4장 검색엔진

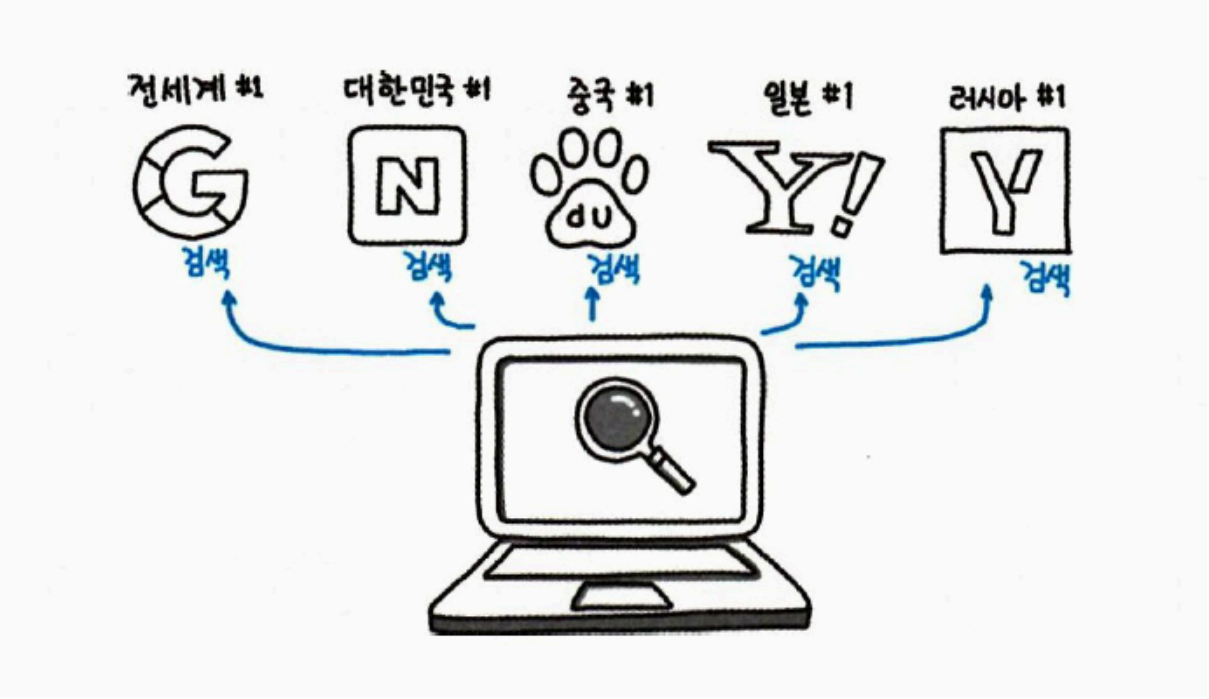
검색엔진의 등장

검색은 현대인의 분실에 꼭 필요한 정보를 찾아주는 가장 핵심적인 역할을담당하고 있음.

늘상 사용하다 보니 인터넷이라는 방대한 공간에 존재하는 수백 조 개의 문서 중에서 내가 찾는 문서를 골라서 찾아주는 엄청난 작업이 라는 사실을 가끔씩 잊어버리곤 함.

