



Information Retrieval

목차



- 정보검색 개념
- 자연언어와 정보검색
- 적합성 및 적합성 판단
- 검색대상 정보 유형
- 정보 검색 응용 및 서비스
- 정보검색시스템 구성
- 정보요구와 질의
- 질의 문서 유사도
- 용어빈도수, 문헌빈도수, 역문헌빈도수
- 질의문서유사도 수식
- 색인 단위
- 문서 색인
- 문서 검색
- 검색 모델

데이터베이스 검색

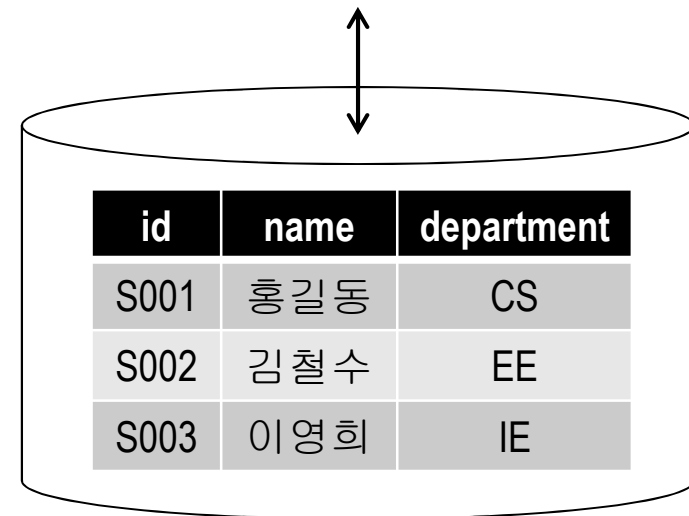
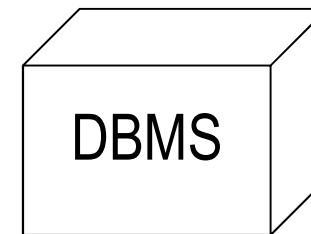
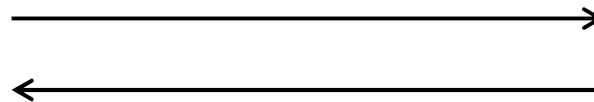
데이터베이스 검색

학생정보검색

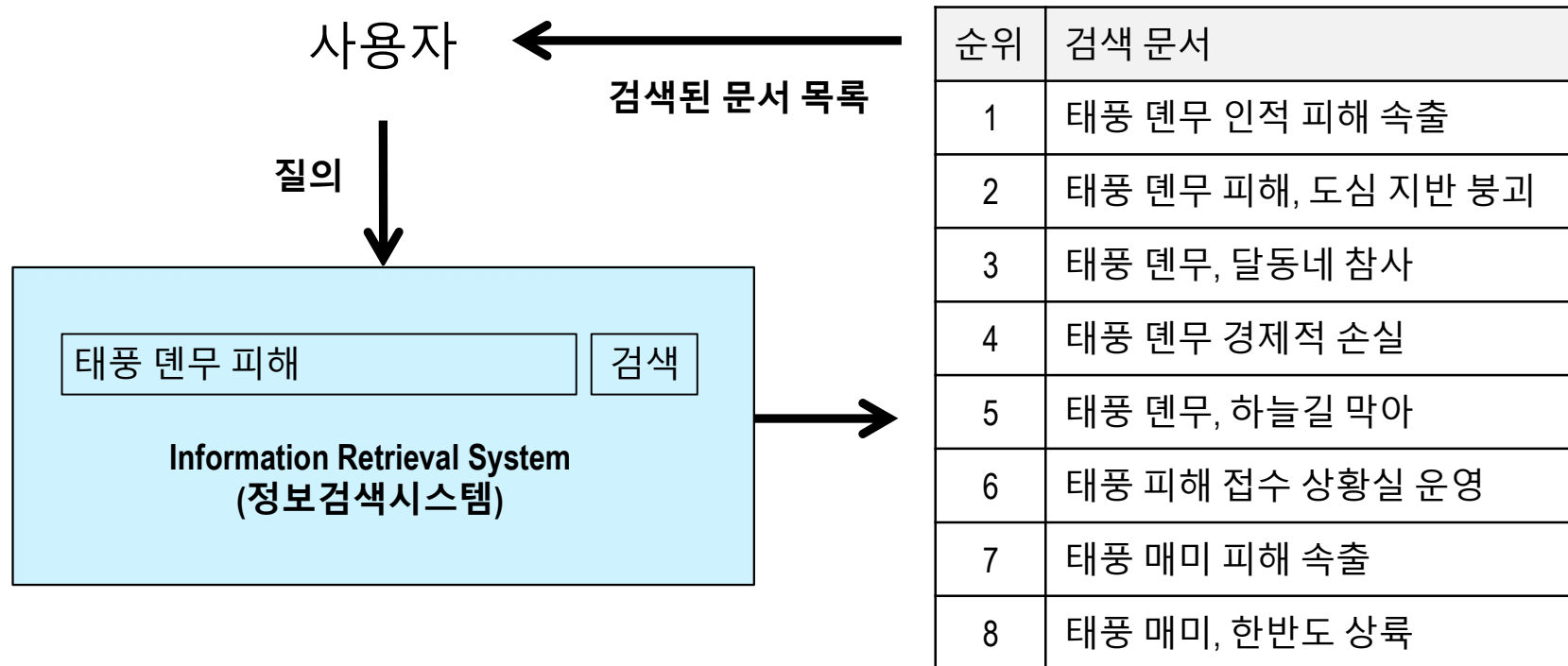
학과 ▼ CS

검색

```
SELECT id, name  
FROM Student  
WHERE department='CS'
```



Information retrieval (정보 검색)



데이터베이스 검색 vs. 정보검색

데이터베이스 검색

- 질의 형식
 - ◆ SQL
- 검색 대상
 - ◆ 구조적 정보

SELECT id, name FROM STUDENT WHERE department='CS'



id	name	department
S001	홍길동	CS
S002	김철수	EE
S003	이영희	IE

정보검색

- 질의 형식
 - ◆ 자연어
- 검색 대상
 - ◆ 비구조적 정보

질의: WHO 코로나19 평가



세계보건기구(WHO)가 11일 코로나19 전염 상황을 '감염병 세계적 유행 (팬데믹)'으로 규정한 가운데, 외신들은 주요 발병국인 한국과 이탈리아의 대응 방식을 비교하는 등 한국이 보여준 검사 및 치료 방식에 주목하는 기사를 잇달아 내놓고 있다.

