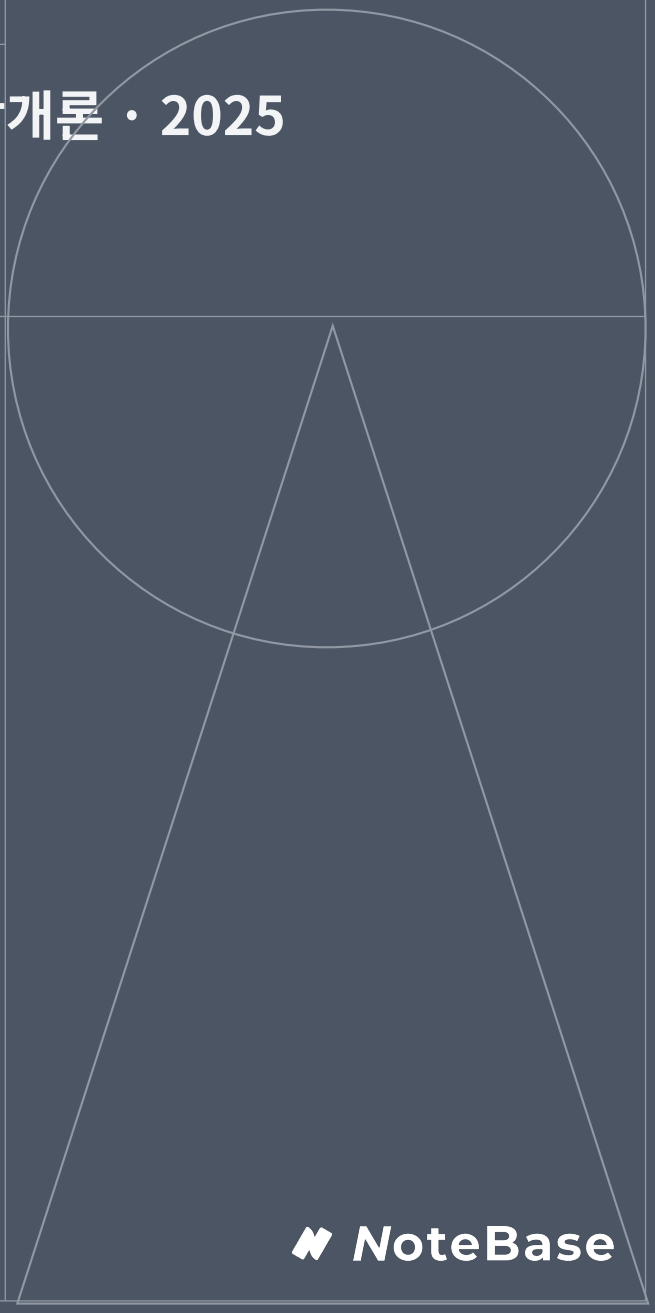
기출OX 해설집

공무원/국회직/9급 · 정보학개론 · 2025



NoteBase

기출OX 해설집

공무원/국회직/9급 · 정보학개론 · 2025

공무원/국회직/9급 · 정보학개론 · 2025

1. 오픈액세스 유형에 대한 설명으로 옳은 것은?

① Bronze OA는 출판사 웹사이트에서만 무료로 논문을 이용할 수 있고 공유와 재사용을 허용하지 않는다.

② Gold OA는 저자가 자신의 프리프린트를 리포지터리에 셀프 아카이빙할 수 있도록 허용한다.

③ Gratis OA는 온라인 접근이 무료이며 재사용도 가능하다.

④ Green OA는 OA 학술지에 수록된 모든 논문에 무료 접근을 허용하고, 논문처리비용을 부과하는 경우가 많다.

⑤ Libre OA는 접근비용 장벽만을 제거한 상태로 무료 온라인 접근을 허용한다.

정답 1번

출제 포인트

이 문제는 오픈액세스를 접근 경로와 재사용 권한의 범위에 따라 구분하는 문제이다.

핵심 개념

Bronze OA: 출판사 웹사이트에서 논문을 무료로 읽을 수 있지만 명시적인 오픈 라이선스가 없어 일반적인 공유·재사용 권한은 부여되지 않는 방식이다.
Gold OA: 학술지 출판 단계에서 출판본을 공개하는 방식이다.
Green OA: 저자가 논문의 원고를 기관·주제 리포지터리 등에 셀프 아카이빙하는 방식이다.
Gratis OA: 가격 장벽만 제거하여 무료로 접근하게 하는 방식이다.
Libre OA: 가격 장벽뿐 아니라 일부 이상의 이용 허가 장벽도 제거하여 라이선스가 허용하는 재사용을 가능하게 하는 방식이다.

NoteBase

선지별 해설

① O — Bronze OA는 출판사 웹사이트에서 논문을 무료로 읽게 하지만 명시적인 오픈 라이선스가 없으므로 일반적인 공유·재사용 권한까지 부여하지 않는다.
② X — 틀린 표현은 ‘Gold OA는 저자가 자신의 프리프린트를 리포지터리에 셀프 아카이빙’한다는 부분이다. 올바르게는 Gold OA는 학술지가 출판본을 공개하는 방식이고, 저자가 원고를 리포지터리에 셀프 아카이빙하는 방식은 Green OA이다.
③ X — 틀린 표현은 ‘Gratis OA는 재사용도 가능하다’는 부분이다. 올바르게는 Gratis OA는 무료 온라인 접근만 보장하며, 별도의 이용 허가가 없다면 재사용 권한까지 보장하지 않는다.
④ X — 틀린 표현은 ‘OA 학술지에 수록된 모든 논문’과 ‘논문처리비용을 부과하는 경우가 많다’는 부분이다. 올바르게는 Green OA는 저자가 원고를 리포지터리에 셀프 아카이빙하여 공개하는 방식이며, 학술지 수록 논문 전체의 공개나 논문처리비용 부과를 뜻하지 않는다.
⑤ X — 틀린 표현은 Libre OA가 ‘접근비용 장벽만’을 제거한다는 부분이다. 올바르게는 접근비용 장벽만 제거한 것은 Gratis OA이고, Libre OA는 가격 장벽과 함께 이용 허가 장벽도 제거하여 라이선스 범위의 재사용을 허용한다.

오답 포인트

②는 Gold와 Green의 공개 경로를 바꾸어 놓았고, ③과 ⑤는 Gratis와 Libre의 재사용 권한을 서로 뒤섞었다. ④의 Green은 OA 학술지 전체 공개가 아니라 저자의 리포지터리 셀프 아카이빙이라는 점이 핵심이다.

다시 나오면 이렇게 판단

출판사가 학술지에서 공개하면 Gold, 저자가 리포지터리에 저장하면 Green으로 판단한다. 무료 열람만 가능하면 Gratis, 명시적 이용 허가로 재사용까지 가능하면 Libre로 구분하며, 출판사 사이트에서 무료이지만 오픈 라이선스가 없으면 Bronze로 판단한다.

암기 포인트

Gold는 출판사 공개, Green은 저자 저장, Gratis는 무료 열람, Libre는 허가된 재사용, Bronze는 출판사 사이트의 무라이선스 무료 열람이다.

NoteBase

2. 대상 영역과 표준 메타데이터 스키마의 연결이 옳지 않은 것은?

① 교육 - GEM, KEM, LOM
② 도서관 - MARC, MODS, TEI Headers
③ 멀티미디어 - MPEG-7, PBCore
④ 박물관 - CCO, CDWA, VRA
⑤ 지리정보 - CSDGM, GILS

정답 5번

출제 포인트

대상 정보자원의 영역과 그 영역에서 사용하는 대표 메타데이터 스키마를 올바르게 연결할 수 있는지를 묻는 문제이다.

핵심 개념

메타데이터 스키마는 정보자원의 속성을 일정한 요소와 규칙에 따라 기술하기 위한 구조이다. 교육 영역에는 GEM·KEM·LOM, 도서관 영역에는 MARC·MODS·TEI Header, 멀티미디어 영역에는 MPEG-7·PBCore, 박물관 영역에는 CCO·CDWA·VRA가 대표적으로 연결된다. 지리정보용 CSDGM과 정부정보 탐색용 GILS는 적용 대상이 서로 다르다.

선지별 해설

① O — GEM, KEM, LOM은 교육자료나 학습객체를 기술하는 교육 분야의 메타데이터 스키마이다.
② O — MARC와 MODS는 서지정보를 기술하며, TEI Header도 전자텍스트의 서지·기술 정보를 담으므로 도서관·문헌정보 영역과 연결된다.
③ O — MPEG-7은 멀티미디어 콘텐츠의 특징을 기술하고, PBCore는 방송·미디어 자원을 기술하므로 멀티미디어 영역과 연결된다.
④ O — CCO, CDWA, VRA는 미술품·문화유산·시각자료를 기술하는 데 쓰이므로 박물관 영역과 연결된다.
⑤ X — 틀린 부분은 지리정보 영역에 GILS를 함께 연결한 것이다. CSDGM은 디지털 지리공간 메타데이터 표준이지만, GILS는 정부정보의 소재를 찾도록 지원하는 Government Information Locator Service이다. 올바르게는 '지리정보 - CSDGM', '정부정보 - GILS'로 구분한다.

오답 포인트

⑤의 CSDGM과 GILS는 모두 정보의 소재를 기술·탐색하는 데 관련되어 보여 같은 영역으로 묶기 쉽다. 그러나 CSDGM의 핵심 대상은 지리공간 데이터이고, GILS의 핵심 대상은 정부정보이다.

