과 해석

주제 1	주제 2	주제 3	주제 4	주제 5	주제6
도서관	이용자	기록	교육	정보	데이터
서비스	이용	관리	인식	자료	논문
공공도서관	평가	기관	프로그램	주제	국내
공간	만족	지원	학교도서관	유형	학술지
운영	요인	구축	독서	디지털	모델
정책	관계	국가	개발	기술	인용
특화	환경	업무	학생	선정	분류
문화	모형	대학	사서	사회	주제
지역	측정	체계	사서교사	콘텐츠	문헌
역할	설문조사	국내	운영	한국	네트워크

파라미터 튜닝 결과를 바탕으로 K=6, 반복수 1000으로 LDA를 실시한 결과  
6개의 주제에서 주요하게 나타나는 10개의 단어는 다음과 같음

# LDA 토픽 모델링 및 결과 해석

주제 1	주제 2	주제 3	주제 4	주제 5	주제6
도서관	이용자	기록	교육	정보	데이터
서비스	이용	관리	인식	자료	논문
공공도서관	평가	기관	프로그램	주제	국내
공간	만족	지원	학교도서관	유형	학술지
운영	요인	구축	독서	디지털	모델
정책	관계	국가	개발	기술	인용
특화	환경	업무	학생	선정	분류
문화	모형	대학	사서	사회	주제
지역	측정	체계	사서교사	콘텐츠	문헌
역할	설문조사	국내	운영	한국	네트워크