출처: 대한민국정책브리핑, <http://www.korea.kr/news/policyNewsView.do?newsId=148870283>

Reference: <https://nlp.stanford.edu/IR-book/pdf/10xml.pdf>

구조적(정형), 비구조적(비정형), 반구조적 데이터



Structured data

● Relational database

id	name	department
S001	홍길동	CS
S002	김철수	EE
S003	이영희	IE

Unstructured data

● Text

세계보건기구(WHO)가 11일 코로나19 전염 상황을 '감염병 세계적 유행(팬데믹)'으로 규정한 가운데, 외신들은 주요 발병국인 한국과 이탈리아의 대응 방식을 비교하는 등 한국이 보여준 검사 및 치료 방식에 주목하는 기사를 잇달아 내놓고 있다.

출처: 대한민국정책브리핑, <http://www.korea.kr/news/policyNewsView.do?newsId=148870283>

Semi-structured data

● XML documents

```
<?xml version="1.0"?>
<book>
  <title>정보검색의 개념과 구현 원리</title>
  <year>2019</year>
  <author>홍길동</author>
  <publisher>경성출판사</publisher>
  <memo>정보검색의 개념, 색인 및 검색 모듈의 구현 원리를 다룬다.</memo>
</book>
```

정보검색



✚ 정보검색(information retrieval)이란 대량의 정보 모음으로부터 사용자의 **정보요구(information need)**에 **적합(relevant)**한 자료를 찾는 것이다

- 대량의 정보 모음

- ◆ 예) 대량의 문서 집합

- 예) 한국어 웹페이지 문서의 전체 집합

- 개별 문서는 **자연어** 텍스트로 표현되어 있음

- 정보요구(information need)

- ◆ 사용자가 찾으려고 하는 정보

- **자연어** 텍스트(단어, 구, 문장 등)로 표현됨

- 정보요구에 적합한

- ◆ 정보요구를 만족하는(satisfying)

자연언어(natural language)

✚ 자연언어(자연어, 언어)

- 한국어, 영어, 중국어 등 사람들이 사용하는 언어
- 말, 글 등

✚ 텍스트(text)

- 단어(word), 구(phrase), 문장(sentence) 등 자연언어의 문자 표현
- 글

✚ 정보검색 관점에서의 특징

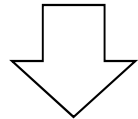
- 동의어(synonym)
 - ◆ 같은 의미를 갖는 다른 표현
 - 자동차, 차, 차량, 카
- 동형이의어(homograph), 다의어(polyseme)
 - ◆ 다른 의미를 갖는 동일 표현
 - 사과(apple), 사과(apology)

자연어와 정보검색: 동의어, 다의어 문제



자동차 매매

검색



문서 1

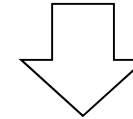
자동차 매매 시 주의점을 알려드립니다.

문서 2

현재 판매 가능한 차량 모델 및 가격은 다음과 같습니다.

사과 재배

검색



문서 1

소비자연합에서는 제조사에 공식 사과를 요청했다.

문서 2

귀농인을 위한 사과 재배 방법은 다음과 같습니다.

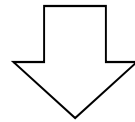
자연어와 정보검색:

자연어이해(Natural Language Understanding, NLU)



- 질의 및 문서 내용 이해
- 자연어 텍스트 이해 (문장 분할, 형태소분석, 품사태깅, 구문분석, 의미분석)
- 정보검색을 위해 질의 및 문서의 자연어이해가 반드시 필요한가?

자동차 매매 주의점 검색



문서 1

자동차 수리 시 주의점을 알려드립니다.

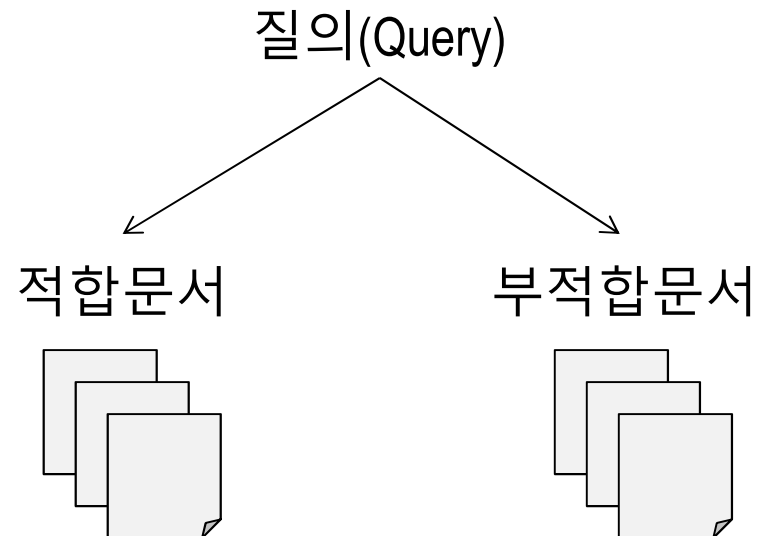
문서 2

이 사이트에서는 자동차 매매건은 취급하지 않습니다.

적합성(Relevance)

✚ 적합성이란 무엇인가?

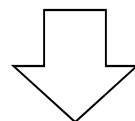
- 어떤 문서가 질의에 적합한 문서인가?
- 질의에 대해 적합문서와 부적합문서를 구분하는 기준은?



적합성 판단 예시 (1/4)

- *적합성의 정도*
- *적합, 부적합 vs. 매우 적합, 적합, 부분 적합, 부적합*

자동차 매매 주의점 검색



문서 1 자동차 매매 시 다음 사항들을 주의해 주시기 바랍니다..