NoteBase

다시 나오면 이렇게 판단

약어를 대상 자원과 연결하여 판단한다. 학습객체는 LOM, 서지자료는 MARC·MODS, 멀티미디어는 MPEG-7·PBCore, 문화유산·시각자료는 CCO·CDWA·VRA, 지리공간 데이터는 CSDGM, 정부정보는 GILS로 구별한다.

암기 포인트

CSDGM은 지리공간, GILS는 정부정보이다.

NoteBase

3. 동시출현분석에 대한 설명으로 옳지 않은 것은?

① 학문의 지적 구조 파악에 널리 활용된다.
② 동시출현단어분석, 공저자분석 등이 포함된다.
③ 주제 영역의 식별이나 관련성 분석에 활용된다.
④ 문헌 간의 관계를 분석하기 위해 동시인용빈도를 활용한다.
⑤ 연구자 간의 협력 관계를 조사·분석하기 위한 도구로 활용된다.

정답 4번

출제 포인트

동시출현분석의 대표 기법과 각 기법이 분석하는 대상 및 관계를 구별하는 문제이다.

핵심 개념

동시출현분석은 같은 문헌에 함께 나타나는 단어·저자 등의 빈도와 연결 관계를 분석하여 지식 구조를 파악하는 계량정보학적 방법이다. 동시출현단어분석은 단어·주제의 관련성을, 공저자분석은 연구자 사이의 공동연구·협력 관계를 파악한다. 반면 동시인용분석은 후속 문헌에서 두 문헌이나 두 저자가 함께 인용되는 빈도를 이용하는 인용분석이다.

선지별 해설

① O — 동시출현 관계를 통해 핵심 주제, 연구 흐름, 지적 연계 구조를 드러낼 수 있으므로 학문의 지적 구조 파악에 활용된다.
② O — 동시출현단어분석과 공저자분석은 단어나 저자가 함께 나타나는 관계를 이용하는 대표적인 동시출현분석 기법이다.
③ O — 함께 자주 등장하는 단어의 연결을 분석하면 주제 영역을 식별하고 주제 사이의 관련성을 파악할 수 있다.
④ X — 틀린 부분은 '동시인용빈도를 활용한다'이다. 동시인용빈도를 이용해 문헌이나 저자의 지적 관련성을 파악하는 것은 동시인용분석이라는 인용분석에 해당한다. 동시출현분석에 관한 설명으로 고치면 '같은 문헌에 함께 출현한 단어나 저자의 빈도와 관계를 분석한다'가 된다.
⑤ O — 공저자분석은 논문을 함께 작성한 연구자들의 연결 관계를 이용하여 연구자 간 협력 관계와 협력 네트워크를 조사·분석한다.

오답 포인트

④의 동시인용은 이름에 '동시'가 들어가 동시출현분석과 혼동하기 쉽다. 그러나 단어·저자의 공동 출현을 세는 동시출현분석과, 후속 문헌에서 문헌·저자가 함께 인용된 횟수를 세는 동시인용분석을 구별해야 한다.

NoteBase

다시 나오면 이렇게 판단

무엇이 어디에 함께 나타나는지를 확인한다. 같은 문헌 안의 단어·저자이면 동시출현분석이고, 후속 문헌의 참고문헌에 함께 인용된 문헌·저자이면 동시인용분석이다.

암기 포인트

동시단어는 주제, 공저자는 협력, 동시인용은 인용 관계이다.

NoteBase

4. 텍스트 마이닝 분석기법으로 옳지 않은 것은?

① 감성분석
② 군집분석
③ 메타분석
④ 토픽모델링
⑤ 단어빈도분석

정답 3번

출제 포인트

텍스트 마이닝에 직접 활용되는 분석기법과 여러 연구 결과를 통계적으로 종합하는 메타분석을 구별하는 문제이다.

핵심 개념

텍스트 마이닝은 대량의 비정형 텍스트에서 단어의 빈도, 감성, 문서 간 유사성, 잠재 주제 등의 패턴과 정보를 추출하는 분석이다.
메타분석은 여러 독립 연구에서 얻은 효과크기 등 통계 결과를 종합하여 전체적인 결론을 도출하는 연구 통합 기법이다.

선지별 해설

① O — 감성분석은 텍스트에 나타난 긍정·부정 등의 감정이나 태도를 판별하므로 텍스트 마이닝 분석기법에 해당한다.
② O — 군집분석은 문서나 단어를 유사도에 따라 여러 집단으로 묶어 텍스트의 구조와 유형을 파악하는 데 활용된다.
③ X — '메타분석'은 텍스트 자체에서 패턴을 추출하는 분석기법이 아니다. 올바르게는 여러 독립 연구의 효과크기 등 통계 결과를 종합하는 연구 통합 기법이다.
④ O — 토픽모델링은 문서 집합에 함께 등장하는 단어의 분포를 이용하여 잠재된 주제를 찾아내는 텍스트 마이닝 분석기법이다.
⑤ O — 단어빈도분석은 단어나 표현의 출현 횟수를 계산하여 핵심어와 주요 경향을 파악하는 기초적인 텍스트 마이닝 분석기법이다.

오답 포인트

③의 '메타'라는 표현 때문에 대규모 텍스트를 분석하는 기법으로 오인하기 쉽다. 그러나 메타분석의 분석 대상은 개별 텍스트의 단어나 주제가 아니라 여러 연구에서 산출된 통계 결과이다.

NoteBase

다시 나오면 이렇게 판단

분석 대상이 문서·단어·문장이고 그 안의 빈도, 감성, 유사성, 주제를 찾으면 텍스트 마이닝으로 판단한다. 여러 연구의 효과크기나 통계 결과를 합치면 메타분석으로 판단한다.

암기 포인트

텍스트 마이닝은 텍스트 속 패턴 추출, 메타분석은 여러 연구 결과의 통계적 종합이다.

NoteBase

5. 전조합 색인과 후조합 색인에 대한 설명으로 옳지 않은 것은?

① 대부분의 컴퓨터 검색 시스템은 키워드를 사용한 후조합 색인을 채택하고 있다.

② 후조합 색인인 키워드 색인은 한 세트의 용어들로 색인어가 구성되는 색인 방식이다.

③ 전조합 색인은 색인어가 탐색 이전에 색인 작성자에 의하여 조합된 것이며, 열거식 분류가 대표적이다.

④ 전조합 색인은 복합 주제를 하나의 색인어로 표현하여 한 번의 검색으로 직접 접근할 수 있도록 한다.

⑤ 후조합 색인은 문헌의 내용을 여러 개의 단위 개념으로 색인하고, 탐색할 때 각 단위 개념을 조합하여 적합한 문헌을 검색할 수 있도록 하는 색인 방식이다.

정답 2번

출제 포인트

전조합 색인과 후조합 색인을 색인어의 조합 시점과 검색 방식으로 구별하고, 키워드 색인의 뜻을 정확히 아는지를 묻는 문제이다.

핵심 개념

전조합 색인은 색인 작성자가 탐색 전에 여러 개념을 미리 조합하여 복합 주제를 하나의 색인어로 표현하는 방식이다. 열거식 분류가 대표적이며, 미리 조합된 색인어로 직접 접근한다.
후조합 색인은 문헌의 주제를 여러 단위 개념으로 나누어 각각 색인하고, 검색할 때 이용자가 필요한 개념들을 조합하는 방식이다. 컴퓨터 검색에 적합하다.
키워드 색인은 표제나 본문에서 추출한 핵심어를 검색의 접근점으로 사용하는 색인이다.

NoteBase

선지별 해설

① O — 컴퓨터 검색 시스템에서는 검색 단계에서 여러 색인어를 자유롭게 조합할 수 있는 후조합 색인이 널리 활용된다.
② X — 틀린 부분은 '키워드 색인은 한 세트의 용어들로 색인어가 구성되는 색인 방식'이라는 설명이다. 키워드 색인은 표제나 본문에서 추출한 핵심어를 접근점으로 삼는 방식이며, 후조합 색인의 핵심은 여러 단위 개념을 독립적으로 색인한 뒤 검색할 때 조합하는 데 있다.
③ O — 전조합 색인은 탐색 전에 색인 작성자가 개념을 미리 조합한다. 열거식 분류는 복합 주제를 사전에 조합하여 표현하므로 대표적인 전조합 방식이다.
④ O — 전조합 색인은 복합 주제를 미리 하나의 조합된 색인어로 표현하므로, 그 색인어를 한 번 검색하여 해당 주제에 직접 접근할 수 있다.
⑤ O — 후조합 색인은 문헌의 내용을 여러 단위 개념으로 나누어 색인하고, 탐색 단계에서 필요한 단위 개념을 조합하여 적합한 문헌을 찾는 방식이다.

오답 포인트

②의 '키워드'라는 말만 보고 후조합 색인의 정의와 같은 것으로 판단하기 쉽다. 그러나 키워드 색인은 핵심어를 어디에서 얻어 접근점으로 삼는지를 나타내고, 전조합·후조합은 여러 개념을 언제 조합하는지를 나타낸다.

다시 나오면 이렇게 판단

조합 시점을 확인한다. 색인 작성 때 미리 조합하면 전조합이고, 단위 개념을 따로 색인한 뒤 검색할 때 조합하면 후조합이다. 키워드 색인은 표제나 본문에서 추출한 핵심어를 접근점으로 삼는 방식으로 판단한다.

암기 포인트

전조합은 작성 전에 미리 결합, 후조합은 검색할 때 나중에 결합, 키워드 색인은 표제·본문의 핵심어이다.

