

# 검색기술 트렌드와 네이버의 연구현황

김상범 / Director of Search Quality  
2018. 5.

**NAVER**

# DISCLAIMER

본 발표내용에는  
개인적인 의견이나 주장이 포함되어 있을 수 있으며  
이는 회사의 공식적 입장과는 무관합니다.

# Outline

- 전세계 의미있는 검색서비스 7개
- 검색서비스를 만드는 과정
- 검색엔진 핵심모듈 : Matching & Ranking
- Take Home Message

# 전세계 의미있는 검색서비스 7개

- Google

| Search Engines - 2015 Stats |        |       |             |       |                      |
|-----------------------------|--------|-------|-------------|-------|----------------------|
| Country                     | Leader | Share | Runner-Up   | Share | Internet Penetration |
| Argentina                   | Google | 92%   | Yahoo       | 3%    | 75.0%                |
| Australia                   | Google | 94%   | Bing        | 4%    | 89.6%                |
| Brazil                      | Google | 95%   | Others      | 6%    | 54.2%                |
| Canada                      | Google | 87%   | Yahoo       | 6%    | 92.5%                |
| China                       | Baidu  | 55%   | Qihoo 360   | 28%   | 49.5%                |
| France                      | Google | 92%   | Yahoo       | 4%    | 83.3%                |
| Germany                     | Google | 94%   | Bing        | 2%    | 88.6%                |
| Hong Kong                   | Google | 73%   | Yahoo       | 24%   | 80.5%                |
| India                       | Google | 96%   | Others      | 4%    | 28.3%                |
| Indonesia                   | Google | 96%   | Others      | 4%    | 28.5%                |
| Italy                       | Google | 95%   | Yahoo       | 2%    | 58.5%                |
| Japan                       | Google | 57%   | Yahoo Japan | 40%   | 90.6%                |
| Malaysia                    | Google | 93%   | Yahoo       | 4%    | 67.5%                |
| Mexico                      | Google | 94%   | Bing        | 3%    | 49.2%                |
| The Netherlands             | Google | 94%   | Bing        | 2%    | 95.7%                |
| The Philippines             | Google | 89%   | Yahoo       | 7%    | 43.0%                |
| Poland                      | Google | 97%   | Others      | 3%    | 66.9%                |
| Russia                      | Yandex | 58%   | Google      | 34%   | 61.4%                |
| Saudi Arabia                | Google | 94%   | Yahoo       | 2%    | 65.9%                |
| Singapore                   | Google | 92%   | Yahoo       | 6%    | 82.0%                |
| South Africa                | Google | 93%   | Bing        | 4%    | 49.0%                |
| South Korea                 | Naver  | 77%   | Daum        | 20%   | 92.3%                |
| Spain                       | Google | 95%   | Yahoo       | 2%    | 74.8%                |
| Sweden                      | Google | 94%   | Bing        | 3%    | 94.8%                |
| Thailand                    | Google | 98%   | Others      | 2%    | 34.9%                |
| Turkey                      | Google | 96%   | Yandex      | 2%    | 56.7%                |
| United Arab Emirates        | Google | 94%   | Yahoo       | 2%    | 93.2%                |
| United Kingdom / UK         | Google | 90%   | Bing        | 5%    | 89.8%                |
| United States               | Google | 72%   | Bing        | 21%   | 87.9%                |
| Vietnam                     | Google | 92%   | Bing        | 4%    | 48.3%                |

# 전세계 의미있는 검색서비스 7개

- Bing : 인터넷 익스플로러를 기반으로 Google 추격?
  - <https://www.wired.com/2011/02/bing-copies-google/>



Google Catches Bing Copying; Microsoft

SHARE



22



RYAN SINGEL BUSINESS 02.01.11 02:31 PM

## GOOGLE CATCHES BING COPYING; MICROSOFT SAYS 'SO WHAT?'

# 전세계 의미있는 검색서비스 7개

- Yahoo

## 야후 3분기 실적 부진...구글과 검색엔진 제휴

송고시간 | 2015/10/21 11:43

f t G+ BAND |  + -

(서울=연합뉴스) 정선미 기자 = 미국 포털업체 야후가 구글과 검색 엔진 제휴를 통해 실적 부진의 돌파구 마련에 나섰다.

20일(현지시간) 영국 일간 파이낸셜타임스(FT)에 따르면 야후는 이날 실적을 발표하는 자리에서 구글과 제휴를 통해 검색 결과와 검색 광고를 공급받게 될 것이라고 밝혔다.

2009년 마이크로소프트(MS)와 검색 엔진 Bing을 10년간 자사검색에 쓰기로 한 야후는 지난 4월 계약 변경을 통해 검색 결과의 51%는 계속 Bing으로부터 받게 될 예정이다.



Adobe Flash Player을(를) 사용하려면 클릭하세요.

나머지 49% 가운데 일부가 구글 검색 결과를 받게 되지만 정확한 비중은 알려지지 않았다.

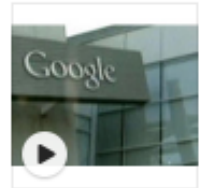
# 전세계 의미있는 검색서비스 7개

- Baidu

## 구글, 중국 본토서 철수

[앵커멘트] 사전 검열 문제로 중국 정부와 갈등을 빚어온 미국의 검색 엔진 구글이 중국 본토에서 철수하고 홍콩에서 우회적으로 서비스를 제공하겠다고 밝혔습니다. 중국 내 비판이 잇따르는 가운데 미국 정부도 실망감을 나타냈습니다. 박신윤 기자의 보도입니다. [리포트] 구글이 결국 중국...

2010.03.23. | YTN | 네이버뉴스



## 바이두 1Q 순익 두배 급증 '땡큐 구글'

바이두의 이 같은 선전은 구글이 중국 정부의 인터넷 검열에 반대하며 중국 시장을 철수한 데 따른 반사이익으로 풀이된다. 크레디트 스위스 그룹의 월러스 청은 "중국내 구글의 광고주들이 대부분 바이두로 이동한 것으로 추정된다"며 "이에 바이두의 광고비 단가가 상승했다"며 순익...

2010.04.29. | 아시아경제 | 네이버뉴스

## 구글 이어 야후도 중국서 짐싼다

구글에 이어 야후마저 철수하면서 중국 인터넷 시장은 바야흐로 외국계 기업이 설 자리가 없는 '무덤'이 되고 있다. 반면 알리바바와 텐센트, 바이두 등 현지 인터넷 기업들은 거대한 내수를 발판으로 고속 성장을 계속하고 있다. 18일(현지시간) 월스트리트저널(WSJ)은 소식통을 인용해...

2015.03.19. | 비즈니스워치



# 전세계 의미있는 검색서비스 7개

- Yandex : Google 이 추격중 ( 2년동안 23%차 → 8%차 )

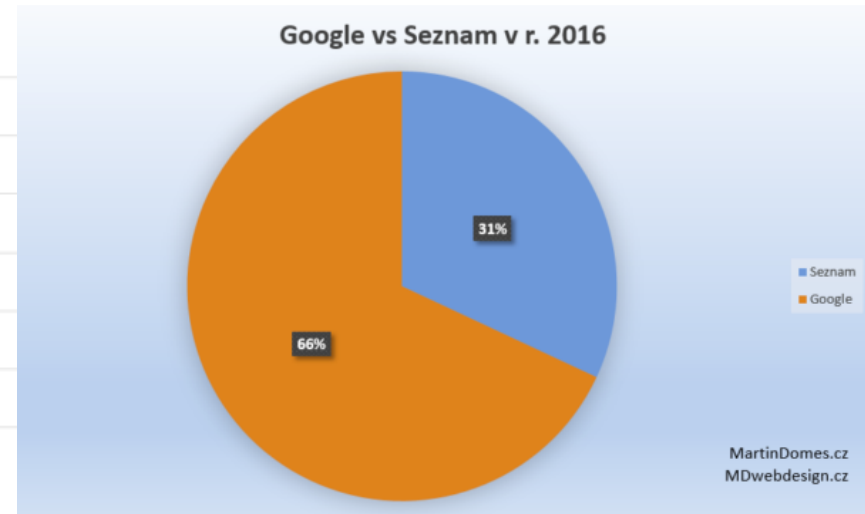
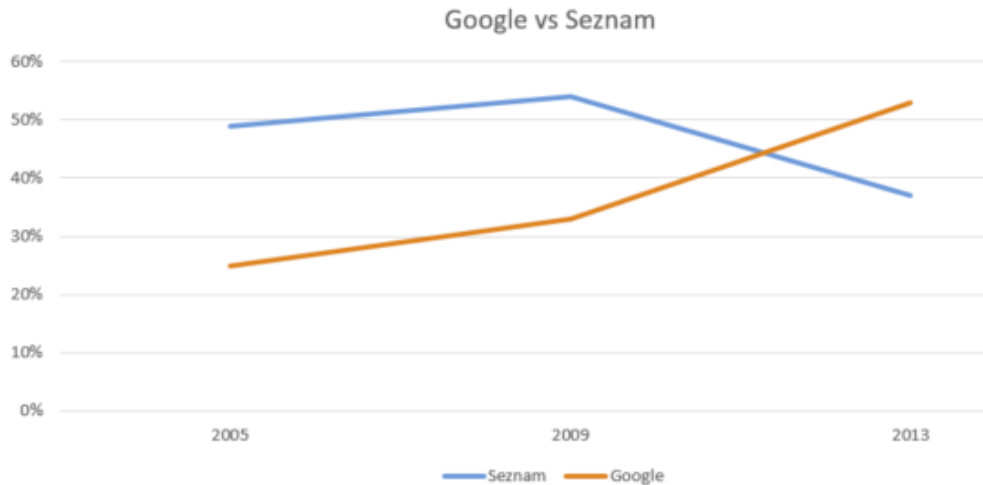
|                                     |        |                                |       |                |       |                |       |
|-------------------------------------|--------|--------------------------------|-------|----------------|-------|----------------|-------|
| << Sep. 15                          |        | October 2015                   |       |                |       | Nov. 15 >>     |       |
| report: from search engines         |        | by days   by weeks   by months |       |                |       |                |       |
| values:                             |        | October 2015                   |       | September 2015 |       | at the average |       |
| average daily / <u>summarized</u>   |        |                                |       |                |       | for 3 months   |       |
| <input checked="" type="checkbox"/> | Yandex | 95,747,996                     | 57.5% | 88,350,993     | 57.4% | 87,944,514     | 57.3% |
| <input checked="" type="checkbox"/> | Google | 58,042,645                     | 34.9% | 53,679,060     | 34.9% | 53,637,708     | 35.0% |

|                                     |        |                                |       |                |       |                |       |
|-------------------------------------|--------|--------------------------------|-------|----------------|-------|----------------|-------|
| << Sep. 17                          |        | October 2017                   |       |                |       | Nov. 17 >>     |       |
| report: from search engines         |        | by days   by weeks   by months |       |                |       |                |       |
| values:                             |        | October 2017                   |       | September 2017 |       | at the average |       |
| average daily / summarized          |        |                                |       |                |       | for 3 months   |       |
| <input checked="" type="checkbox"/> | Yandex | 59,732,021                     | 52.0% | 55,293,894     | 51.5% | 55,305,245     | 51.6% |
| <input checked="" type="checkbox"/> | Google | 50,258,202                     | 43.8% | 47,429,267     | 44.2% | 47,134,364     | 44.0% |



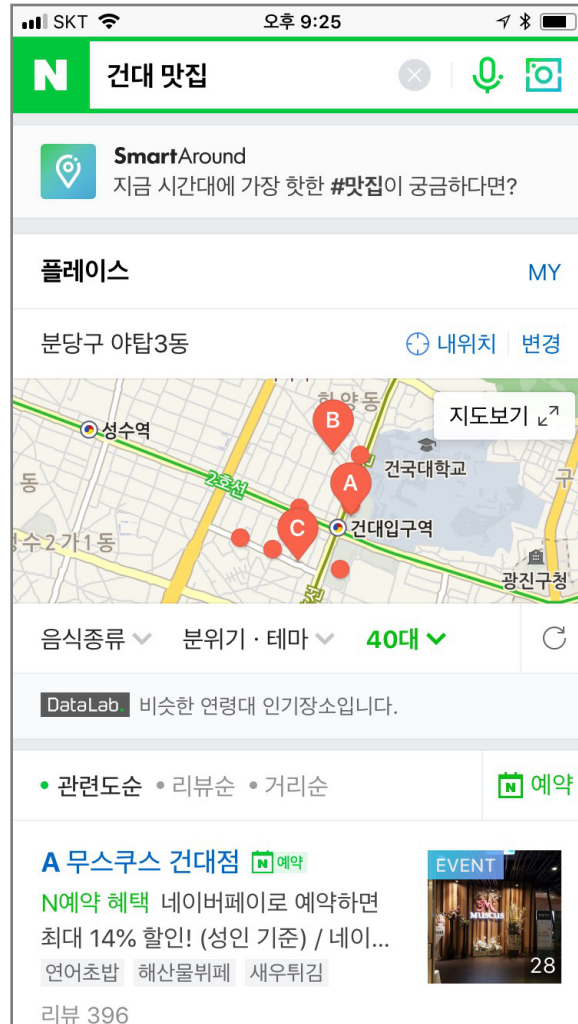
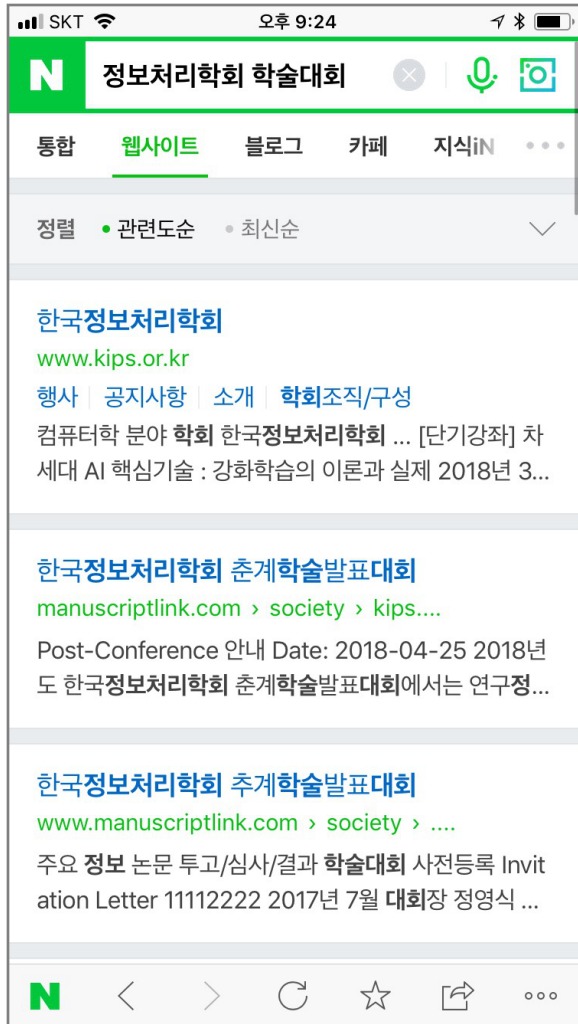
# 전세계 의미있는 검색서비스 7개