〈표 2〉 문헌정보학 주제분류표

영역	하위주제	세부주제	영역	하위주제	세부주제	영역	하위주제	세부주제
문헌정보학 기초	도서관 역사	도서관과 사회	도서관 경영(계속)	경영관리	조직관리	정보학(계속)	디지털도서관	이론연구
법령/기준	도서관 기준	도서관법/지아법	의사결정	조직 일반	조직 커뮤니케이션	이용연구 일반	이용자 인식/요구	정보이용행태
연구	비교 문헌정보학	연구동향	홍보/마케팅	도서관 서비스	정보서비스 일반	인터넷 일반	웹사이트 설계/평가	웹사이트 설계/구축
이론 및 철학	문헌정보학 이론	철학/사상	전문성 일반	도서관 유형별 서비스	공공/국가도서관	전문가시스템	전문가시스템 일반	전문가시스템 평가
전문성	교육	사서직	독서교육/지도	독서교육/지도	독서교육/지도	정보정책	정책 기법/전략	정책시스템
출판/유통/지적자유	출판/유통	지적자유	서비스 평가	서비스 평가	서비스 평가	정책연구	정책연구	정책연구
전문성 일반	전문성 일반	전문성 일반	열람/대출 봉사	열람/대출 봉사	열람/대출 봉사	시스템/정책 효율성 평가	시스템/정책 효율성 평가	시스템/정책 효율성 평가
도서관 건물 및 설비	도서관 건물	도서관 설비	자료조직	자료조직	자료조직	정보정책 일반	정보정책 일반	정보정책 일반
도서관 체계	도서관 체계	도서관 체계	분류	분류	분류	소프트웨어	소프트웨어	소프트웨어
도서관 및 정보정책	도서관 및 정보정책	도서관 및 정보정책	도서관 사정/실태조사	도서관 사정/실태조사	도서관 사정/실태조사	인공지능	인공지능	인공지능
도서관 사정/실태조사	도서관 사정/실태조사	도서관 사정/실태조사	도서관 사정 일반	도서관 사정 일반	도서관 사정 일반	인공지능 일반	인공지능 일반	인공지능 일반
도서관 사정 일반	도서관 사정 일반	도서관 사정 일반	인공지능 일반	인공지능 일반	인공지능 일반	정보기술일반	정보기술일반	정보기술일반
도서관 통계/연구	도서관 통계/연구	도서관 통계/연구	사서교육	사서교육	사서교육	정보일반	정보일반	정보일반
실태 조사	실태 조사	실태 조사	주제분석	주제분석	주제분석	패턴/문자 인식	패턴/문자 인식	패턴/문자 인식
도서관 협동/자원공유	도서관 협동/자원공유	도서관 협동/자원공유	색인/목록	색인/목록	색인/목록	워드웨어	워드웨어	워드웨어
도서관 경영	도서관 경영	도서관 경영	시소러스	시소러스	시소러스	정보유통	정보유통	정보유통
경영관리	경영관리	경영관리	주제명표표	주제명표표	주제명표표	정보이론	정보이론	정보이론
경영 기법/전략	경영 기법/전략	경영 기법/전략	주제분석 일반	주제분석 일반	주제분석 일반	정보시스템	정보시스템	정보시스템
도서관 경영/지식경영	도서관 경영/지식경영	도서관 경영/지식경영	메타데이터	메타데이터	메타데이터	데이터 구조/설계	데이터 구조/설계	데이터 구조/설계
도서관 기획/확성화	도서관 기획/확성화	도서관 기획/확성화	관측	관측	관측	데이터베이스 일반	데이터베이스 일반	데이터베이스 일반
도서관 평가	도서관 평가	도서관 평가	MARC	MARC	MARC	데이터베이스 평가	데이터베이스 평가	데이터베이스 평가
인사관리	인사관리	인사관리	고사관록	고사관록	고사관록	사지데이터베이스	사지데이터베이스	사지데이터베이스
자료의 유형	자료의 유형	자료의 유형	목록규칙	목록규칙	목록규칙	원문데이터베이스	원문데이터베이스	원문데이터베이스
전자저널	전자저널	전자저널	온라인명록을 목록(OPAC)	온라인명록을 목록(OPAC)	온라인명록을 목록(OPAC)	이미지데이터베이스	이미지데이터베이스	이미지데이터베이스
장부간행물	장부간행물	장부간행물	웹 자원조직	웹 자원조직	웹 자원조직	하이퍼텍스트	하이퍼텍스트	하이퍼텍스트
학위논문	학위논문	학위논문	전자통제	전자통제	전자통제	컴퓨터네트워크	컴퓨터네트워크	컴퓨터네트워크
연속간행물	연속간행물	연속간행물	관측일반	관측일반	관측일반	사지학	사지학	사지학
웹 자원	웹 자원	웹 자원	정보학	정보학	정보학	사지학 일반	사지학 일반	사지학 일반
자료유형 일반	자료유형 일반	자료유형 일반	계량정보학	계량정보학	계량정보학	고사학	고사학	고사학
장서개발/관리	장서개발/관리	장서개발/관리	계량정보학 일반	계량정보학 일반	계량정보학 일반	체계사지학(목록학)	체계사지학(목록학)	체계사지학(목록학)
보존/세분/수리	보존/세분/수리	보존/세분/수리	언론분석/인용색인	언론분석/인용색인	언론분석/인용색인	형태사지학(관문학)	형태사지학(관문학)	형태사지학(관문학)
서고관리	서고관리	서고관리	학술커뮤니케이션	학술커뮤니케이션	학술커뮤니케이션			
수사/등록/교환/납본	수사/등록/교환/납본	수사/등록/교환/납본	자료화	자료화	자료화	출판	출판	출판
자료의 유형별 관리	자료의 유형별 관리	자료의 유형별 관리	자동화 일반	자동화 일반	자동화 일반	전자출판	전자출판	전자출판
장서개발/정책	장서개발/정책	장서개발/정책	도서관업무별 자동화	도서관업무별 자동화	도서관업무별 자동화	도서/출판	도서/출판	도서/출판
장서관리 일반	장서관리 일반	장서관리 일반	자동화관리/분류	자동화관리/분류	자동화관리/분류			
장서집결	장서집결	장서집결	자동색인/목록	자동색인/목록	자동색인/목록	기록관리학	기록관리학	기록관리학
장서평가	장서평가	장서평가	자동화효율성 평가	자동화효율성 평가	자동화효율성 평가			
폐기	폐기	폐기						

주제에서 주요하게 나타나는 단어들과 문헌정보학 주제분류표를 대조하는 것으로, 각 주제들이 문헌정보학의 어떤 분야에 대한 연구인지 확인

# LDA 토픽 모델링 및 결과 해석 - 주제 1



주제 1은 도서관 체계, 도서관 경영에 해당하는 주제이고,  
도서관의 지속과 발전을 위한 연구가 주를 이루고 있음

# LDA 토픽 모델링 및 결과 해석 - 주제 1



고전적인 지역 도서관 현황 관리 및 체계 연구부터,  
빅데이터 활용 도서관 정책 시행 현황 분석 연구까지 매우 다양하게 연구되고 있음.