NoteBase

6. IFLA가 제안한 메타데이터의 핵심요소 중에서 FRBR의 검색, 식별, 선택, 획득 기능에 전부 사용되는 것만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. 날짜
ㄴ. 저자
ㄷ. 주제
ㄹ. 자원명
ㅁ. 자원식별자

① ㄱ, ㅁ
② ㄴ, ㄹ
③ ㄷ, ㅁ
④ ㄱ, ㄷ, ㅁ
⑤ ㄴ, ㄷ, ㄹ

정답 2번

출제 포인트

IFLA의 FRBR 이용자 과업인 검색·식별·선택·획득 네 기능에 모두 공통으로 사용되는 메타데이터 핵심요소를 구별하는 문제이다.

핵심 개념

FRBR의 이용자 과업은 원하는 자원을 찾는 검색, 찾은 자원이 원하는 것인지 확인하는 식별, 적합한 자원을 고르는 선택, 실제로 접근하거나 입수하는 획득으로 구분된다. 이 네 기능에 모두 사용되는 핵심요소는 저자와 자원명이다. 날짜·주제·자원식별자는 일부 과업을 지원하지만 네 기능 모두에 공통되는 요소는 아니다.

NoteBase

선지별 해설

① X — ㄱ 날짜와 ㅁ 자원식별자는 네 기능 모두에 공통으로 사용되는 요소가 아니다. 틀린 조합 'ㄱ, ㅁ'은 'ㄴ, ㄹ', 즉 저자와 자원명으로 고쳐야 한다.
② O — ㄴ 저자와 ㄹ 자원명은 검색·식별·선택·획득의 네 기능에 모두 사용되는 핵심요소이므로 옳다.
③ X — ㄷ 주제와 ㅁ 자원식별자는 네 기능 모두에 공통되는 조합이 아니다. 틀린 조합 'ㄷ, ㅁ'은 'ㄴ, ㄹ'로 고쳐야 한다.
④ X — ㄱ 날짜, ㄷ 주제, ㅁ 자원식별자는 모두 네 기능 공통요소가 아니며 저자와 자원명도 빠져 있다. 틀린 조합 'ㄱ, ㄷ, ㅁ'은 'ㄴ, ㄹ'로 고쳐야 한다.
⑤ X — ㄴ 저자와 ㄹ 자원명은 맞지만 ㄷ 주제는 네 기능 모두의 공통요소가 아니다. 틀린 조합 'ㄴ, ㄷ, ㄹ'에서는 ㄷ을 제외하여 'ㄴ, ㄹ'로 고쳐야 한다.

오답 포인트

주제와 자원식별자도 검색이나 식별 등에 중요한 요소이므로 모든 이용자 과업에 쓰인다고 오인하기 쉽다. 이 문항은 '어느 기능에 사용되는가'가 아니라 '네 기능 전부에 공통으로 사용되는가'를 묻는다는 점이 함정이다.

다시 나오면 이렇게 판단

FRBR 핵심요소 문제에서는 먼저 검색·식별·선택·획득의 네 과업을 확인한 뒤, 보기의 각 요소가 일부 과업이 아니라 네 과업 모두에 해당하는지를 판단한다. 네 기능 공통요소는 저자와 자원명이다.

암기 포인트

FRBR 네 기능 공통요소는 저자와 자원명이다.

NoteBase

7. OAIS 참조모델의 보존기술정보 유형으로 옳은 것만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. 맥락정보
ㄴ. 참조정보
ㄷ. 출처정보
ㄹ. 행위주체자정보
ㅁ. 관리메타데이터정보
ㅂ. 기술메타데이터정보

① ㄱ, ㄴ, ㄷ
② ㄱ, ㄷ, ㅁ
③ ㄱ, ㄹ, ㅁ
④ ㄴ, ㄷ, ㄹ, ㅂ
⑤ ㄴ, ㄷ, ㅁ, ㅂ

정답 1번

출제 포인트

OAIS 참조모델에서 보존기술정보(PDI)에 속하는 정보 유형과 일반적인 메타데이터 항목을 구별하는 문제이다.

핵심 개념

보존기술정보는 디지털 객체를 장기간 이해하고 신뢰하며 식별할 수 있도록 콘텐츠정보를 보완하는 정보이다.
주요 유형에는 참조정보, 출처정보, 맥락정보, 고정정보, 접근권한정보가 있다.
맥락정보: 콘텐츠정보가 다른 정보나 환경과 맺는 관계를 설명한다.
참조정보: 식별자 등 콘텐츠정보를 명확히 식별하는 정보를 제공한다.
출처정보: 콘텐츠정보의 기원, 이력, 소유·보관 경로와 보존 과정에서 일어난 변경을 설명한다.
행위주체자정보, 관리메타데이터정보, 기술메타데이터정보는 다른 메타데이터 체계에서 활용될 수 있으나, 이 문항이 묻는 OAIS 보존기술정보의 명명된 유형은 아니다.

NoteBase

선지별 해설

① O — ㄱ 맥락정보, ㄴ 참조정보, ㄷ 출처정보는 모두 OAIS 보존기술정보의 유형이므로 옳은 조합이다.
② X — ㄱ과 ㄷ은 맞지만 ㅁ 관리메타데이터정보가 잘못 포함되었고 ㄴ 참조정보가 빠졌다. 올바른 조합은 ㄱ, ㄴ, ㄷ이다.
③ X — ㄱ은 맞지만 ㄹ 행위주체자정보와 ㅁ 관리메타데이터정보는 이 분류의 유형이 아니다. ㄹ과 ㅁ을 ㄴ 참조정보와 ㄷ 출처정보로 고치면 ㄱ, ㄴ, ㄷ이 된다.
④ X — ㄴ과 ㄷ은 맞지만 ㄹ 행위주체자정보와 ㅂ 기술메타데이터정보가 잘못 포함되었으며 ㄱ 맥락정보가 빠졌다. 올바른 선택은 ㄱ, ㄴ, ㄷ이다.
⑤ X — ㄴ과 ㄷ은 맞지만 ㅁ 관리메타데이터정보와 ㅂ 기술메타데이터정보는 이 분류의 유형이 아니며 ㄱ 맥락정보가 빠졌다. 올바른 선택은 ㄱ, ㄴ, ㄷ이다.

오답 포인트

행위주체자정보, 관리메타데이터정보, 기술메타데이터정보도 기록관리에서 쓰일 법한 표현이어서 보존기술정보의 공식 유형처럼 보이는 점이 함정이다. 이 문항에서는 일반적인 메타데이터 항목이 아니라 OAIS가 명명한 보존기술정보 유형인지를 판단해야 한다.

다시 나오면 이렇게 판단

OAIS 보존기술정보 유형을 묻는 경우 참조·출처·맥락·고정·접근권한의 다섯 범주에 해당하는지를 먼저 확인한다. 보기 중 이 범주에 직접 해당하는 ㄱ, ㄴ, ㄷ만 남기면 된다.

암기 포인트

OAIS 보존기술정보는 '참·출·맥·고·접', 즉 참조·출처·맥락·고정·접근권한이다.

NoteBase

8. 한국과학기술정보연구원(KISTI)에서 운영하는 연구데이터 공유·관리 플랫폼은?

① DataON
② dCollection
③ OAK
④ RISS
⑤ ScienceON

정답 1번

출제 포인트

KISTI가 운영하는 여러 서비스 가운데 연구데이터의 공유·관리 기능에 특화된 플랫폼을 구별하는 문제이다.

핵심 개념

DataON은 KISTI가 구축·운영하는 국가연구데이터플랫폼으로, 연구데이터의 등록·관리·검색·공유·활용을 지원한다. 반면 ScienceON은 논문·특허·국가 R&D 보고서·연구데이터·분석 서비스·연구 인프라를 연계하는 과학기술 지식인프라 통합 서비스이다. dCollection과 RISS는 한국교육학술정보원(KERIS)의 학술정보 서비스이고, OAK는 국립중앙도서관이 추진하는 오픈액세스 지식정보 서비스이다.

선지별 해설

① O — DataON은 KISTI의 국가연구데이터플랫폼으로, 연구데이터의 등록·관리와 공유·활용을 지원하므로 정답이다.
② X — dCollection은 대학에서 생산한 학위논문 등 학술정보를 수집·유통하는 KERIS의 디지털 학술정보 유통체계이다. 틀린 표현은 KISTI의 연구데이터 공유·관리 플랫폼으로 본 부분이며, 올바르게는 DataON이다.
③ X — OAK(Open Access Korea)는 국립중앙도서관이 추진하는 오픈액세스 지식정보 서비스이다. 틀린 표현은 KISTI의 연구데이터 공유·관리 플랫폼으로 본 부분이며, 올바르게는 DataON이다.
④ X — RISS는 KERIS가 운영하는 학술연구정보서비스이다. 틀린 표현은 KISTI의 연구데이터 공유·관리 플랫폼으로 본 부분이며, 올바르게는 DataON이다.
⑤ X — ScienceON은 KISTI가 운영하지만, 논문·특허·보고서·연구데이터와 분석·연구 인프라를 한곳에 연계하는 과학기술 지식인프라 통합 서비스이다. 틀린 부분은 연구데이터 공유·관리에 특화된 플랫폼명으로 ScienceON을 고른 것이며, 올바르게는 DataON이다.

오답 포인트

ScienceON도 KISTI가 운영하고 연구데이터를 제공하므로 ⑤가 가장 강한 함정이다. 그러나 문장의 핵심은 단순한 운영 기관이 아니라 '연구데이터 공유·관리 플랫폼'이라는 서비스의 주된 기능이다.