정보 폭발의 시작

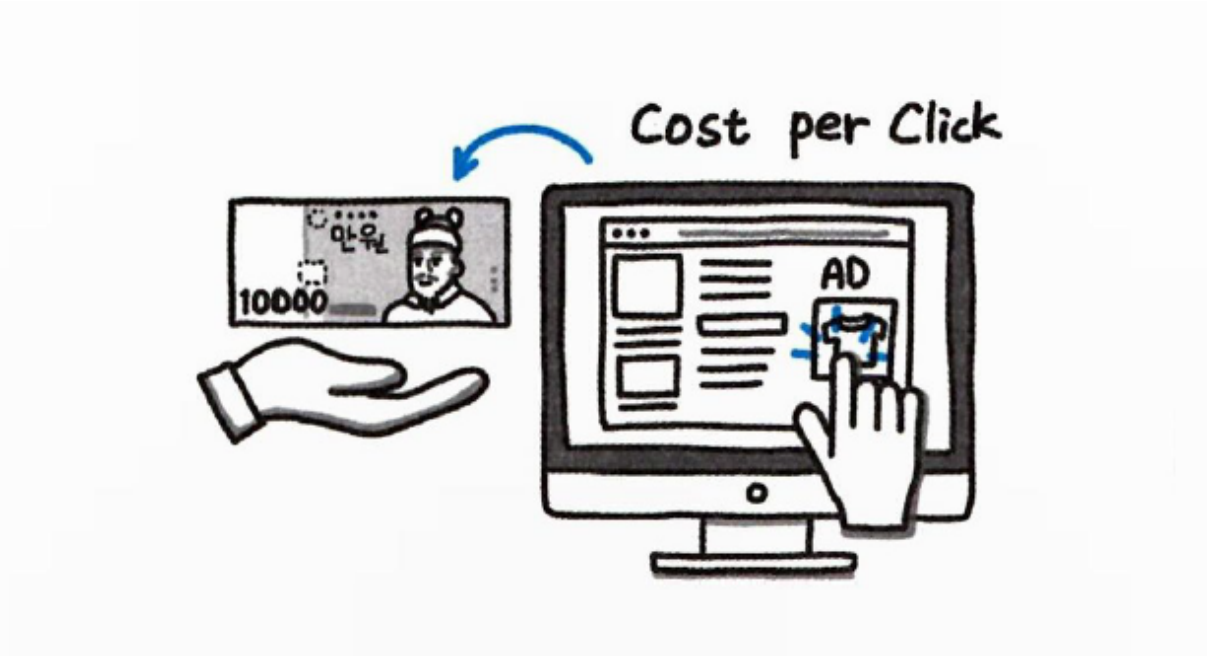
- 인터넷이 커지면서 **사람이 직접 정보를 찾는 방식은 한계**에 도달
- 검색엔진 탄생