- Seznam : 체코의 검색서비스. 하락세.
  - 2011년경부터 역전. 현재 7:3으로 구글 우세



# 전세계 의미있는 검색서비스 7개

## • 네이버



# 웹검색이 뉴스검색, 블로그검색과 다른점

|             | 뉴스/블로그        | 웹                             |
|-------------|---------------|-------------------------------|
| 검색대상의 양과 위치 | 대략 알 수 있음     | 얼마나 많은 문서가 어디에 있는지 파악 자체가 어려움 |
| 검색컨텐츠의 품질   | 일정부분 관리 가능    | 관리 불가능                        |
| 사용자가 원하는 문서 | 여러 좋은 문서 중 몇개 | 오직 그 웹페이지                     |

# 웹검색이 뉴스검색, 블로그검색과 다른점

N
롯데타워 전망대 가격

정렬
• 관련도순
• 최신순
• 오래된순

**롯데월드·타워전망대·아쿠아리움, 수험생 우대 혜택 연장**  
 서울 잠실에 위치한 롯데월드 운영 사업장인 롯데월드 어드벤처, 롯데월드 아쿠아리움, 롯데월드타워 전망대 서...  
 2017.11.21. | 쿠키뉴스



[롯데월드, 수능탈출 우대, ...](#)
[일간스포츠](#)
[네이버뉴스](#)  
[롯데월드, '수능 대박' 프로모션 15일...](#)
[한국스포츠경제](#)

**[비즈 르포] "두바이보다 낫다"...롯데월드타워 전망대, 개장 첫 주말 2만명 가까이 몰려**  
 전망대까지 올라가는 엘리베이터까지 약 1시간 20분의 대기 시간이 걸린다는 직원의 설명을 듣자 비로소 대기 없...  
 2017.04.11. | 조선비즈 | 네이버뉴스



**롯데월드 어드벤처 · 아쿠아리움 · 전망대, 수험생 할인 우대**  
 롯데월드타워 전망대 서울스카이가 그동안 수능 준비에 지쳐있던 수험생을



N
롯데타워 전망대 가격

정렬
• 관련도순
• 최신순

**<서울 데이트 가볼만한곳> "잠실 롯데타워 전망대"**  
 오랫동안 주말에가볼만한곳 올려보아 요 ㅎㅎ 아직까지도 핫플레이스죠~ "...  
 "건강과 함께 떠나는 힐... | 2017.11.22. SE (3.0)



**롯데월드타워 전망대. 서울스카이. 입장권가격.주차권**  
 ㅎㅎㅎㅎ 전망대 입구. 지하 1층 입니다. 롯데월드타워인데서울 스카이.....  
 Sienna's Diary | 어제



**잠실 롯데타워 전망대 가격정보**  
 잠실 롯데타워 전망대 가격 정보 주말에 서울 나들이 제대로 했네요 인터...  
 해밀이와 해밀맘의 성장... | 2017.09.08. SE (3.0)



**[서울스카이] 롯데타워 전망대 / 가**



N
롯데타워 전망대 가격

정렬
• 관련도순
• 최신순

보기
 • 전체
 • 사이트
 • 웹문서

**SEOUL SKY**  
[seoulsky.lotteworld.com](http://seoulsky.lotteworld.com)  
 SEOUL SKY 소개 개인정보처리방침 이용약관 이메일 주소무단수집거부 영상정보처리 기기운영/관리방침 사이트맵 공지사항 인재채용 서울특별시 송파구 올림픽...

**롯데월드타워**  
[www.lwt.co.kr](http://www.lwt.co.kr) > tower > main > main.d....  
[타워 시설안내](#) | [쇼핑&다이닝](#) | [이용안내](#)  
 롯데월드타워의 타워스토리, 타워시설, 엔터테인먼트, 쇼핑&다이닝, 이용안내등의 정보를 제공합니다.

**롯데월드타워 서울스카이**  
[mticket.lotteworld.com](http://mticket.lotteworld.com) > web > pdt > ....  
 홈페이지 바로가기 이용요금 안내 일반티켓 : 관람일과 관람시간 사전지정 구분 가격 연령 어른 27,000원 만 13세 이상 어린이 24,000원 만... 이용관련 [전망대 입...](#)

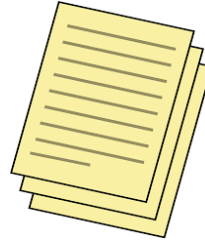
# 검색서비스를 개발하는 과정

- 먼저 가이드라인에 따른 기계학습용 학습데이터를 구축하고,

① 네이버가 지향하는  
좋은 검색결과에 대해 논의



② 검색평가 가이드라인



Ex)  
사용자가 입력한 질의에 대해  
입력시점에서 중요한 내용을 많이 포함하  
고 있으면 좋은 문서

사용자가 입력한 질의에 대해  
입력시점에서 중요한 내용은 별로 없고  
이상한 곳으로 연결되는 링크만 있으면  
나쁜 문서

④ 학습용 데이터집합



③ 가이드라인에 따른 평가



Ex) 2017년 11월 13일 오후 6시에 입력  
된 “페미니스트” 라는 질의에 대하여

doc0001 : 5점  
doc0002 : 2점  
doc0003 : 3점  
doc0004 : 1점  
...

# 검색서비스를 개발하는 과정

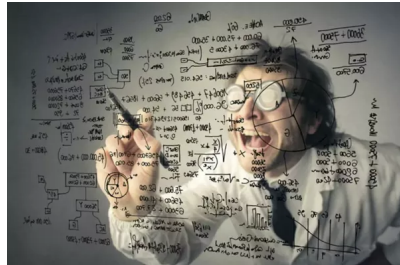
- 기계학습알고리즘을 이용하여 검색랭킹모듈을 학습

Ex)  
질의가 제목에 매치된 빈도  
제목 전체에서 질의와 매치된 비율  
질의가 본문에 매치된 빈도  
질의내 단어별 본문매치빈도의 분산  
해당 사이트의 신뢰도  
해당 도메인의 남은 유효기간

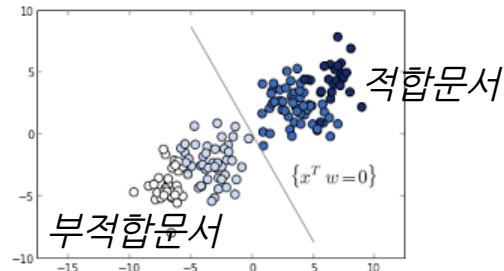
뉴스검색학습용 데이터집합



## ⑤ 엔지니어들의 시그널 발굴



## ⑥ 실험데이터생성 및 기계학습 실행



서비스적용여부  
판단

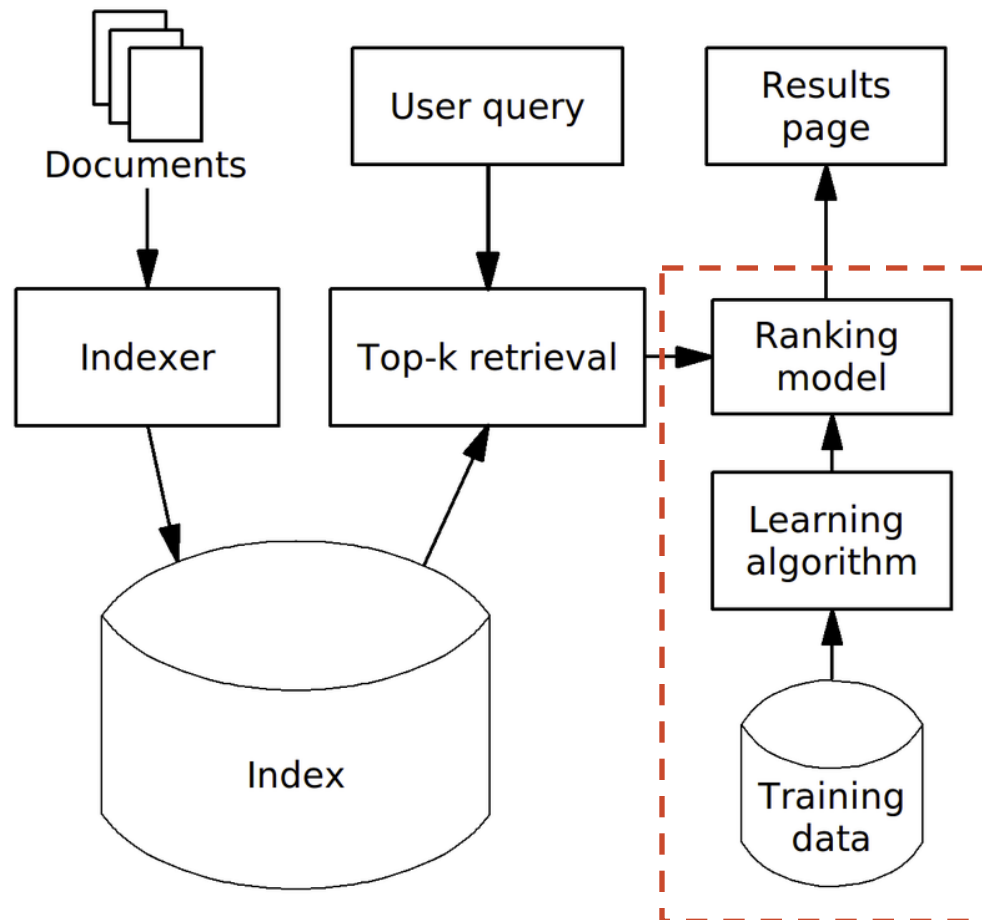
A/B online test

새로운 모델

# 랭킹모듈을 “학습”한다?

- Learning-to-Rank

- [https://en.wikipedia.org/wiki/Learning\\_to\\_rank](https://en.wikipedia.org/wiki/Learning_to_rank)



# 전형적인 기계학습의 예

- 만일 아래와 같이 입력에 대해 출력값을 내주는

| x1 | x2 | f   |
|----|----|-----|
| 0  | 0  | 0.0 |
| 0  | 1  | 0.3 |
| 1  | 0  | 0.7 |
| 1  | 1  | 1.0 |

- 함수  $f$ 의  $w1, w2$  를 구하려면?

- $f = x1 * w1 + x2 * w2$



# 전형적인 기계학습의 예

- 만일 아래와 같이 입력에 대해 출력값을 내주는

| x1 | x2 | f   |
|----|----|-----|
| 0  | 0  | 0.0 |
| 0  | 1  | 0.3 |
| 1  | 0  | 0.7 |
| 1  | 1  | 1.0 |

- 함수  $f$ 의  $w1, w2$  를 구하려면?

$$- f = x1 * w1 + x2 * w2$$



0.3



0.7

# 전형적인 기계학습의 예

- 만일 아래와 같이 입력에 대해 출력값을 내주는

| x1 | x2 | f   |
|----|----|-----|
| 0  | 0  | 0.0 |
| 0  | 1  | 0.3 |
| 1  | 0  | 0.7 |
| 1  | 1  | 1.0 |

- 함수  $f$ 의  $w1$ ,  $w2$  를 구하려면?

$$f = x1 * w1 + x2 * w2$$

↓  
0.3

↓  
0.7

| x1  | x2  | f    |
|-----|-----|------|
| 0.1 | 0.1 | 0.1  |
| 0.9 | 0.1 | 0.66 |
| 0.1 | 0.9 | 0.34 |
| 0.6 | 0.6 | 0.6  |

# 기계학습을 구성하는 두 파트

- 벡터로 만들어진 학습데이터

| x1 | x2 | f   |
|----|----|-----|
| 0  | 0  | 0.0 |
| 0  | 1  | 0.3 |
| 1  | 0  | 0.7 |
| 1  | 1  | 1.0 |

- 학습알고리즘

- 신경망, 결정트리, 로지스틱회귀분석 등...
- 모두 오픈되어 있고 각 기업은 다양한 실험을 통해 적합한 알고리즘을 찾고 조금씩 튜닝

# 검색랭킹 기계학습을 구성하는 두 파트

- 벡터로 만들어진 검색랭킹용 학습데이터

| x1 | x2 | f   |
|----|----|-----|
| 0  | 0  | 0.0 |
| 0  | 1  | 0.3 |
| 1  | 0  | 0.7 |
| 1  | 1  | 1.0 |

# 검색랭킹 기계학습을 구성하는 두 파트

- 벡터로 만들어진 검색랭킹용 학습데이터

| x1 | x2 | f   |
|----|----|-----|
| 0  | 0  | 0.0 |
| 0  | 1  | 0.3 |
| 1  | 0  | 0.7 |
| 1  | 1  | 1.0 |

$$\frac{\text{number of title matched words in user query}}{\text{number of words in user query}}$$

$$\frac{\text{number of title matched words in user query}}{\text{number of words in title}}$$

# 검색랭킹 기계학습을 구성하는 두 파트

- 벡터로 만들어진 검색랭킹용 학습데이터

| x1 | x2 | f   |
|----|----|-----|
| 0  | 0  | 0.0 |
| 0  | 1  | 0.3 |
| 1  | 0  | 0.7 |
| 1  | 1  | 1.0 |

$$\frac{\text{number of title matched words in user query}}{\text{number of words in user query}}$$

$$\frac{\text{number of title matched words in user query}}{\text{number of words in title}}$$