# LDA 토픽 모델링 및 결과 해석 - 주제 2



주제 2는 정보서비스 이용과 관련한 연구에 해당하는 주제이고,  
이용자의 만족도 평가, 품질에 대한 평가 등의 연구가 주를 이루고 있음

# LDA 토픽 모델링 및 결과 해석 - 주제 2



또한, 해당 주제는 도서관 이용, 학술정보 서비스 등  
다양한 정보서비스에 대한 연구 포함하고 있음

# LDA 토픽 모델링 및 결과 해석 - 주제 3



주제 3는 기록관리학과 관련된 주제이며,  
기록관리학은 인간활동의 증거와 기억으로서의 기록을 연구하는 학문임

# LDA 토픽 모델링 및 결과 해석 - 주제 3



해당 주제에서는 기록을 적절하게 관리·생산하고 가치가 있는 기록을 보존하여  
활용할 수 있도록 체계를 구축을 위한 연구가 진행되고 있음



# LDA 토픽 모델링 및 결과 해석 - 주제 4



주제 4는 도서관 및 정보 서비스에 대한 연구에 해당하는 주제이고,  
주로 도서관 유형별 서비스에 관한 연구와 서비스 평가 연구가 활발하게 연구되고 있음

# LDA 토픽 모델링 및 결과 해석 - 주제 4



또한, 주제 4의 도서관 서비스 연구는  
도서관 이용자의 교육과 정보 리터러시에 대한 연구도 포함하고 있음

# LDA 토픽 모델링 및 결과 해석 - 주제 5



주제 5는 자료조직 분야의 분류에 대한 연구와 디지털 도서관, 서지학에 대한 연구들이 복합적으로 나타나는 주제인 것으로 확인되었음

# LDA 토픽 모델링 및 결과 해석 - 주제 5



이는 최근 서지학 연구들이 디지털화 되는 과정에서 자료분류 연구나 도서관의 디지털화에 대한 연구들의 중간다리 역할을 했기 때문으로 예상

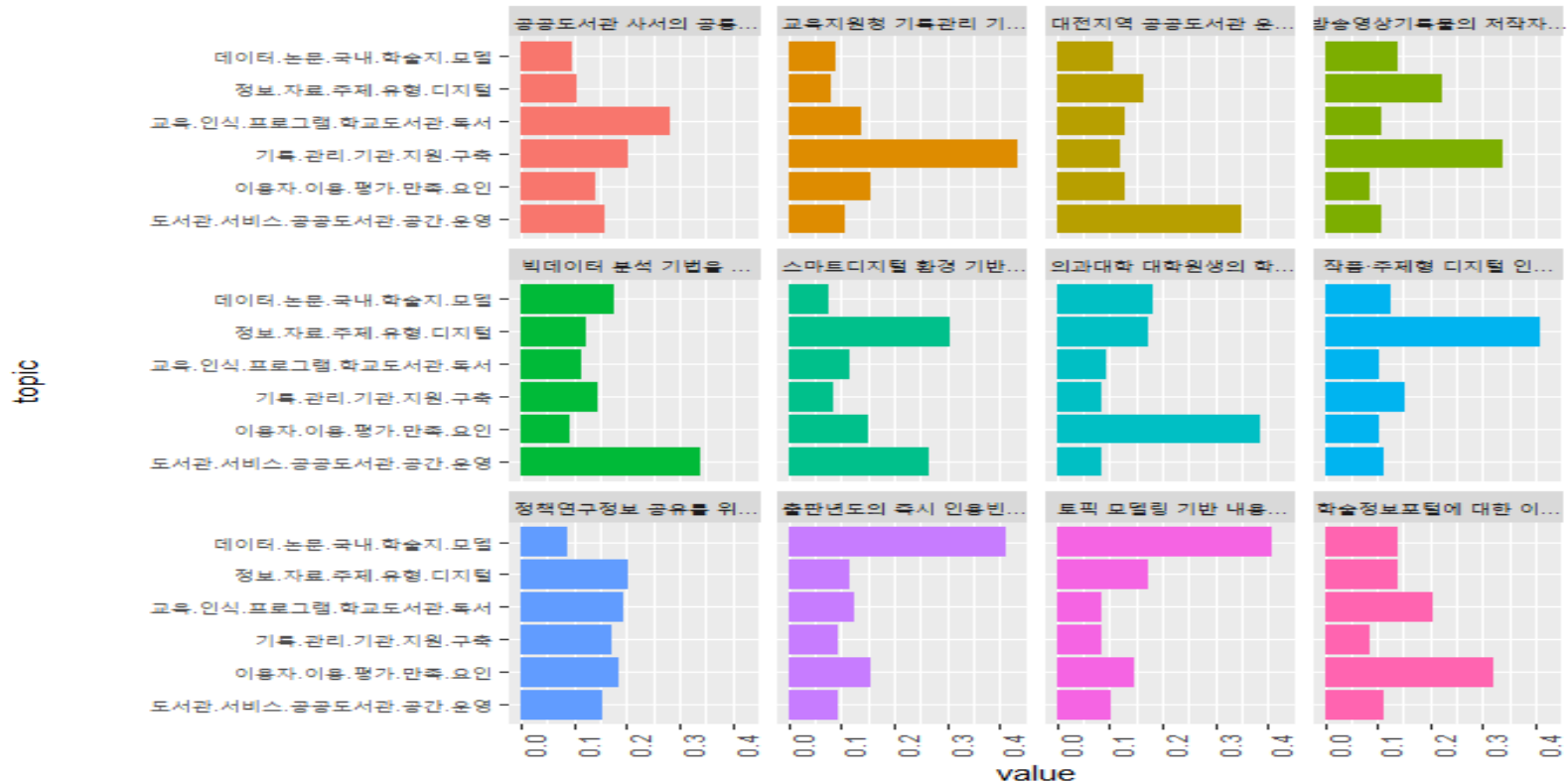
# LDA 토픽 모델링 및 결과 해석 - 주제 6



주제 6은 자료조직분야의 주제분석 연구와 메타데이터의 연구에 해당하는 주제이고, 본 연구와 같은 메타 연구의 경우에도 주제 6에 속하는 경향을 보임