검색엔진은 왜 돈을 많이 벌까?

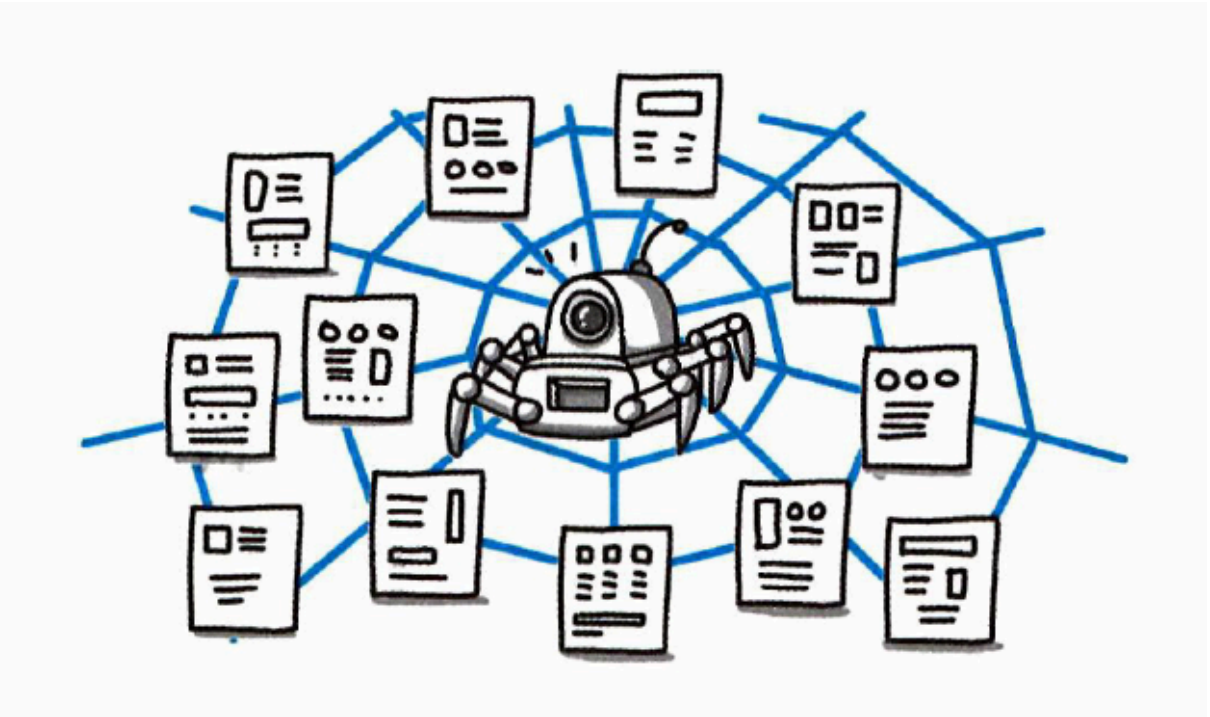
- 사람들은 검색할 때 이미 **목적**이 있음
- 해당 목적과 관련된 광고가 노출