랭킹시그널(비밀유지)

# 랭킹시그널을 비밀로 하는 이유

- 좋은 랭킹시그널을 발견하는것은 엔지니어의 몫
- 랭킹시그널을 공개하면, 그 랭킹시그널은 빠른 시일 내에 무력화되므로, 새로운 시그널을 개발해야 함
- 엔지니어들은 점점 힘들어지고, 좋은 서비스를 만들기는 점점 어려워짐

# 랭킹시그널을 비밀로 하는 이유

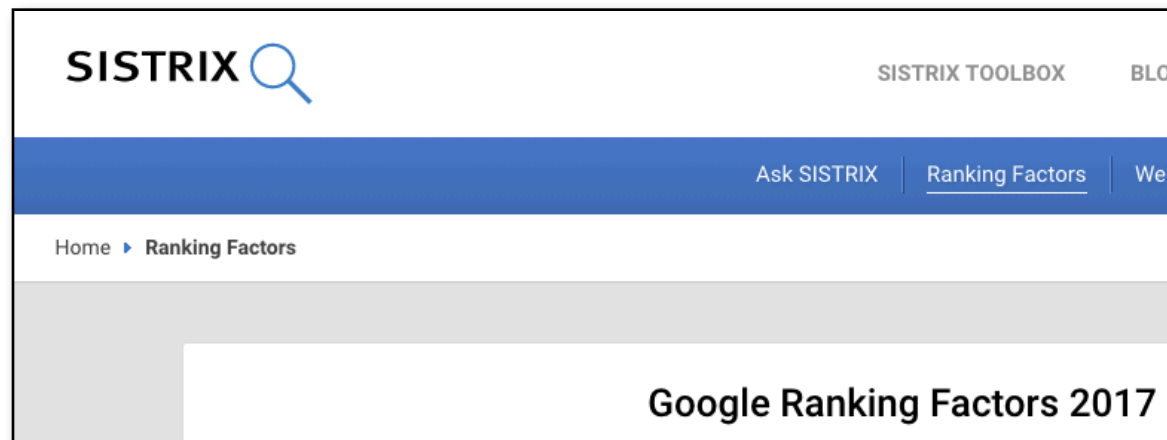
## Schmidt: Listing Google's 200 Ranking Factors Would Reveal Business Secrets

- 설리번 : 200개정도 시그널을 쓴다고 했는데 왜 얘기 안해주나?
- 슈미트 : 계속 바뀌고 ... 우리가 힘들어지니까 안된다
- 설리번 : 뭐가 중요한지 이런건 아니더라도 그냥 리스트만..
- 슈미트 : 리스트도 안된다
- 설리번 : 한 50개는 계속 안변하고 쓰지 않냐?
- 슈미트 : 아무튼 비밀이다
- 블룸버그기자 : 너무 개방적이지 않은 모습 아닌가?
- 슈미트 : 그럼 블룸버그는 얼마나 개방적인지 같이 얘기해볼까? ...
- .....





# 랭킹시그널을 체계적 방식으로 추정 : SEO Business

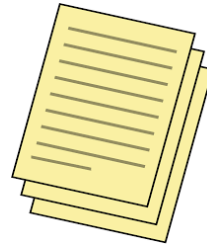


# 검색서비스를 개발하는 과정 : wrap up

① 네이버가 지향하는  
좋은 뉴스검색결과에 대해 논의



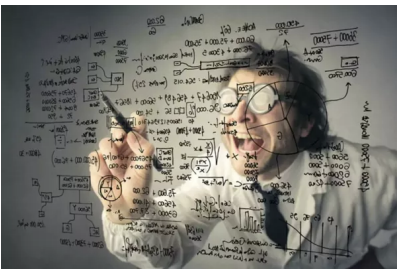
② 검색평가 가이드라인



④ 학습용 데이터집합



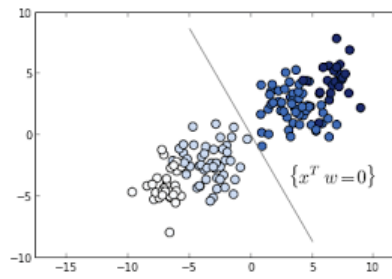
⑤ 엔지니어들의 시그널 발굴



③ 가이드라인에 따른 평가



⑥ 실험데이터생성 및 기계학습 실행



서비스적용여부  
판단

A/B online test

새로운 모델



# 랭킹성능은 어떻게 구하고 실제로 어느정도인가?

- 이상적인 랭킹 : 5, 4, 3, 2, 1
  - $DCG = 5 + 4/\log_2(3) + 3/\log_2(4) + 2/\log_2(5) + 1/\log_2(6) = 10.27$
  - $nDCG = 10.27 / 10.27 = 1$
- 만일 A방법으로 한 랭킹이 5, 4, 3, 1, 2 라면...
  - $DCG = 5 + 4/\log_2(3) + 3/\log_2(4) + 1/\log_2(5) + 2/\log_2(6) = 10.23$
  - $nDCG = 10.23/10.27 = 0.996$
- 만일 B방법으로 한 랭킹이 4, 5, 3, 2, 1 라면...
  - $DCG = 4 + 5/\log_2(3) + 3/\log_2(4) + 1/\log_2(5) + 2/\log_2(6) = 8.27$
  - $nDCG = 8.27/10.27 = 0.805$

# 랭킹성능은 어떻게 구하고 실제로 어느정도인가?

- Yahoo에서 주관한 Learning-to-rank challenge (2011년)

|          | Validation |         | Test    |         |
|----------|------------|---------|---------|---------|
|          | ERR        | NDCG    | ERR     | NDCG    |
| BM25F-SD | 0.42598    | 0.73231 | 0.42853 | 0.73214 |
| RankSVM  | 0.43109    | 0.75156 | 0.43680 | 0.75924 |
| GBDT     | 0.45625    | 0.78608 | 0.46201 | 0.79013 |

<http://proceedings.mlr.press/v14/chapelle11a/chapelle11a.pdf>

# 랭킹성능은 어떻게 구하고 실제로 어느정도인가?

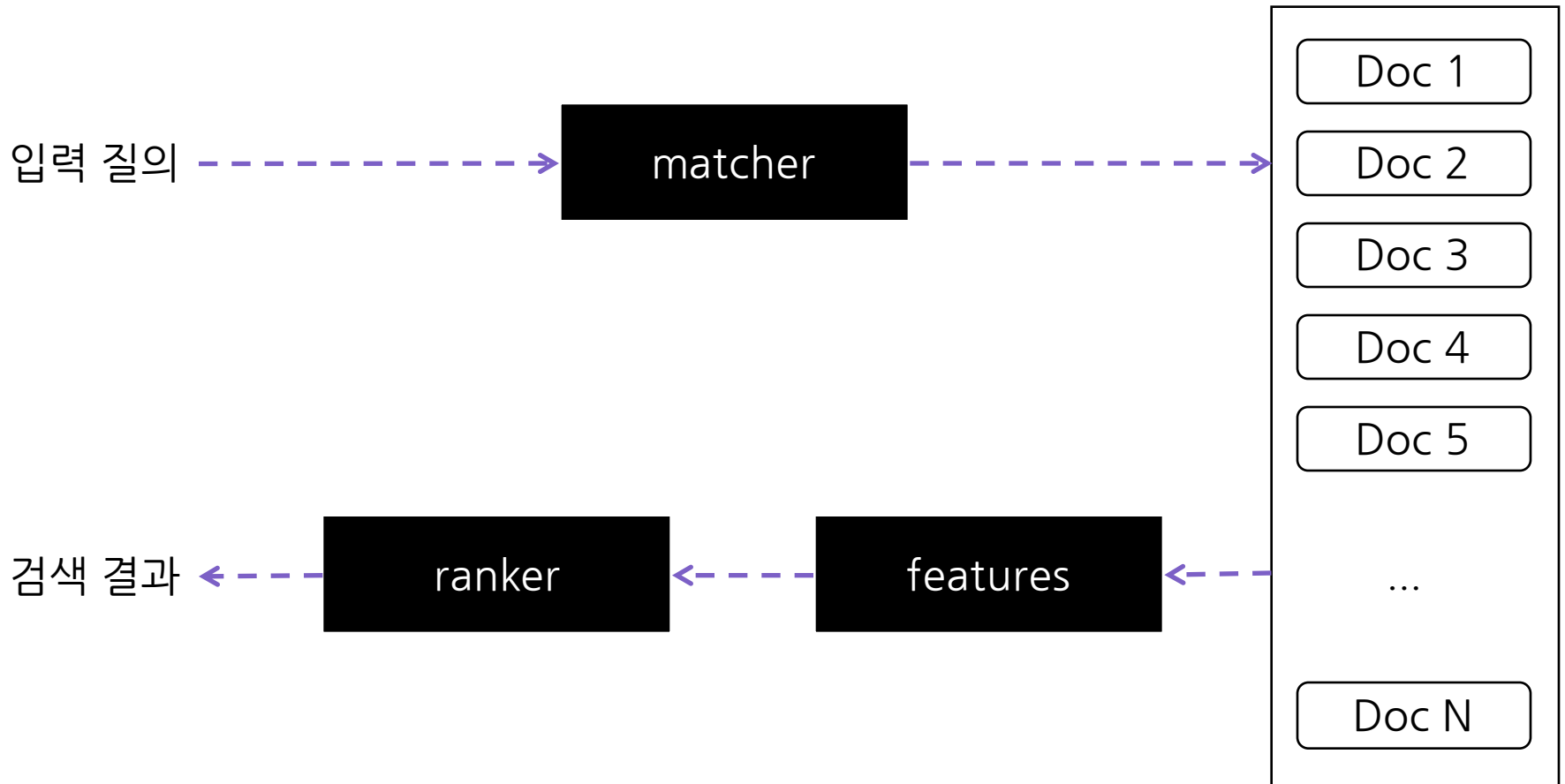
- Yahoo에서 주관한 Learning-to-rank challenge (2011년)

|          | Validation |         | Test    |         |
|----------|------------|---------|---------|---------|
|          | ERR        | NDCG    | ERR     | NDCG    |
| BM25F-SD | 0.42598    | 0.73231 | 0.42853 | 0.73214 |
| RankSVM  | 0.43109    | 0.75156 | 0.43680 | 0.75924 |
| GBDT     | 0.45625    | 0.78608 | 0.46201 | 0.79013 |

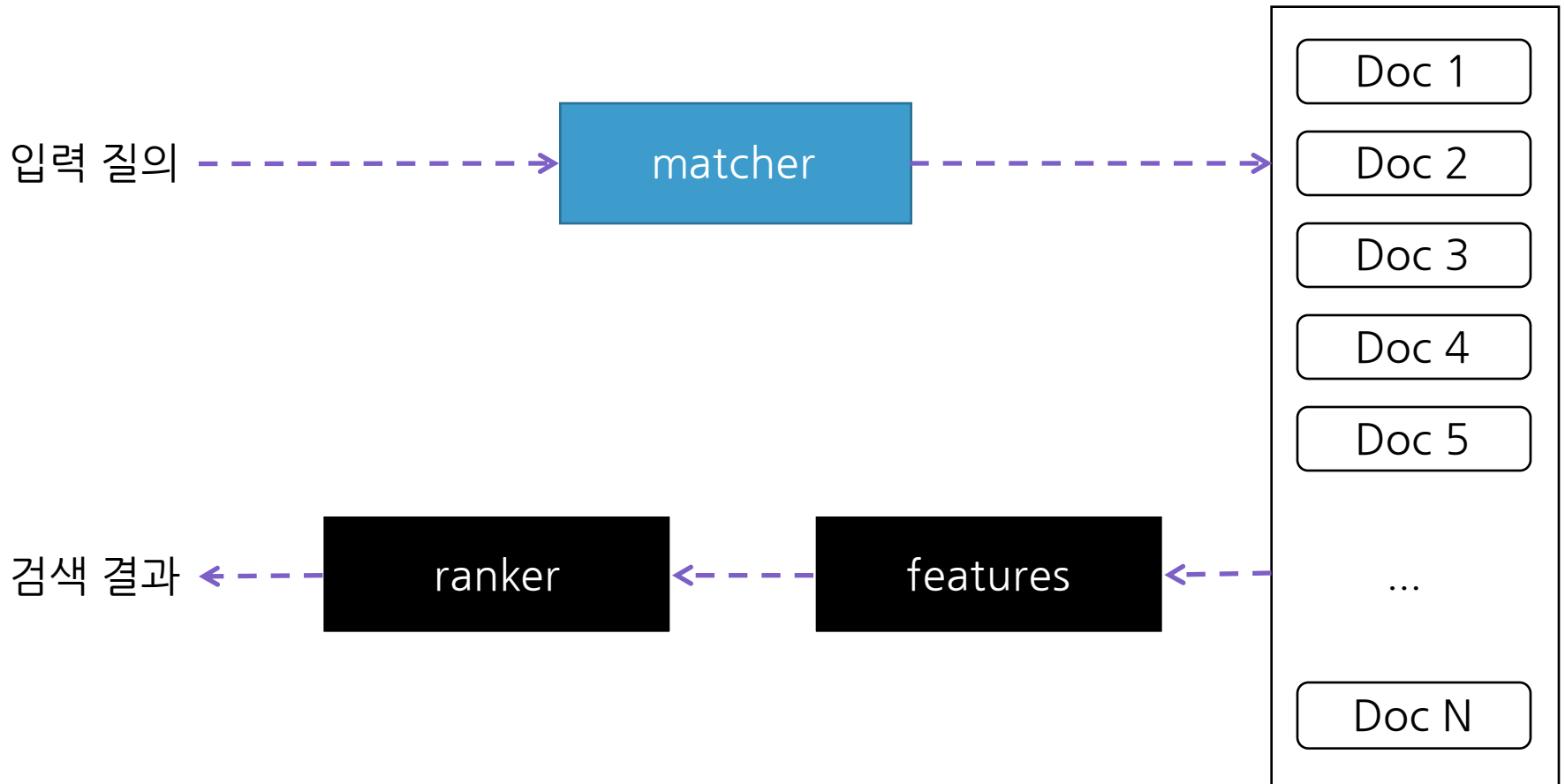
<http://proceedings.mlr.press/v14/chapelle11a/chapelle11a.pdf>