# $\theta$ 를 통한 문서 별 주제 가중치 확인



논문들 중 몇 개를 뽑아 각 문서 별 주제 가중치를 확인해본 결과,  
대체적으로 한 주제에 대한 가중치가 높게 나오는 것을 확인 할 수 있었음

# θ를 통한 문서 별 주제 가중치 확인

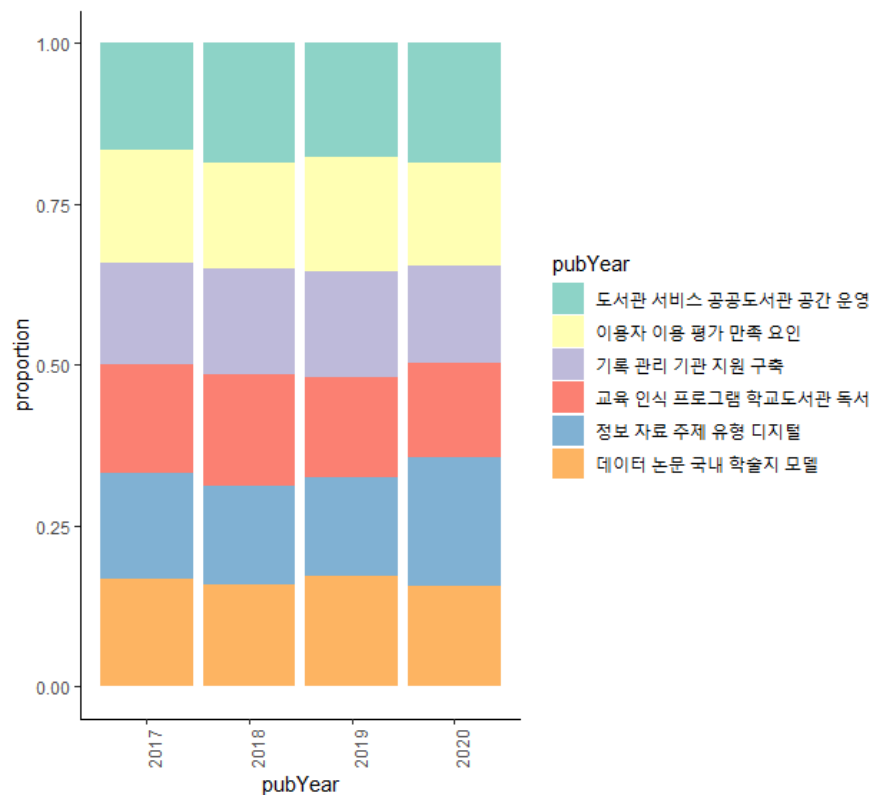


몇몇 논문의 경우에는 주제에 대한 가중치가 고르게 나타나는 경우도 존재

→ 이는 문헌정보학이 가지는 융복합적인 성격으로 인한 것으로 예상함.



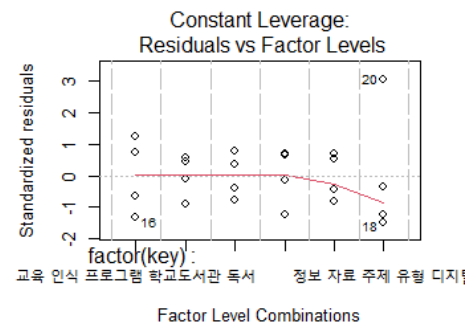
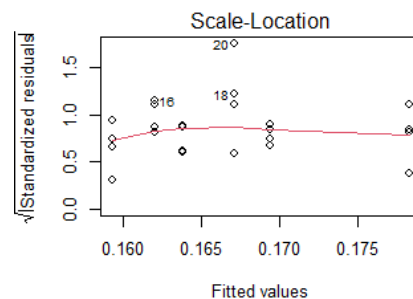
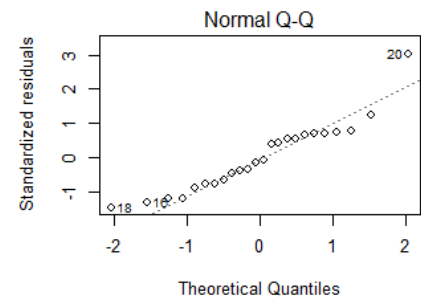
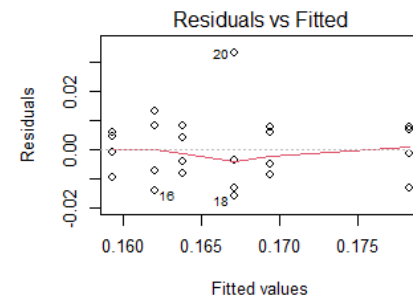
# 연도별 문헌정보학 내 주제 변화 분석



