추천시스템 week 2: Non-personalized, Streotype Based Recommeders

와이빅타 16기 김주은

1. Nonpersonalized and Stereotyped Recommendation

- 왜 non-personalized 추천을 이용하는가?
  - 유저에 대한 충분한 정보가 없을 때 (새로운 유저)
  - 간단한 연산에 비해 결과가 충분한 대표성을 가짐
  - personalization이 불가능한 경우
  - 대중적인 주제에 대한 추천일 경우

- 추천의 역사
  - 인쇄물: 책, 영화, 음악 리뷰; 미쉐린 가이드; The Negro Motorist Green-Book; → editorially selected
  - 전체 데이터의 평균: Zagat, Billboard, E-commerce, ...
  - Weak Personalization: 지역/나이/성별/국적 등 유저에 대해서 알 수 있는 최소한의 정보만을 가지고 차별화된 추천을 진행하는 것. **steoreotype-based recommenders** 이라고도 하는데, 소비와 관련되지 않은 기타 정보를 통해 implicit/explicit rating을 모아 추천을 하는 시스템을 말한다.

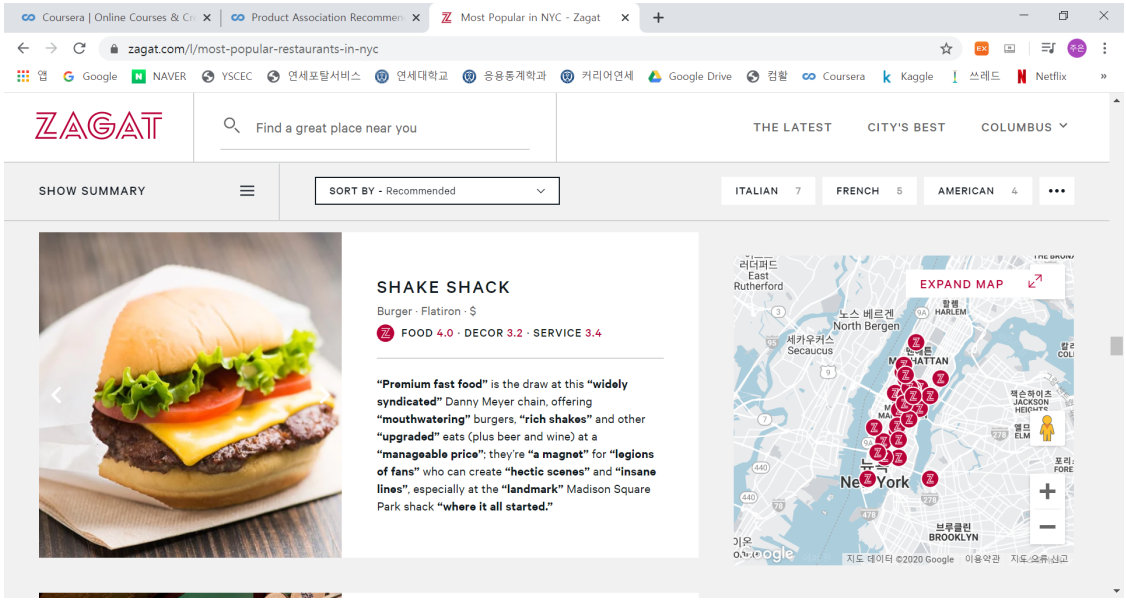
2. Summary Statistics

- Zagat의 경우(전세계적으로 유명한 음식점 추천 시스템)

  : 주변 지인들이 레스토랑에 팁을 얼마나 주었는지와 그 레스토랑에 대한 평가를 수집한 것이 시초

  Rating = {0, 1, 2, 3} 중에 고르도록 하고, 이를 평균내서 10을 곱한 정수값(30점 만점)이 음식/분위기/서비스/가격 분야별 점수가 되는 시스템.

  현재 Zagat의 인터페이스는 다음과 같이 바뀌었다. (5점 만점의 음식/분위기/서비스, $ 개수로 표시되는 가격)