랭킹이 이상한건 조작이 아니라  
기술력의 한계때문으로 이해해주셔야 합니다

# 검색엔진 핵심모듈 : Matching and Ranking

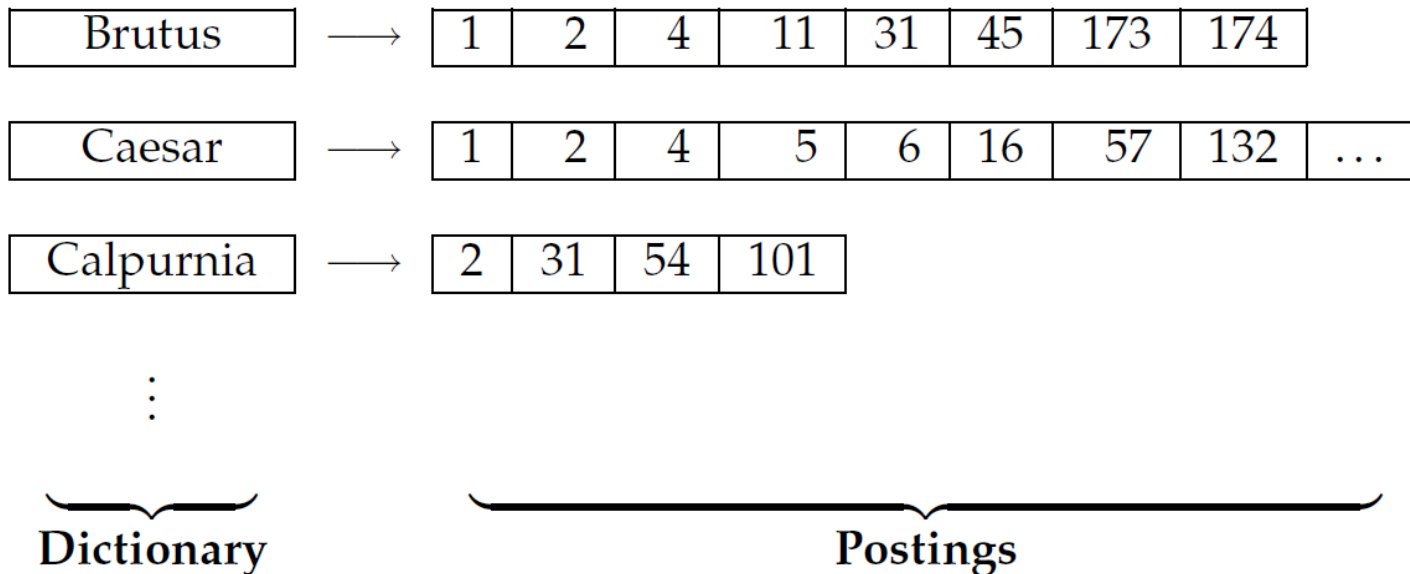


# 검색엔진 핵심모듈 : Matching and Ranking



# Matching

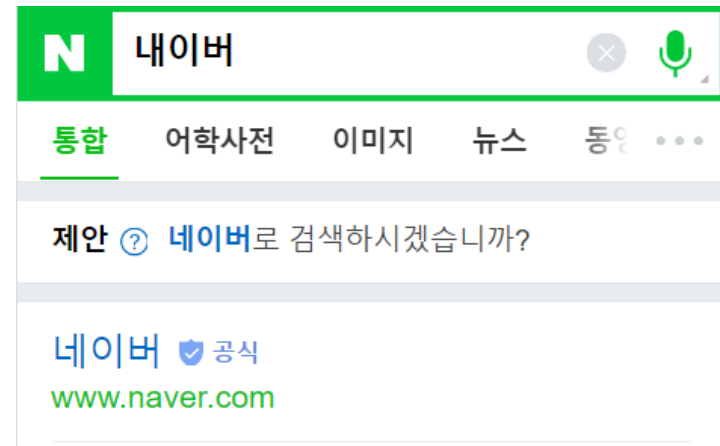
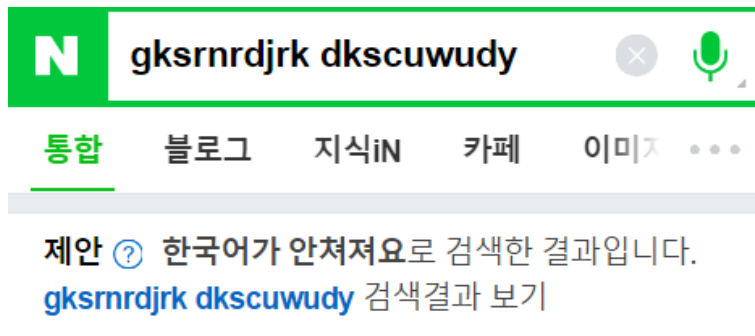
- 사용자가 입력한 질의와 관련이 있는 문서를 일단 모두 추려내는 일로, 색인(Indexing)에 의해 가능
  - 1952년, Taube 등에 의해 처음 제안됨. 그 전까지는, 많은 문서들은 **카테고리 코드**를 부착하는 식으로 관리되었음.
  - 그 당시는 매우 급진적인 아이디어로 인식되었으나, 현재는 너무 당연한 구조로 받아들여짐.





# Matching

- 현재까지는 단어기반 매칭이 주류
- 단어기반 매칭은 형태론적 분석, 동의어처리가 도전과제
  - university / universities, rewritten / rewriting
  - 대학생선교회 / 대학주변교회
  - 이탈리아 / 이태리, UN / United Nations, to be or not to be
- 그 외에 질의오류교정 등의 기술이 서비스에 적용됨



# Semantic Matching

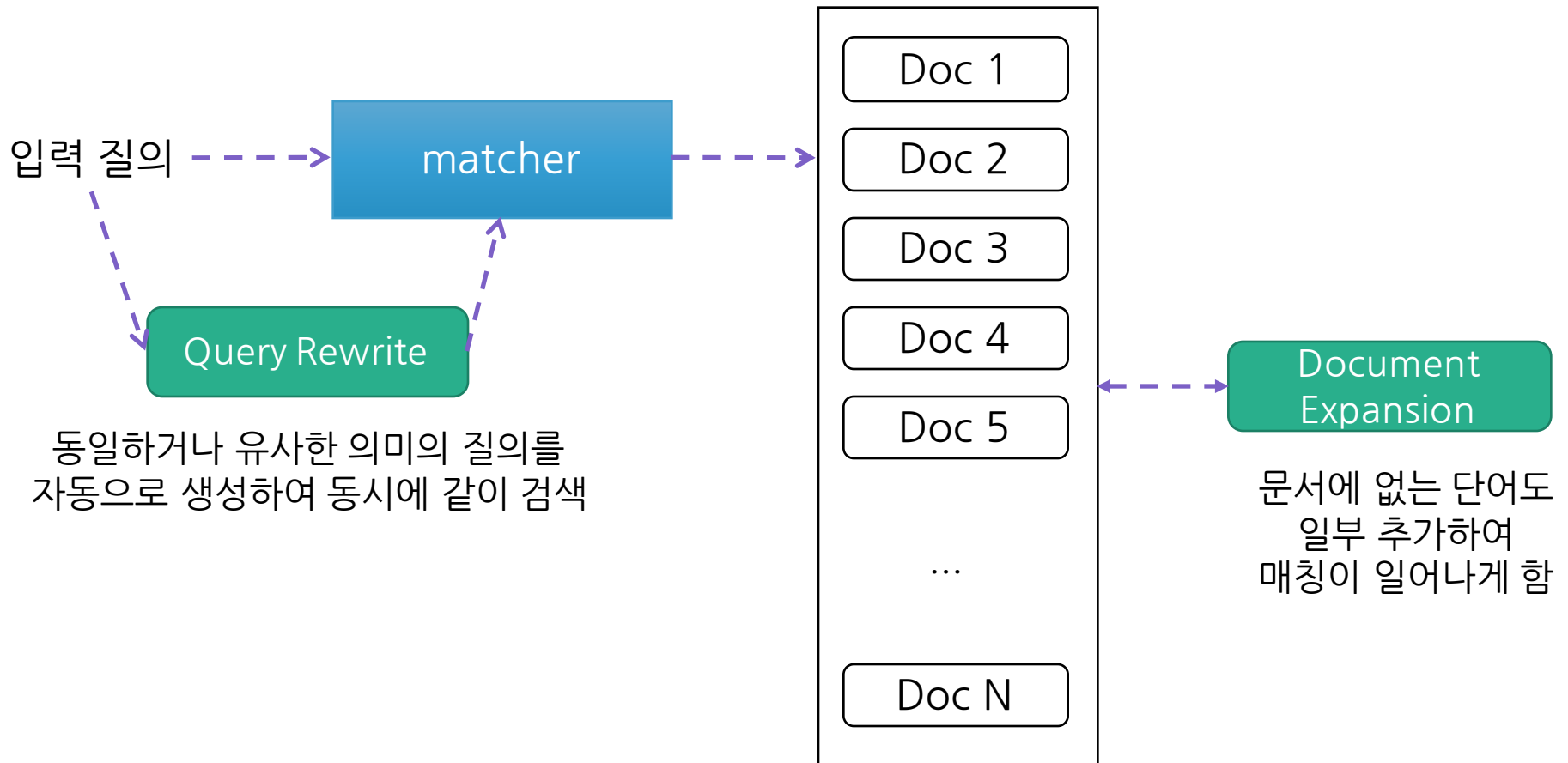
- 질의단어를 조금 다르게 입력해도 잘 검색이 될 수 있도록!

**N** 롯데월드 이용요금

**N** 롯데월드 입장료

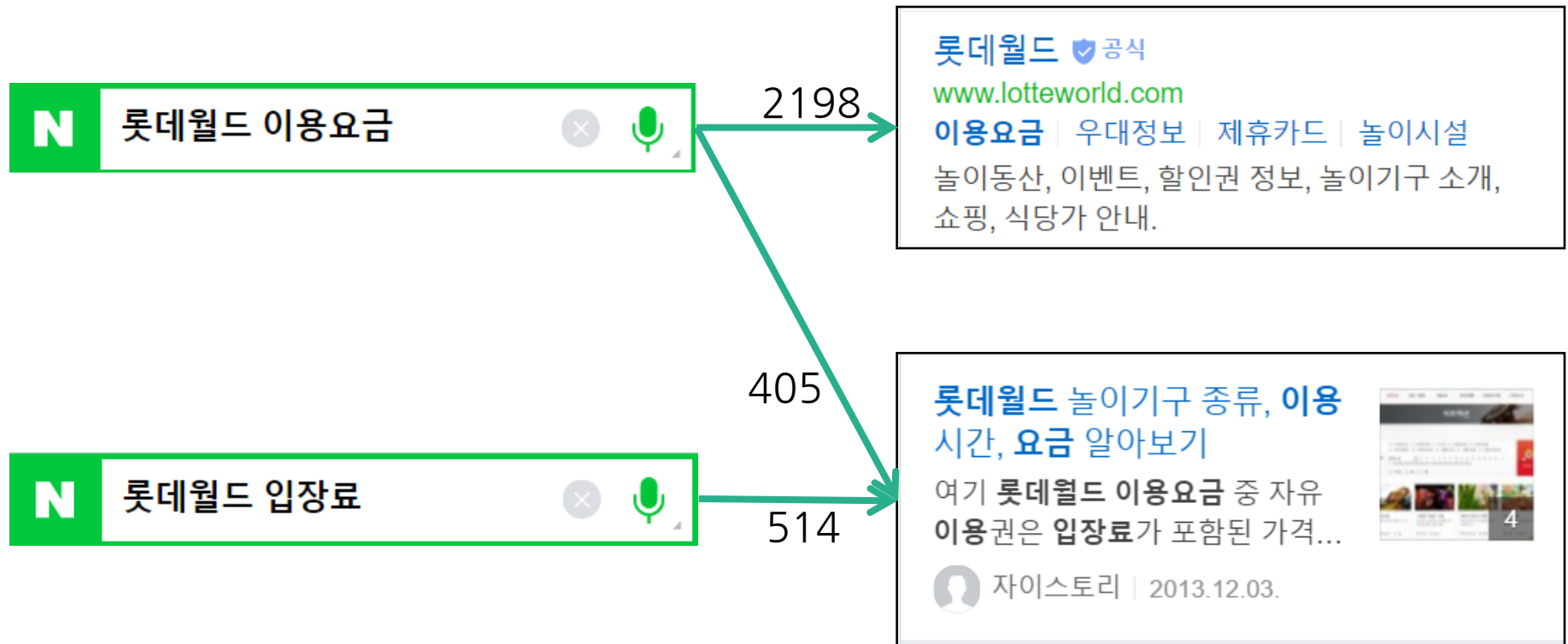
롯데월드  공식  
[www.lotteworld.com](http://www.lotteworld.com)  
이용요금 | 우대정보 | 제휴카드 | 놀이시설  
놀이동산, 이벤트, 할인권 정보, 놀이기구 소개,  
쇼핑, 식당가 안내.

