문제 해결을 위한 머신러닝 오픈 플랫폼, 카카오 아레나

글 | 최은필 reyna.c@kakaocorp.com IT 업계의 에반젤리스트(evangelist)를 자처하며 빠르게 변화하는 시대에 그 누구보다 부지런히 움직이며, 좋은 것들을 나누며 살고 있습니다. 재즈 힙합(Jazz Hip-hop)을 즐겨 들으며, 언젠가 추리소설을 쓰겠다는 원대하지만 소박한 꿈을 가지고 있습니다. 전자공학을 전공했지만 실험보다는 글쓰기를 더 좋아하며, 오랜 시간 '전략, 기술, 기획'이라는 단어들과 함께하였습니다. 전 세계 수많은 사람들의 마음을 움직여보고 싶다는 커다란 꿈을 가지고, 하루하루 즐겁고 재미있게 살고 있습니다.

지난 2018년 11월 카카오는 머신러닝으로 문제를 해결하고 최적의 솔루션을 찾는 대회를 시작했다. 이 대회는 '카카오 아레나'라는 공간을 통해 오픈됐고, 이곳은 머신러닝 대회를 상시 개설하고 운영할 수 있는 플랫폼이다.*1

첫 번째 대회는 쇼핑몰에 등록된 상품의 텍스트, 이미지 정보 등을 활용해 카테고리 분류의 정확도를 높이는 '쇼핑몰 상품 카테고리 분류'를 주제로 진행되고 있다. 머신러닝 기술을 활용해 다음(Daum) 쇼핑에 존재하는 수억 개의 상품을 더 정확하게 카테고리화할 수 있는 기술을 만드는 것을 목표로 한다.

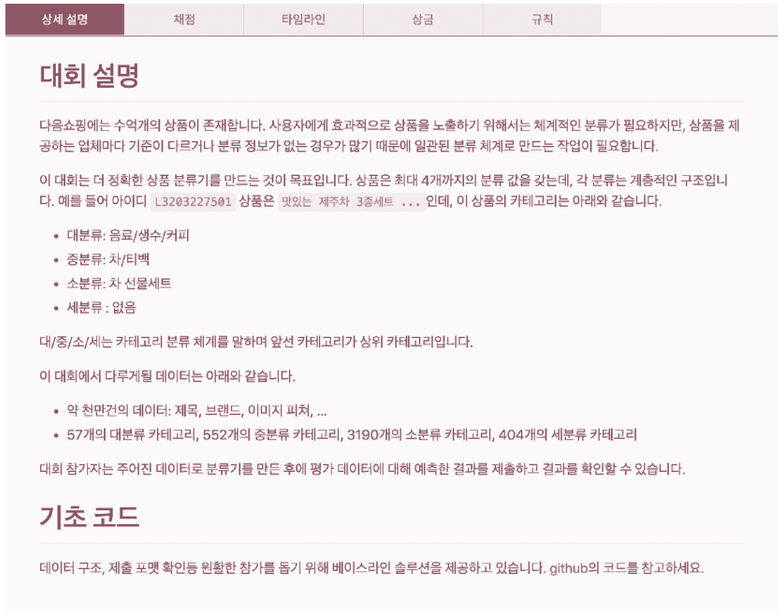
이 대회를 위해 카카오는 다음 쇼핑에 누적된 상품 이름, 브랜드, 이미지 등 약 1000만 건에 달하는 데이터와 4000건 이상의 카테고리 데이터를 오픈했다.

[그림 1] 카카오 아레나 홈페이지 화면*1

'카카오 아레나'의 첫 번째 대회인 '쇼핑몰 상품 카테고리 분류'에 대해 간단히 살펴보면, 올 11월 7일부터 2019년 1월 7일까지 약 2달간 일정으로 진행되며 최종 제출 마감일 일주일 뒤인 1월 14일 순위가 공개될 예정이다. 대회의 채점 및 심사 기준을 보면, 제출된 결과물의 예측도와 실제 카테고리와 매칭되는 정확도로 평가하며, 예측도가 높은 참가자 순서로 수상 순위를 결정한다. 11월 31일 현재 약 430팀이 대회에 참가했으며, [그림 3]과 같은 중간 결과들을 보여주고 있다.

*1 참고 | https://arena.kakao.com

[그림 2] 쇼핑몰 상품 카테고리 분류 대회 요강*1



상세 설명 | 채점 | 타임라인 | 상금 | 규칙

대회 설명

다음쇼핑에는 수억개의 상품이 존재합니다. 사용자에게 효과적으로 상품을 노출하기 위해서는 체계적인 분류가 필요하지만, 상품을 제공하는 업체마다 기준이 다르거나 분류 정보가 없는 경우가 많기 때문에 일관된 분류 체계로 만드는 작업이 필요합니다.

이 대회는 더 정확한 상품 분류기를 만드는 것이 목표입니다. 상품은 최대 4개까지의 분류 값을 갖는데, 각 분류는 계층적인 구조입니다. 예를 들어 아이디 `L3203227501` 상품은 `맛있는 제주차 3종세트 ...` 인데, 이 상품의 카테고리는 아래와 같습니다.