연도별로 전체 문서들에서 나타난 주제들의 가중치의 평균을  
막대그래프로 시각화하면 다음과 같음

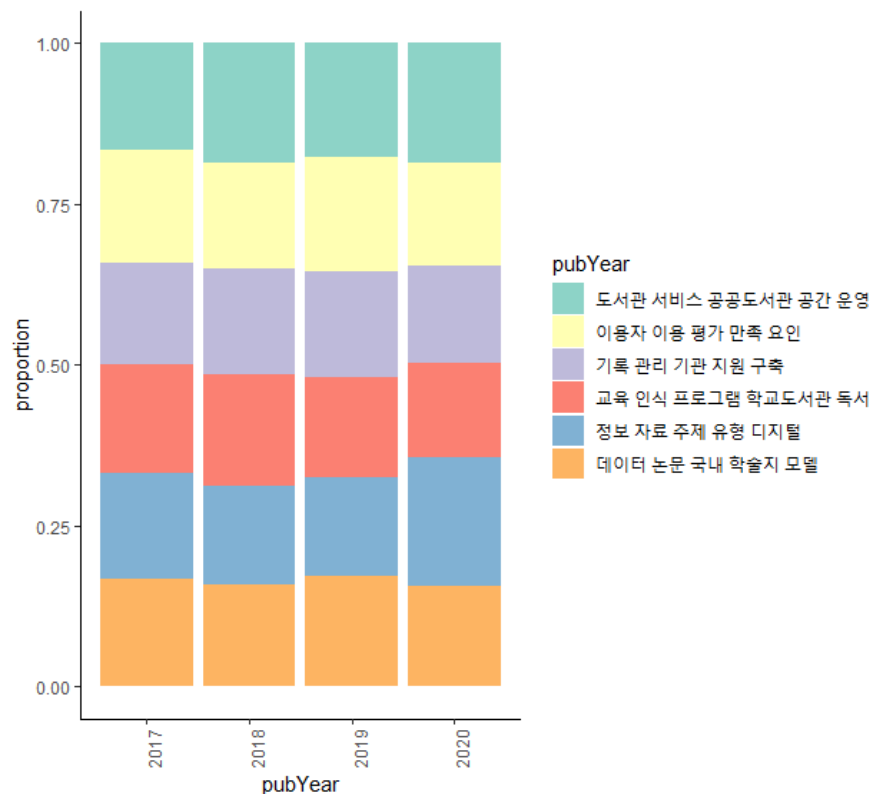
# 연도별 문헌정보학 내 주제 변화 분석

	DF	SS	MS	F	Pr(>F)
주제	5	0.000910	0.000182	1.173	0.36
Error	18	0.002792	0.000155	3.218380	0.000393
Total	23	0.003702			