# Semantic Matching @ NAVER



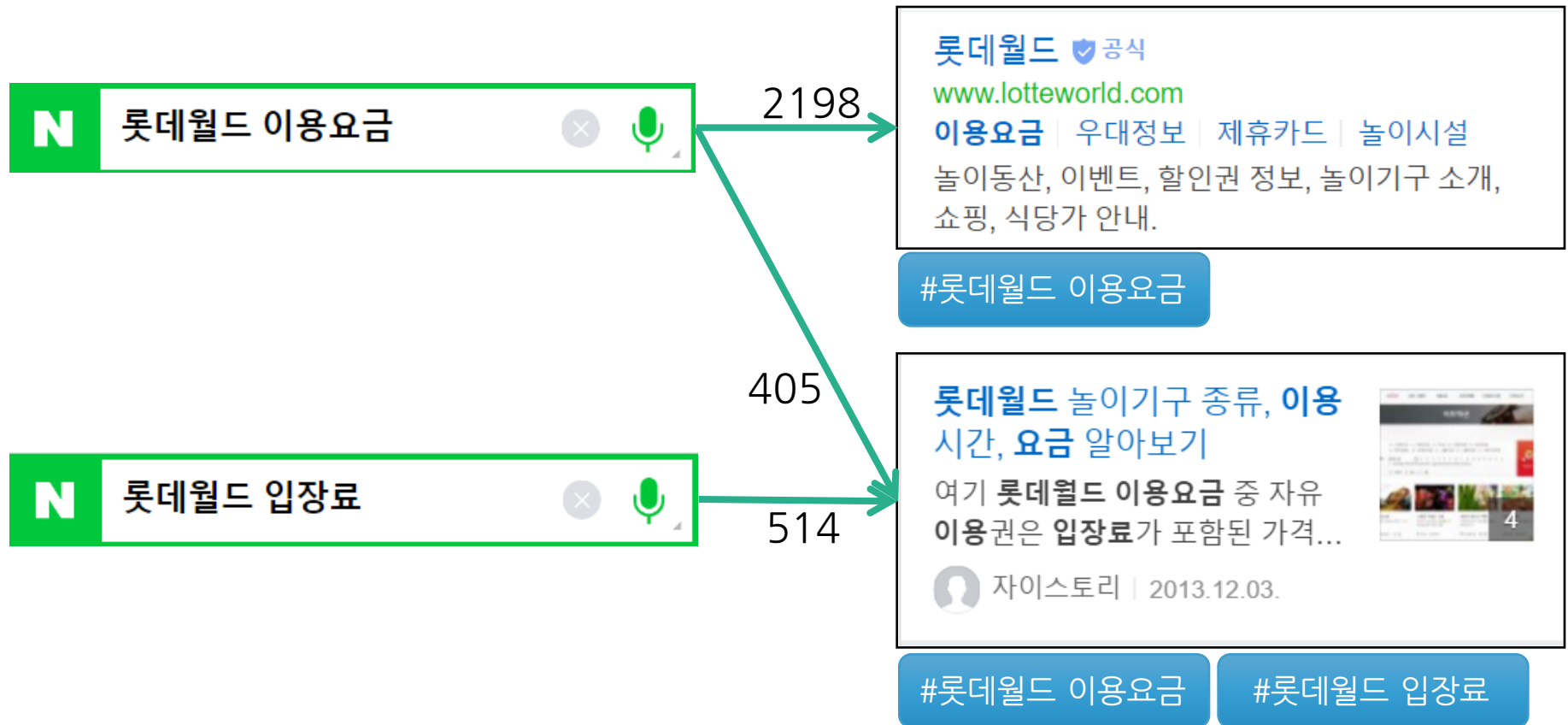
# Document Expansion

- QueryText = 사용자가 문서를 클릭했을 때 이용한 쿼리



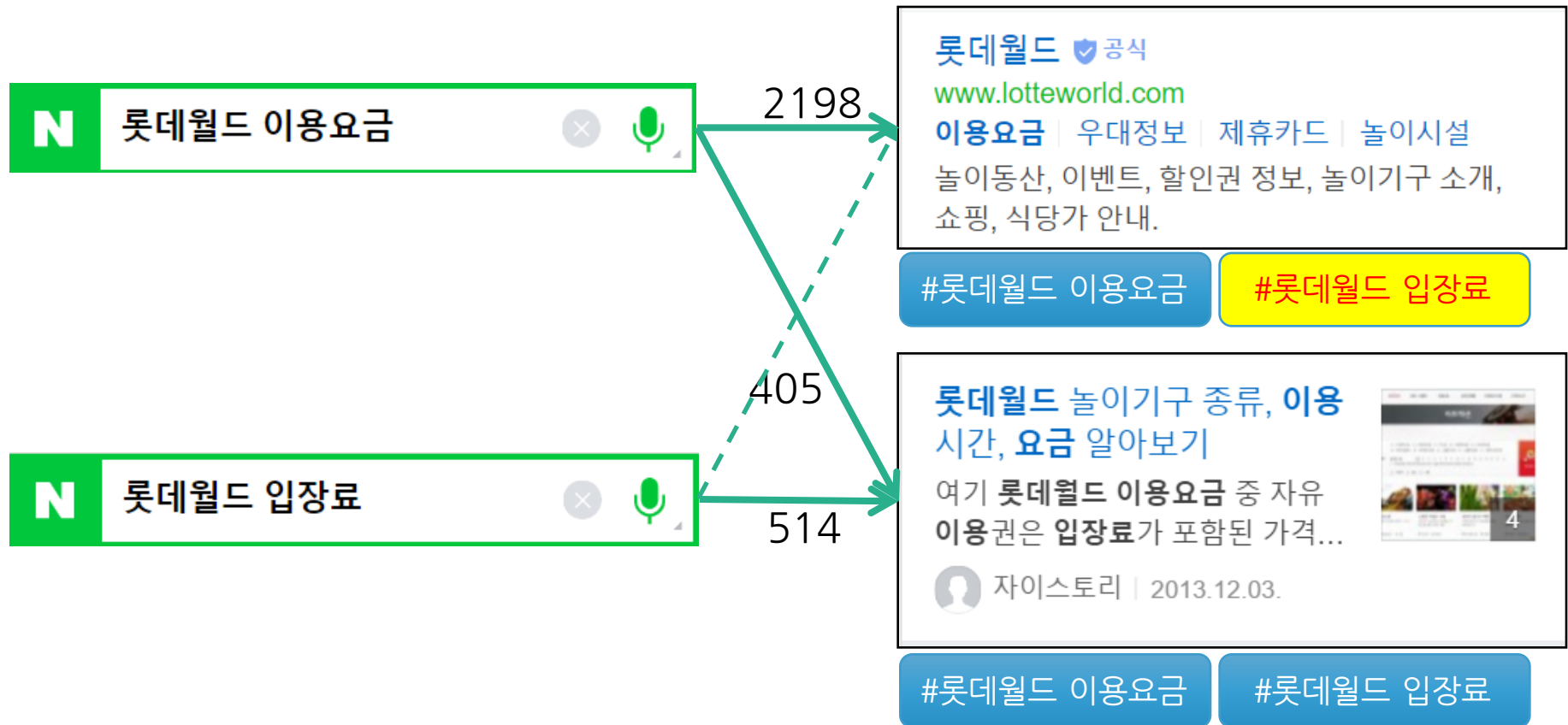
# Document Expansion

- QueryText는 검색에서 중요한 시그널로 사용됨



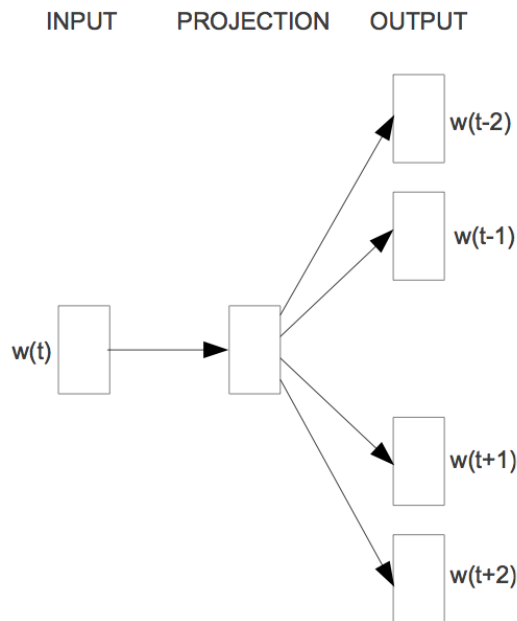
# Document Expansion

- 클릭그래프에서 이웃간에는 관련성이 있지 않을까?



# Document Expansion

- 질의와 문서를 노드, 클릭발생을 간선으로 하여 그래프를 구성
- 그래프의 모든 노드를 벡터로 임베딩
- word2vec과의 analogy :

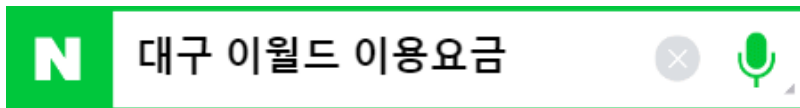


word2vec 에서의 단어~주변단어  
→ node2vec에서의 노드~주변노드

클릭은 매일 엄청나게 발생하므로,  
incremental 한 embedding 진화가 중요

# Document Expansion

- 앞의 방식은 클릭이 어느 정도 존재해야 가능하다는 문제가  $\pi$

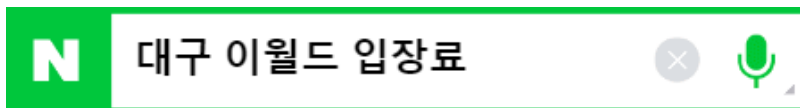


이월드 공식

[www.eworld.kr](http://www.eworld.kr)

[이용요금](#) | [할인정보안내](#) | [제휴사할인](#)

유럽식 도시공원 **이월드** 83타워, 아이스링크, 매직아트, 푸드폴리탄, 두루온, 83그릴바이애슬리...



대구 이월드 입장료만 내고 사진찍으며 데이트했네요

대구 이월드 입장료만 내고 사진찍으며 데이트했네요 항상 놀이공원으로 놀러가면 자유이용권을..





# Document Expansion

- 자동번역기법을 응용해보자!

- Click Graph Embedding에 의해 발견된 유사질의쌍을 바탕으로

|              |             |
|--------------|-------------|
| 롯데월드 이용요금    | 롯데월드 입장료    |
| 에버랜드 이용요금    | 에버랜드 입장료    |
| 에코랜드 이용요금 할인 | 에코랜드 입장료 할인 |

- 자동번역기법(SMT/NMT)을 활용하여 유사질의를 생성

|             |     |
|-------------|-----|
| 대구 이랜드 이용요금 | ??? |
|-------------|-----|

# Document Expansion

- 자동번역기법을 응용해보자!

- Click Graph Embedding에 의해 발견된 유사질의쌍을 바탕으로

---

|              |             |
|--------------|-------------|
| 롯데월드 이용요금    | 롯데월드 입장료    |
| 에버랜드 이용요금    | 에버랜드 입장료    |
| 에코랜드 이용요금 할인 | 에코랜드 입장료 할인 |

---

- 자동번역기법(SMT/NMT)을 활용하여 유사질의를 생성

---

|             |            |
|-------------|------------|
| 대구 이랜드 이용요금 | 대구 이랜드 입장료 |
|-------------|------------|

---

- “수줍게” 문서에 넣어보고... 클릭이 발생하면... CGE 로 선순환
- 클릭이 발생하지 않으면... negative example

# 잠깐!

- 왜 딥러닝 기반 번역이 났을까?



|              |             |
|--------------|-------------|
| 롯데월드 이용요금    | 롯데월드 입장료    |
| 에버랜드 이용요금    | 에버랜드 입장료    |
| 에코랜드 이용요금 할인 | 에코랜드 입장료 할인 |

|             |            |
|-------------|------------|
| 대구 이랜드 이용요금 | 대구 이랜드 입장료 |
|-------------|------------|

|          |         |
|----------|---------|
| 렌터카 이용요금 | 렌터카 ??? |
|----------|---------|