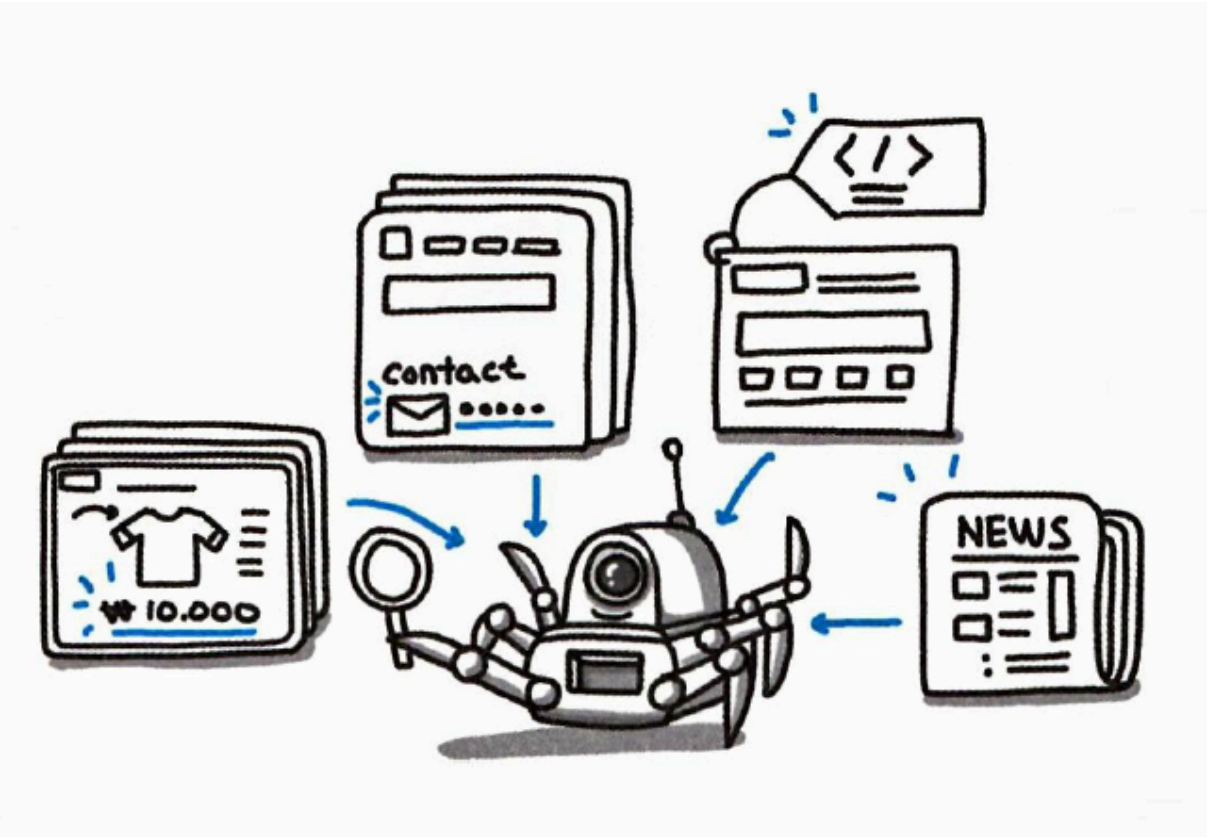
- 대표적인 검색광고 모델이자 검색엔진의 수익 모델



검색엔진은 문서를 어떻게 모을까?

- 검색엔진은 웹을 돌아다니는 로봇(크롤러)을 보냄
- 링크를 따라가며 페이지를 수집
- 이렇게 모은 문서가 **검색엔진의 재료**



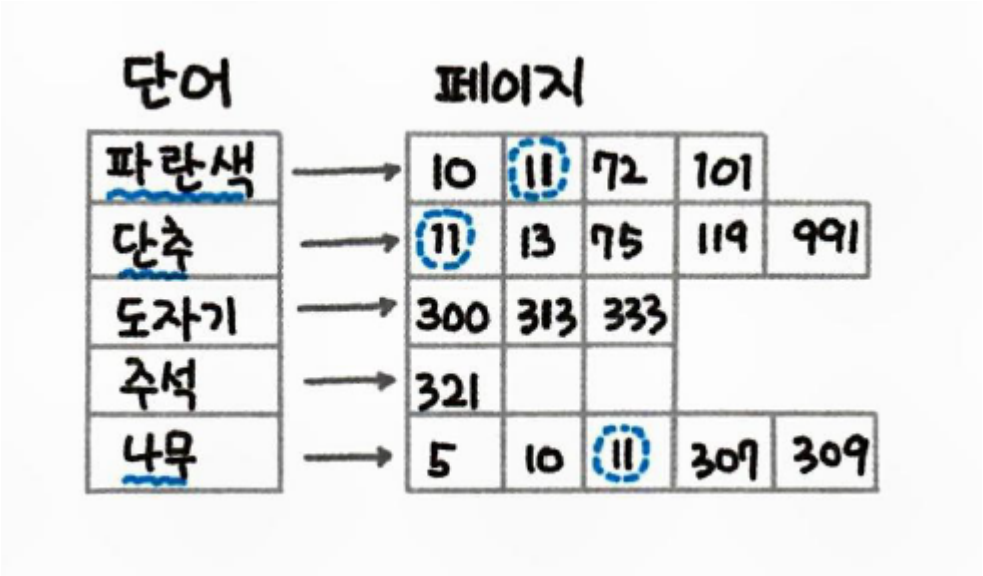


검색엔진은 어떻게 검색할까?

3단계 구조

1. **수집(Crawl)**
2. **정리(Index)**
    - 문서를 단어 단위로 쪼개서 색인
3. **검색(Rank)**
    - 사용자 쿼리에 맞는 문서를 정렬