NoteBase

다시 나오면 이렇게 판단

운영 기관과 핵심 기능을 함께 대응한다. KISTI와 연구데이터의 등록·관리·공유가 결합되면 DataON, KISTI와 과학기술 지식인프라의 연계·통합이 결합되면 ScienceON으로 판단한다.

암기 포인트

연구데이터는 DataON, 과학기술 지식인프라 통합은 ScienceON이다.

NoteBase

9. 섀넌(Shannon)의 정보이론에 대한 설명으로 옳은 것만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. 정보량과 엔트로피는 반비례한다.

ㄴ. 정보의 단위를 비트(bit)로 표기한다.

ㄷ. 전달되는 정보를 선택한 이후 불확실성은 강화된다.

ㄹ. 정보량이 많다는 것은 무작위성과 불확실성이 크다는 것을 의미한다.

ㅁ. 정보의 의미론적 측면을 배제하고 효율적인 정보 전송에 관한 문제에 집중한다.

① ㄱ, ㄴ, ㄷ

② ㄱ, ㄹ, ㅁ

③ ㄴ, ㄷ, ㄹ

④ ㄴ, ㄹ, ㅁ

⑤ ㄷ, ㄹ, ㅁ

정답 4번

출제 포인트

섀넌의 정보이론에서 정보량·엔트로피·불확실성의 관계와 정보의 단위 및 의미론 배제라는 특징을 구별하는 문제이다.

핵심 개념

정보량: 어떤 사건이 발생했다는 사실이 주는 놀라움의 크기이며, 발생확률이 낮을수록 커진다.
엔트로피: 정보원이 만들어 내는 평균 정보량이자 메시지를 알기 전 불확실성의 정도이다. 엔트로피가 클수록 무작위성과 불확실성이 크다.
정보의 수신: 전달된 정보를 알게 되면 불확실성이 감소한다.
비트(bit): 밑이 2인 로그를 사용할 때의 정보량 단위이다.
섀넌의 정보이론: 정보의 의미나 가치보다 부호화와 전송의 효율을 수학적으로 다룬다. 따라서 옳은 설명은 ㄴ, ㄹ, ㅁ이다.

NoteBase

선지별 해설

① X
ㄱ과 ㄷ이 틀렸다. 엔트로피는 평균 정보량이므로 정보량과 반비례한다고 할 수 없고, 정보를 선택하거나 수신하면 불확실성은 강화되는 것이 아니라 감소한다. 올바르게 고치면 ㄴ, ㄹ, ㅁ의 조합이다.

② X
ㄱ이 틀렸다. 정보량과 엔트로피는 반비례 관계가 아니라 엔트로피가 정보원의 평균 정보량을 나타내는 관계이다. ㄹ과 ㅁ은 옳지만, 올바른 조합은 ㄴ, ㄹ, ㅁ이다.

③ X
ㄷ이 틀렸다. 전달된 정보를 선택하여 알게 된 뒤에는 불확실성이 강화되는 것이 아니라 감소한다. ㄷ을 ㅁ으로 바꾸면 옳은 조합인 ㄴ, ㄹ, ㅁ이 된다.

④ O
ㄴ은 정보량의 단위가 비트라는 설명이고, ㄹ은 엔트로피가 클수록 무작위성과 불확실성이 크다는 설명이며, ㅁ은 섀넌의 정보이론이 의미론을 배제하고 효율적 정보 전송에 집중한다는 설명이므로 모두 옳다.

⑤ X
ㄷ이 틀렸다. 정보를 받은 뒤에는 불확실성이 감소한다. ㄹ과 ㅁ은 옳으며, ㄷ을 ㄴ으로 바꾸면 옳은 조합인 ㄴ, ㄹ, ㅁ이 된다.

오답 포인트

ㄱ의 '반비례'와 ㄷ의 '불확실성은 강화된다'가 함정이다. 엔트로피는 평균 정보량이며, 정보의 획득은 불확실성을 줄인다는 방향을 반대로 읽지 않아야 한다.

다시 나오면 이렇게 판단

'알기 전'에는 엔트로피가 불확실성의 크기를 나타내고, '정보를 받은 뒤'에는 불확실성이 감소한다고 판단한다. 단위가 비트이고 의미가 아닌 전송 효율을 다룬다는 설명도 섀넌 정보이론의 핵심 특징으로 판별한다.

암기 포인트

섀넌 정보이론은 비트로 측정하고, 엔트로피는 평균 정보량·불확실성이며, 정보 획득은 불확실성을 줄이고, 의미보다 전송을 다룬다.

NoteBase

10. 단어빈도와 역문헌빈도에 대한 설명으로 옳지 않은 것은?

① 색인어 가중치는 단어빈도와 역문헌빈도의 값을 조합하여 계산한다.

② 역문헌빈도는 문헌빈도가 낮은 단어일수록 높은 가중치가 부여된다.

③ 단어빈도는 문헌집단의 크기, 텍스트 길이 등을 고려하여 정규화한 빈도를 사용하기도 한다.

④ 단어빈도는 발생빈도가 높은 특정 단어가 해당 문헌의 주제를 잘 나타내는 것으로 판단하는 근거가 된다.

⑤ A라는 용어가 전체 문헌 5개 중에서 4개의 문헌에 출현할 때 계산식의 로그를 적용하지 않을 경우 A의 역문헌빈도는 0.8이다.

정답 5번

출제 포인트

이 문제는 단어빈도(TF)와 역문헌빈도(IDF)의 의미를 구별하고, 역문헌빈도에서 전체 문헌 수와 해당 단어가 출현한 문헌 수의 비율을 올바르게 계산할 수 있는지를 묻는다.

핵심 개념

단어빈도(TF)는 한 문헌 안에서 특정 단어가 얼마나 자주 나타나는지를 나타내며, 그 단어가 해당 문헌의 내용을 대표하는 정도를 평가하는 데 활용된다.

역문헌빈도(IDF)는 여러 문헌에 흔히 나타나는 단어의 가중치는 낮추고, 적은 수의 문헌에 나타나는 희소한 단어의 가중치는 높이는 값이다. 전체 문헌 수를 N, 해당 단어가 출현한 문헌 수를 df라고 하면 기본 형태는 N/df이고, 흔히 여기에 로그를 적용하여 $\log(N/df)$로 사용한다.

TF는 문헌 내부의 중요도를, IDF는 문헌집단 전체에서의 희소성을 반영하며, 두 값을 결합한 TF-IDF가 색인어 가중치로 널리 사용된다.

NoteBase

선지별 해설

① O — 색인어 가중치는 한 문헌 안의 출현 정도를 나타내는 TF와 문헌집단에서의 희소성을 나타내는 IDF를 결합하여 계산할 수 있다.
② O — 문헌빈도 df가 낮을수록 N/df가 커지므로 더 높은 역문헌빈도와 가중치가 부여된다.
③ O — 단순 출현 횟수는 문헌 길이 등의 영향을 받으므로, 문헌집단의 규모나 텍스트 길이 등을 고려해 정규화한 단어빈도를 사용하기도 한다.
④ O — 특정 단어가 한 문헌에서 자주 나타나면 그 문헌의 주제를 잘 나타낼 가능성이 있으므로, 단어빈도는 주제 대표성을 판단하는 근거가 된다.
⑤ X — 틀린 표현은 '역문헌빈도는 0.8이다'이다. 0.8은 해당 단어가 출현한 문헌의 비율 4/5일 뿐 역문헌빈도가 아니다. 로그를 적용하지 않는 기본 역문헌빈도는 $N/df = 5/4 = 1.25$이다. 올바르게는 'A의 역문헌빈도는 1.25이다'라고 해야 한다.

오답 포인트

⑤는 문헌빈도 비율 $df/N = 4/5 = 0.8$과 역문헌빈도 $N/df = 5/4 = 1.25$를 뒤바꾼 선지이다. '역'이라는 말에 맞게 전체 문헌 수를 해당 단어가 출현한 문헌 수로 나누어야 한다.

다시 나오면 이렇게 판단

TF는 '한 문헌 안에서 얼마나 자주 나오는가', IDF는 '전체 문헌 중 얼마나 드물게 나오는가'로 구별한다. 로그가 없는 IDF 계산에서는 N/df의 순서를 먼저 확인하며, df/N로 계산한 값은 역문헌빈도가 아니라 출현 문헌의 비율이다.

암기 포인트

TF는 문헌 안의 빈도, IDF는 문헌집단에서의 희소성이다. IDF의 기본 방향은 N/df이므로 드문 단어일수록 값이 커진다.

NoteBase

11. 자연어 색인과 통제어 색인에 대한 설명으로 옳은 것만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. 통제어 색인은 용어 사용의 모호성을 해결할 수 있다.

ㄴ. 통제어 색인은 상위어, 하위어 등 관련 개념에 대한 검색이 용이하다.

ㄷ. 자연어 색인은 관련 개념에 대한 검색이 누락되는 것을 최소화하기 쉽다.

ㄹ. 자연어 색인은 소셜 태깅(social tagging)과 같이 새로운 개념에 대한 표현이 용이하다.

① ㄱ, ㄴ

② ㄴ, ㄹ

③ ㄷ, ㄹ

④ ㄱ, ㄴ, ㄹ

⑤ ㄱ, ㄴ, ㄷ, ㄹ

정답 4번

출제 포인트

통제어 색인과 자연어 색인의 장단점을 동의어·계층 검색·신개념 표현의 측면에서 구분하는 문제이다.