# 잠깐!

- 단어의 의미가 벡터로 임베딩되는 특징때문

롯데월드 이용요금

롯데월드 입장료

에버랜드 이용요금

에버랜드 입장료

에코랜드 이용요금 할인

에코랜드 입장료 할인

대구 이랜드 이용요금

대구 이랜드 입장료

렌터카 이용요금

렌터카 이용료

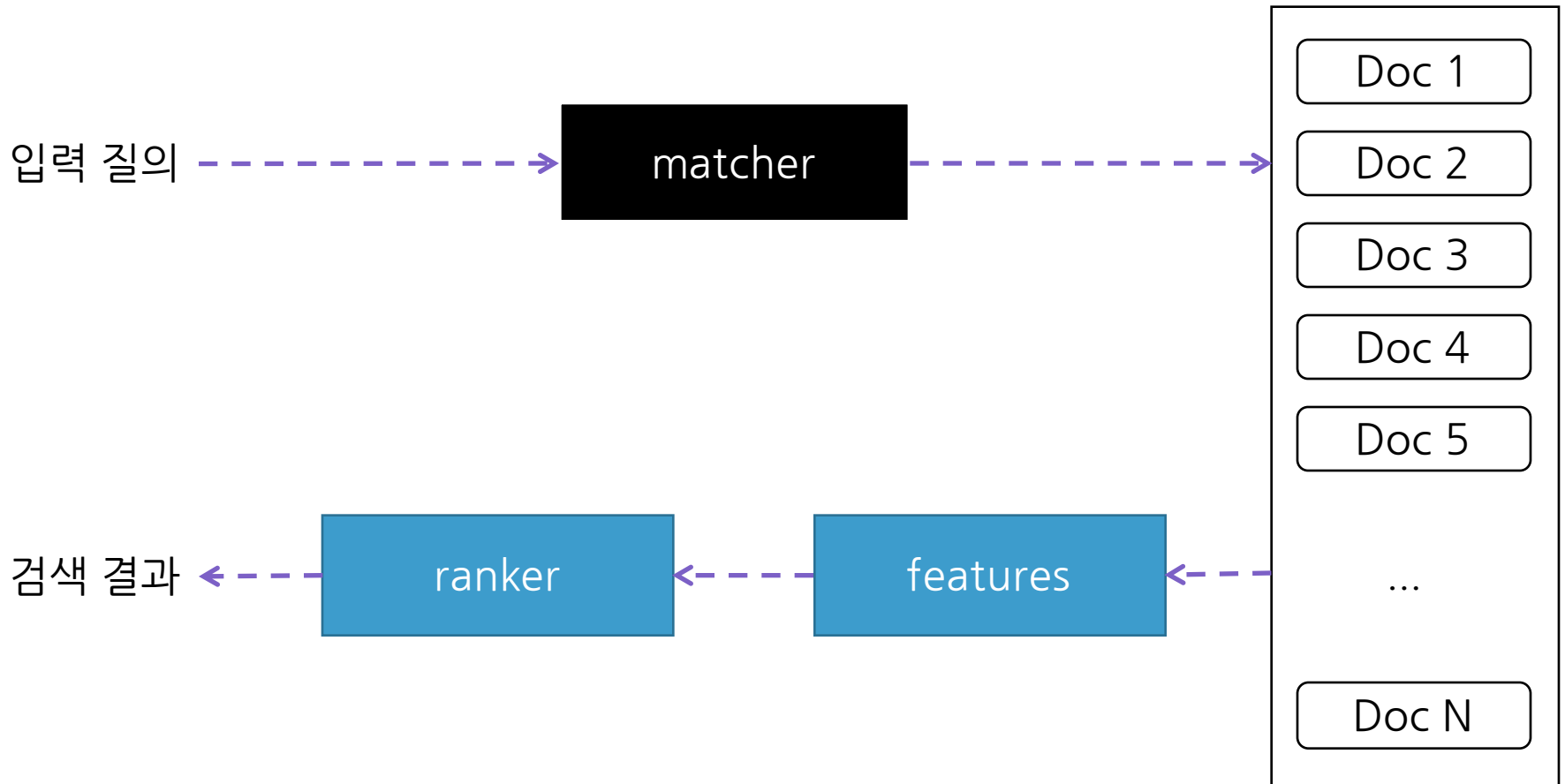
파리 지하철 이용요금

파리 지하철 이용료

한강 유람선 이용요금

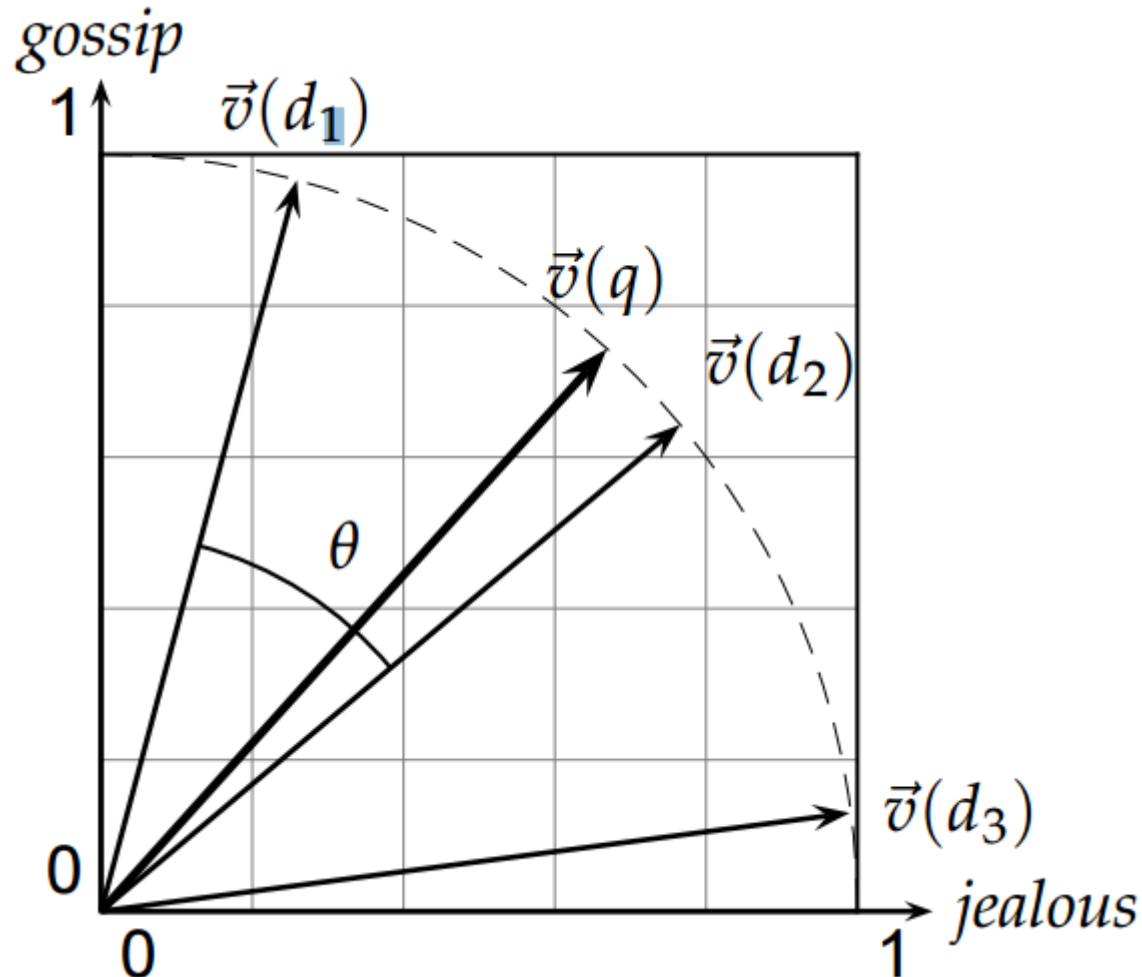
한강 유람선 이용료

# 검색엔진 핵심모듈 : Matching and Ranking



# Ranking : Old Approaches

- Vector Space Model (1975년 )



# Ranking : Old Approaches

- Probabilistic Ranking Model (1994년, a.k.a BM25)

$$P(\text{rel} | d, q) \propto_q \frac{P(\text{rel} | d, q)}{P(\overline{\text{rel}} | d, q)} \quad (2.1)$$

$$= \frac{P(d | \text{rel}, q)}{P(d | \overline{\text{rel}}, q)} \frac{P(\text{rel} | q)}{P(\overline{\text{rel}} | q)} \quad (2.2)$$

$$\propto_q \frac{P(d | \text{rel}, q)}{P(d | \overline{\text{rel}}, q)} \quad (2.3)$$

$$\approx \prod_{i \in V} \frac{P(TF_i = tf_i | \text{rel}, q)}{P(TF_i = tf_i | \overline{\text{rel}}, q)} \quad (2.4)$$

$$\begin{aligned} w_i^{\text{BM25}}(tf) &= \frac{tf'}{k_1 + tf'} w_i^{\text{RSJ}} \\ &= \frac{tf}{k_1 \left( (1 - b) + b \frac{dl}{avdl} \right) + tf} w_i^{\text{RSJ}} \end{aligned}$$

1. 한계효용의 법칙
2. 여러 단어 질의에 대한 and/or 제어

# Ranking : Old Approaches

- Language Model based IR (1998)

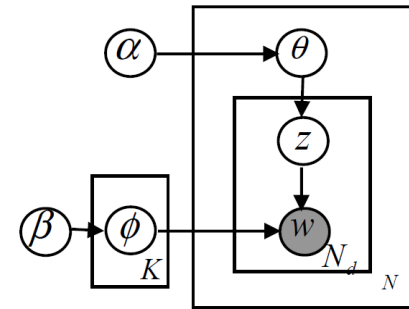
$$P(d|q) \propto P(d) \prod_{t \in q} ((1 - \lambda)P(t|M_c) + \lambda P(t|M_d))$$

✓ Simple smoothing

$$\hat{P}(t|d) = \frac{\text{tf}_{t,d} + \alpha \hat{P}(t|M_c)}{L_d + \alpha}$$

✓ LDA (2006)

$$P(w | D) = \lambda \left( \frac{N_d}{N_d + \mu} P'(w | D) + \left(1 - \frac{N_d}{N_d + \mu}\right) P'(w | \text{coll}) \right) \\ + (1 - \lambda) \left( \sum_{t=1}^K \frac{n_{-i,j}^{(w_i)} + \beta_{w_i}}{\sum_{v=1}^V (n_{-i,j}^{(v)} + \beta_v)} \times \frac{n_{-i,j}^{(d_i)} + \alpha_{z_i}}{\sum_{t=1}^T (n_{-i,t}^{(d_i)} + \alpha_t)} \right) \quad (9)$$





# 2000년대 초반 학교 밖 세상에는...

- 인터넷에 데이터가 폭발적으로 늘어나기 시작
  - 야후, 구글, 네이버, 다음 등 검색서비스 업체들이 자리잡음
  - 사람들이 검색서비스를 통해 인터넷에서 무언가를 뒤지기 시작
  - 데이터 ( 문서, 클릭로그 ) 가 쌓임
- 
- 기계학습 전공자들은 데이터를 보면 **흥분**함

# 검색은 전형적인 기계학습문제 아닌가!

- 실제로 좋은 검색결과를 만들기 위해 많은 시그널(피쳐)을 사용
  - 질의에 있는 단어가 얼마나 많이, 또 얼마나 가깝게 나타났는가?
  - 많은 사람들이 보았거나 링크를 걸었는가?
  - 다른 사람이 어떤 단어로 링크를 걸었는가?
  - 문서를 작성한 사람이 과거에 스팸작성자로 경고조치 됐었는가?
  - 언제 만들어진 문서인가?
  - 몇 명이 조회한 문서인가?
  - 문서의 제목과 본문의 관련성은 높은가?
- (노이즈는 좀 있지만) 비교적 대규모의 학습집합을 구축하게 됨
  - 클릭여부 와 그 위치(랭크)
  - Last Click
  - Quick Close

# Learning-to-Rank Overview

- Feature vector 생성의 예

$q_1$

$x_1^{(1)}$

$x_2^{(1)}$

$\vdots$

$x_m^{(1)}$

$y^{(1)}$

= “한국 경제 전망”

= [ 0.66 , 2 , 0.08 , 1 , 3 , 1 , 0 , 4.2 , 25 , 0.43 ]

질의 단어 중 제목에 나타난 단어의 비율

제목에 출현한 질의 단어의 총 합

질의 단어별 “본문출현빈도/본문길이”의 합

질의 단어 중 본문에 나타난 단어의 비율

질의의 단어수

전문정보를 찾는 질의인가?

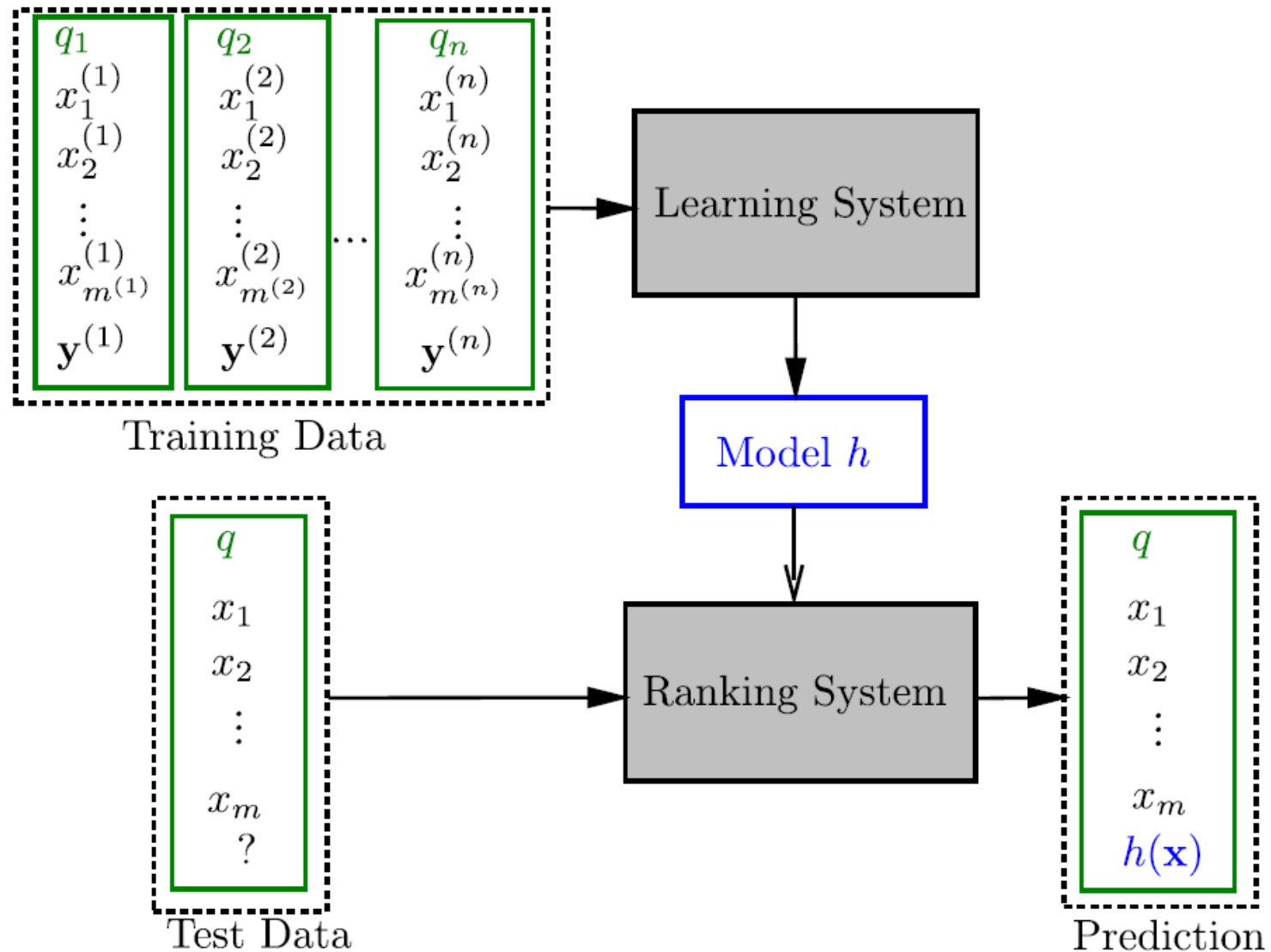
질의에 연예인 이름이 포함되어 있나?