각 주제별로 주제 가중치의 평균에 차이가 없는 것 같아 ANOVA를 진행하였고,  
주제간의 평균 주제 가중치 차이가 없는 것으로 나타났음

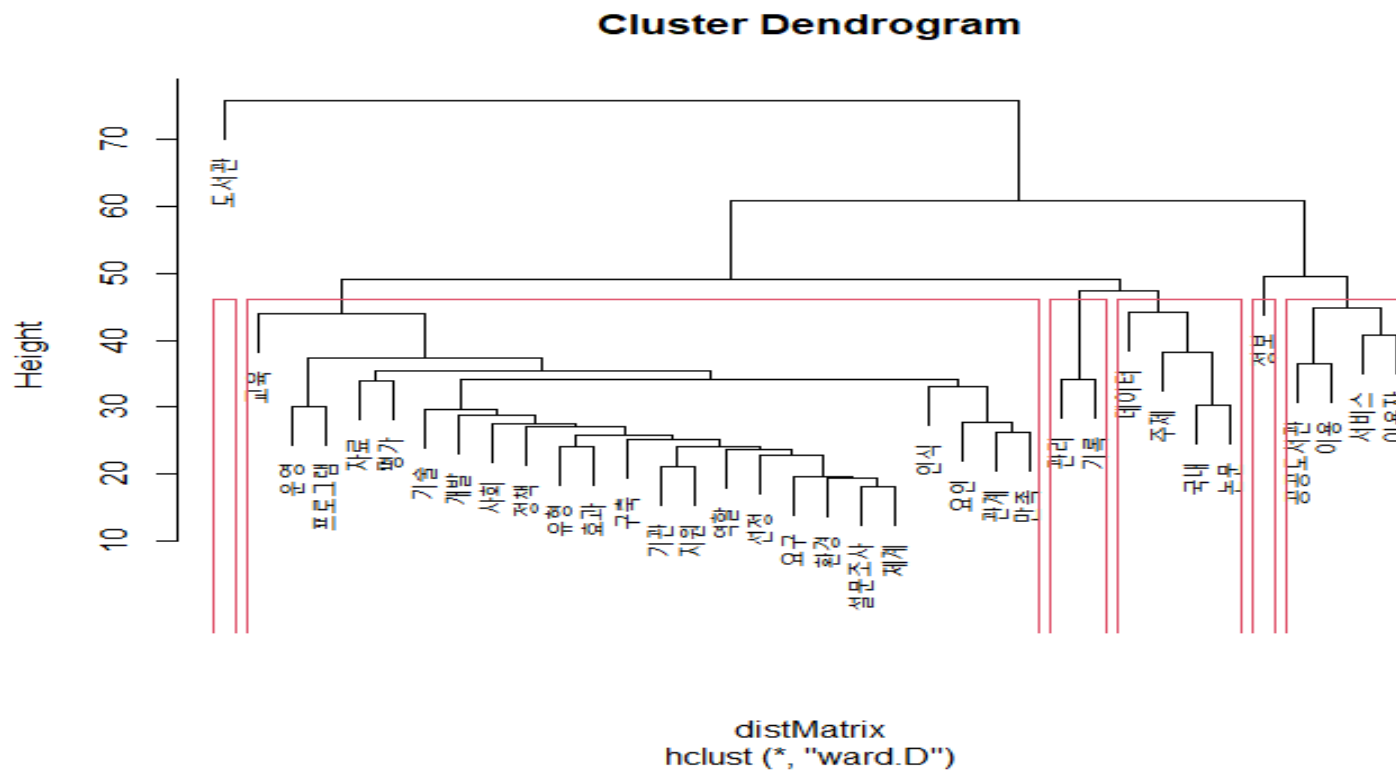
# $\theta$ 를 통한 문서 별 주제 가중치 확인



	주제1	주제2	주제3	주제4	주제5	주제6
2017	0.1651	0.1752	0.1582	0.1700	0.1633	0.1678