핵심 개념

통제어 색인은 미리 정한 표준 용어로 문헌의 주제를 표현하는 방식이다. 동의어를 하나의 표준어로 모으고 상위어·하위어 같은 용어 관계를 구조화하므로 일관된 검색에 유리하다.

자연어 색인은 문헌에 실제로 쓰인 단어나 이용자가 자유롭게 붙인 표현을 그대로 활용하는 방식이다. 새로운 개념을 빠르게 반영하기 쉽지만, 동의어와 표현 차이를 통제하지 않으므로 관련 자료의 검색 누락이 생길 수 있다.

따라서 보기에서는 ㄱ, ㄴ, ㄹ이 옳고 ㄷ이 틀리다.

NoteBase

선지별 해설

① X — ㄱ과 ㄴ은 옳지만 옳은 ㄹ이 빠진 조합이다. ‘ㄱ, ㄴ’은 ‘ㄱ, ㄴ, ㄹ’로 고쳐야 한다.
② X — ㄴ과 ㄹ은 옳지만 옳은 ㄱ이 빠진 조합이다. ‘ㄴ, ㄹ’은 ‘ㄱ, ㄴ, ㄹ’로 고쳐야 한다.
③ X — ㄷ은 틀리고, 옳은 ㄱ과 ㄴ이 빠져 있다. ‘ㄷ, ㄹ’은 ‘ㄱ, ㄴ, ㄹ’로 고쳐야 한다. 자연어 색인은 표현의 다양성 때문에 관련 개념의 검색 누락을 최소화하기 어렵다.
④ O — ㄱ, ㄴ, ㄹ은 모두 옳다. 통제어 색인은 동의어를 통제하고 계층 관계 검색을 지원하며, 자연어 색인은 소셜 태깅처럼 새 개념을 자유롭게 표현하기 쉽다.
⑤ X — ㄷ까지 옳다고 본 부분이 틀리다. ‘ㄱ, ㄴ, ㄷ, ㄹ’은 ‘ㄱ, ㄴ, ㄹ’로 고쳐야 한다.

오답 포인트

ㄷ의 ‘검색 누락을 최소화’라는 표현이 함정이다. 자연어 색인은 최신 표현을 잘 받아들이지만, 같은 개념이 여러 동의어나 변형된 표현으로 나타나면 검색어와 일치하지 않는 자료가 누락될 수 있다. 신개념 표현의 용이성과 검색 일관성은 같은 장점이 아니다.

다시 나오면 이렇게 판단

동의어 통제·상위어와 하위어의 관계·일관된 검색은 통제어 색인의 장점으로 판단한다. 최신 용어·자유로운 표현·소셜 태깅은 자연어 색인의 장점으로 판단하되, 동의어 때문에 생기는 검색 누락은 자연어 색인의 약점으로 판단한다.

암기 포인트

통제어는 ‘표준화와 관계 검색’, 자연어는 ‘자유로운 신개념 표현과 표현 불일치 위험’이다.

NoteBase

12. 다음의 계량서지법칙에 대한 설명으로 옳지 않은 것은?

> n편의 논문을 생산한 저자의 수는 한 편의 논문을 생산한 저자의 수의 약 $1/n^2$이다. 그리고 한 편의 논문을 생산한 저자의 수는 논문을 생산한 모든 저자의 수의 약 60%이다.

① 로트카(Lotka)의 역제곱 법칙으로도 불린다.

② 1926년 로트카가 화학자와 물리학자의 논문 건수를 조사하여 분석하였다.

③ 다양한 학문 분야에서 저자의 생산성을 분석하고 예측하는 데 활용되고 있다.

④ 학문 분야나 연구 방식에 따라 위 법칙의 적용 가능성에 대해 다양한 의견이 존재한다.

⑤ 위 법칙에 따르면 전체 논문 생산 저자의 수가 2,000명인 경우 2편의 논문을 생산한 저자의 수는 약 500명이다.

정답 5번

출제 포인트

로트카의 역제곱 법칙을 이해하고, 전체 저자 수로부터 특정 논문 편수를 생산한 저자 수를 계산하는 문제이다.

조건 정리

주어진 값: 전체 논문 생산 저자는 2,000명이고, 1편의 논문을 생산한 저자는 전체 저자의 약 60%이다.

구해야 할 값: 2편의 논문을 생산한 저자 수이다.

주의할 조건: 2편 저자 수는 전체 저자 수의 $1/2^2$이 아니라, 1편 저자 수의 $1/2^2$이다.

사용할 공식

$A_n = \frac{A_1}{n^2}$

여기서 A_1은 1편을 생산한 저자 수이고, A_n은 n편을 생산한 저자 수이다. 제시문이 n편 저자 수를 1편 저자 수의 약 $1/n^2$로 제시하므로 이 공식을 사용한다.

계산 과정

1. 1편을 생산한 저자 수는 $2{,}000 \times 0.6 = 1{,}200$명이다.
2. 2편을 생산한 저자 수는 $A_2 = 1{,}200/2^2$이다.
3. 따라서 $A_2 = 1{,}200/4 = 300$명이므로 약 500명이라는 설명은 틀리다.

NoteBase

선지 판단

① O — 논문 생산량이 n편인 저자 수가 1편인 저자 수에 비해 $1/n^2$로 감소하므로 로트카의 역제곱 법칙이라고 한다.
② O — 로트카는 1926년에 화학자와 물리학자의 논문 생산 건수를 조사·분석하여 이 법칙을 제시하였다.
③ O — 로트카의 법칙은 저자 생산성의 분포를 설명하는 계량서지 법칙이므로 여러 학문 분야에서 저자의 생산성을 분석하고 예측하는 데 활용된다.
④ O — 학문 분야의 출판 관행과 연구 방식이 서로 다르므로 역제곱 관계가 어느 분야에나 동일하게 적용되는지에 관해서는 다양한 견해가 존재한다.
⑤ X — '2편의 논문을 생산한 저자의 수는 약 500명'이 틀렸다. 1편 저자는 전체의 60%인 1,200명이고, 2편 저자는 그 수의 $1/2^2$이므로 올바르게는 약 300명이다.

오답 포인트

⑤에서 60%를 적용해 구한 1,200명을 기준으로 다시 역제곱 비율을 적용해야 한다. 전체 2,000명에 임의의 비율을 곱하거나, $1/2^2$의 분모를 2로만 계산하면 오답이 된다.

다시 나오면 이렇게 판단

먼저 전체 저자 수에 60%를 적용해 A_1을 구한 뒤, 목표 논문 편수를 n으로 놓고 $A_n = A_1/n^2$를 적용한다. 마지막으로 문제가 옳은 것을 묻는지 옳지 않은 것을 묻는지 확인한다.

암기 포인트

로트카 법칙은 '1편 저자가 약 60%, n편 저자는 1편 저자의 $1/n^2$'로 정리된다.

NoteBase

13. OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting)에 대한 설명으로 옳지 않은 것은?

① SP(Service Provider)와 DP(Data Provider) 사이에서 작동한다.

② MARC가 아닌 더블린 코어를 채택하고 있어 상대적으로 저비용이다.

③ 메타데이터를 수확하지 않는 분산검색 방식에 비해 최신의 정보를 검색할 수 있다.

④ HTTP 프로토콜을 이용한 요청과 XML 레코드의 응답이라는 단순한 형식으로 커뮤니케이션한다.

⑤ 통합검색을 위해 수확검색 방식을 사용하고 있으며, 기술 장벽이 낮고 저장소 간 정보를 쉽게 공유할 수 있다.

정답 3번

출제 포인트

OAI-PMH의 메타데이터 수확 구조와 분산검색 방식의 차이를 묻는 문제이다.

핵심 개념

OAI-PMH는 데이터 제공자(DP)가 공개한 메타데이터를 서비스 제공자(SP)가 주기적으로 수확하여 통합 서비스에 활용하도록 하는 상호운용 프로토콜이다.
수확검색은 메타데이터를 미리 모아 둔 뒤 검색하므로 빠른 통합검색이 가능하지만, 수확 주기 사이의 변경 사항은 즉시 반영되지 않을 수 있다. 반면 분산검색은 검색 시점에 여러 저장소에 질의하므로 최신 정보 반영에는 상대적으로 유리하지만 응답 지연과 저장소별 장애의 영향을 받기 쉽다.

선지별 해설

① O — OAI-PMH에서는 DP가 메타데이터를 제공하고 SP가 이를 수확하여 서비스를 구성하므로 두 주체 사이에서 작동한다.
② O — OAI-PMH는 상호운용을 위한 기본 메타데이터 형식으로 MARC가 아니라 단순 더블린 코어를 지원하며, 구조와 구현이 비교적 단순하여 도입 비용이 낮다.
③ X — 틀린 표현은 '메타데이터를 수확하지 않는 분산검색 방식에 비해 최신의 정보를 검색할 수 있다'이다. 수확검색은 마지막 수확 이후의 변경 사항이 다음 수확 때까지 반영되지 않을 수 있으므로, 최신성은 검색 시점에 각 저장소에 직접 질의하는 분산검색 방식이 상대적으로 유리하다고 고쳐야 한다.
④ O — OAI-PMH는 HTTP를 통해 요청하고 XML 형식의 응답을 주고받는 비교적 단순한 통신 구조를 사용한다.
⑤ O — 여러 저장소의 메타데이터를 수확해 한곳에서 검색하게 하므로 통합검색에 적합하고, 프로토콜의 기술적 장벽이 비교적 낮아 저장소 간 정보를 공유하기 쉽다.