문서가 포함된 사이트의 Site Authority

문서의 나이(Age)

문서의 품질(Quality)

# Learning-to-Rank Overview



# Ranking SVM

- Problem Definition
  - Input space:  $X$
  - Ranking function  $f : X \rightarrow R$
  - Ranking:  $x_i \succ x_j \Leftrightarrow f(x_i; w) > f(x_j; w)$
  - Linear ranking function:  $f(x; w) = \langle w, x \rangle$   
 $\langle w, x_i - x_j \rangle > 0 \Leftrightarrow f(x_i; w) > f(x_j; w)$
  - Transforming to pairwise classification:

$$(x_i - x_j, z), \quad z = \begin{cases} +1 & x_i \succ x_j \\ -1 & x_j \succ x_i \end{cases}$$

# Ranking SVM

- Solution

$$\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i$$

$$z_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i \quad i = 1, \dots, l$$

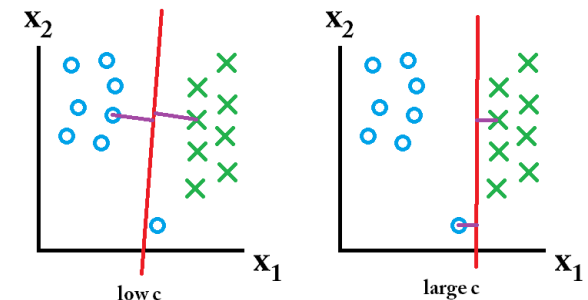
$$\xi_i \geq 0$$



$$\min_w \sum_{i=1}^l \left[ 1 - z_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \right]_+ + \lambda \|w\|^2$$

$$[s]_+ = \max(0, s) \quad \lambda = \frac{1}{2C}$$

Slack variable 의 도입  
 → “틀린 정도” 도 모델링



C는 overfitting을 제어하는 hyperparameter

# Ranking SVM : 한계

- Problems
  - Error 에도 그 중요도가 있는데 반영을 못한다  
(검색의 특수성을 반영한 평가척도를 직접 최적화하지는 못함)
  - Query별 labeled 문서수에 따라 bias가 생길 수 있다

# Ranking SVM : 한계

- Problems

- Error 에도 그 중요도가 있는데 반영을 못한다  
(검색의 특수성을 반영한 평가척도를 직접 최적화하지는 못함)

|     |                     |
|-----|---------------------|
| 최적  | 3 2 2 2 1 1 1 1 1 1 |
| 모델1 | 3 2 2 1 2 1 1 1 1 1 |
| 모델2 | 2 3 2 2 1 1 1 1 1 1 |

- Query별 labeled 문서수에 따라 bias가 생길 수 있다



# Ranking SVM : 한계

- Problems

- Error 에도 그 중요도가 있는데 반영을 못한다  
(검색의 특수성을 반영한 평가척도를 직접 최적화하지는 못함)

|     |                     |
|-----|---------------------|
| 최적  | 3 2 2 2 1 1 1 1 1 1 |
| 모델1 | 3 2 2 1 2 1 1 1 1 1 |
| 모델2 | 2 3 2 2 1 1 1 1 1 1 |

- Query별 labeled 문서수에 따라 bias가 생길 수 있다

Q1 : 3 2 2 1 1 1 1

→  $2 + 4 + 8 = 14$  개의 학습용 pairwise data 생성

Q2 : 3 3 2 2 2 1 1 1 1 1

→  $6 + 10 + 15 = 31$  개의 학습용 pairwise data 생성

# Ranking SVM : 한계

- Problems

- Error 에도 그 중요도가 있는데 반영을 못한다  
(검색의 특수성을 반영한 평가척도를 직접 최적화하지는 못함)

|     |                     |
|-----|---------------------|
| 최적  | 3 2 2 2 1 1 1 1 1 1 |
| 모델1 | 3 2 2 1 2 1 1 1 1 1 |
| 모델2 | 2 3 2 2 1 1 1 1 1 1 |

- Query별 labeled 문서수에 따라 bias가 생길 수 있다

Q1 : 3 2 2 1 1 1 1

→  $2 + 4 + 8 = 14$  개의 학습용 pairwise data 생성

Q2 : 3 3 2 2 2 1 1 1 1 1

→  $6 + 10 + 15 = 31$  개의 학습용 pairwise data 생성

개선된 RankSVM이나 Listwise 접근법 등 다양한 연구가 진행됨

# 더 근본적인 한계1 : “One size NEVER fits all !”

**N** 증권거래세



**N** 썰전



**N** 탄핵



**N** 여권 사진 규정



**N** 새학기 친구사귀기



“모든 유형의 질의에 걸맞는  
단 하나의 랭킹 함수는  
존재하기 어려움”

# 더 근본적인 한계2 : “랭킹”만 잘하면 되나?

**N** 인공지능

지식백과 어린이 2 | 학생 3

**인공지능** artificial intelligence, 人工知能  
인간의 학습능력과 추론능력, 지각능력, 자연언어의 ...  
인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해능력 등을 컴퓨터 프로그램으로 실현한 기술. 인간의 지...  
두산백과



체스는 이미 인간이 컴퓨터에 정복당한 영역 중 하나이며, 인공지능 기술은 빠르게 발전하고 있다. 2016년 3월 12...  
네이버캐스트



지식백과 더보기 →

**N** 리니지2 레볼루션

통합 공략 게시판 카페 이미지 ...

사이트 ? 사이트 신고

**리니지2 레볼루션 - 넷마블**  
[l2.netmarble.net](http://l2.netmarble.net)  
리니지2 레볼루션 공식 홈페이지 - 모바일로 탄생한 최초의 리니지를 만나보세요!

**리니지2 레볼루션 공식카페**   
[cafe.naver.com/l2mobile](http://cafe.naver.com/l2mobile)  
리니지2 레볼루션 공식카페, 공지사항, 이벤트안내, 게임가이드, 게임영상 등 수록.

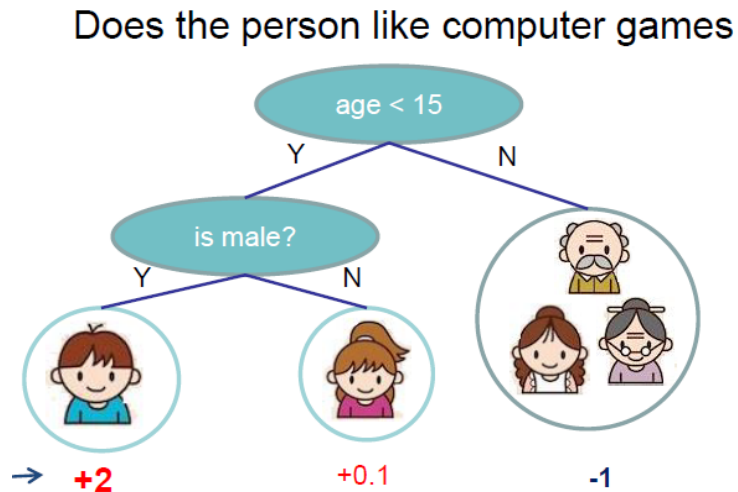
**리니지2 레볼루션 인벤**  
[l2mobile.inven.co.kr](http://l2mobile.inven.co.kr)  
리니지2 레볼루션 인벤은 리니지2 레볼루션과 관련된 모든 게임 정보를 제공합니다. 뉴스, 프리뷰, 리뷰, 공...

사이트 더보기 →

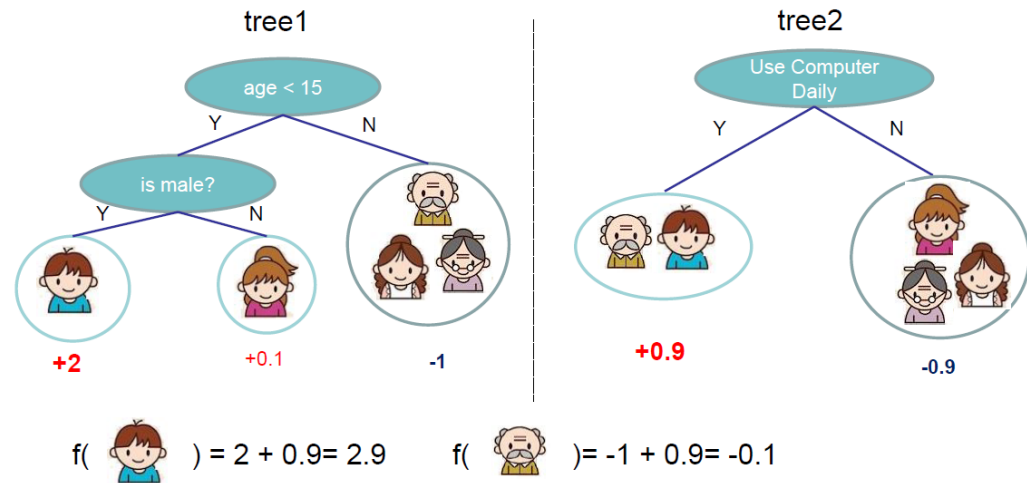
“랭킹뿐 아니라, 그 결과가 얼마나 적합한지도 알아야 함”

# Ranking and Regression : GBRT

- Regression Tree



- Regression Tree Ensemble



- Model: assuming we have K trees

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F}$$

- Objective

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Training loss

Complexity of the Trees

SVM, NN, LR같이  
학습방법이 잘 연구되어온  
Numerical vector/matrix기반  
classifier가 아니라서...  
학습방법 자체가 큰 연구토픽

# Ranking and Regression : GBRT

- GBRT (Gradient Boosted Regression Tree)

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$

...

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad \leftarrow \text{New function}$$

Model at training round t

Keep functions added in previous round

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i) \\ &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant \end{aligned}$$

Goal: find  $f_t$  to minimize this

$$= \sum_{i=1}^n \left( y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)) \right)^2 + \Omega(f_t) + const$$

$$= \sum_{i=1}^n \left[ 2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2 \right] + \Omega(f_t) + const$$

# Ranking and Regression : GBRT

- GBRT (Gradient Boosted Regression Tree)

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i)$$

Model at traini

$$Obj^{(t)} =$$

$$=$$

$$=$$

$$= \sum_{i=1}^n \left[ 2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2 \right] + \Omega(f_t) + const$$



on

minimize this

st

# Thankful open sources!

## Getting started: an Example Problem

You will find an example problem at

[http://download.joachims.org/svm\\_light/examples/example3.tar.gz](http://download.joachims.org/svm_light/examples/example3.tar.gz)

It consists of 3 rankings (i.e. queries) with 4 examples each. It also contains a file with 4 test examples. Unpack the archive with

```
gunzip -c example3.tar.gz | tar xvf -
```

This will create a subdirectory `example3`. To run the example, execute the commands:

```
svm_rank_train example3/train.dat example3/model
svm_rank_classify example3/test.dat example3/model
```

The output  
so, you will  
do not have  
equivalent

```
svm_learn
```

Note the di

It can also b  
equivalent c  
are misorde  
to the traini

```
>>> import numpy as np
>>> from sklearn.metrics import mean_squared_error
>>> from sklearn.datasets import make_friedman1
>>> from sklearn.ensemble import GradientBoostingRegressor

>>> X, y = make_friedman1(n_samples=1200, random_state=0, noise=1.0)
>>> X_train, X_test = X[:200], X[200:]
>>> y_train, y_test = y[:200], y[200:]
>>> est = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1,
...     max_depth=1, random_state=0, loss='ls').fit(X_train, y_train)
>>> mean_squared_error(y_test, est.predict(X_test))
5.00...
```

```
svm_rank_classify example3/train.dat example3/model example3/predictions.train
```



# Thankful open sources!

## Getting started: an Example Problem

You will find an example problem at

[http://download.joachims.org/svm\\_light/examples/example3.tar.gz](http://download.joachims.org/svm_light/examples/example3.tar.gz)

It consists of 3 rankings (i.e. queries) with 4 examples each. It also contains a file with 4 test examples. Unpack the archive with

```
gunzip -c example3.tar.gz | tar xvf -
```

This will create a subdirectory `example3`. To run the example, execute the commands:

```
svm_rank  
svm_rank
```

The output  
so, you will  
equivalent

일단 샘플 데이터로 돌려보고,  
괜찮다 싶으면 우리 데이터로도 돌려보면서 감 잡고,  
이거다 싶으면 좀 더 개념을 파보고,  
필요하면 코드도 분석하고 수정해서 쓰거나 새로 만들기도 하고...

```
>>> X, y = make_friedman1(n_samples=1200, random_state=0, noise=1.0)  
>>> X_train, X_test = X[:200], X[200:]
```

