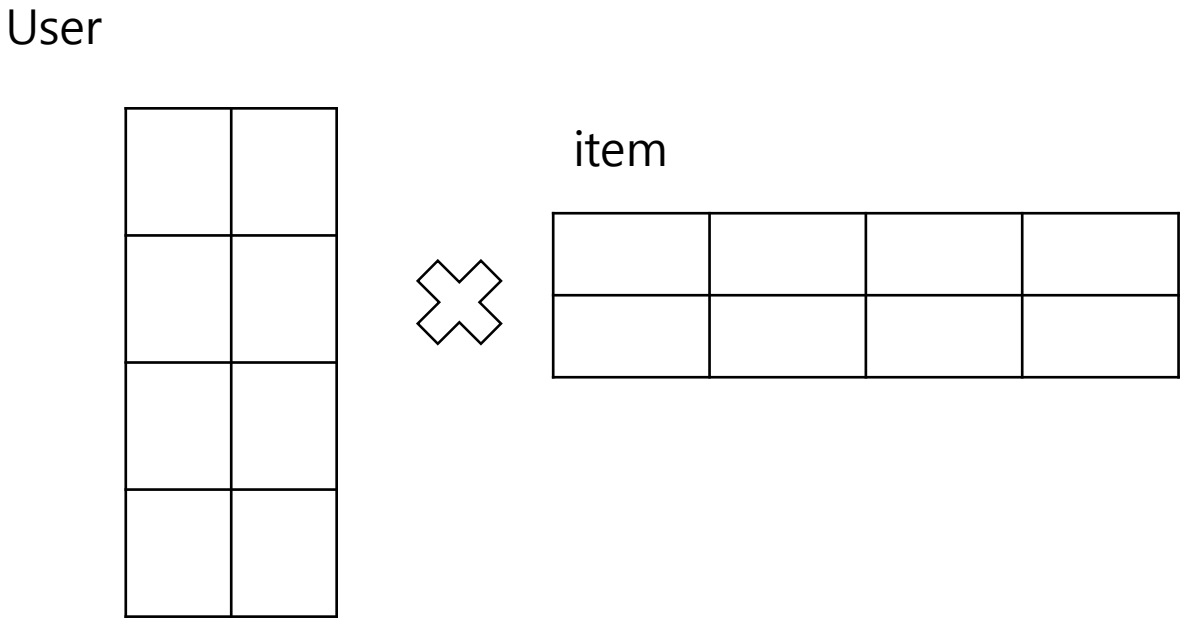
기업에 기반한 수강 과목 추천 시스템
with 학생과 기업의 관계 파악
(강원대학교)

강원대학교 박재헌

Neural collaborative filtering 소개

- MF (Matrix Factorizaion)의 행렬분해 방식을 Mlp 로 교체한 것

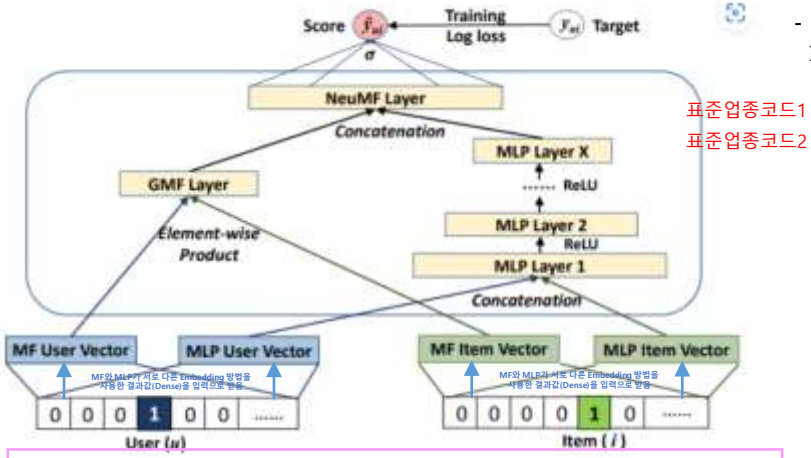
	Item1	Item2	Item3	item4
User1	1	0	0	1
User2	1	1	1	0
User3	0	0	1	1
user4	1	1	1	1



직종별 교과 추천 Architecture

- **NCF 모델** : NCF는 GMF와 MLP를 앙상블한 모델이다. GMF는 MF에 non-linear activation function을 더해 표현력을 높인 MF이고, MLP는 임베딩된 userId와 itemId를 받아 0~1의 $\hat{y}$를 구해주는 Multi-Layer-Perceptron이다.