NoteBase

오답 포인트

③은 수확검색이 미리 모은 메타데이터를 대상으로 빠르게 검색된다는 장점을 '최신성'의 장점으로 바꿔 놓았다. 검색 속도와 정보의 최신성은 서로 다른 기준이다.

다시 나오면 이렇게 판단

'미리 수확하여 저장'이라는 표현이 보이면 통합검색과 검색 효율은 장점이지만 실시간 최신성은 약점이라고 판단한다. '검색할 때 각 저장소에 직접 질의'하면 최신성은 높을 수 있지만 속도와 안정성은 불리할 수 있다고 구분한다.

암기 포인트

OAI-PMH는 DP의 메타데이터를 SP가 수확한다. 수확검색은 통합·효율, 분산검색은 최신성에 상대적으로 유리하다.

NoteBase

14. 계량서지 지표에 대한 설명으로 옳지 않은 것은?

① CNCI(Category Normalized Citation Impact)는 실제 인용 횟수를 동일한 문서 유형, 출판 연도, 주제 영역에 속하는 문서의 예상 인용률로 나누어 정규화한 지수이다.

② Eigenfactor는 하나의 학술지가 다른 학술지에 비해 상대적으로 어느 정도의 영향력을 가지는지 평가하는 데 사용된다.

③ JIF(Journal Impact Factor)는 일정 기간 특정 학술지에 게재된 논문의 평균 피인용 횟수를 의미한다.

④ SJR(SCImago Journal Rankings)은 SCOPUS 내 정보를 활용한 학술지 영향력 지수이다.

⑤ SNIP(Source Normalized Impact per Paper)는 학술 논문이 소셜 미디어, 언론 보도 등에 노출된 정도를 기반으로 계산되는 사회적 영향력 지수이다.

정답 5번

출제 포인트

이 문제는 CNCI, Eigenfactor, JIF, SJR, SNIP이 각각 무엇을 측정하고 어떤 자료를 이용하는지 구별하는 문제이다.

핵심 개념

CNCI는 논문의 실제 피인용 횟수를 같은 문서 유형·출판 연도·주제 분야의 기대 피인용 횟수와 비교한 정규화 지수이다.
Eigenfactor는 학술지 간 인용망에서 인용의 출처와 영향력까지 반영하여 학술지의 전체 영향력을 평가한다.
JIF는 일정 기간 해당 학술지에 실린 논문이 받은 평균 피인용 횟수를 나타낸다.
SJR은 Scopus의 인용 정보를 바탕으로 인용한 학술지의 위신까지 가중하여 학술지 영향력을 평가한다.
SNIP는 학문 분야마다 다른 인용 관행과 인용 가능성을 보정한 논문당 영향력 지수이다.

NoteBase

선지별 해설

① O — CNCI는 실제 피인용 횟수를 동일한 문서 유형·출판 연도·주제 분야의 기대 피인용 횟수로 나누어 분야와 조건의 차이를 정규화한다.
② O — Eigenfactor는 학술지 인용 네트워크를 이용하여 한 학술지가 다른 학술지들과 비교해 지니는 상대적 영향력을 평가한다.
③ O — JIF는 정해진 기간에 특정 학술지의 논문들이 받은 피인용 횟수를 해당 논문 수로 나눈 평균 피인용 지표이다.
④ O — SJR은 Scopus 수록 정보를 활용하며, 인용 횟수뿐 아니라 인용한 학술지의 위신도 반영하는 학술지 영향력 지수이다.
⑤ X — 틀린 표현은 '소셜 미디어, 언론 보도 등에 노출된 정도를 기반으로 계산되는 사회적 영향력 지수'이다. SNIP는 소셜 미디어 노출도가 아니라 해당 주제 분야의 인용 가능성과 인용 관행을 보정하여 계산하는 논문당 정규화 인용 영향력 지수이다. 소셜 미디어·언론 보도 등의 관심도를 반영하는 것은 알트메트릭스 계열 지표에 가깝다.

오답 포인트

⑤는 SNIP의 'Source Normalized'를 사회적 노출의 정규화로 바꾸어 놓은 함정이다. SNIP의 정규화 대상은 소셜 미디어 반응이 아니라 학문 분야별 인용 환경이다.

다시 나오면 이렇게 판단

지표가 무엇을 비교하고 어떤 자료를 쓰는지 확인한다. CNCI는 개별 논문의 기대 인용 대비 실제 인용, Eigenfactor는 학술지 인용망의 전체 영향력, JIF는 학술지 논문의 평균 피인용, SJR은 Scopus 기반 위신 가중, SNIP는 분야별 인용 관행 보정으로 판단한다. 소셜 미디어·뉴스 노출이 나오면 알트메트릭스를 떠올린다.

암기 포인트

CNCI는 기대 대비 실제, Eigenfactor는 인용망, JIF는 평균 인용, SJR은 Scopus·위신, SNIP는 분야 보정, 사회적 관심은 알트메트릭스이다.

NoteBase

15. 다음은 단어빈도를 제시한 문헌과 용어의 행렬이다. 주어진 불리언 질의에 따라 검색되는 문헌을 모두 고르면?

문헌 \ 용어	data	media	literacy	reading
d1	3	4	2	4
d2	2	2	1	0
d3	1	0	3	0
d4	0	1	0	2

불리언 질의

(data AND media) OR (literacy AND (NOT reading))

① d1, d2

② d2, d3

③ d3, d4

④ d1, d2, d3

⑤ d1, d2, d4

정답 4번

출제 포인트

이 문제는 문헌별 단어빈도 표에 불리언 연산자 AND, OR, NOT을 적용하여 검색되는 문헌의 합집합을 구하는 문제이다.

자료 읽는 순서

1. (data AND media)는 data와 media의 빈도가 모두 1 이상인 문헌을 찾는다.
2. (literacy AND (NOT reading))은 literacy의 빈도가 1 이상이고 reading의 빈도가 0인 문헌을 찾는다.
3. 두 조건은 OR로 연결되므로 어느 한쪽이라도 만족하는 문헌을 모두 합친다.

핵심 근거

첫째 조건 (data AND media)는 data와 media가 각각 3, 4인 d1과 2, 2인 d2가 만족한다. 둘째 조건 (literacy AND (NOT reading))은 literacy가 존재하면서 reading이 0인 d2와 d3가 만족한다. 두 결과의 합집합은 d1, d2, d3이다. d4는 data가 0이고 literacy도 0이므로 어느 조건도 만족하지 않는다.

NoteBase

선지별 해설

① X — d1과 d2는 검색되지만, 표에서 literacy가 3이고 reading이 0인 d3도 둘째 조건을 만족한다. ‘d1, d2’가 아니라 ‘d1, d2, d3’이 올바른 결과이다.
② X — d2와 d3은 둘째 조건을 만족하지만, 표에서 data가 3이고 media가 4인 d1도 첫째 조건을 만족한다. ‘d2, d3’이 아니라 ‘d1, d2, d3’이 올바른 결과이다.
③ X — d3은 둘째 조건을 만족하지만 d4는 data가 0이고 literacy가 0이어서 검색되지 않는다. 또한 d1과 d2가 빠져 있다. ‘d3, d4’가 아니라 ‘d1, d2, d3’이 올바른 결과이다.
④ O — d1과 d2는 (data AND media)를 만족하고, d2와 d3은 (literacy AND (NOT reading))을 만족한다. OR로 합치면 d1, d2, d3이 된다.
⑤ X — d4는 data가 0이고 literacy가 0이므로 두 조건을 모두 만족하지 않으며, 둘째 조건을 만족하는 d3이 빠져 있다. ‘d1, d2, d4’가 아니라 ‘d1, d2, d3’이 올바른 결과이다.

오답 포인트

NOT reading을 ‘reading을 제외한 모든 문헌’으로만 보고 literacy 조건을 함께 확인하지 않거나, OR를 두 검색 결과의 공통부분으로 잘못 처리하기 쉽다. NOT reading은 reading의 빈도가 0이라는 뜻이고, 최종 OR는 두 조건의 합집합이다.

다시 나오면 이렇게 판단

괄호 안의 AND 조건을 각각 따로 계산한 뒤 OR로 합집합을 만든다. 단어빈도 표에서는 해당 용어의 빈도가 1 이상이면 존재하고, NOT이 붙은 용어의 빈도는 0이어야 한다.

암기 포인트

AND는 교집합, OR는 합집합, NOT 용어는 빈도 0으로 판단한다.

NoteBase

16. 정보 탐색전략에 대한 설명으로 옳지 않은 것은?

① 블록 만들기 방법을 사용하여 복잡한 탐색 과제를 단순하게 분해할 수 있다.

② 특정 패싯 우선탐색 방법은 다수의 개념을 포함하는 탐색에 적용되므로 초보자에게 유용하다.

③ 눈덩이 굴리기 방법은 탐색 결과가 양적으로 증가하는 것을 의미하며 재현율 향상을 목적으로 한다.

④ 점진 분할 방법은 넓은 개념으로 탐색을 시작한 후에 다양한 제한 기법을 사용하여 탐색 범위를 좁혀가는 방법이다.

⑤ 신속/편의 지향법은 탐색기능을 사용하지 않고 탐색하려는 용어 하나를 입력하거나 스페이스로 구분한 몇 개의 용어를 한꺼번에 입력하여 검색하는 방법이다.

정답 2번

출제 포인트

정보 탐색전략별 탐색의 출발점, 검색 범위의 변화, 적합한 이용자라는 차이를 구별하는 문제이다.