- 대분류: 음료/생수/커피
- 중분류: 차/티백
- 소분류: 차 선물세트
- 세분류 : 없음

대/중/소/세는 카테고리 분류 체계를 말하며 앞선 카테고리가 상위 카테고리입니다.

이 대회에서 다루게될 데이터는 아래와 같습니다.

- 약 천만건의 데이터: 제목, 브랜드, 이미지 피처, ...
- 57개의 대분류 카테고리, 552개의 중분류 카테고리, 3190개의 소분류 카테고리, 404개의 세분류 카테고리

대회 참가자는 주어진 데이터로 분류기를 만든 후에 평가 데이터에 대해 예측한 결과를 제출하고 결과를 확인할 수 있습니다.

기초 코드

데이터 구조, 제출 포맷 확인등 원활한 참가를 돕기 위해 베이스라인 솔루션을 제공하고 있습니다. github의 코드를 참고하세요.

[그림 3] 대회 참가 팀 및 순위 현황(2018년 12월 26일 기준)*1

공개 리더보드 | 파이널 리더보드

#	△1d	팀명	점수	제출 횟수	마지막 제출
1	-	제로콜라 $	1.049394	35	44분 전
2	▲1	선동한 $	1.038339	9	2시간 전
3	▼1	송골매 $	1.038263	71	14시간 전
4	-	딥멱성 $	1.030391	93	3일 전
5	-	baseline $	1.029824	41	3일 전
6	-	박준우 $	1.027933	11	11일 전
7	▲3	제로	1.026555	22	15시간 전
8	-	¯_(ツ)_/¯	1.024888	44	33분 전
9	▼2	댕댕이조하	1.022980	3	하루 전
10	▼1	soonbot	1.017511	1	하루 전
11	▲7	S2Bstar	1.008519	62	17시간 전
12	▼1	kweonwooj	0.993584	21	9일 전
13	▼1	김윤수	0.989139	7	하루 전
14	▼1	cleancoder	0.968501	9	10일 전
15	▼1	DBM	0.968426	3	2일 전
16	▼1	import this	0.966273	1	5일 전
17	▼1	neptune	0.966117	3	2일 전
18	▲2	Cockroach54	0.952282	15	14시간 전
19	▼2	orange	0.951889	2	2일 전
20	▲2	24시간이모자라	0.926444	14	7일 전

지금까지 설명한 내용은 조금은 심심한 1차원적 정보들이다. 지면상으로 찾아볼 수 있는 내용들 외에 '카.카.오.아.레.나'라는 여섯 글자 뒤에 숨겨진 이야기들을 알아보기 위해 플랫폼을 기획하고 대회를 준비한 카카오 담당 크루를 만나봤다.

Q. 많은 사람이 이번 대회에 대해 많이 궁금해하는데, 카카오 아레나 대회 심사와 제출 결과물 평가에서 가장 중요하게 보는 것은 무엇인가요?

A. 아무래도 정답에 가장 가까운 결과가 높은 점수를 받게 될 텐데, 스팸 필터링과 정확도가 가장 우선시될 것으로 생각합니다. 가장 먼저 오픈한 대회가 다음 쇼핑에 적용하기 위한 것으로 현재 쇼핑몰을 자동화하기 위한 것이 주된 목적입니다.

Q. 비슷한 대회로 알파벳(Alphabet)에서 개최하는 캐글(Kaggle)을 들 수 있을 것 같은데, 국내외에 어떤 유사한 대회들이 있나요?

A. 아무래도 국내외 통틀어서 가장 유명하고 제대로 기획된 대회가 캐글이라 생각됩니다. 국내에도 책 추천, 코드 스프린트 등의 대회가 있지만, 카카오처럼 '아레나'라는 플랫폼까지 지원한 곳은 없습니다. 대회들마다 성격이 다르기도 하지만 제대로 만들어진 대회를 찾기가 굉장히 어렵습니다. 그나마 캐글이 명확한 방향성과 목적의식을 가지고 운영되고 있습니다. 실제 '카카오 아레나'가 지향하는 방향도 캐글과 비슷합니다.

Q. 기존 대회와 '카카오 아레나'의 차이점이 '플랫폼'이라고 말씀했는데 특징과 차별성에 대해 더 자세하게 말씀해 줄 수 있나요?

A. 무엇보다 대회 디자인을 잘했다고 자신 있게 말씀드릴 수 있을 것 같습니다. 잘 기획된 '머신러닝 대회'라고 하면 2가지 결정적 요소가 필요합니다. 첫째, 양질의 데이터가 제공되어야 합니다. 데이터의 경우 라이센스 문제가 복잡하게 얽혀 있기 때문에 라이센스 확보가 되지 않으면 제공할 수 있는 데이터에 한계가 있고, 결과물 역시 실용성에서 한계를 가질 수밖에 없습니다.