<NCF 모델 구성 *확인필요 >



- User Latent Vector : $P_{M\times k} = \{p_{uk}\}$
※ 산업분류코드-학생 one-hot encoding 결과

- Item Latent Vector : $Q_{N\times k} = \{q_{ik}\}$
※ 학생-과목 one-hot encoding 결과

One-hot encoding

One-hot encoding

입력 데이터 : STID(학번), GRUP_CD(그룹화된 과목코드), BZC_CD(산업분류코드)

- 학습 결과(y_{ui}) : 표준업종코드- 과목 간 임베딩 벡터 + 후처리 : 추천결과를 교육과정과 비교 또는 이수영역별로 필터하여 추천
※ u : 산업분류코드, k : 학생, i : 과목

표준업종코드1
표준업종코드2

산업분류코드1

K개 과목 추천

<사용자 입·출력>

- **입력 : 표준업종코드**(예: 표준업종코드1)
- **출력 : K개 추천 과목**(예: 벡터가 큰 순서대로 K개 출력)

산업분류코드 1
(One Hot encoding vector)

0	0	1	0	0	....

산업분류코드 2
(One Hot encoding vector)

0	1	0	0	0	....

과목 1
(One Hot encoding vector)

0	0	1	0	0	....

과목 2
(One Hot encoding vector)

0	0	0	0	1	....

Data processing

- 데이터는 학번, 성별 , 단과대, 이수년도, 과목코드, 성적, 재학상태, 취업년도, 취업명,산업코드 로 이루어져 있음

	STID	SEX	CCD	RK	YY	HK	GRUP_CD	S_AVG	HCD	HIRED_YY	CO	BZC_CD
267944	20011499	2.0	110555225.0	1	2001	1	003585	1.5	chj79001	2010.0	강원대학교(삼척캠퍼스)	85302.0
267945	20011499	2.0	110555225.0	1	2001	1	100014	3.0	chj79001	2010.0	강원대학교(삼척캠퍼스)	85302.0
267946	20011499	2.0	110555225.0	1	2001	1	006808	3.0	chj79001	2010.0	강원대학교(삼척캠퍼스)	85302.0
267947	20011499	2.0	110555225.0	1	2001	1	005575	2.0	chj79001	2010.0	강원대학교(삼척캠퍼스)	85302.0
267948	20011499	2.0	110555225.0	1	2001	1	006809	3.5	chj79001	2010.0	강원대학교(삼척캠퍼스)	85302.0

Data processing

- **학습 데이터**

  - 실제 학습 시 사용되는 컬럼(추정)*확인필요 : STID(학번), GRUP_CD(그룹화된 과목코드), BZC_CD(표준업종코드)

	학생1	학생2	학생3	학생4
표준업종코드1	1	1	0	0
표준업종코드2	0	1	1	1

→ STID

→ BZC_CD

	과목 1	과목2	과목3	과목4
학생1	1	0	1	1
학생2	0	0	0	1

→ GRUP_CD

→ STID

GRUP_CD

	과목1	과목2	과목3	과목4
표준업종코드1	1	0	0	1
표준업종코드2	0	1	1	1

→ 최종 input

→ BZC_CD

Data processing

- 입력 데이터 *확인필요 :

① 산업분류코드- one-hot encoding / ② 과목 one-hot encoding 결과

산업분류코드 1 : [1,0,0,0,…0]　　과목 1: [1,0,0,0,…0]

산업분류코드 2 : [0,1,0,0,…0]　　과목 2 : [1,0,0,0,…0]

- 산업분류코드 수만큼 one-hot encoding 의 차원이 나온다.

- 과목 수만큼 one-hot encoding 차원이 나온다.

③ one-hot encoding을 진행한 산업분류코드와 과목을 embedding layer에 넣어 dense하게 만듦

user input : embedding layer에 들어간 one – hot encoding vector는 embedding layer의 차원 개수만큼 차원을 갖는다

-> 산업분류코드 1 [2.4, 3.5 , ….] (차원은 embedding 차원에 맞춰짐)

Item input : embedding layer에 들어간 one – hot encoding vector는 embedding layer의 차원 개수만큼 차원을 갖는다

-> 과목 2 [9.1 ,3.6,…] (차원은 embedding 차원에 맞춰짐)