문서 2 이 사이트에서는 자동차 매매 및 수리 시 주의점을 알려드립니다.
다음은 자동차 매매 시 주의점 목록입니다.
가. 마일리지
나. 사고 유무
다음은 자동차 수리 시 주의점 목록입니다.
가. 제동장치
나. 오일류

적합성 판단 예시 (2/4)

- 서로 다른 사용자(나이, 경험, 지식, 전문성 등)의 적합성 기준

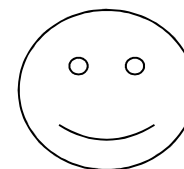
초등학생



자율주행차 현황

검색

자동차연구소 책임연구원

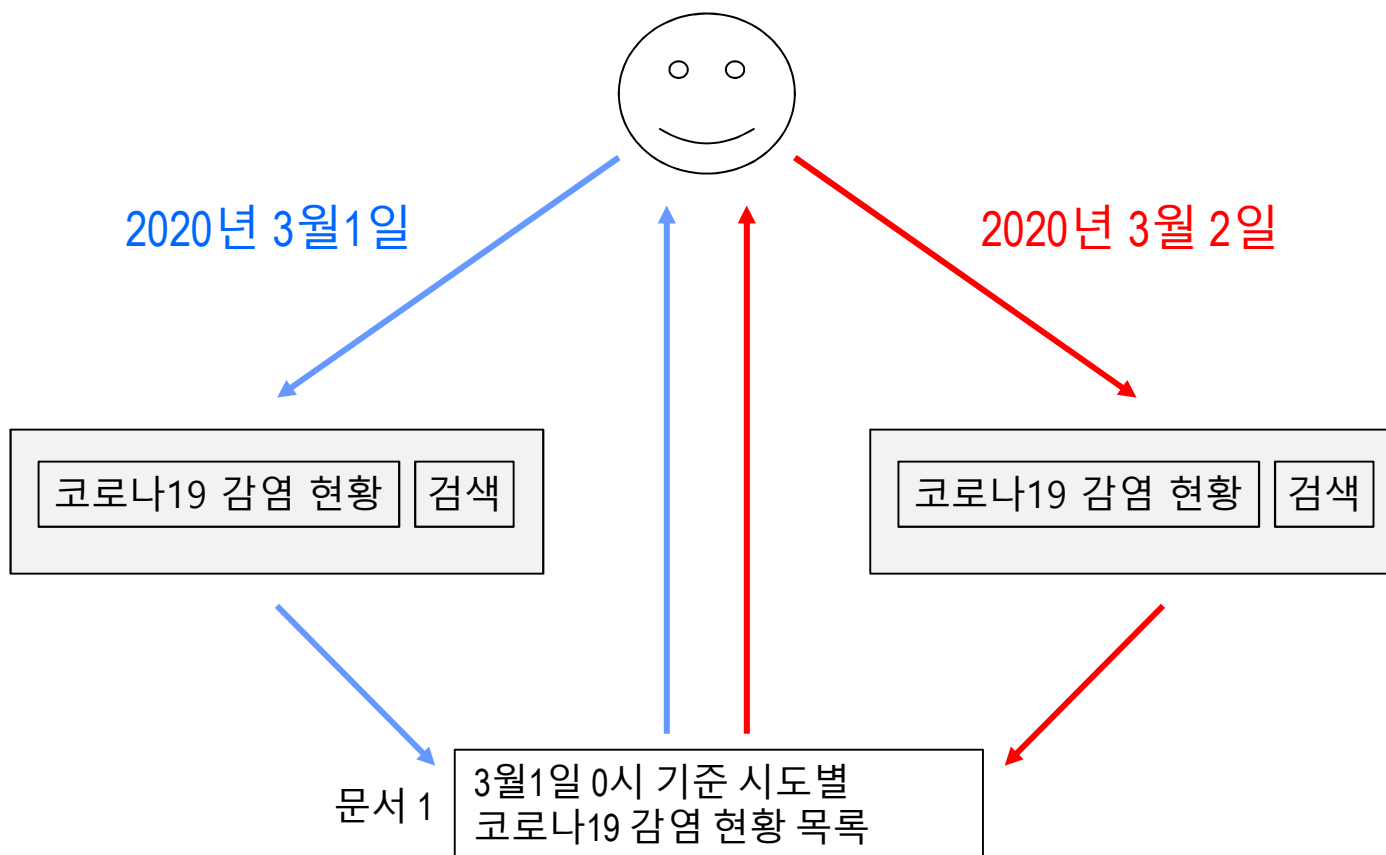


자율주행차 현황

검색

적합성 판단 예시 (3/4)

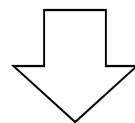
- 서로 다른 시점에서의 적합성 기준



적합성 판단 예시 (4/4)

- 서로 다른 순위 지점에서의 적합성 기준

코로나19 감염 현황 검색



- | | |
|---------|----------------------------------|
| 1 순위 문서 | 시도별 코로나19 감염 현황 목록 (미래경성방송사 제공) |
| 2 순위 문서 | 시도별 코로나19 감염 현황 목록 (글로벌경성방송사 제공) |
| 3 순위 문서 | 시도별 코로나19 감염 현황 목록 (세계경성방송사 제공) |

검색 대상 정보 유형



정보 유형

- 텍스트
 - ◆ Text retrieval
- 오디오
 - ◆ Sound retrieval
- 이미지
 - ◆ Image retrieval
- 비디오
 - ◆ Moving picture retrieval

검색 대상 텍스트 유형



텍스트 유형







- 웹 문서
 - ◆ Web search
- 뉴스
 - ◆ News search
- 특허
 - ◆ Patent retrieval
- 학술문헌(논문, 보고서)
 - ◆ Academic search service
- 트윗(Tweet)
 - ◆ Tweet retrieval
- 책
 - ◆ Book search

텍스트 단위

텍스트 단위

- Document (문서)
 - ◆ Document retrieval
- Chapter (장)
- Paragraph (단락)
- Sentence (문장)
- Phrase (구)
- Word (단어)
- Character (문자)

검색시스템 구분: 검색 문서 규모

-  Web search
-  Domain-specific search
 -  Patent search
 -  Academic search
-  Enterprise search
-  Desktop search

