연구과제명 :사용자 감정 상태에 맞는 개인 맞춤 향수 추천 서비스

|  |  |  |  |
| --- | --- | --- | --- |
| Skt kobert 모델을 사용해 감성 분석 모델을 학습을 진행하려고 한다.  AI Hub의 감정 분류를 위한 대화 음성 데이터셋의 텍스트 데이터셋과 감성 대화 말뭉치를 이용해 감성 분석 모델 훈련을 진행하겠다.  -데이터셋의 구조  -감정 분류를 위한 대화 음성 데이터셋의 택스트 데이터셋  -감정의 종류 : 슬픔, 행복, 화남, 공포, 놀람, 중립, 혐오  -감정의 개수 : 14000개, 4548개, 11635개, 4131개, 1755개, 3262개, 4660개  -감성 대화 말뭉치 데이터셋  -감정의 종류 : 분노, 기쁨, 불안, 당황, 슬픔, 상처  -감정의 개수 : 9160개, 6126개, 9320개, 8756개, 9125개, 9143개  감성 대화 말뭉치 데이터셋이 감정별로 데이터셋의 개수가 더 균등하고 개수도 많이 분포되어 있다.  따라서 감성 대화 말뭉치 데이터셋으로 먼저 학습을 진행했다.  -데이터셋 정제  1.감정 대분류에서 ‘상처’를 삭제했다.  -우리가 개발하고자 하는 모델과는 관련이 없기 때문에 상처를 제외했다.  -test accuracy가 약 0.6484 정도 나왔다.  -생각보다 낮게 나와서 데이터를 정제를 다시 했다.  2.각 감정 데이터의 개수를 6천개로 통일했다  -데이터의 개수를 고르게 분포하도록 하여 효과적으로 학습이 되도록 했다.  -test accuracy가 약 0.6617 정도 나왔다.  -전과 동일하게 나와서 데이터 정제를 다시 했다.  3.감정대분류와 감정소분류를 비교했다.  -감정소분류가 더 정확한 경우가 많아 데이터셋의 개수를 줄이고  감정 라벨링의 정확도를 높이고자 데이터를 정제했다.  -test accuracy가 약 0.7447 정도 나왔다.  -정확성이 전보다 많이 올라갔지만 조금 더 정확한 데이터셋이 필요하다는 것을 알게되었다.  감성 대화 말뭉치 데이터셋이 데이터는 균등하지만 문장에 대한 감정 라벨링의 신뢰성이 떨어진다는 것을 알게 되었다. 따라서 감정 분류를 위한 대화 음성 데이터셋의 텍스트 데이터셋을 사용해보기로 했다.  -데이터셋 정제  1.’중립’, ’혐오’ 삭제했다.  -우리가 개발하고자 하는 모델과는 관련이 없기 때문에 삭제했다.  -학습  1. 5가지 감정에 대해서 10epoch로 진행했다.  2, 4epoch일 때 validation loss가 0.1987로 가장 낮게 나왔다.  3. 그 때 train accuracy는 0.9475, validation accuracy는 0.9312가 나왔다.  Epoch가 10일 때 validation 데이터셋을 이용해 각 클래스 별 정확도를 구해보았다.  슬픔 : 0.9408, 행복 : 0.9397, 화남 : 0.9511, 공포 : 0.9281, 놀람 : 0.9238  높은 정확도를 보이는 것을 알 수 있다. 하지만 데이터셋이 고르지 않고 놀람의 경우 1755개로 아주 적은 개수이므로 데이터셋을 목적에 맞게 신뢰성 높게 정제하려고 한다. | | | |
| **작성자:김정현** | **일자:2023/09/19** | **확인자** | **일자** |
| **서명:김정현** | **서명** |