실험

- User -> embedding dim의 차원을 가짐
- item -> embedding dim의 차원을 가짐
- MF에서 도 MF user vector 과 MF item vector를 뽑음
- [User x item](내적) ->output -> Linear(embedding dims, 1)(sigmoid) -> output
- 내적을 해서 값이 높아지면 벡터끼리 유사도를 가지고 있기 때문에, user와 item의 유사도가 높은 지 파악이 가능

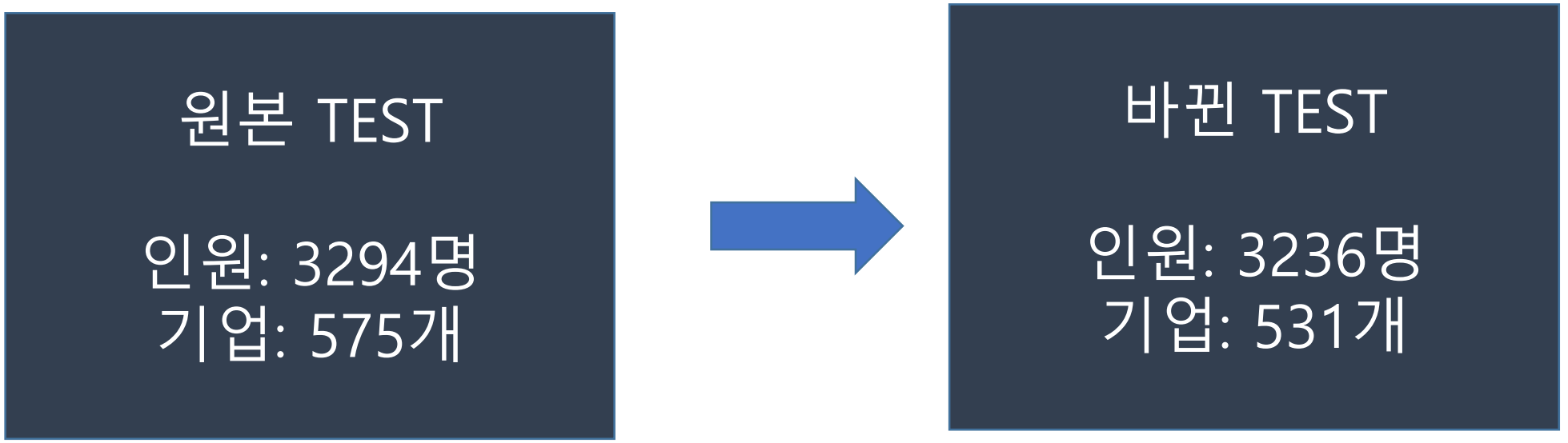
검증

Train set 기업수 : 878

Test set 기업수 : 575

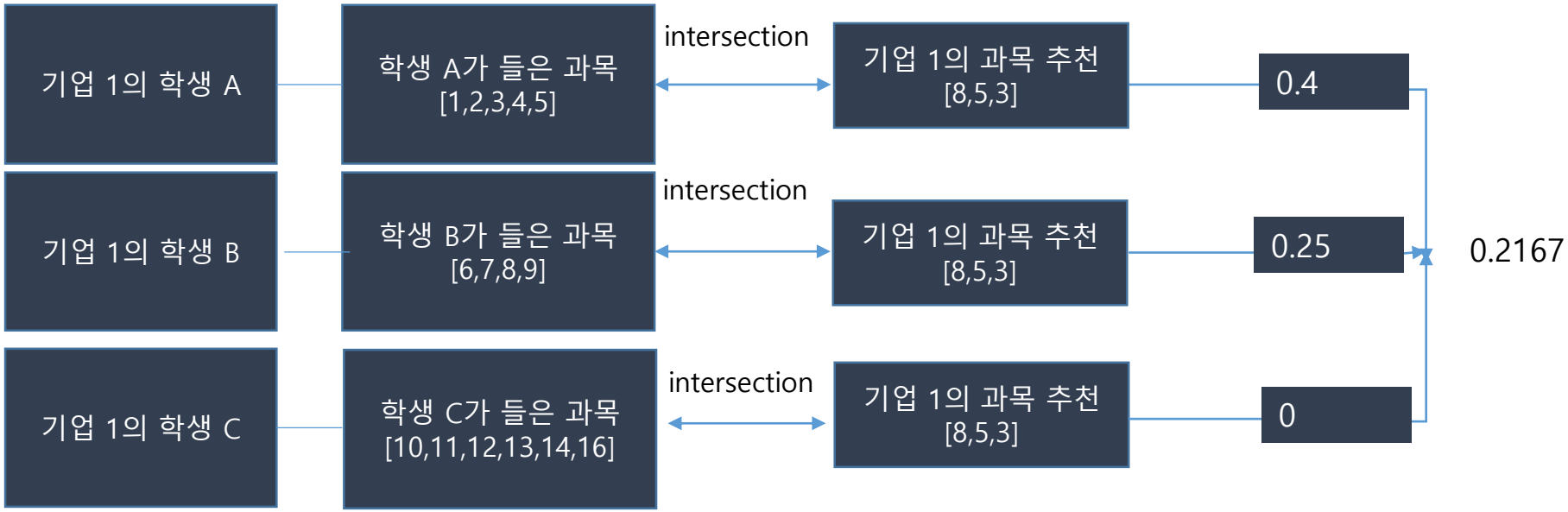
Train set에는 있지만 Test set에 없는 기업의 수: 44개

총 531개의 기업을 가진 3236명을 대상으로 test 진행



기업 추천 검증