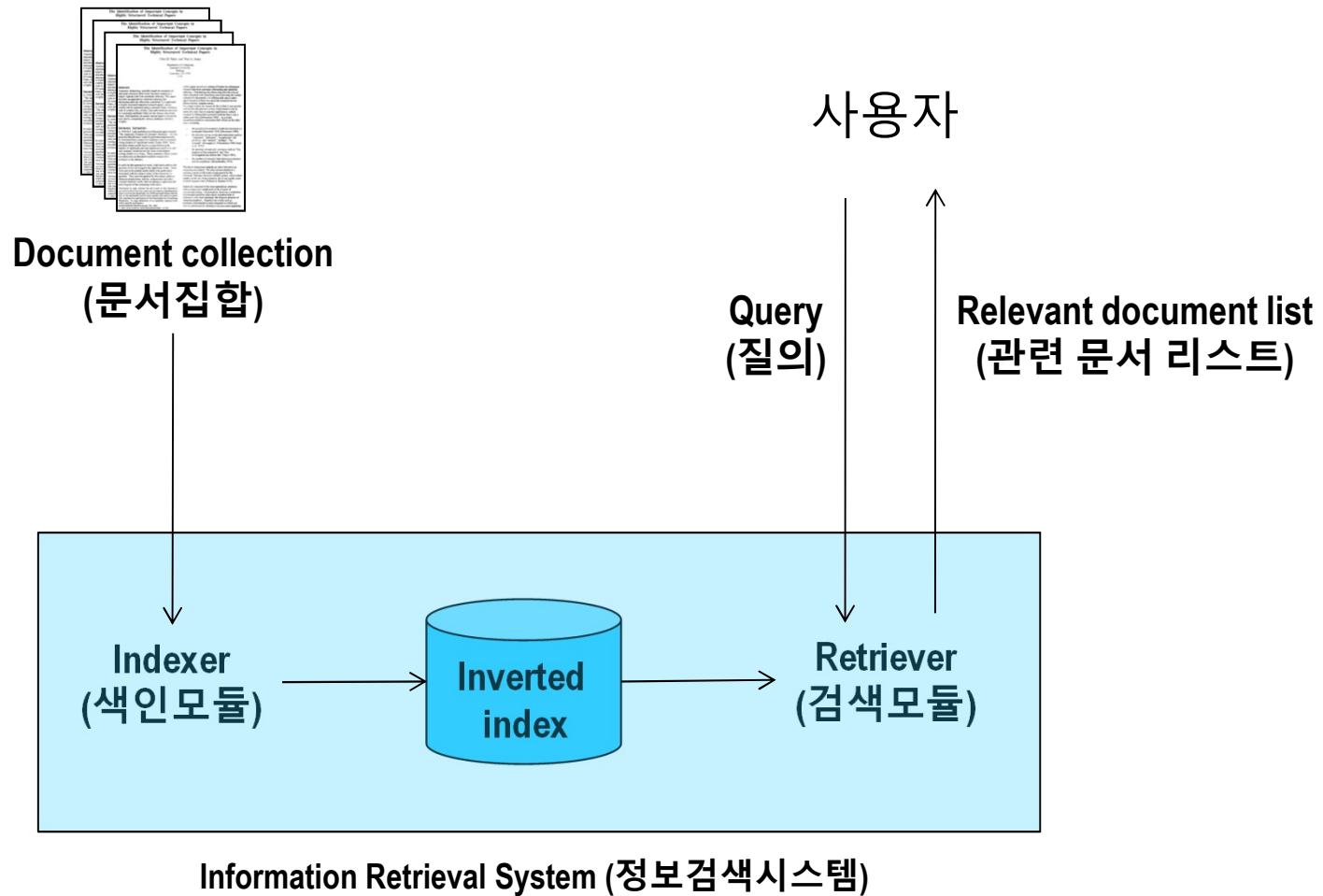
검색 응용 및 서비스



- + Information filtering
 - Recommender system (추천시스템)
- + Question-answering system (질의응답시스템)
- + Cross-language information retrieval (교차언어정보검색)
- + Document classification (문서분류)
- + Document clustering (문서군집화)

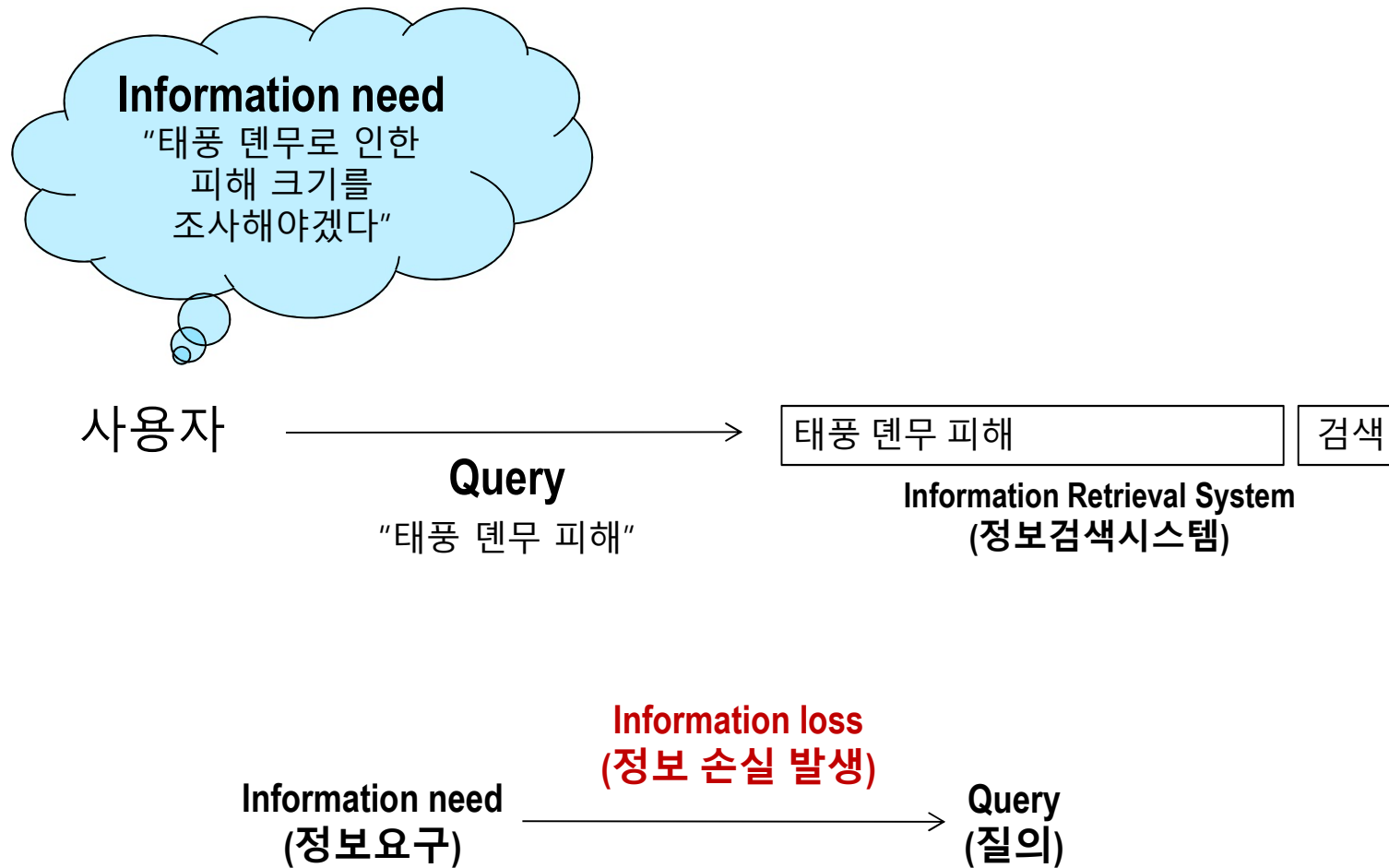
Information retrieval:

Information retrieval system (IR system)



Information retrieval:

Information need & Query



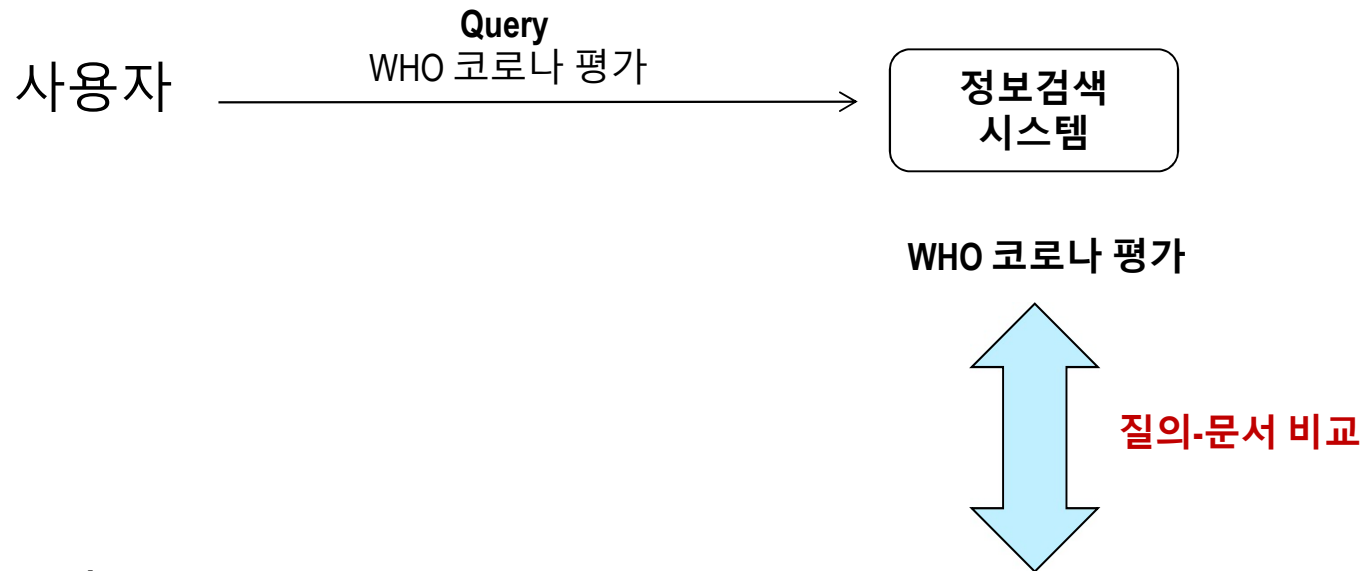
Query 예



예

- Q = 세종대왕
- Q = 경성대학교
- Q = 통영
- Q = 경성대학교 근처 맛집
- Q = 부산에서 여수까지 거리는?
- Q = 췌장암 치료 방법
- Q = 다문화 가정 지원을 위한 정부 차원의 정책
- Q = 불법 체류자 인권 변호사

Information retrieval:
질의-문서 비교



Document (문서)

WHO 팬데믹 선언과 외신의 주목...“한국, 코로나19 대응 모범 사례”

세계보건기구(WHO)가 11일 코로나19 전염 상황을 '감염병 세계적 유행(팬데믹)'으로 규정한 가운데, 외신들은 주요 발병국인 한국과 이탈리아의 대응 방식을 비교하는 등 한국이 보여준 검사 및 치료 방식에 주목하는 기사를 잇달아 내놓고 있다. 특히 한국은 충분한 키트와 검사의 정확성을 보유하고 있다며, 대규모 검사 방식을 다른 나라의 '모범 사례'로 평가했다.

출처: 대한민국정책브리핑, <http://www.korea.kr/news/policyNewsView.do?newsId=148870283>

질의-문서 유사도

✚ 적합성

- 어떤 문서가 질의에 적합한 문서인가?
- 질의에 대해 적합문서와 부적합문서를 구분하는 기준은?

✚ 질의-문서 유사도 (Query-document similarity)

- 검색시스템 내부에서 계산되는 질의와 문서 간 관련성의 정도
 - ◆ 검색(retrieval) 모듈
 - 질의-문서 유사도에 따라 문서집합 내 문서들을 순위화
 - 문서집합 내 각 문서와 질의와의 유사도 계산 고속 수행 필요
 - ◆ 색인(indexing) 모듈
 - 효율적 검색을 위해 문서집합을 미리 가공해 두는 과정
- 예
 - ◆ $\text{Sim}(Q, D1) = 0.3$
 - ◆ $\text{Sim}(Q, D2) = 0.9$
 - ◆ $\text{Sim}(Q, D3) = 0.7$

질의-문서 유사도 계산 연습

- 다음 각 질의 및 문서 집합에 대해 문서 순위를 결정해 보시오
- 당신의 문서 순위 결정 기준은 무엇인가?
 - 당신이 고안한 질의-문서 유사도 계산 수식을 적어 보시오

Example 1

Q: 부산

문서 집합(색인 용어)

D1 = [부산 서울]
D2 = [부산 부산]

Example 2

Q: 부산

문서 집합(색인 용어)

D1 = [부산]
D2 = [부산 서울]

Example 3

Q: 부산 서울

문서 집합(색인 용어)

D1 = [부산]
D2 = [서울]
D3 = [부산 울산]
D4 = [부산 경남]
D5 = [부산 창원]