- **98%의 소비자가 그냥저냥 행복한 아이템 vs 90%의 소비자가 극도로 행복한 아이템**

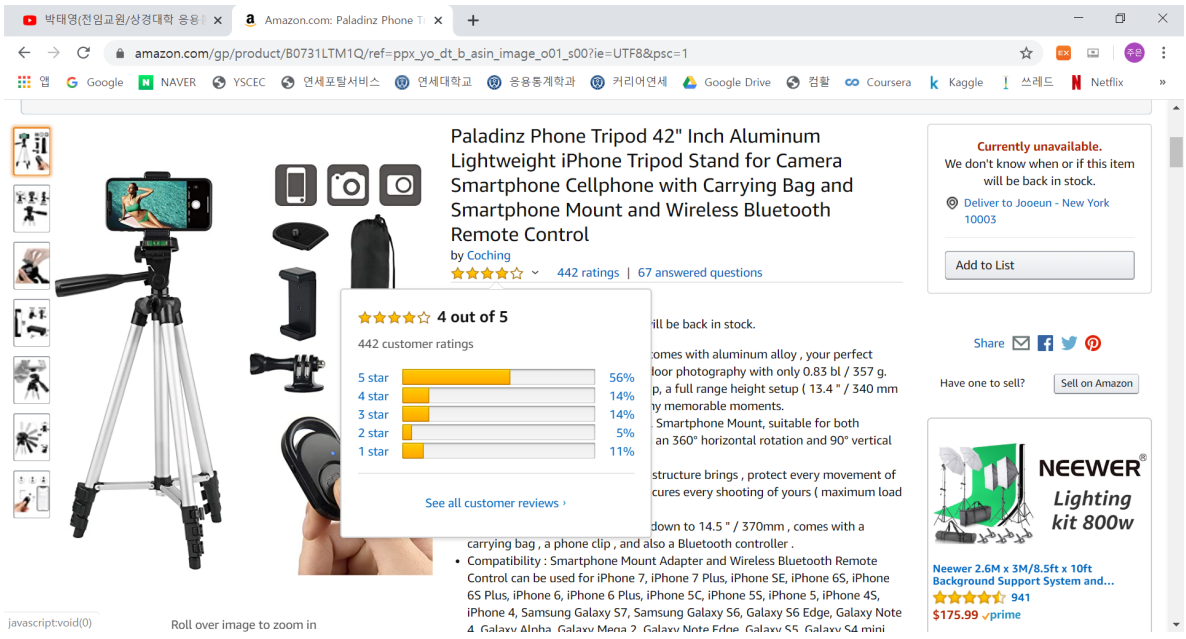
  무엇을 선택할 것인가? 호불호 갈리는 선택지와 대다수가 만족한 선택지 중 선택하기 위해서는 좋은 평가를 한 집단과 나의 취향의 유사성을 판단해보아야 한다.

  → 대중성은 분명 중요한 평가 기준이지만, *단순 평균은 충분히 좋은 추천을 하기에 부족하다*

  - 점수별 평가인원의 비율, 사용자 평점의 정규화, 평가자의 신뢰성 파악 등의 극복 방안
  - 최대한 많은 데이터를 수집하는 것이 좋음

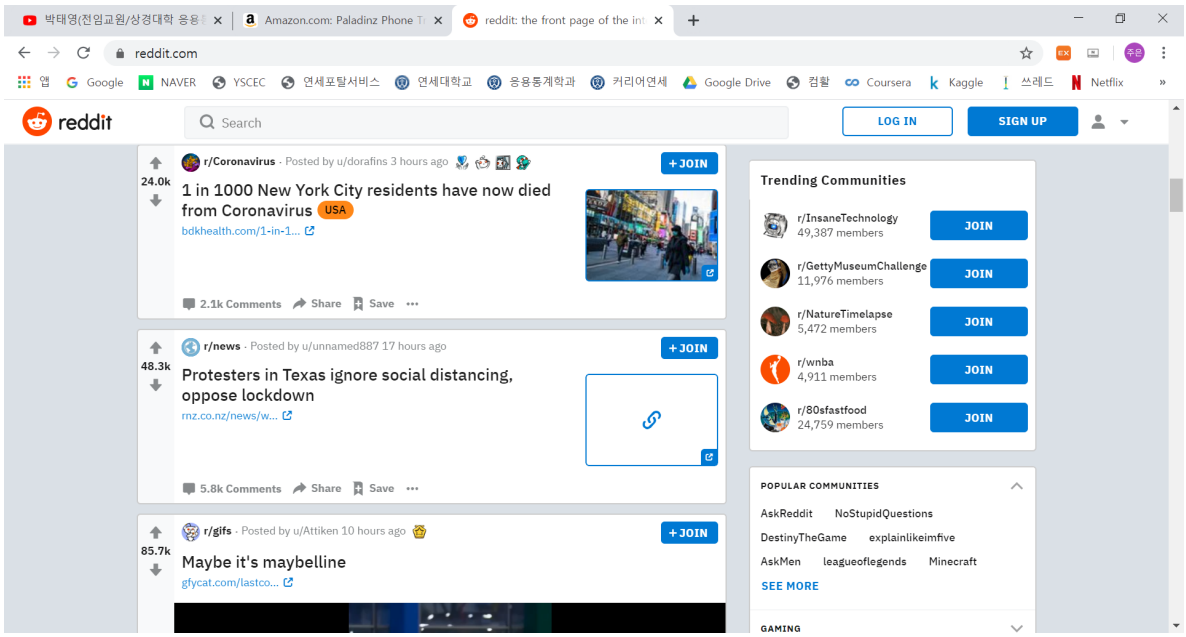
- 우리 교수님에게 15세 소녀가 좋아하는 노래를 추천해준다면? 또는 아이스크림 소스로 케찹을 추천해준다면?ㅠㅠㅠ