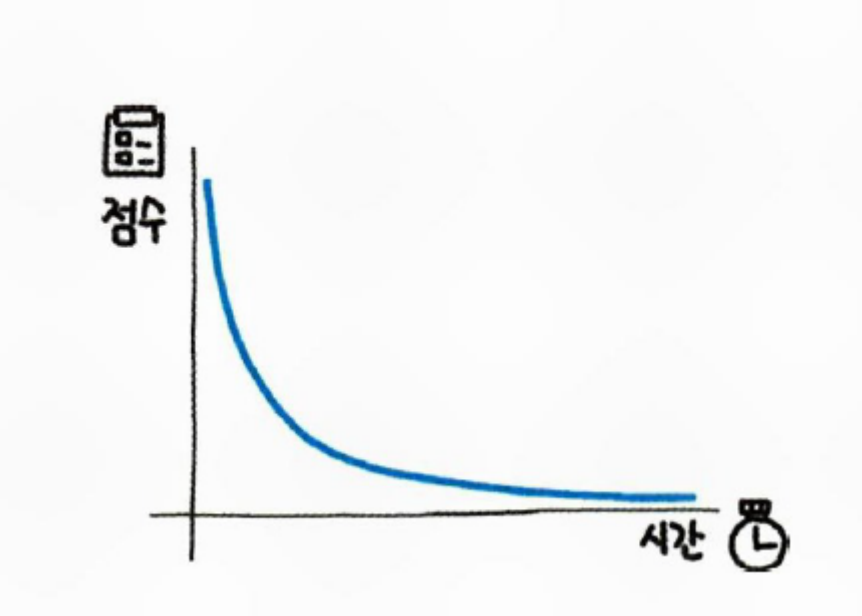
진짜 기술력은 랭킹



랭킹의 핵심, 품질 좋은 문서란?

검색엔진이 보는 기준

- **최신성**: 오래된 정보 vs 방금 나온 정보



- **품질**: 신뢰할 수 있는가?
- **관련성**: 질문과 얼마나 잘 맞는가?

단순히 “많이 나온 단어”가 아니라 **의도에 맞는 문서**가 중요

페이지랭크(PageRank)

구글의 출발점

- 링크 = 추천
- 많이, 그리고 **좋은 사이트로부터** 링크를 받으면 점수 상승

아이디어는 단순

> “사람들이 많이 추천한 문서는 좋은 문서다”

이 개념이 구글을 검색 1위로 올려놓음

쿼리에 딱 맞는 문서 찾기

TF–IDF & BM25

TF–IDF

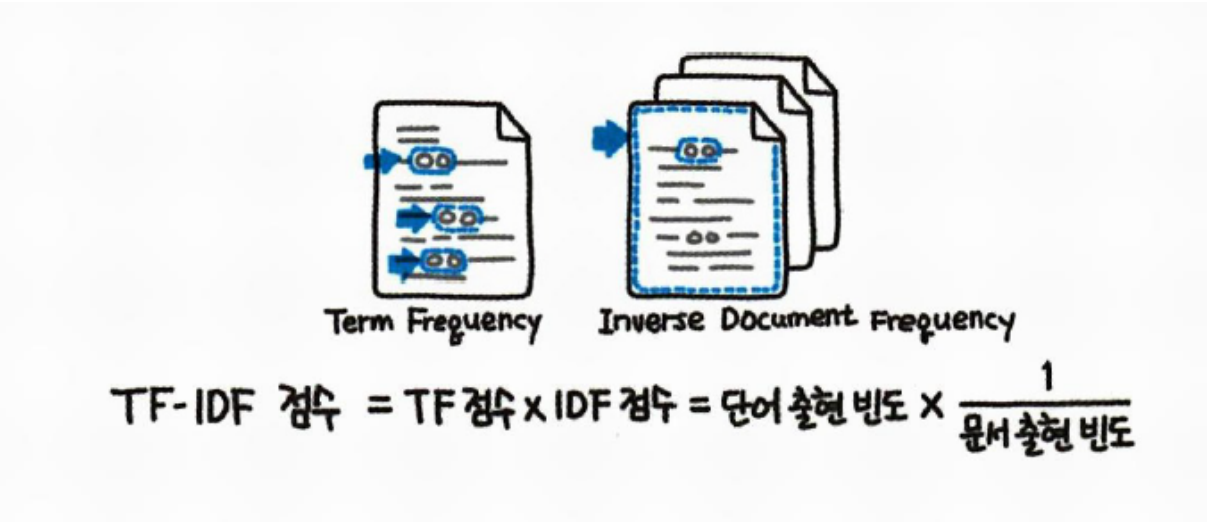
- **TF (Term Frequency)**

문서 안에서 단어가 얼마나 자주 등장?