2018	0.1853	0.1645	0.1639	0.1751	0.1513	0.1596
2019	0.1768	0.1770	0.1652	0.1548	0.1539	0.1721
2020	0.1859	0.1608	0.1498	0.1479	0.1999	0.1555

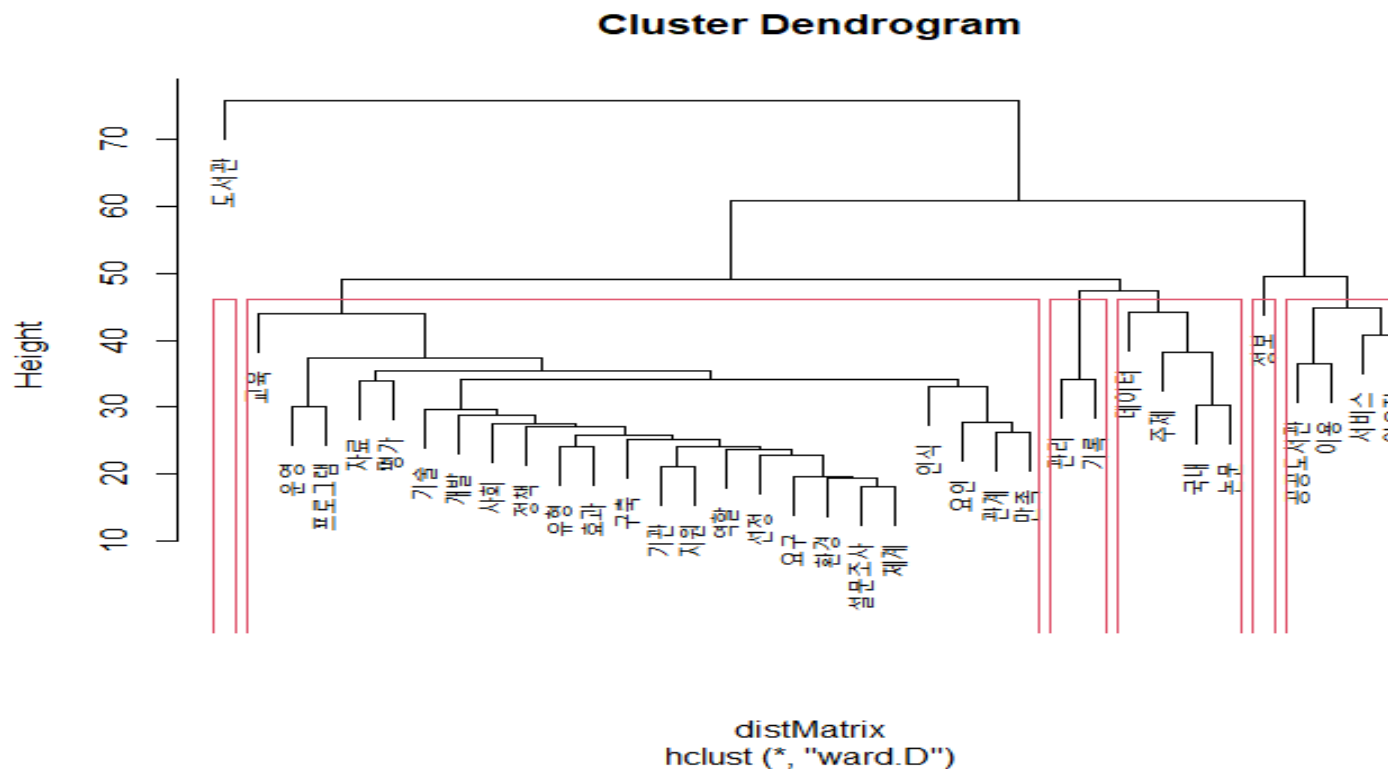
전반적으로 모든 주제가 골고루 연구되고 있으며, 2020년을 기점으로 도서관과 도서관의 디지털화에 대한 연구가 활발해졌음을 확인 가능함