Recall @ k

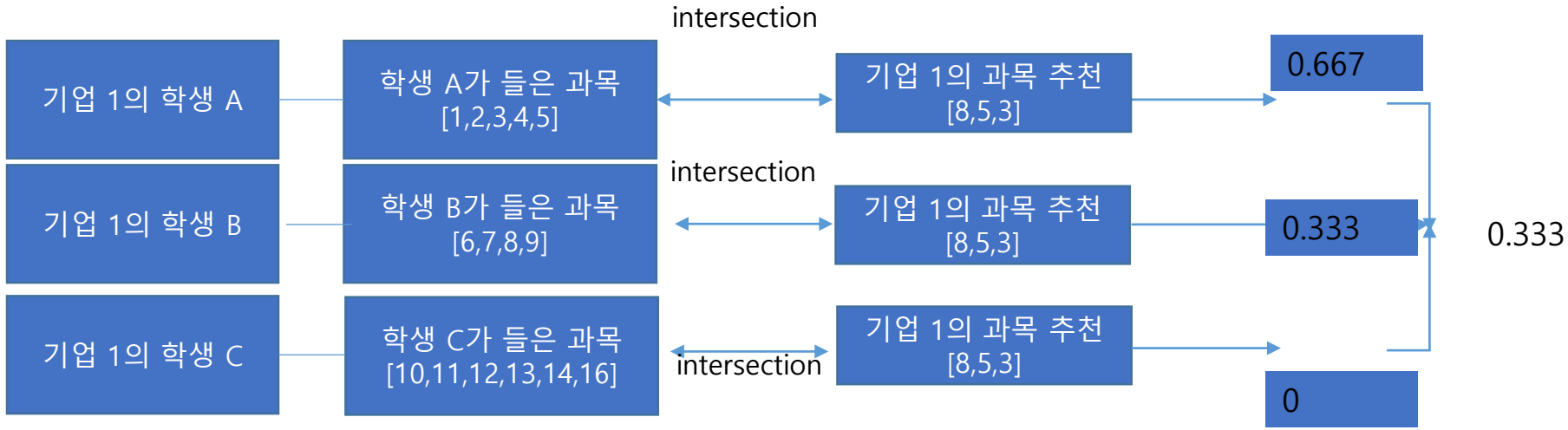


$$\text{Recall} = \frac{(\text{학생들이 들은 과목} \cap \text{추천한 과목})\text{의 수}}{\text{학생들이 들은 과목 수}}$$

기업 당 추천하는 과목k개를 뽑은 후 그 기업에 간 각 학생이 수강한 과목t개 중 k에서 겹치는 r개를 뽑은 후 r/t로 나눈 후 모두 더해서 학생 수 만큼 나눠 평균을 낸다

기업 추천 검증

Presicion@k



$$\text{Presicion@k} = \frac{(\text{학생들이 들은 과목} \cap \text{추천한 과목})\text{의 수}}{\text{추천한 과목 의 수}}$$

기업 당 추천하는 과목k개를 뽑은 후 그 기업에 간 각 학생이 수강한 과목t개 중 k에서 겹치는 r개를 뽑은 후 r/k로 나눈 후 모두 더해서 학생 수 만큼 나눠 평균을 낸다