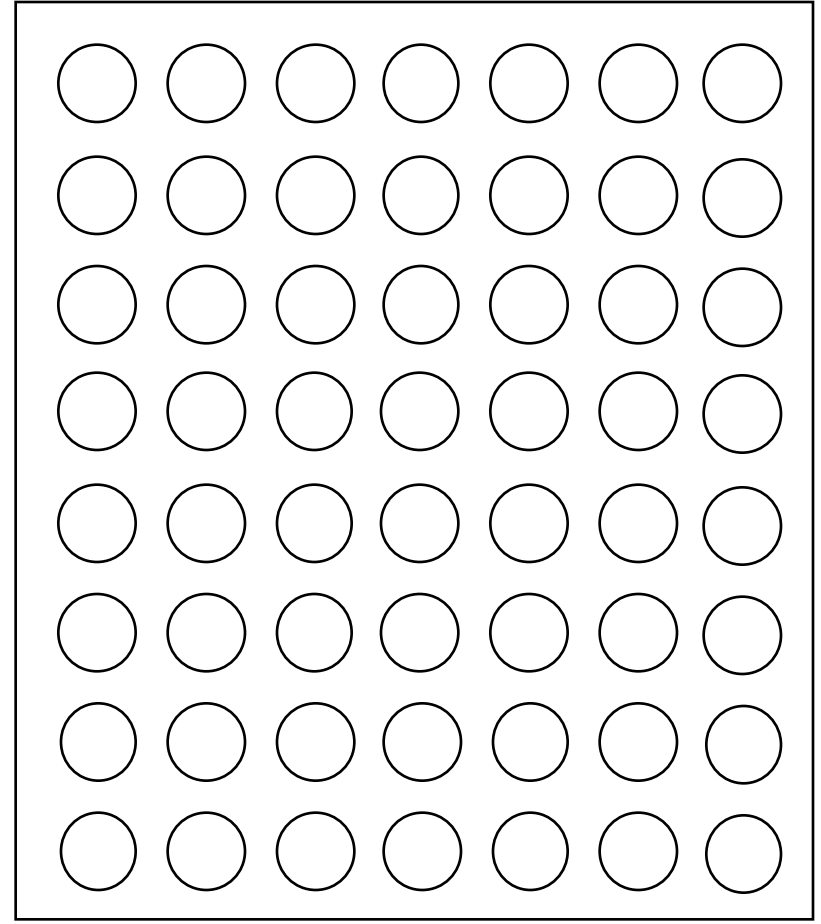
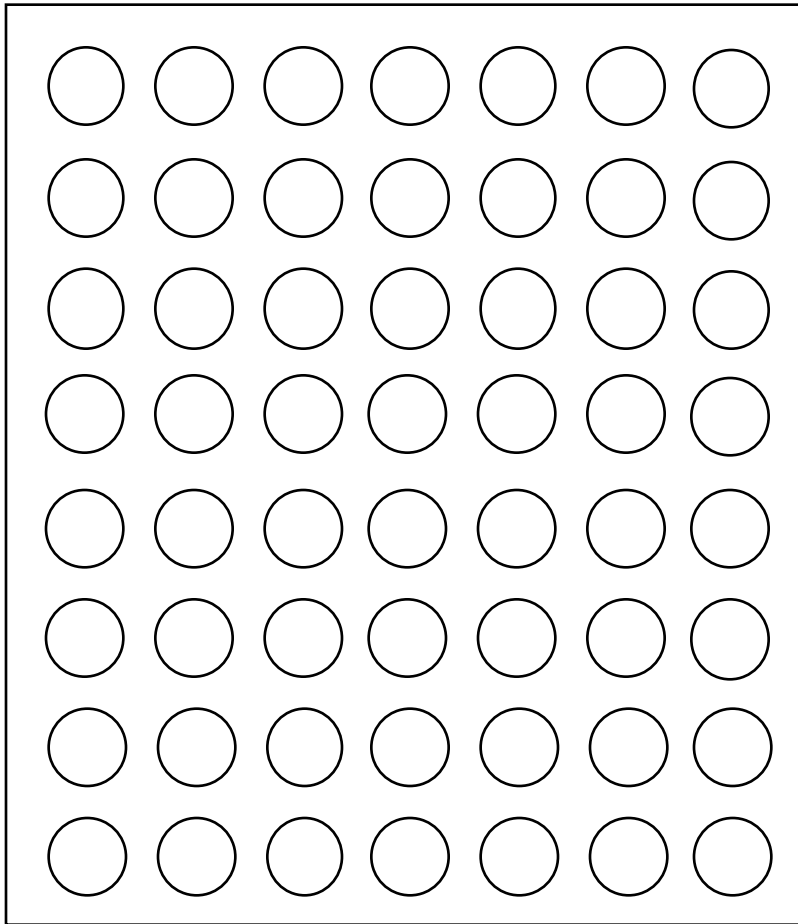
Note the di  
It can also b  
equivalent c  
are misorde  
to the traini

```
>>> y_train, y_test = y[:200], y[200:]  
>>> est = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1,  
...     max_depth=1, random_state=0, loss='ls').fit(X_train, y_train)  
>>> mean_squared_error(y_test, est.predict(X_test))  
5.00...
```

```
svm_rank_classify example3/train.dat example3/model example3/predictions.train
```

# 새로운 시도들

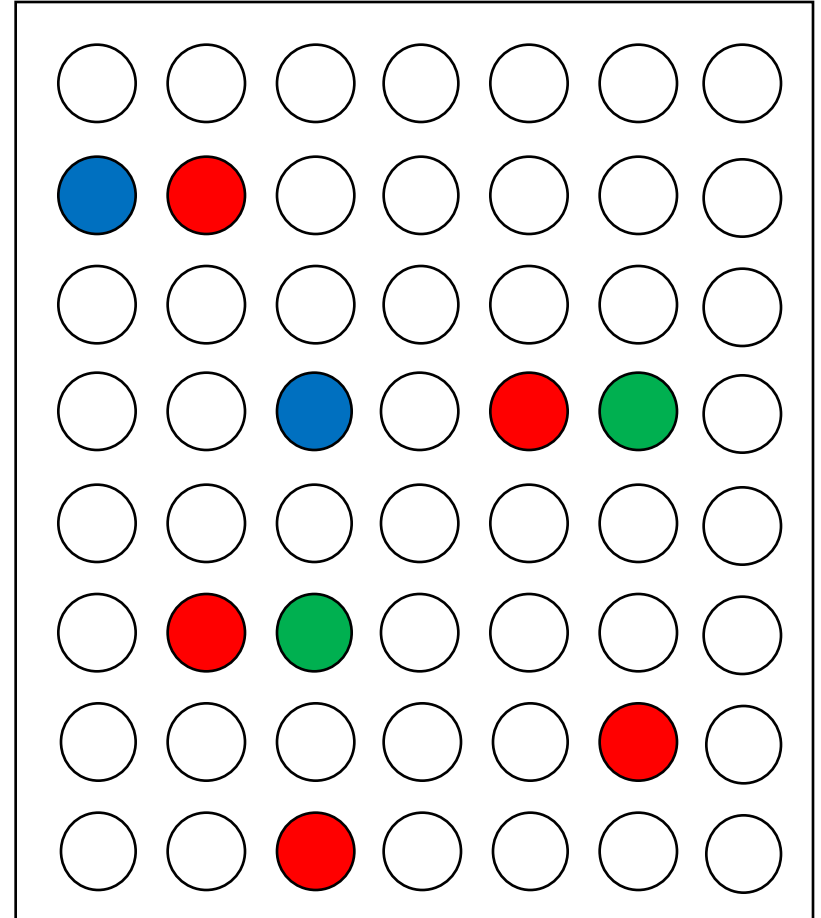
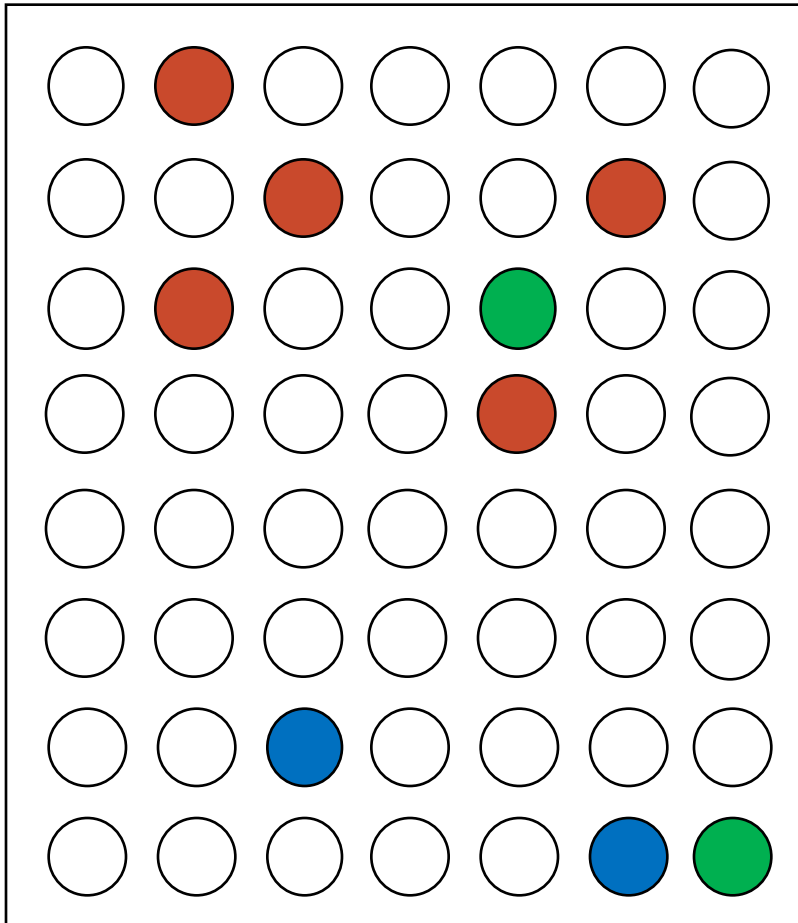
질의 :   



Bhaskar Mitra, Fernando Diaz and Nick Craswell. *Learning to Match using Local and Distributed Representations of Text for Web search*, WWW 2017

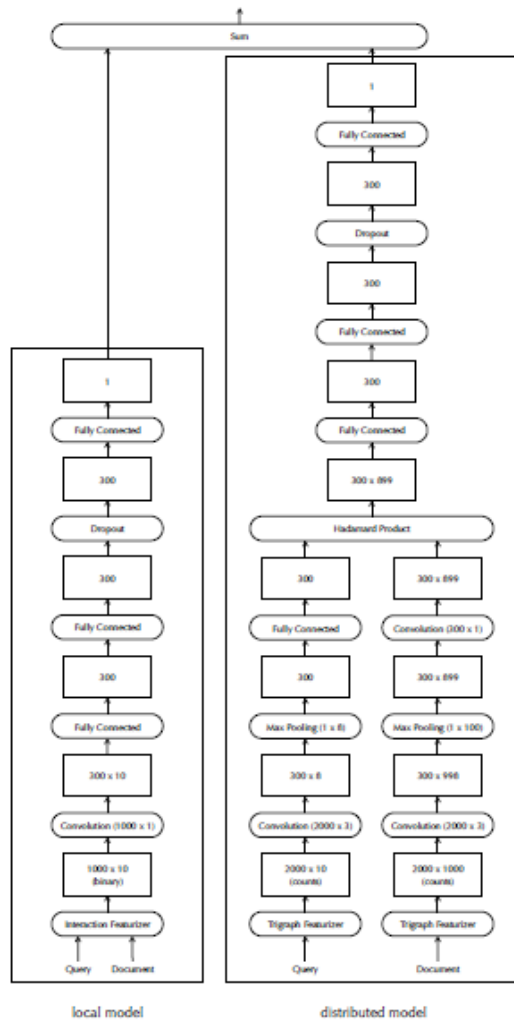
# 새로운 시도들

질의 : ● ● ●



Bhaskar Mitra, Fernando Diaz and Nick Craswell. *Learning to Match using Local and Distributed Representations of Text for Web search*, WWW 2017

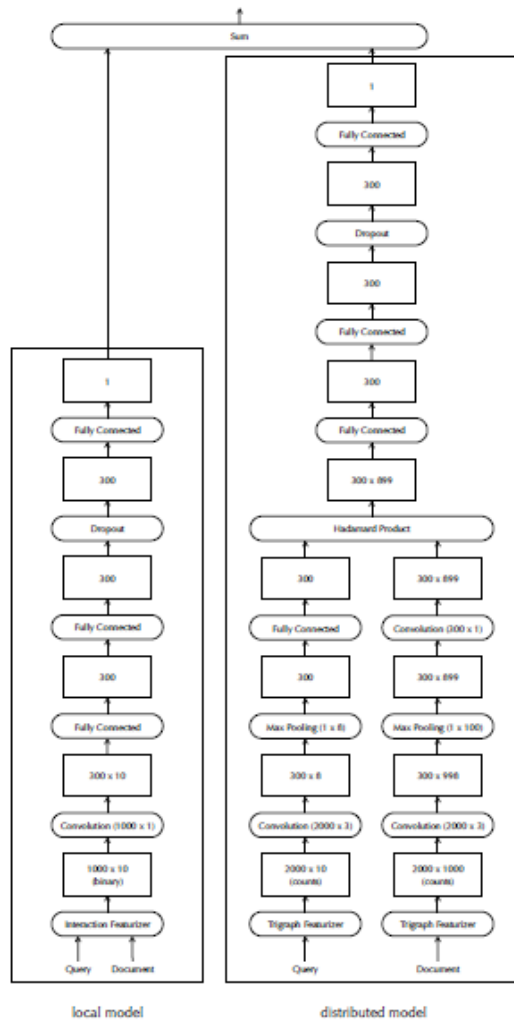
# 새로운 시도들



|                             | NDCG@1      | NDCG@10     |
|-----------------------------|-------------|-------------|
| <b>Non-neural baselines</b> |             |             |
| LSA                         | 31.9        | 62.7        |
| BM25                        | 34.9        | 63.3        |
| DM                          | 35.0        | 63.4        |
| QL                          | 34.9        | 63.4        |
| <b>Neural baselines</b>     |             |             |
| DRMM                        | 35.6        | 65.1        |
| DSSM                        | 34.3        | 64.4        |
| CDSSM                       | 34.3        | 64.0        |
| DESM                        | 35.0        | 64.7        |
| <b>Our models</b>           |             |             |
| Local model                 | 35.0        | 64.4        |
| Distributed model           | 35.2        | 64.9        |
| Duet model                  | <b>37.8</b> | <b>66.4</b> |

Bhaskar Mitra, Fernando Diaz and Nick Craswell. *Learning to Match using Local and Distributed Representations of Text for Web search*, WWW 2017

# 새로운 시도들



|                             | NDCG@1      | NDCG@10     |
|-----------------------------|-------------|-------------|
| <b>Non-neural baselines</b> |             |             |
| LSA                         | 31.9        | 62.7        |
| BM25                        | 34.9        | 63.3        |
| DM                          | 35.0        | 63.4        |
| QL                          | 34.9        | 63.4        |
| <b>Neural baselines</b>     |             |             |
| DRMM                        | 35.6        | 65.1        |
| DSSM                        | 34.3        | 64.4        |
| CDSSM                       | 34.3        | 64.0        |
| DESM                        | 35.0        | 64.7        |
| <b>Our models</b>           |             |             |
| Local model                 | 35.0        | 64.4        |
| Distributed model           | 35.2        | 64.9        |
| Duet model                  | <b>37.8</b> | <b>66.4</b> |

Bhaskar Mitra, Fernando Diaz and Nick Craswell. *Learning to Match using Local and Distributed Representations of Text for Web search*, WWW 2017

# Take Home Message

- 전세계에서 검색서비스를 제공하는 회사는 얼마 안남았고, 네이 버는 살아남기 위해 불철주야 열심히 노력하고 있습니다.
- 검색에서 가장 중요하고도 비밀스러운 부분은 어떤 랭킹시그널 을 사용하는지 입니다.
- 사용자가 적당히 검색창에 입력해도 잘 검색될 수 있도록, 정확 하게 단어가 일치하지 않아도 잘 매치시킬 수 있는 방법을 연구 하고 있습니다.
- 끊임없이 개발되고 있는 랭킹알고리즘들을 섭렵하고 실험하며 서비스에 맞게 변형시키고 적용하는 일을 많은 엔지니어들이 하 고 있습니다.

# Take Home Message

We are hiring!

<http://recruit.navercorp.com>

감사합니다