연구과제명 :사용자 감정 상태에 맞는 개인 맞춤 향수 추천 서비스

|  |  |  |  |
| --- | --- | --- | --- |
| Skt kobert 모델을 사용해 감성 분석 모델을 학습을 진행했다.  데이터셋으로는 AI Hub의 감정 분류를 위한 대화 음성 데이터셋의 텍스트 데이터셋을 사용하였다.  데이터셋을 다운받았을 때 3가지 파일로 나뉘어져 있었다.  3가지 파일을 합쳐 하나의 csv 파일로 만들었다.  -데이터셋의 구조  -감정 분류를 위한 대화 음성 데이터셋의 택스트 데이터셋  -감정의 종류 : 슬픔, 행복, 화남, 공포, 놀람, 중립, 혐오  -감정의 개수 : 14000개, 4548개, 11635개, 4131개, 1755개, 3262개, 4660개  이 7가지 감정 중에서 우리가 필요로 하는 슬픔, 행복, 화남 공포, 놀람 5가지를 뽑아 학습을 진행했다.  이전에는 max\_length의 길이를 50으로도 해보았으나 문장의 길이는 학습의 정확도에 영향을 미친다고 해서 모든 문장을 확실하게 학습하게 하기 위해 가장 긴 문장의 길이를 구하였다.  가장 긴 문장의 길이가 137자여서 max\_length를 137로 해서 학습을 진행했다.  그리고 이전에는 optimizer를 AdamW를 사용했으나 이번에는 RAdam을 사용해 진행하였다.  둘 다 adam의 변형이지만 Radam은 적응적 학습률을 도입했다. 초기 학습률을 높게 설정하고 학습 진행에 따라 자동으로 조절해준다. 우리의 데이터셋 더 학습을 도와줄 수 있는 optimizer를 고르기 위해 이번에는 RAdam으로 진행하였다. 그리고 이전에는 데이터셋을 train과 test으로 나눌 때 단순하게 나누었다.  -이전 코드  -train, test = train\_test\_split(df, test\_size=0.20, random\_state=42)  이렇게 단순하게 전체를 8대2로 나누어 진행하니 validation loss 의 결과가 이상하게 나왔다.  그래프가 극소점의 형태를 띄지 않고 epoch가 진행되어도 비슷한 값이 나왔다.  따라서 이번에는 좀 다르게 진행을 하였다.  우리가 선택한 데이터셋과 같이 라벨 당 데이터의 개수가 고르지 않을 경우 단순하게 train과 test를 나누는 것은 학습에 영향을 미친다는 것을 알 수 있었다.  그래서 라벨 당 train과 test를 8대2로 나누어주었다.  -수정한 코드  -train, test = train\_test\_split(df, test\_size=0.20, stratify=df[‘label’], random\_state=42)  이렇게 나누어주고 이전에는 epoch를 10까지만 진행하였다.  이번에는 조금 더 전체적으로 확인할 수 있도록 epoch를 20까지 진행하였다.    위 그림과 같은 그래프를 볼 수 있었다. Epoch 6에서 loss가 0.2313으로 가장 낮게 나왔다.  그리고 그 때 validation 정확도는 0.9339로 높은 정확도를 얻을 수 있었다. | | | |
| **작성자:김정현** | **일자:2023/09/28** | **확인자** | **일자** |
| **서명:김정현** | **서명** |