- **IDF (Inverse Document Frequency)**

흔한 단어일수록 중요도 ↓

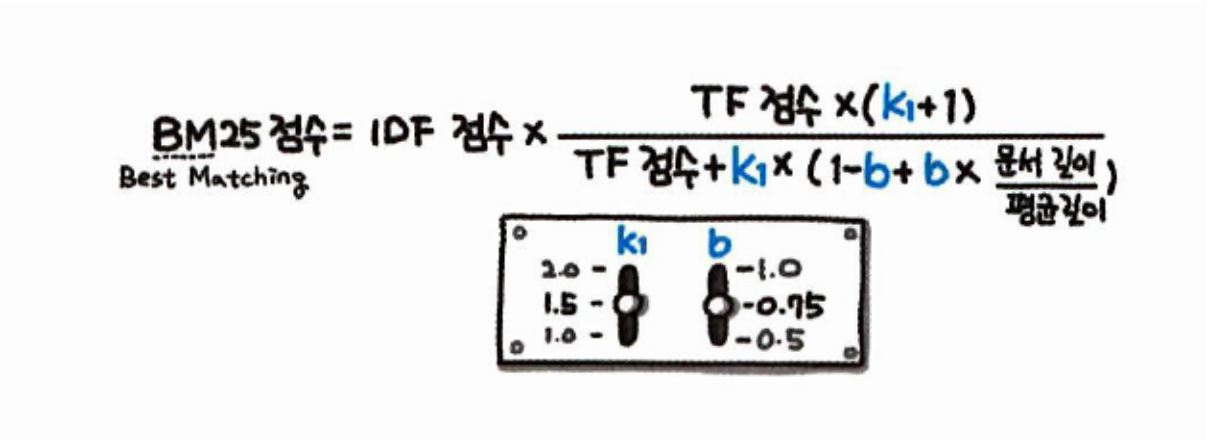
의미 있는 단어에 가중치 부여



BM25

- TF–IDF를 더 똑똑하게 개선한 공식
- 문서 길이, 단어 반복 과다 문제 해결

현대 검색엔진의 기본 뼈대



A/B 테스트

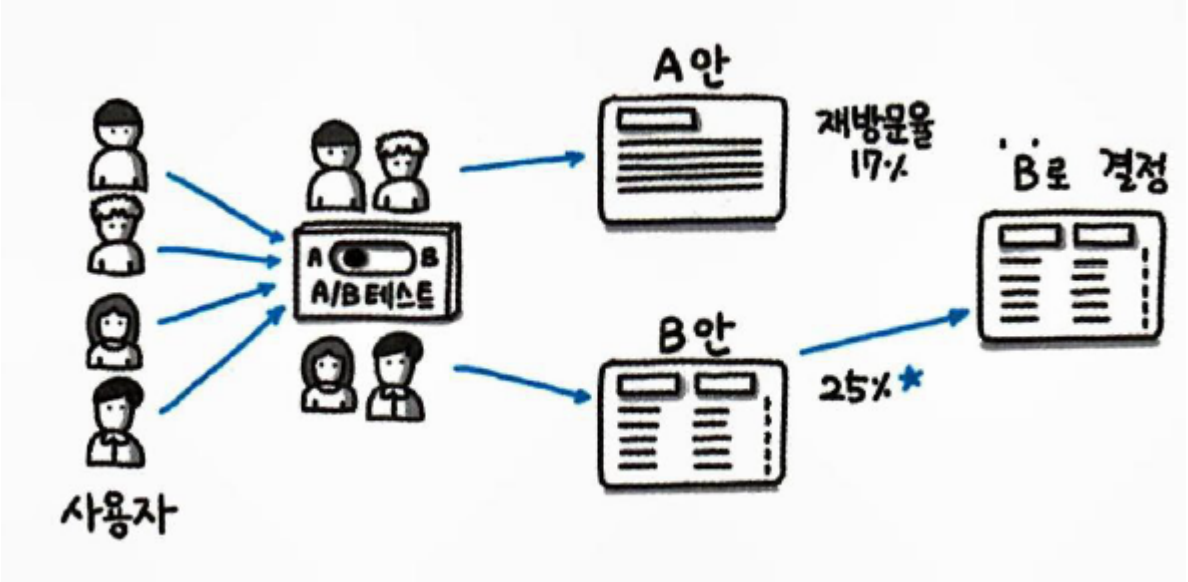
검색 품질은 어떻게 개선될까?