질의-문서 유사도 계산 연습: Example 1

- 다음 질의에 대해 문서 순위를 결정해 보시오
- 당신의 문서 순위 결정 기준은 무엇인가?
 - 당신이 고안한 질의-문서 유사도 계산 수식을 적어 보시오

Extreme case of Example 1

Q: 부산

문서 집합(색인 용어)

D1 = [부산 서울]
D2 = [부산 부산]
D3 = [부산 부산 부산]
D4 = [부산 부산 부산 부산]
D5 = [부산 부산 부산 부산 부산]
D6 = [부산 부산 부산 부산 부산 부산]
D7 = [부산 부산 부산 부산 부산 부산 부산]

질의-문서 유사도 계산 연습: Example 2



- 다음 질의에 대해 문서 순위를 결정해 보시오
- 당신의 문서 순위 결정 기준은 무엇인가?
 - 당신이 고안한 질의-문서 유사도 계산 수식을 적어 보시오

Extreme case of Example 2

Q: 부산

문서 집합(색인 용어)

D1 = [부산]
D2 = [부산 서울]
D3 = [부산 서울 대전]
D4 = [부산 서울 대전 광주]
D5 = [부산 서울 대전 광주 대구]
D6 = [부산 서울 대전 광주 대구 창원]
D7 = [부산 서울 대전 광주 대구 창원 제주]

질의-문서 유사도 계산 연습: Example 3



- 다음 질의에 대해 문서 순위를 결정해 보시오
- 당신의 문서 순위 결정 기준은 무엇인가?
 - 당신이 고안한 질의-문서 유사도 계산 수식을 적어 보시오

Extreme case of Example 3

Q: 부산 서울

문서 집합(색인 용어)

D1 = [부산]
D2 = [서울]
D3 = [부산 울산]
D4 = [부산 경남]
D5 = [부산 창원]

...

D1000 = [부산 제주]

- D2를 제외한 모든 문서가 부산을 포함하고 있다고 가정
- D2 문서에만 서울이 출현했다고 가정

Term Frequency (용어빈도수)

✚ Term frequency (TF)

- 특정한 하나의 문서 내에 출현한 특정한 용어의 출현 횟수

D99 한국, 한국, 한국, 미국, 미국, 중국

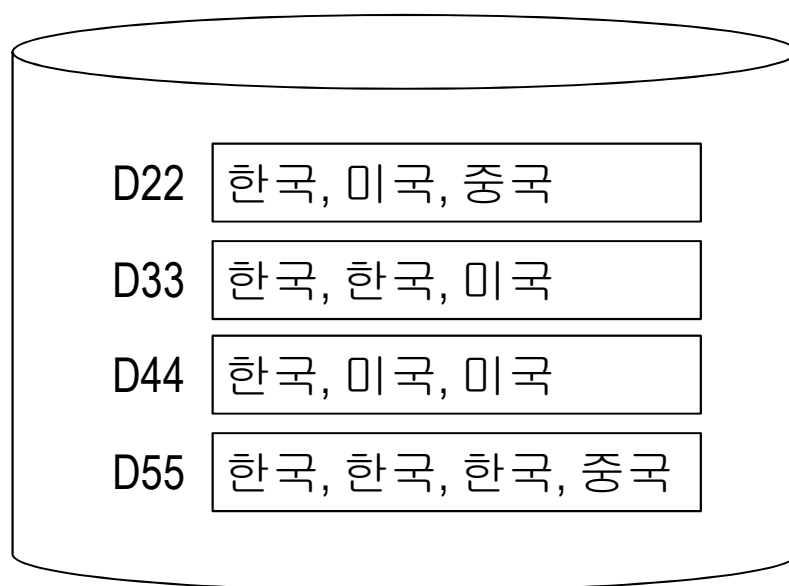
● 예)

- ◆ $tf(D99, \text{한국})=3$
- ◆ $tf(D99, \text{미국})=2$
- ◆ $tf(D99, \text{중국})=1$

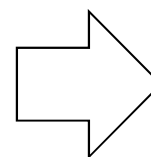
Document Frequency (문헌빈도수)

Document frequency (DF)

- 특정한 용어가 출현한 문서의 개수



문서 집합 (a collection of documents)

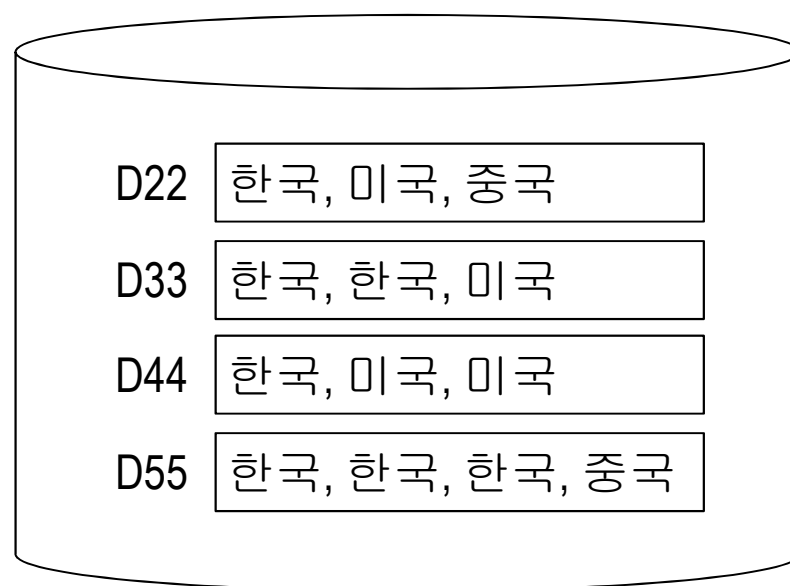


$df(\text{한국})=4$
 $df(\text{미국})=3$
 $df(\text{중국})=2$

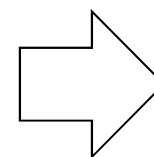
Collection Frequency (컬렉션 빈도수)

Collection frequency (CF)

- 특정한 용어가 문서 집합 전체에서 출현한 횟수



문서 집합 (a collection of documents)



$cf(\text{한국})=7$
 $cf(\text{미국})=4$
 $cf(\text{중국})=2$

Inverse Document Frequency (역문헌빈도수)