  → 따라서, **개인의 특성**과 **맥락**에 대한 고려가 필요하다. 즉, **personalization**이 필요하다



- 아마존의 예시: non-personalized recommendation은 적절한 곳에 사용되었을 때 효과적일 수 있다.
- non-personalized recommendation의 가장 좋은 방법: 평가 개수, 평균, 분포를 모두 보여주는 것
- 순위를 매기려면: threshold를 정하고 이보다 낮은 평가의 비율로 점수를 매길 수 있음

- 이 때 non-personalized recommendation을 하던 Zagat에 생긴 문제
  - **Self-selection bias**
    - 평가의 압축화(compression in ratings): 그저 그런 레스토랑들이 좋은 점수를 받게 되고, 훌륭한 레스토랑들이 그저 그런 점수를 받게 됨
    - 사람들이 본인의 취향에 맞는 레스토랑만 계속 가게 됨. 작년에 갔는데 별로여서 낮은 평점을 준 식당에 올해는 가지 않으므로 평가를 할 기회가 없고 그에 따라서 해당 레스토랑의 평점이 올라가게 되는 현상
  - 평가하는 유저들의 **다양화**
    - 취향이 다양해지면서 단순 평균으로는 대부분의 사람이 2점을 준 레스토랑과 절반은 3점 절반은 1점을 준 레스토랑을 구분할 수 없다는 문제에 직면하게 되었다.



- Reddit(뉴스 모아보기 서비스)의 예시
  - non-personalized, 유저의 vote (UP or DOWN) 을 통해 순위를 매기는 방식 (simple display)
  - simple display의 문제점: **5점(5 UPs + 0 DOWNs) vs 5점(5000 UPs + 4995 DOWNs)**

- ranking: 어떤 아이템을 위에 놓을것인가?
  - 단순 평점으로 나열할 수는 없다. 왜냐하면 *1명이 평가한 5점*이 *500명이 평가한 4.8점*보다 낫다고 판단할 수 없기 때문에!
  - 또한 histogram(평점, 평가한 사람 수, 비율)을 일일이 보여주기에는 직관적이지 못함
  - 따라서 **damped mean** 이용

$$\frac{\sum_u r_{ui} + km}{n + k}$$

여기서 k는 **전체 평점의 영향력(strength of evidence)**을 의미한다. k가 클수록 전체 평점의 중요도(영향력)가 커지는 방식.

$$r_{ui} : \text{user ratings}$$
$$n : \text{number of ratings}$$
$$km : \text{ k ratings of global mean}$$

따라서 damped mean을 이용하면 많은 사람이 평가한 아이템은 k가 n에 비해 상대적으로 작기 때문에 전체 평점의 영향을 상대적으로 적게 받고, 적은 사람이 평가한 아이템은 k가 n에 비해 상대적으로 크기 때문에 전체 평점에 많은 영향을 받아 소수의 좋은 평점이 아이템을 과대평가하거나 과소평가하는 경우를 방지해준다.

- **신뢰구간(Confidence Interval)** 이용
  - 통계적 추론을 이용해서 특정 수치의 신뢰구간을 통해 유저가 아이템을 좋게 또는 나쁘게 평가할 것이라는 예측을 한다. 관련된 정보(evidence)가 많을수록 신뢰구간의 폭이 좁아지므로 더 정확한 예측을 할 수 있다.
  - 이 때 신뢰구간의

    Lower Bound를 기준으로 추천: 보수적인 관점에서 안전한 예측을 할 수 있다.