# 계층적 클러스터링을 통한 단어의 계층적 구조 확인



계층적 클러스터링을 통해 단어들 간의 계층적인 관계를 살펴본 결과,  
대체로 단어의 출현 빈도에 따라 단어들이 묶인 것을 확인 가능함

# 계층적 클러스터링을 통한 단어의 계층적 구조 확인



자주 등장하는 '도서관'이라는 단어나, 어떤 기법에 대한 단어를 제외하면,  
대체로 LDA 토픽 모델링 결과와 비슷하게 나타나는 것을 확인할 수 있음

→ 토픽 모델링 결과와 문헌정보학 분야의 단어들의 계층적인 구조는 서로 밀접한 관련이 있다!

## Part 4.

### 연구의 의의 및 한계



# 연구의 한계

## 데이터셋의 한계

전처리 과정에서 몇몇 논문들이 제외되면서  
전체 문헌정보학 분야에 대한 조사가  
이루어지지 않게 되었음

이 과정에서 한국문헌정보학회지의  
2021년 논문을 모두 제외하게 되어  
연도별 흐름 분석의 경우  
연구 문제를 불완전하게  
해결할 수 밖에 없었음

## R(KoNLP)의 한계

NIADic이 많은 형태소를 분해할 수 있음에도  
문헌정보학에서 주로 사용되는 학술용어를  
분류하는데 한계가 있었음

KoNLP의 형태소 분해 능력에 한계가 있었음

토픽모델링 성능을 불용어 처리로만  
보완할 수 밖에 없었음

# 연구의 의의



최근 5개년간 문헌정보학 분야의 연구를 대상으로 주제 개수를 더 적게하여 분석  
→ 최신 연구 트렌드를 기존 선행 연구 대비 더 빠르게 확인 가능





→ 국내 연구동향 분석에 새로운 방법론 제시

The background of the slide is a stylized illustration of a library. It features rows of bookshelves filled with books of various colors (red, blue, green, yellow, and white). The shelves are arranged in a grid-like pattern, with some sections having open shelves and others having closed doors. The overall color palette is soft and pastel, with a light pinkish-beige background. The text '감사합니다!' is centered over the middle of the image.

# 감사합니다!

분석용 코드 및 자세한 내용은 6/5(일)까지 iCampus 열린게시판에 업로드 예정입니다~