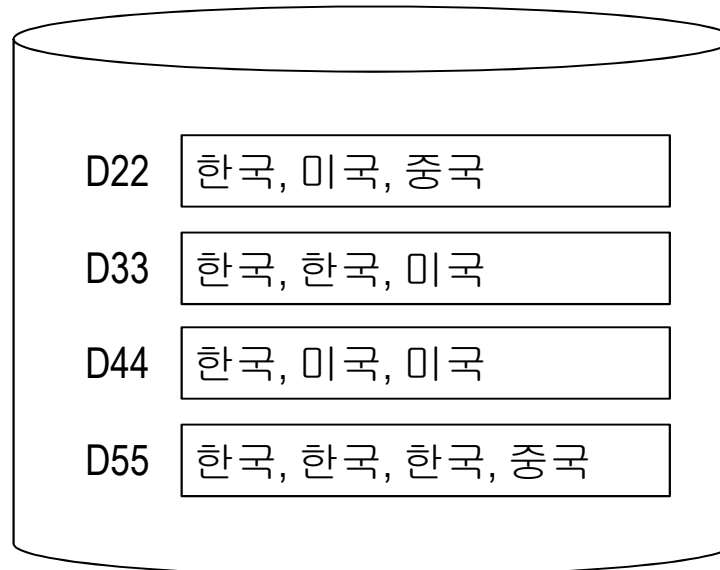


✚ Inverse document frequency (idf)

- $idf(t) = \frac{N}{df(t)}$

N: 문서 집합 내 총 문서의 개수

- $idf(t) = \log\left(1 + \frac{N}{df(t)}\right)$



문서 집합 (a collection of documents)

$$idf(t) = \log\left(1 + \frac{N}{df(t)}\right)$$

$$idf(\text{한국}) = \log(1+4/4)$$

$$idf(\text{미국}) = \log(1+4/3)$$

$$idf(\text{중국}) = \log(1+4/2)$$

질의-문서 유사도 수식 예시

✚ 문서 D에서의 용어 t의 가중치(중요도, weight)

- $w(t, D) = \frac{tf(t, D) \times idf(t)}{Length(D)}$

✚ 질의-문서 유사도 (Query-Document Similarity)

- $sim(Q, D) = \sum_{i=1}^n w(q_i, D) = \sum_{i=1}^n \frac{tf(q_i, D) \times idf(q_i)}{Length(D)}$

- 예

- ◆ Q = [부산, 부산, 여행]

- $q_1 = \text{부산}, q_2 = \text{부산}, q_3 = \text{여행}$

- ◆ $D_7 = [\text{부산}, \text{여행}, \text{여행}, \text{광안리}, \text{해운대}]$

- ◆ $idf(\text{부산})=4.7, idf(\text{여행})=1.6$

- ◆ $sim(Q, D_7) = \frac{1 \times 4.7}{5} + \frac{1 \times 4.7}{5} + \frac{2 \times 1.6}{5}$

질의-문서 유사도: IDF



Q: 특허

문서 집합(색인 용어)

D1 = [특허 자동차 투명 와이퍼]

D2 = [특허 우산 접이 버튼]

D3 = [특허 신발 표면 방수]

D4 = [특허 자동차 안전 도어]

D5 = [특허 안경 투명 코걸이]

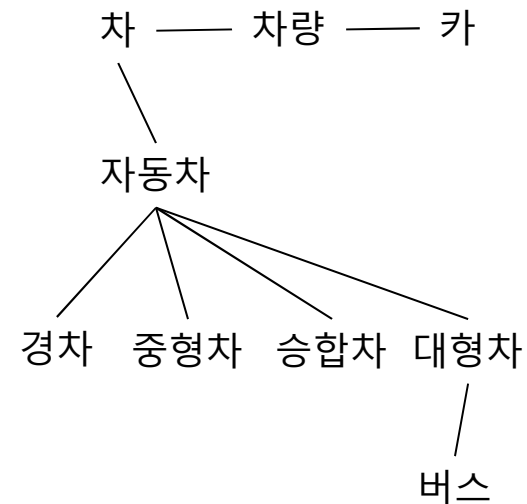
질의-문서 유사도: 유의어 사전



Q: 자동차 도어

문서 집합(색인 용어)

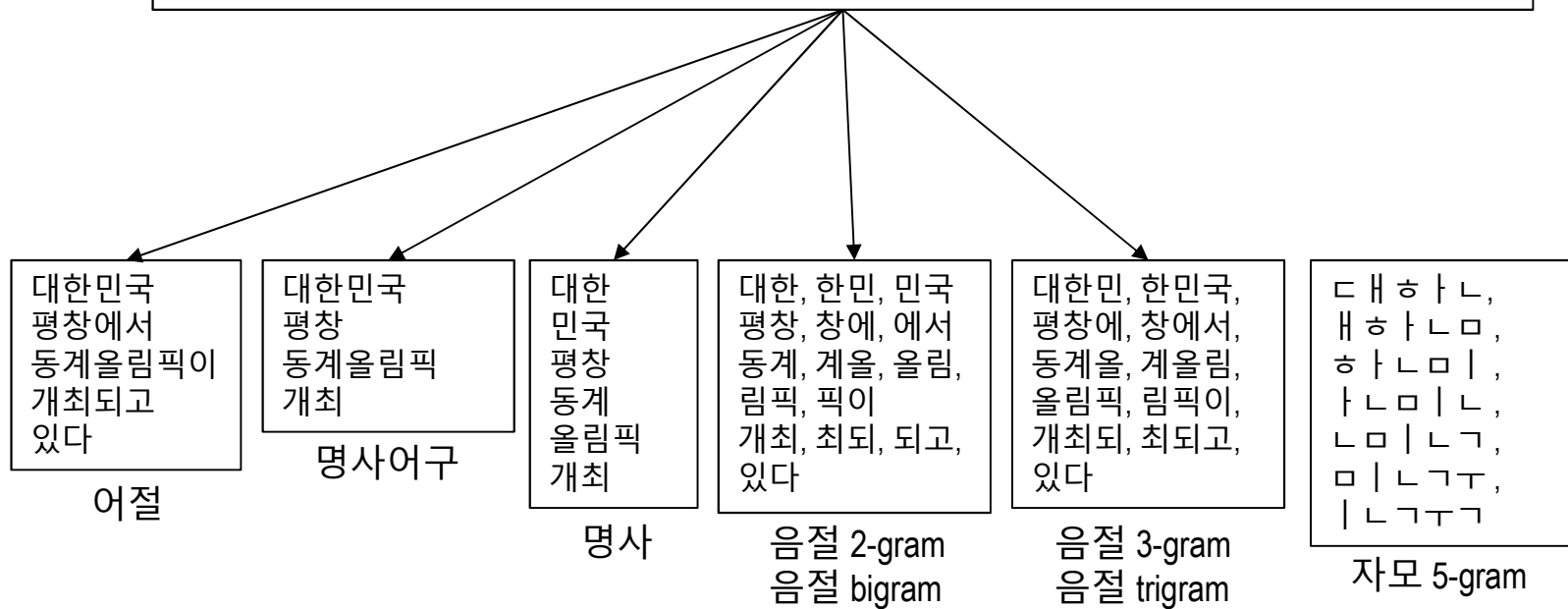
D1 = [자동차 투명 와이퍼]
D2 = [차량 안전벨트]
D3 = [카 안전 도어]
D4 = [경차 속도 저감 장치]
D5 = [승합차 자동 도어]



