분류 알고리즘 🡪

초록

우리은행 오픈 API중 거래내역조회로 응답 받은 TRN\_TXT 의 값인 String을 보고 카테고리를

나눌 수 있을까?

🡺

사용처 분류 기준 정의를 2030 세대에 특화된 우리만의 분류기준으로 독자적으로 정의했습니다.

2030 세대의 지출이 집중되는 배달, 온라인 쇼핑, 오프라인 쇼핑, 택시, 편의점, 서비스 구독, 운동 등을 소분류에서 꺼내서 독자적인 카테고리로 정의를 하였으며, 이후 고정비 지출이나 구독 서비스 관리 부분도 고려했습니다. 이 외에도

등으로 구분을 하였습니다.

25가지로 상세히 카테고리로 구분을 짓고 나니, String으로 받은 값으로는 구분하기가 복잡하다는 것을 알게 되었고, 그에 대한 해결책으로 여러가지 오픈 API와 크롤링으로 긁어온 텍스트에서

TF점수를 계산하고 상위 10개에 해당하는 것들에 대해서 또 다시 분류를 해주었습니다.

또한 오픈 API와 텍스트 마이닝을 통해 유용한 정보를 추출하는 방식의 단점을 보완하기 위해서 첫 layer로 String의 특징을 가지고 분류를 하였습니다.

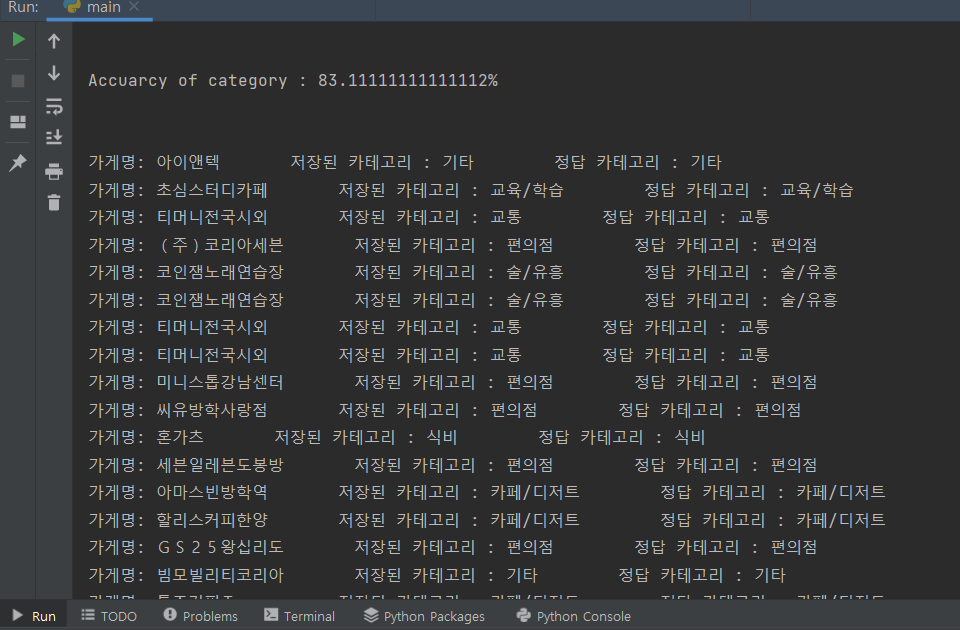
분류 알고리즘을 다시 상세히 하자면,

세 단계의 layer가 있습니다.

1. 첫번째 layer에서는 카테고리를 구분하기에 명확한 카테고리 순으로 String으로 처리를 해줬습니다. 예를 들어 ‘우아한 형제들’, ‘넷플릭스’ 처럼 명확한 카테고리 들은 첫 layer에서 거르도록 했습니다.
2. 두번째 layer에서는 네이버 Search 지역검색 API를 이용해서 미리 분류된 카테고리를 이용하여 저희가 구분한 기준에 맞게 분류를 하였습니다.
3. 두번째 layer에서 검색이 되지 않는 String이나 검색이 잘못 된 경우에는 해당 String을 네이버 웹문서와 블로그에서 검색을 하여 가져온 text들을 전처리 하고 나서 단어나 어절의 등장횟수로 TF점수를 매겨 키워드를 추출 하였습니다. 그렇게 해서 나온 키워드를 가지고 마지막 분류를 해주어서 정확도를 높였습니다.

900개의 데이터 셋을 가지고 정확도를 구해보았습니다.

우리은행 홈페이지에서 거래내역조회를 뽑게 되면 기재내용으로 나오는 TRN\_TXT string이 완전하게가 아닌 짤려서 나오기에 정확도를 높이는 데 한계가 있었습니다.



정확도의 경우는 83% 정도이며 지금까지 설명한 분류 방법이 외에도 검색에 있어서 정확성을 높이기 위해 String에서 유효한 값으로 분할 혹은 재조합으로 검색을 하거나 검색해서 나온 키워드의 정확성을 높이는 방법으로 현재의 정확도를 높일 수 있을 것 같습니다.

파이썬 KoNLPy 패키지 안의 Twitter 클래스를 이용하였습니다.

Twitter는 트위터에서 만든 오픈소스 한국어 처리기이고 KoNLPy 패키지의 다양한 클래스 중에서 Twitter를 쓴 이유는 속도가 빠르며, 다른 클래스를 사용하여도 정확도의 차이가 크지 않기에 써본 경험이 잇는 Twitter를 사용하게 되었습니다.

아래는 ‘돈수백강서점’이라는 String을 가지고 저희가 분류해 둔 카테고리에 맞게 분석해 보겠습니다.

돈수백강서점의 경우는 명확한 String이 없으며, 네이버지도에 검색을 하여도 나오지 않습니다.

그럴 경우는 네이버 블로그와 네이버 웹문서에서 검색한 text들을 긁어 와서 전처리를 하고 형태소 분석기를 이용하여 키워드를 추출하였고 키워드 에서 식비와 관련 된 것들이 나왔으므로

‘돈수백강서점’의 카테고리는 ‘식비’입니다.