둘째, 대회의 목적이 명확해야 합니다. 가끔씩 왜 개최하는지 모를 정도의 대회들이 여기저기서 개최됩니다. 목적이 분명하지 않기 때문인데, 대회 결과물을 활용해 경제적 이익을 얻거나 학문적으로 깊이 있는 연구를 목적으로 하는 등 결과 활용에 대한 명확한 계획이 있어야 합니다. 하지만 기업의 경제성과 학문적 성격 등이 혼재되어 실패한 대회들의 사례도 많았습니다.

Q. 국내외 여러 기업에서 비슷한 대회들을 개최하고 있는데 각각의 특징과 '카카오 아레나'만의 특징이 있을까요?

A. 앞에서 잠깐 말씀드린 것처럼 '카카오 아레나'는 이런 대회에 반드시 필요한 2가지 모두를 명확하게 정의하여 준비하고 기획한 플랫폼입니다. 비즈니스 관점에서 실제 필요한 사항들을 대회 목적으로 설정했고, 필요한 데이터는 모두 공개했습니다. 단순 1회성에 그치는 대회가 아닌 카카오에 필요한 것들을 다 같이 고민하고 해결점을 찾아보자는 취지에서 기획한 것입니다.

보통 이런 대회를 개최하다 보면 가장 문제되는 점이 대회에서 활용할 수 있는 데이터가 없다는 것입니다. 실제 기업에서 활용하는 데이터가 아닌, 단순 데이터를 활용할 경우 이론적 연구는 가능하지만 실제와는 거리가 있는 결과들을 얻을 수밖에 없습니다. 이것은 데이터 사이언스 전반의 문제인데, 실제와 차이가 있는 데이터를 사용한다는 것 자체가 대회 디자인을 잘못한 게 아닌가라고 생각합니다. 저 역시 많은 대회에 참석해보면서 느낀 한계들이 많았는데, 그런 경험을 바탕으로 '카카오 아레나'를 기획했습니다.

Q. '카카오 아레나'를 시작하면서 주변에서 우려는 없었을까요? 수많은 데이터를 공개해야 하고 어떻게 보면 카카오가 풀지 못하는 문제를 직접 오픈해야 하는 상황이었는데 유쾌하지만은 않았을 것 같습니다.

A. 이쪽 분야 기술은 빠르게 변화합니다. 문제를 쥐고서 내부적으로만 해결하겠다는 것은 편협하게 보일 수 있습니다. 현재 머신러닝 기술들은 대부분 오픈되어 있고, 오픈 마인드 차원에서 접근해야 더 많은 기술이 발전할 수 있을 것으로 생각됩니다. 그런 차원에서 과감하게 내부 데이터도 공개하면서 다 같이 문제를 해결해보자는 취지에서 시작했습니다. 어떻게 보면 '소통의 장'을 만들었다고 볼 수 있습니다. 그리고 카카오에서 만든 플랫폼의 장점을 계속 유지할 생각이며, 대회 주제와 범위는 지속적으로 발전시켜나갈 예정입니다.

'아레나'의 어원은 고대 로마에서 원형 극장 한가운데 모래를 깔아놓은 경기장을 지칭하는 단어로, 현대에 오면서 스탠드 등을 설치하여 중앙을 볼 수 있게 해놓은 경기장 및 공연장 등을 말하는 단어로 널리 쓰이고 있다.*2 '카카오 아레나'가 시작된 계기 그리고 대회의 취지를 살펴보면 카카오의 기업 정신과 유사한 부분들이 많은 것으로 보인다. 온라인상의 '아레나'를 통해 카카오가 보유한 데이터를 공개·공유하고, 참여자들은 이 데이터 소스를 활용해 실질적인 문제를 해결하기

*2 참고 | http://dic.daum.net/word/view.do?wordid=kkw000166304&supid=kku000208550

위한 솔루션을 찾으며, 모든 사람이 데이터와 과정, 결과를 투명하게 공개한다는 원칙을 가지고 있다. 공개와 공유를 원칙으로 하는 카카오의 문화가 자연스럽게 녹아들어 '카카오 아레나'라는 공간이 만들어질 수 있지 않았을까 생각된다.

기업에서 실제 필요한 솔루션을 찾고, 그것을 위해 필요한 데이터는 아낌없이 오픈하고, 결국 모두가 윈윈할 수 있는 방법들을 고민하고, 공유·공개·소통하는 카카오의 문화가 고스란히 녹아든 진정한 오픈 플랫폼이라고 할 수 있다. 기업 입장에서는 자사가 보유한 데이터를 공개하고, 내부적인 문제를 오픈하는 것은 쉽지 않은 결정이다. 그럼에도 내부의 재산이라 할 수 있는 데이터는 물론 대회 결과물을 공개한다는 취지는 오픈 플랫폼 위에서 모두가 머리를 맞대고 문제를 고민하고 해결점을 찾아보는 진정한 오픈 이노베이션(open innovation)의 지향점이 되지 않을까 생각된다.

〈문제 해결을 위한 머신러닝 오픈 플랫폼, 카카오 아레나〉 브런치로 연결되는 QR 코드입니다.