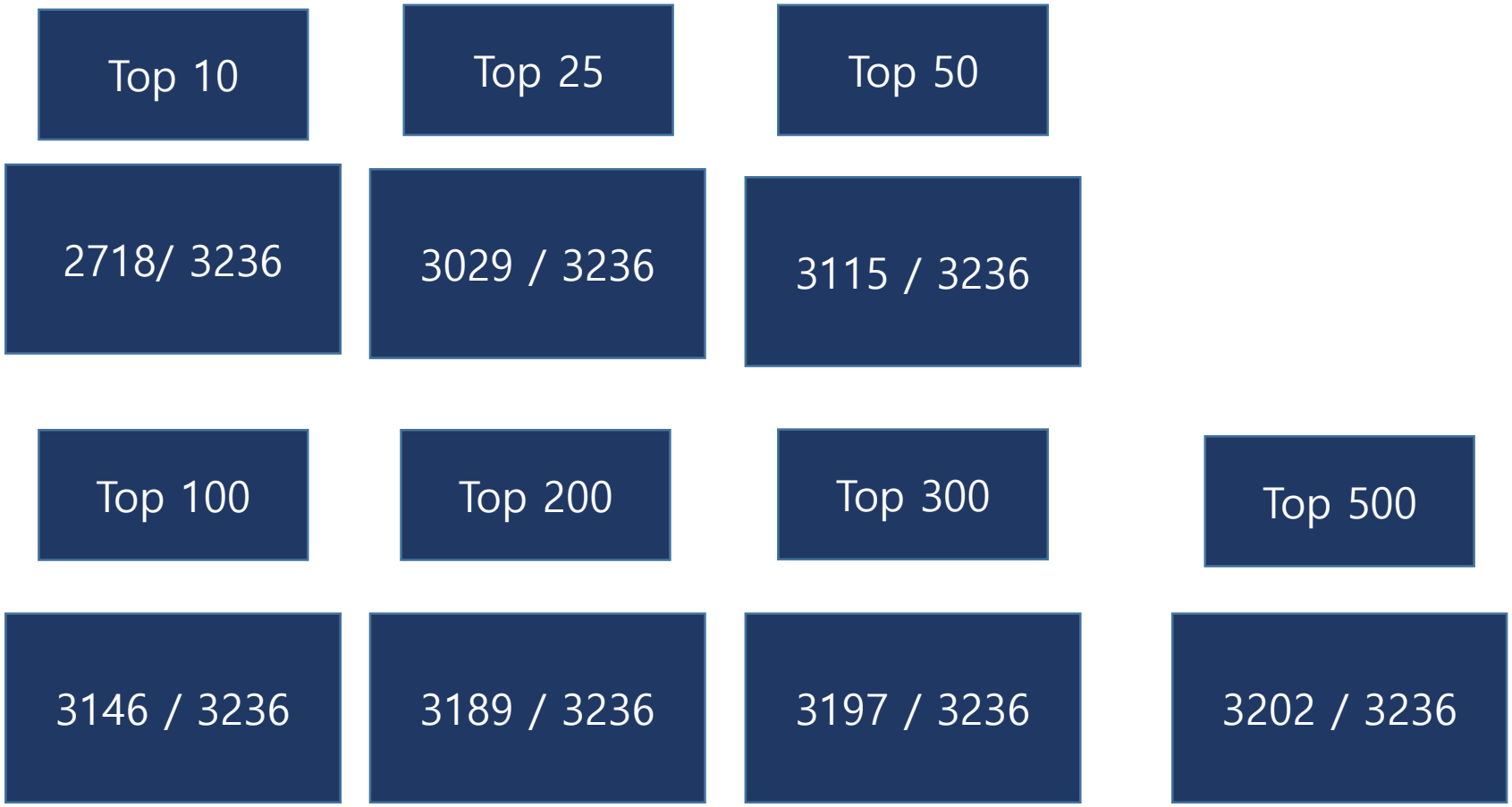
Presicion@k

Top 10	Top 25	Top 50	Top 100	Top 200	Top 300	Top 500
0.19604	0.16723	0.13600	0.09999	0.07009	0.05612	0.04088

Recall @ k

Top 10	Top 25	Top 50	Top 100	Top 200	Top 300	Top 500
0.03758	0.08081	0.13185	0.19377	0.27117	0.32572	0.39365

Hit-Ratio @ k



Top 10	Top 25	Top 50	Top 100	Top 200	Top 300	Top 500
0.83992	0.93603	0.96260	0.97218	0.98547	0.98794	0.99886