핵심 개념

블록 만들기는 주제를 여러 개념 블록으로 나누고 각 블록의 동의어나 관련어를 구성하여 결합하는 전략이다. 특정 패싯 우선탐색은 가장 구체적이고 식별력이 높은 패싯부터 검색한 뒤 필요에 따라 다른 패싯을 추가하는 전략이다. 눈덩이 굴리기는 이미 찾은 적합 문헌의 참고문헌·인용 관계 등을 따라 관련 문헌을 확장하는 전략이다. 점진 분할은 넓게 시작한 검색에 조건을 차례로 더하여 결과 범위를 좁히는 전략이다. 신속·편의 지향법은 복잡한 탐색기능 없이 자연어식으로 하나 또는 몇 개의 검색어를 바로 입력하는 간편 탐색이다.

NoteBase

선지별 해설

① O — 블록 만들기는 복잡한 탐색 과제를 개념별 블록으로 나누어 각 블록의 용어를 구성한 뒤 결합하므로, 복잡한 과제를 체계적으로 단순화할 수 있다.

② X — 틀린 표현은 '다수의 개념을 포함하는 탐색에 적용되므로 초보자에게 유용하다'이다. 특정 패싯 우선탐색은 여러 개념을 모두 블록화하는 방식이 아니라 가장 구체적이고 검색 결과를 잘 제한하는 패싯을 먼저 골라야 하므로, 주제와 검색 시스템에 익숙한 탐색자에게 적합하다. 올바르게는 '가장 구체적인 패싯부터 검색하고 필요에 따라 다른 패싯을 추가하는 전략으로, 특정 패싯의 식별력을 판단할 수 있는 숙련된 탐색자에게 유용하다'이다.

③ O — 눈덩이 굴리기는 한 건의 적합 문헌에서 참고문헌·인용 문헌·관련 문헌으로 탐색을 계속 확장하므로 검색 결과가 늘어나며 재현율 향상을 지향한다.

④ O — 점진 분할은 넓은 개념으로 시작한 뒤 검색어, 연산자, 제한 조건 등을 단계적으로 추가하여 결과 집합과 탐색 범위를 좁혀 가는 방식이다.

⑤ O — 신속·편의 지향법은 고급 탐색기능을 세밀하게 사용하지 않고 하나의 검색어 또는 여러 검색어를 한꺼번에 입력하여 빠르고 간편하게 검색하는 방식이다.

오답 포인트

②는 '다수의 개념'이라는 말 때문에 블록 만들기와 혼동하게 한다. 특정 패싯 우선탐색의 핵심은 개념의 수가 아니라 여러 패싯 가운데 검색을 가장 강하게 제한할 패싯을 먼저 선택하는 것이며, 이 판단은 초보자보다 숙련자에게 더 적합하다.

다시 나오면 이렇게 판단

전략 이름을 탐색 결과의 변화와 연결한다. 개념별 묶음은 블록 만들기, 가장 구체적인 면부터 시작하면 특정 패싯 우선탐색, 관련 문헌을 따라 확장하면 눈덩이 굴리기, 넓게 시작해 조건을 추가하며 축소하면 점진 분할, 검색어만 곧바로 입력하면 신속·편의 지향법이다.

암기 포인트

블록은 나누어 결합하고, 특정 패싯은 구체적인 것부터 시작하며, 눈덩이는 넓히고, 점진 분할은 좁히고, 신속·편의는 바로 입력한다.

NoteBase

17. 벡터공간모델에 대한 설명으로 옳은 것만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. 문헌과 질의에 나타난 용어를 벡터의 차원으로 표현한다.

ㄴ. 불리언모델이 가지는 이진 가중치에 대한 대안으로 나온 검색모델이다.

ㄷ. 설튼(Salton)이 대수론(algebra theory)을 이론적 근거로 하여 만든 모델이다.

ㄹ. 벡터 간 사이각의 코사인 값이 클수록 벡터 간 유사도가 크다는 것을 나타낸다.

ㅁ. 문헌의 색인어와 질의어가 각각 가중치를 가지며, 질의와 부분적으로 일치하는 문헌을 검색할 수 있다.

① ㄱ, ㄴ, ㄹ

② ㄱ, ㄹ, ㅁ

③ ㄴ, ㄷ, ㄹ

④ ㄱ, ㄴ, ㄷ, ㅁ

⑤ ㄱ, ㄴ, ㄷ, ㄹ, ㅁ

정답 5번

출제 포인트

벡터공간모델의 표현 방식, 이론적 배경, 유사도 계산, 가중치와 부분일치 검색의 특성을 종합하여 옳은 진술을 모두 가려내는 문제이다.

핵심 개념

벡터공간모델은 문헌과 질의를 색인어로 이루어진 다차원 공간의 벡터로 표현하고, 두 벡터의 방향이 얼마나 가까운지를 코사인 유사도로 계산하여 검색 결과의 관련도를 순위화하는 정보검색모델이다. 각 색인어와 질의어에는 중요도를 반영한 가중치를 줄 수 있으므로, 질의와 일부 용어만 일치하는 문헌도 유사도에 따라 검색할 수 있다. 불리언모델이 정확한 논리 일치를 중심으로 검색하는 방식이라면, 벡터공간모델은 가중치와 유사도를 이용해 부분일치와 순위검색을 지원한다.

NoteBase

선지별 해설

① X
ㄱ·ㄴ·ㄹ만 고른 조합으로, 옳은 ㄷ과 ㅁ이 빠져 있다. ㄷ은 벡터공간모델이 솔턴에 의해 대수론을 이론적 근거로 제안된 모델이라는 설명이고, ㅁ은 색인어와 질의어의 가중치를 이용하여 부분적으로 일치하는 문헌도 검색할 수 있다는 설명이므로 모두 포함해야 한다.

② X
ㄱ·ㄹ·ㅁ만 고른 조합으로, 옳은 ㄴ과 ㄷ이 빠져 있다. 벡터공간모델은 불리언모델의 이진적·정확일치 검색을 보완하는 대안으로 제안되었으며, 솔턴이 대수론을 이론적 근거로 만든 모델이므로 ㄴ과 ㄷ도 포함해야 한다.

③ X
ㄴ·ㄷ·ㄹ만 고른 조합으로, 옳은 ㄱ과 ㅁ이 빠져 있다. 벡터공간모델은 문헌과 질의의 용어를 벡터의 차원으로 표현하고, 색인어·질의어 가중치로 부분일치 문헌도 검색하므로 ㄱ과 ㅁ도 포함해야 한다.

④ X
ㄱ·ㄴ·ㄷ·ㅁ만 고른 조합으로, 옳은 ㄹ이 빠져 있다. 코사인 값은 두 벡터가 이루는 각의 코사인으로서 값이 클수록 방향이 가까우므로 문헌 벡터와 질의 벡터의 유사도가 크다는 뜻이다. 따라서 ㄹ도 포함해야 한다.

⑤ O
ㄱ은 벡터 표현, ㄴ은 불리언모델의 대안이라는 등장 배경, ㄷ은 솔턴과 대수론이라는 이론적 기반, ㄹ은 코사인 유사도의 해석, ㅁ은 가중치와 부분일치 검색을 각각 바르게 설명한다. 따라서 ㄱ·ㄴ·ㄷ·ㄹ·ㅁ을 모두 고른 ⑤가 정답이다.

오답 포인트

ㄹ의 코사인 값을 벡터 사이의 각도와 혼동하면 틀리기 쉽다. 벡터 사이의 각도는 작을수록 유사하지만, 그 각의 코사인 값은 클수록 유사하다. 또한 벡터공간모델은 단순히 용어를 벡터로 바꾸는 데 그치지 않고, 용어 가중치와 유사도 계산을 통해 부분일치 문헌까지 순위화한다.

다시 나오면 이렇게 판단

벡터공간모델의 선지는 '문헌·질의의 벡터화 → 용어 가중치 부여 → 코사인 유사도 계산 → 부분일치 문헌의 순위화' 순서로 연결하여 판단한다. 코사인 값이 클수록 유사하고, 불리언모델과 달리 일치 여부만을 0과 1로 단절하여 판단하지 않는다는 점을 확인한다.

암기 포인트

벡터공간모델은 '가중치 벡터, 코사인 유사도, 부분일치 순위검색'으로 기억한다.

NoteBase

18. 기계학습 방법에 대한 설명으로 옳지 않은 것은?

① 비지도학습은 입력 값만을 활용하여 학습하는 방식이다.
② 지도학습은 입력 값과 출력 값을 함께 활용하여 학습하는 방식이다.
③ 딥러닝은 인공신경망 구조를 활용하여 데이터 표현을 학습한다.
④ 클러스터링은 동질적인 데이터를 하나의 그룹으로 묶는 대표적인 지도학습 방법이다.
⑤ 강화학습은 현재의 상태에서 누적 보상이 최대가 되는 행동을 하도록 하는 학습 방법이다.

정답 4번

출제 포인트

지도학습·비지도학습·딥러닝·클러스터링·강화학습의 정의와 서로의 포함 관계를 구별하는 문제이다.

핵심 개념

지도학습은 입력과 정답에 해당하는 출력(레이블)의 관계를 학습하는 방식이다. 비지도학습은 정답 레이블 없이 입력 데이터의 구조나 패턴을 찾는 방식이다. 클러스터링은 유사한 데이터를 같은 집단으로 묶는 대표적인 비지도학습 방법이다. 딥러닝은 여러 층으로 이루어진 인공신경망을 이용해 데이터의 특징과 표현을 학습한다. 강화학습은 상태에서 행동을 선택하고 받은 보상을 바탕으로 누적 보상을 최대화하는 정책을 학습한다.