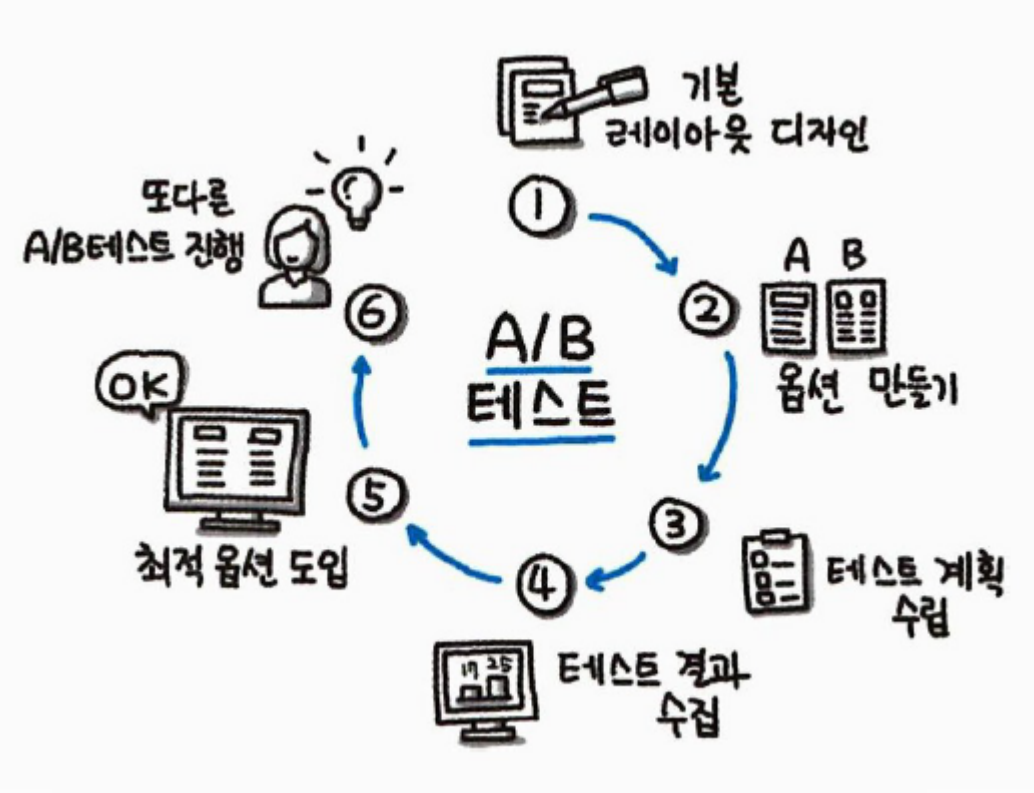
- 검색 알고리즘 A vs B
- 사용자 반응 비교
    - 클릭률

- 체류 시간

- **숫자로 더 좋은 검색을 선택**

검색 품질은 감이 아니라 **실험의 결과**





점점 더 똑똑해지는 검색

- 과거: 단어 일치
- 현재: 문장의 의미, 의도 파악
- 앞으로: 질문에 **답을 생성**하는 검색