    Upper Bound를 기준으로 추천: risky하지만 맞았을 경우에는 훨씬 더 정확한 추천이 가능해진다.
- **시간**의 경과가 중요한 요소로 작용하는 경우 (ex. 뉴스)
- Hacker News의 예시
  - $$\frac{(U-D-1)^{\alpha}}{(t_{now}-t_{post})^{\gamma}} \times p$$

    여기서 이 식은 net upvote를 age로 **polynomially decay**하는 인자를 넣어서 각각 초반에 받은 upvote와 최근 뉴스에 더 많은 가중치를 부여해서 **최근**에 올라온 뉴스이면서 **초반에 많은 UP** 표를 받은 뉴스를 상단에 배치하도록 점수를 계산할 수 있다. 이러한 효과를 주기 위해서 알파는 1보다 작은 값(0.8)으로, 감마는 1보다 큰 값(1.8)으로 설정한다. 그렇게 하면 net upvote는 우하향하고, age는 우상향하는 그래프를 그리면서 결국 게시물의 rank는 초반에 급격하게 떨어지다가 일정 기간이 지나면 사실상 큰 변화가 없게 된다.

$$
\begin{gathered}
U : \text{number of UP votes} \\
D : \text{number of DOWN votes} \\
\rightarrow (U-D-1) : \text{net upvotes} \\
t_{now}-t_{post} : \text{age of the news} \\
\alpha, \gamma : \text{polynomial decay factors} \\
\text{giving more weight to early votes and reducing the effect of age respectively} \\
p : \text{penalty given to specific kinds of posts}
\end{gathered}
$$

- Reddits의 예시
  - net upvote와 시간(age)에 대한 항들이 독립적으로 존재하고 각자 계산되어서 더해진다. UP vote보다 DOWN vote를 더 많이 받은 뉴스는 아예 (log1)0점 + (sign(U-D)=-1)음수값을 부여해서 상단에 절대 나올 수 없도록 해놓았다.

$$\log_{10} max(1, |U-D|) + \frac{sign(U-D)t_{post}}{45,000}$$

  - 첫번째 항: 초반에 받은 표들에게 많은 가중치를 부여하기 위해서 log를 씌운 결과로 첫 10표가 그 11번째부터 100번째 표와 같은 영향을 미치는 효과가 나타난다.
  - 두번째 항: t가 뉴스가 게시된 시간이므로 최신 뉴스일수록 t가 커짐 → 최근에 게시되었고 그와 동시에 UP>DOWN 인 뉴스일수록 decay factor이 크게 나타나게 된다.

- 위와 같이 특정 요소의 영향력 / 아이템의 분류 / 기타 등등 도메인 특성들을 수식을 통해 페널티(decay factor)를 부과하는 방식으로 제어할 수 있다.

3. Demographics and Related Approaches

- **Demographic**이란?
  - *인구통계*; 즉 나이, 성별, 인종 등등 유저의 소비와 직접적인 관련이 없는 유저와 관련된 특성(분류)

- **왜** 인구통계를 추천에 고려해야할까?
  - **대중성**과 **취향**은 꼭 일치하지 않는다!
  - 따라서 가장 인기가 많은(대중성이 있는) 아이템을 모두에게 추천하기보다는 **최소한의 인구통계 정보**를 가지고 추천하면 조금이나마 더 취향을 저격할 수 있을 것이라는 발상에서 착안하게 되었다.

- 그렇다면 인구통계를 **어떻게** 추천에 반영할 수 있을까?
  - **전처리**: 먼저 전처리가 필요하다. 예를 들자면, 나이를 추천의 기본 정보로 이용하기 위해서는 정확히 23살의 사람들이 좋아하는 아이템을 모든 23살에게 추천하려면 적은 표본으로 복잡한 모델을 설계해야하고, 소수의 표본만을 가지고 하기 때문에 정확성이 떨어질 수 있다. 따라서 나이에 *구간을 설정*하여 일정한 나이대의 사람들의 선택을 반영한 추천을 진행하는 것이 좋다. 또한 우편번호 등의 정보를 통해 주거환경, 기본소득, 대중 집단의 성향 등에 대한 추론이 가능한 것처럼 인구통계적 정보를 기반으로 한 *추론을 통한 그룹화*도 가능하다.
  - **상관성 분석**: 다음으로는 추천의 기준으로 삼으려는 인구통계적 요소가 실제 소비 성향과 유사성을 보이는지 확인해야 한다. 산점도나 상관계수 등을 분석해서 데이터 간에 차별화가 가능한 구간이나 요소들을 찾아볼 수 있다.
  - 이 과정을 통해 추천의 기준으로 삼을만한 요소를 발견했다면,
    1. summary statistics를 인구 요소에 따라 구분한다.
    2. 인구통계량을 기반으로 선호하는 아이템을 예측할 수 있는 회귀분석(선형 회귀, 로지스틱 회귀 등)을 진행한다.
  - 충분한 **가정사항** 필요: 단순한 전체 인구의 선호도와 차별화 되었는지, 새로운 유저들에게 일반화 해도 될만큼 신뢰성이 있는지, 변수들 간의 상관성이 있는지 등등
  - 인구통계량이 유용하다면, 유저들로부터 **충분한 양의 데이터**를 확보하는 것이 중요하다.