연구과제명 :사용자 감정 상태에 맞는 개인 맞춤 향수 추천 서비스

* Kobert 학습 및 웹사이트 크롤링

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Skt kobert 모델을 사용해 감성 분석 모델을 학습을 진행했다.  데이터셋으로는 AI Hub의 감정 분류를 위한 대화 음성 데이터셋의 텍스트 데이터셋을 사용하였다.  -데이터셋의 구조  -감정 분류를 위한 대화 음성 데이터셋의 택스트 데이터셋  -감정의 종류 : 슬픔, 행복, 화남, 공포, 놀람, 중립, 혐오  -감정의 개수 : 14000개, 4548개, 11635개, 4131개, 1755개, 3262개, 4660개  이 7가지 감정 중에서 우리가 필요로 하는 슬픔, 행복, 화남 공포, 놀람 5가지를 뽑아 학습을 진행했다.  Train data, Test data, Validation data를 4:1:1로 나누었다.   |  |  |  |  | | --- | --- | --- | --- | |  | **Train Data** | **Test Data** | **Validation Data** | | **슬픔** | **8960개** | **2800개** | **2240개** | | **행복** | **2910개** | **910개** | **728개** | | **화남** | **7447개** | **2327개** | **1861개** | | **공포** | **2644개** | **826개** | **661개** | | **놀람** | **1123개** | **351개** | **281개** |   .  위 표와 같이 데이터를 나누어주었다.  이전 주에서 진행해서 6 epoch에서 가장 낮은 loss를 보였다.  그 때 정확도는 train accuracy가 0.9621, validation accuracy가 0.9339가 나왔다.  그리고 이번주에는 향수 데이터를 뽑아내기 위해 웹사이트 크롤링을 진행했다.  파이썬을 이용해 크롤링을 진행하였고 BeautifulSoup과 requests를 이용해 크롤링을 했다.  <https://basenotes.com/> 사이트에서 크롤링을 진행하였다.  첫번째로 모든 개별 향수의 url을 텍스트 파일로 저장하였다.  두번째로 그 파일을 여러 개의 파일로 분할하였다.  마지막으로 크롤링을 하여 각 향수의 탑노트, 미들노트, 베이스노트를 엑셀에 저장하였다.  그러나 우리가 기대했던 데이터와 다르게 나왔다.  기대했던 데이터는 향수 하나에 여러 향들이 들어있으나 실제로 크롤링한 데이터에는 향 데이터가 조금밖에 들어있지 않았다.  향수의 향 개수가 많을수록 데이터가 손실되는 것을 발견했고 그 오류를 수정하기 위해 노력하고 있다.  그리고 MF 알고리즘을 더 정확하게 파악하기 위해 실제로 영화 평점 데이터를 이용해 추천 알고리즘을 돌려보았다. 유저와 영화 데이터와 영화와 평점 데이터를 pivot\_table을 이용해 행을 userId 열을 영화로 해서 각 항에는 유저의 영화 평점을 넣었고 평점을 매기지 않은 항에 대해서는 0을 채워 넣었다. from sklearn.decomposition import TruncatedSVD  Sklearn에서 제공하는 TruncatedSVD를 이용해 잠재요인의 개수를 12로 설정하고 추천 알고리즘을 영화 평점에 맞춰 돌려보았다.  실제로 확인해본 결과 영화를 추천해주기는 하지만 NULL값을 0으로 대체했을 때 0을 하나의 값으로 적용돼서 0에 대해서도 예측값이 도출되기 때문에 정확하게 추천되지 않았다.  다음주에는 그것을 개선한 알고리즘을 돌려볼 예정이다. | | | |
| **작성자:김정현** | **일자:2023/10/05** | **확인자** | **일자** |
| **서명:김정현** | **서명** |

연구과제명 :사용자 감정 상태에 맞는 개인 맞춤 향수 추천 서비스

* 웹 크롤링를 이용한 데이터 수집

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Skt kobert 모델을 사용해 감성 분석 모델을 AI Hub의 감정 분류를 위한 대화 음성 데이터셋의 텍스트 데이터셋으로 학습시켰다.  -데이터셋의 구조  -감정 분류를 위한 대화 음성 데이터셋의 택스트 데이터셋  -감정의 종류 : 슬픔, 행복, 화남, 공포, 놀람  -감정의 개수 : 14000개, 4548개, 11635개, 4131개, 1755개  Train data, Test data, Validation data를 4:1:1로 나누었다.   |  |  |  |  | | --- | --- | --- | --- | |  | **Train Data** | **Test Data** | **Validation Data** | | **슬픔** | **8960개** | **2800개** | **2800개** | | **행복** | **2910개** | **910개** | **910개** | | **화남** | **7447개** | **2327개** | **2327개** | | **공포** | **2644개** | **826개** | **826개** | | **놀람** | **1123개** | **351개** | **351개** |   지난 주의 결과에 따라서 validation loss가 가장 낮은 6 epoch로 학습을 진행하기로 결정했다.  정확도는 train accuracy가 0.9621, validation accuracy가 0.9339가 나왔다.  tokenizer = KoBertTokenizer.from\_pretrained('monologg/kobert')  이 코드에서 갑자기 오류가 나서 코드가 진행이 되지 않았다.  미리 훈련된 토크나이저를 불러오는 코드인데 오류가 나서 수정하기 위해 검색중에 있다.  그리고 이번주에는 향수 데이터를 뽑아내기 위해 웹사이트 크롤링을 진행했다.  저번주에는 파이썬을 이용해 크롤링을 진행하였고 BeautifulSoup과 requests를 이용해 크롤링을 했다.  그치만 정적으로 크롤링을 할 경우 자바 스크립트로 이루어진 사이트에서는 크롤링이 진행이 되지 않았다.  따라서 selenium이라는 툴을 이용해 크롤링을 진행하였다.  먼저 셀레니움에 대해 알아보았다.  셀레니움이란 웹 상에서 정적인 페이지를 탐색하는데 사용하던 beautifulsoup 같은 패키지가 하지 못하는 동적인 크롤링을 지원한다. 따라서 우리의 현 상황과 정확하게 맞아 떨어졌기 때문에 사용하게 되었다.  셀레니움을 이용해 직접 사이트에 들어가 조작하는 것처럼 크롤링을 진행할 수 있었다.  그치만 크롤링을 진행할 때 for나 while 문을 이용해 크롤링을 하는 경우가 많은데 생각보다 웹 사이트에서 템소의 일관성이 없는 경우가 많아 크롤링이 제대로 되지 않았다. 따라서 일관성이 좋은 사이트를 선택하는 것도 하나의 방법이라고 생각이 들었다.  그리고 크롤링을 진행할 때, mf 알고리즘의 경우 유저 당 향수를 평가한 평점이 여러 개가 있어야 신뢰성이 올라가는데 올리브영 사이트의 향수의 경우에는 여러 개의 향수를 크롤링 해도 사용자가 평점을 하나만 내리는 경우가 많았다. 따라서 이 부분을 해결하기 위해 향수 사이트를 옮겨서 진행하려고 한다. 향수만 전문적으로 다루는 사이트가 있는데 그 곳에도 향수, 유저, 평점 등의 데이터들이 더 자세하게 표현되어 있다. 하지만 그 사이트 역시 xpath가 일관성이 없는 문제가 있어 데이터 수집에 어려움을 겪고 있다.  하지만 웹 크롤링 문제는 해결 가능한 문제라고 생각이 들었고 우리가 사용하는 추천 알고리즘에 맞게 어떤 데이터를 수집할 것인지에 대한 고민만 해결하면 될 것 같다.  추천 알고리즘은 지난 주와 같이 svd를 이용한 행렬분해를 사용해보고 있다. | | | |
| **작성자:김정현** | **일자:2023/10/12** | **확인자** | **일자** |
| **서명:김정현** | **서명** |