선지별 해설

① O — 비지도학습은 정답 레이블 없이 입력 데이터 자체에서 숨은 구조나 패턴을 학습하므로 옳다.
② O — 지도학습은 입력 값과 이에 대응하는 출력 값, 즉 정답 레이블을 함께 이용해 예측 관계를 학습하므로 옳다.
③ O — 딥러닝은 다층 인공신경망 구조를 활용하여 데이터의 특징과 표현을 계층적으로 학습하므로 옳다.
④ X — 틀린 표현은 '대표적인 지도학습 방법'이다. 클러스터링은 데이터의 유사성을 기준으로 동질적인 대상을 하나의 그룹으로 묶는 대표적인 비지도학습 방법이다.
⑤ O — 강화학습은 현재 상태에서 행동을 선택하고 환경에서 받은 보상을 바탕으로 장기적인 누적 보상이 최대가 되도록 학습하므로 옳다.

오답 포인트

④는 '동질적인 데이터를 하나의 그룹으로 묶는다'는 클러스터링의 기능을 정확히 제시한 뒤, 이를 '지도학습'이라고 잘못 분류했다. 기능 설명이 맞더라도 학습 유형까지 확인해야 한다.

NoteBase

다시 나오면 이렇게 판단

정답 레이블이 있으면 지도학습, 레이블 없이 데이터의 구조를 찾으면 비지도학습, 유사한 대상을 묶으면 클러스터링, 행동과 보상의 상호작용으로 정책을 배우면 강화학습으로 판단한다. 인공신경망의 여러 층을 통한 표현 학습이 강조되면 딥러닝이다.

암기 포인트

지도학습은 입력+정답, 비지도학습은 입력만, 클러스터링은 비지도학습, 강화학습은 행동+보상이다.

NoteBase

19. 정보학의 이론과 기술 발전에 영향을 준 학자의 주요 업적에 대한 설명으로 옳지 않은 것은?

① 토브(Taube)는 유니텀에 불리언 논리를 결합하여 정보검색에 적용하는 기초를 마련하였다.

② 스팍 존스(Spärck Jones)는 단어의 가중치 방법으로 역문헌빈도를 이용하는 방식을 제안하였다.

③ 스몰(Small)은 텍스트에 출현한 단어의 통계적 특성에 의해 색인어를 선정하는 방식을 처음 제시하였다.

④ 설튼(Salton)은 SMART Information Retrieval System을 이용하여 벡터공간모델, 적합성 피드백 등을 실험하였다.

⑤ 가필드(Garfield)는 인용색인 기반의 학술논문 정보서비스 기관인 ISI(Institute for Scientific Information)를 설립하였으며, 계량서지학과 계량과학의 창시자 중 한 사람이다.

정답 3번

출제 포인트

정보검색과 계량정보학의 주요 학자를 대표 업적과 정확히 연결하는 문제이다. 특히 단어의 출현 빈도를 이용한 자동색인 연구는 스몰이 아니라 룬의 업적이라는 점을 구별해야 한다.

핵심 정리

토브: 유니텀과 조합색인 — 단일 개념어인 유니텀을 조합하여 문헌을 검색하는 좌표색인 체계를 발전시켰다.
스팍 존스: 역문헌빈도 — 여러 문헌에 흔한 단어의 가중치는 낮추고 드문 단어의 가중치는 높이는 방식을 제안하였다.
룬: 단어 빈도와 자동색인 — 문헌에 나타난 단어의 통계적 출현 특성을 이용하여 색인어를 선정하는 방식을 제시하였다.
스몰: 동시인용 분석 — 두 문헌이 후속 문헌들에서 함께 인용되는 정도를 이용하여 문헌 간 주제적 관련성을 분석하였다.
솔튼: SMART 시스템 — 벡터공간모델과 적합성 피드백 등 현대 정보검색의 핵심 기법을 실험하였다.
가필드: 인용색인과 ISI — 과학 문헌의 인용 관계를 검색·분석하는 체계를 발전시키고 ISI를 설립하였다.

NoteBase

선지별 해설

① O — 토브는 유니텀을 조합하는 좌표색인 체계를 제안하였고, 검색어의 논리적 결합을 정보검색에 적용하는 기반을 마련하였다.
② O — 스팍 존스는 한 문헌에 나온 단어가 전체 문헌집합에서 얼마나 드문지를 반영하는 역문헌빈도를 단어 가중치에 활용하였다.
③ X — 틀린 부분은 '스몰이 단어의 통계적 특성에 의해 색인어를 선정하는 방식을 처음 제시하였다'는 연결이다. 올바르게는 룬이 단어의 출현 빈도 등 통계적 특성을 이용한 자동색인 방식을 제시하였으며, 스몰은 동시인용 분석과 연결된다.
④ O — 솔튼은 SMART Information Retrieval System을 통해 벡터공간모델, 적합성 피드백 등 정보검색 기법을 개발하고 실험하였다.
⑤ O — 가필드는 인용색인을 기반으로 학술정보 서비스를 제공한 ISI를 설립하였고, 인용 분석과 계량서지학·계량과학의 발전에 크게 기여하였다.

출제자가 노린 함정

③은 '통계적 특성'이라는 말이 계량정보학자인 스몰과 어울려 보이도록 만든 선지이다. 그러나 단어의 출현 빈도를 이용한 색인어 선정은 룬, 문헌이 함께 인용되는 관계의 분석은 스몰로 구분해야 한다.

다시 나오면 이렇게 판단

학자 이름을 보면 연구 대상과 대표 키워드를 한 쌍으로 확인한다. 단어 빈도는 룬, 역문헌빈도는 스팍 존스, 동시인용은 스몰, 벡터공간모델과 SMART는 솔튼, 인용색인과 ISI는 가필드로 연결한다.

암기 포인트

룬은 단어 빈도, 스팍 존스는 역문헌빈도, 스몰은 동시인용, 솔튼은 SMART, 가필드는 인용색인·ISI이다.

NoteBase

20. 다음은 동일한 문헌 집단을 대상으로 질의 Q를 사용하여 A 검색 시스템과 B 검색 시스템의 검색 성능을 평가한 결과이다. 질의 Q에 대한 문헌 집단 내의 총 적합 문헌 수가 20건이라고 할 때, ㉠과 ㉡의 값으로 옳은 것은?

구분	검색 결과	재현율	정확률
A 검색 시스템	15	(㉠)%	60%
B 검색 시스템	32	40%	(㉡)%

	㉠	㉡
①	35	35
②	45	25
③	45	35
④	55	25
⑤	55	35

정답 2번

출제 포인트

검색 결과 수, 적합 문헌 수, 전체 적합 문헌 수의 관계를 이용하여 재현율과 정확률을 계산하는 문제이다.

조건 정리

전체 적합 문헌 수는 20건이다.
A 검색 시스템: 검색 결과 15건, 정확률 60%이다.
B 검색 시스템: 검색 결과 32건, 재현율 40%이다.
구해야 할 값은 A의 재현율 ㉠과 B의 정확률 ㉡이다.

사용할 공식

재현율=검색된 적합 문헌 수전체 적합 문헌 수×100
정확률=검색된 적합 문헌 수검색 결과 수×100
두 공식 모두 공통으로 검색된 적합 문헌 수를 사용하므로, 주어진 비율에서 먼저 그 문헌 수를 구한 뒤 나머지 비율을 계산한다.

NoteBase

계산 과정

1. A가 검색한 적합 문헌 수는 15 × 0.60 = 9건이다. 따라서 A의 재현율은 $\frac{9}{20}$ × 100 = 45%이므로 ㉠은 45이다.
2. B가 검색한 적합 문헌 수는 20 × 0.40 = 8건이다. 따라서 B의 정확률은 $\frac{8}{32}$ × 100 = 25%이므로 ㉡은 25이다.
3. 따라서 ¤ = 45, ¤ = 25이다.

선지 판단

① X — ㉠ 35, ㉡ 35는 계산 결과와 모두 다르다. 올바르게는 ㉠ 45, ㉡ 25이다.
② O — ㉠ 45, ㉡ 25는 A의 재현율 45%와 B의 정확률 25%에 정확히 일치한다.
③ X — ㉠ 45는 맞지만 ㉡ 35가 틀리다. B의 정확률은 8/32 × 100 = 25%이므로 올바르게는 ㉠ 45, ㉡ 25이다.
④ X — ㉡ 25는 맞지만 ㉠ 55가 틀리다. A의 재현율은 9/20 × 100 = 45%이므로 올바르게는 ㉠ 45, ㉡ 25이다.
⑤ X — ㉠ 55와 ㉡ 35가 모두 틀리다. 올바르게는 ㉠ 45, ㉡ 25이다.

오답 포인트

재현율과 정확률의 분모를 바꾸어 적용하는 것이 핵심 함정이다. 재현율은 전체 적합 문헌 수 20건을 분모로 삼고, 정확률은 각 시스템의 검색 결과 수를 분모로 삼는다.

다시 나오면 이렇게 판단

먼저 검색된 적합 문헌 수를 구한 뒤, 재현율이면 전체 적합 문헌 수를 분모로, 정확률이면 검색 결과 수를 분모로 놓아 계산한다.

암기 포인트

재현율의 분모는 전체 적합 문헌, 정확률의 분모는 검색 결과이다.

NoteBase