    ex. Facebook: 개인 정보가 많이 노출되어 있는 SNS를 활용해서 (물론 합법이겠죠..?) 유저에 대한 기본적인 정보를 얻고 그에 기반하여 추천을 하는 시스템 등

결론: *Demographics are sometimes valuable, but never perfect.*

즉 인구통계는 데이터와 아이템의 목적과 특성에 따라 효과적일 수 있지만, 언제나 예외가 존재하고 소비에 영향을 미치는 결정적인 요소가 아니기 때문에 추천시스템의 완벽한 기준이 될 수는 없다.

4. Product Association Recommender

- Non-personalized recommender: 맥락에 대한 고려가 부족하다
- 인구통계만으로 부족하다면, 사람들이 소비한 **제품 간의 연관성**에 대해 알아보고 이를 기반으로 추천해보자!

- ephemeral, contextual personalization (**휘발성**, **맥락성**):
  - 소비자 개개인이 아닌 **제품의 특징을 기반으로 개인화**하는 시스템
  - 유저의 **일회성 소비**만을 기준으로 삼기 때문에, 장기적인 **개인에 대한 특성 발견은 불가능** (휘발성)
  - 계산: 수기(manual) → 알고리즘 이용(data mining) 으로 추천시스템이 도입되기 이전에도 시대의 흐름에 따라 변화해왔다.

- **most likely** to buy vs **most *extra* likely** to buy 간에 구별을 하기 위한 목적
  - 먼저 간단한 발상에서 착안했다. X를 산 사람 중에 X와 Y를 모두 산 사람의 비율을 계산하고, 이 비율이 높은 제품 순서대로 추천.

    $$\frac{X \wedge Y}{X}$$

    이 방법을 이용한다면 X를 산 사람들이 Y를 살 가능성이 높다고 할 수 있지만, 그것이 X와 Y 제품 간의 연관성을 설명해주는 의미있는 추론인가?

    단점: 이 수치가 높다고 해서 **X와 Y 간의 연관성**이 높다고 할 수 없다. Y의 소비 비율이 높은 이유가 X와의 연관성 때문이 아니라, 단순히 Y가 대중적으로 인기가 많고 보편적으로 소비되는 제품일 가능성도 존재하기 때문에. 이 방법을 넘어서 X와 Y의 unique link를 찾을 필요가 생겼다.
- Bayes' Law
  - $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

    앞서 살펴보았던 식과 유사하다! 하지만 X가 주어진 조건부 확률이기 때문에 Y를 산 사람 중에 X를 산 사람의 비율을 계산해서 **X의 소비가 Y의 소비를 촉발(trigger)했을 가능성**을 알 수 있다.
  - $$\frac{P(Y|X)}{P(Y)} = \frac{P(X \wedge Y)}{P(X) \times P(Y)} = \frac{P(X|Y)}{P(X)}$$

    여기에서 이 값이 1과 가깝다면 Y에 있어서 X의 영향력이 거의 없는 것이고, 1보다 큰 값이라면 Y가 X에게 맥락적으로 영향을 미칠 가능성이 있다고 판단할 수 있다. 하지만 식을 분해해 보았을 때 위와 같이 나타나는 것을 보아, X가 Y에게 영향을 미치는 것인지, Y가 X에게 영향을 미치는 것인지 관계의 방향성을 확실하게 알 수 없다.

- Beer and diaper story: 대형 마트에서 소비되는 제품 간의 연관성을 분석해보았더니, 맥주와 기저귀를 함께 구매하는 비율이 굉장히 높다는 사실이 밝혀졌던 적이 있다. 그 때 가능했던 추론으로는 기저귀를 열심히 갈면서 육아에 지친 부모들이 피로를 달래기 위해 맥주 한 캔을 한다는 가설과 음주가무를 즐기며 즐겁게 노는 젊은이들이 기저귀를 가는 부모가 될 가능성이 높다는 가설 두 가지가 있는데, 어느 것이 진짜인지는 아무도 알 수 없다. 베이즈 정리를 통한 추론으로는 인과관계의 방향성에 대한 근거를 얻을 수 없기 때문이다.