연구과제명 :사용자 감정 상태에 맞는 개인 맞춤 향수 추천 서비스

* 데이터 수집 및 kobert 학습 완료

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Skt kobert 모델을 사용해 감성 분석 모델을 AI Hub의 감정 분류를 위한 대화 음성 데이터셋의 텍스트 데이터셋으로 학습시켰다.  -데이터셋의 구조  -감정 분류를 위한 대화 음성 데이터셋의 택스트 데이터셋  -감정의 종류 : 슬픔, 행복, 화남, 공포, 놀람  -감정의 개수 : 14000개, 4548개, 11635개, 4131개, 1755개  Train data, Test data, Validation data를 4:1:1로 나누었다.   |  |  |  |  | | --- | --- | --- | --- | |  | **Train Data** | **Test Data** | **Validation Data** | | **슬픔** | **8960개** | **2800개** | **2800개** | | **행복** | **2910개** | **910개** | **910개** | | **화남** | **7447개** | **2327개** | **2327개** | | **공포** | **2644개** | **826개** | **826개** | | **놀람** | **1123개** | **351개** | **351개** |     10 epoch 학습 후 loss graph 이다.  Epoch 6에서 극솟점을 보이고 있다. 따라서 epoch 6으로 학습을 진행하려고 한다.    클래스 0~4는 각각 슬픔, 행복, 화남, 공포, 놀람을 의미한다.  그리고 학습을 완료한 모델을 h5 파일로 저장하여 또 다시 학습할 필요 없이 파일을 불러올 수 있게 만들었다.  From tensorflow.keras.models import load\_model  위 코드를 이용해 저장하고 불러올 수 있었다.  향수 데이터를 얻는 과정은 셀레니움으로 fragrantica 사이트에서 진행을 했다.  하지만 사이트에서 셀레니움의 접근이 제한되어 있어 데이터 수집을 할 수가 없었다.  따라서 다른 방법으로 얻기 위해 basenote 등의 사이트에서 향수 정보에 대한 데이터만 수집을 진행하였다.  하지만 추천 알고리즘을 위해서는 유저 평점 데이터가 필요했는데 검색을 하던 중 깃허브에서 데이터셋을 발견하였다. 유저 평점 데이터와 향수 데이터, 그리고 향수 안에 있는 향 데이터까지 얻을 수 있었다.  향수에 대한 프로젝트를 진행한 사람이 fragrantica와 basenote를 참고하여 만든 데이터셋을 발견하였다.  다음 연구에서는 이 데이터를 이용해 추천 알고리즘을 돌려볼 예정이다. | | | |
| **작성자:김정현** | **일자:2023/10/19** | **확인자** | **일자** |
| **서명:김정현** | **서명** |