색인 단위



대한민국 평창에서 동계올림픽이 개최되고 있다.



질의-문서 유사도: 색인 단위



Q: 경성대학교

문서 집합(색인 용어)

D1 = [경성 대학교]
D2 = [경성대 주차장]
D3 = [경성대학교건학기념관]
D4 = [부산 소재 대학교]
D5 = [전국 대학 지역 분포]

질의-문서 유사도: 색인 단위



Q: 아놀드슈워제네거

문서 집합(색인 용어)

D1 = [아놀드슈와제네거]
D2 = [아놀드슈왈제네거]
D3 = [아놀드슈왈재네거]
D4 = [아놀드슈왈제내거]
D5 = [아놀드슈왈츠제네거]
D6 = [아놀드슈왈쯔제네거]

질의-문서 유사도: 색인 단위

Q: 부산에서 서울로 운행하는 기차

Q': 부산 서울 운행 기차

문서 집합(색인 용어)

D1 = [부산 서울 운행 기차]
D2 = [서울 부산 운행 기차]

Q: 부산에서 서울로 운행하는 기차

Q': 부산에서 서울로 운행하는 기차

문서 집합(색인 용어)

D1 = [부산에서 서울로 운행하는 기차]
D2 = [서울에서 부산으로 운행하는 기차]

질의-문서 유사도: 색인 단위 (의미, 개념)



Q: 사과 배

문서 집합

D1 = [일본 사과 배상 요구]
D2 = [가을 사과 풍년]
D3 = [동해안 배 침몰]
D4 = [유전자 조작 배 논란]

Q: 사과 배

Q': 사과 apple 배 pear

문서 집합

D1 = [일본 사과 apology 배상 요구]
D2 = [가을 사과 apple 풍년]
D3 = [동해안 배 ship 침몰]
D4 = [유전자 조작 배 pear 논란]

Information retrieval:

Document indexing (문서 색인)



Document (문서)

출처: 대한민국정책브리핑, <http://www.korea.kr/news/policyNewsView.do?newsId=148870283>

WHO 팬데믹 선언과 외신의 주목...“한국, 코로나19 대응 모범 사례”

세계보건기구(WHO)가 11일 코로나19 전염 상황을 '감염병 세계적 유행(팬데믹)'으로 규정한 가운데, 외신들은 주요 발병국인 한국과 이탈리아의 대응 방식을 비교하는 등 한국이 보여준 검사 및 치료 방식에 주목하는 기사를 잇달아 내놓고 있다. 특히 한국은 충분한 키트와 검사의 정확성을 보유하고 있다며, 대규모 검사 방식을 다른 나라의 '모범 사례'로 평가했다.

형태소분석 (Morphological analysis)

품사 태깅 (POS tagging)

불용어 제거 (Stop-word removal)

복합명사분해 (Compound noun segmentation)

색인어 추출 (index term extraction)

형태소분석, 품사태깅 데모
<http://kle.postech.ac.kr/demo/>

WHO, 팬데믹, 선언, 외신, 주목, 한국, 코로나, 19, 대응, 모범, 사례, 세계, 보건, 기구, WHO, 11, 코로나, 19, 전염, 상황, 감염병, 세계, 유행, 팬데믹, 규정, 가운데, 외신, 발병국, 한국, 이탈리아, 대응, 방식, 비교, 한국, 검사, 치료, 방식, 주목, 기사, 한국, 키트, 검사, 정확성, 보유, 대규모, 검사, 방식, 나라, 모범, 사례, 평가

Document indexing (문서 색인)



- D1 바그다드 폭탄 테러로 한
국대사관 유리창 깨져
- D2 아프간 자살 테러로 최소 3
명이 사망하고 17명이 부상
하는 등
- D3 뭄바이 테러 관련 한국대
사관에서는 한국인 사망과
부상 피해를 ...

Document Indexing

Inverted index (역파일 색인)

색인 용어	용어 출현 정보
바그다드	D1
폭탄	D1
테러	D1, D2, D3
한국대사관	D1, D3
유리창	D1
아프간	D2
자살	D2
사망	D2, D3
부상	D2, D3
뭄바이	D3
한국인	D3

Document retrieval (문서 검색)



Query: 자살 테러 사망

Document
Retrieval

순위	문서번호	유사도
1	D2	3
2	D3	2
3	D1	1

Inverted index (역파일 색인)

D1 바그다드 폭탄 테러로 한국대사관
유리창 깨져

D2 아프간 자살 테러로 최소 3명이 사
망하고 17명이 부상하는 등

D3 뭄바이 테러 관련 한국대사관에서
는 한국인 사망과 부상 피해를 ...

Document
Indexing

색인 용어	용어 출현 정보
바그다드	D1
폭탄	D1
테러	D1, D2, D3
한국대사관	D1, D3
유리창	D1
아프간	D2
자살	D2
사망	D2, D3
부상	D2, D3
뭄바이	D3
한국인	D3

검색 모델



- + Boolean model
- + Vector-space model
- + Probabilistic model
- + Language model

Reference



- + Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze (2008). Introduction to Information Retrieval. Cambridge University Press.
- + Bruce Croft, Donald Metzler, Trevor Strohman. (2009). Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company.
- + Ricardo Baeza-Yates, Berthier Ribeiro-Neto. (1999). Modern Information Retrieval. Addison-Wesley Publishing Company.