텍스트 분류 프로젝트

- 프로그래밍 언어 및 도구: Python (pandas, numpy, nltk, scikit-learn, xgboost)
- 활용한 기술: 데이터 전처리, 머신러닝

[프로젝트 소개]

- 2024년도 봄 학기 머신러닝 수업 프로젝트
- 의료 진단 차트의 텍스트를 분류 알고리즘 모델을 활용하여 6개의 카테고리 중 한가지로 분류하였습니다.

Step 1: 데이터 전처리

[소문자 변경과 특수문자 제거]

2	114	PROCEDURE: , Newborn circumcision.,INDICATIONS: , Parental preference.,ANESTHESIA:, Dorsal penile nerve block.,DESCRIPTION OF PROCEDURE:, The baby was prepared and draped in a sterile manner. Lidocaine 1% 4 mL without epinephrine was instilled into the base of the penis at 2 o'clock and 10 o'clock. The penile foreskin was removed using a XXX Gomco. Hemostasis was achieved with minimal blood loss. There was no sign of infection. The baby tolerated the procedure well. Vaseline was applied to the penis, and the baby was diapered by nursing staff.

2	114	procedure newborn circumcisionindications parental preferenceanesthesia dorsal penile nerve blockdescription of procedure the baby was prepared and draped in a sterile manner lidocaine 1 4 ml without epinephrine was instilled into the base of the penis at 2 oclock and 10 oclock the penile foreskin was removed using a xxx gomco hemostasis was achieved with minimal blood loss there was no sign of infection the baby tolerated the procedure well vaseline was applied to the penis and the baby was diapered by nursing staff

- 대문자와 소문자가 섞여 있으면 같은 단어를 다르게 인식할 수 있기 때문에, 'The', 'the'와 같은 단어들을 소문자로 통일하여 동일하게 인식하도록 합니다.
- 특수문자는 의미 있는 정보를 제공하지 않는 경우가 많아 제거합니다.

[불용어 **(stop words)** 제거]

2	114	procedure newborn circumcisionindications parental preferenceanesthesia dorsal penile nerve blockdescription of procedure the baby was prepared and draped in a sterile manner lidocaine 1 4 ml without epinephrine was instilled into the base of the penis at 2 oclock and 10 oclock the penile foreskin was removed using a xxx gomco hemostasis was achieved with minimal blood loss there was no sign of infection the baby tolerated the procedure well vaseline was applied to the penis and the baby was diapered by nursing staff

2	114	procedure newborn circumcisionindications parental preferenceanesthesia dorsal penile nerve blockdescription procedure baby prepared draped sterile manner lidocaine 1 4 ml without epinephrine instilled base penis 2 oclock 10 oclock penile foreskin removed using xxx gomco hemostasis achieved minimal blood loss sign infection baby tolerated procedure well vaseline applied penis baby diapered nursing staff

- 불용어는 의미 있는 정보를 제공하지 않기 때문에 제거합니다.
- 불용어를 제거함으로써 모델이 더 의미 있는 단어들에 집중하게 되어 분류 작업에 더 잘 맞는 패턴과 관계를 쉽게 파악할 수 있습니다.

[표제화 **(lemmatization)**]

2	114	procedure newborn circumcisionindications parental preferenceanesthesia dorsal penile nerve blockdescription procedure baby prepared draped sterile manner lidocaine 1 4 ml without epinephrine instilled base penis 2 oclock 10 oclock penile foreskin removed using xxx gomco hemostasis achieved minimal blood loss sign infection baby tolerated procedure well vaseline applied penis baby diapered nursing staff

2	114	procedure newborn circumcisionindications parental preferenceanesthesia dorsal penile nerve blockdescription procedure baby prepare drape sterile manner lidocaine 1 4 ml without epinephrine instill base penis 2 oclock 10 oclock penile foreskin remove use xxx gomco hemostasis achieve minimal blood loss sign infection baby tolerate procedure well vaseline apply penis baby diapered nurse staff

- 환자가 호소하는 증상, 치료받은 기록과 같은 행동에 관련된 텍스트 데이터이기 때문에 동사 형태로 표제화 합니다.
- 다양한 형태로 변형된 단어들을 기본 동사 형태로 일관되게 처리하여 중복된 뜻의 단어들과 노이즈를 제거하여 모델의 성능을 향상 시킬 수 있습니다.

[벡터라이징 **(TF-IDF)**]

index	Sl. No.	transcription
0	480	cc orthostatic lightheadednesshx 76 yo male complained several month generalized weakness malaise two week history progressively worsening orthostatic dizziness dizziness worsened moving upright position addition complained intermittent throbbing holocranial headache worsen positional change past several week lost 40 pound past year denied recent fever sob cough vomiting diarrhea hemoptysis melena hematochezia bright red blood per rectum polyuria night sweat visual change syncopal episodeshe 100 packyear history tobacco use continued smoke 1 2 pack per day history sinusitisexam bp 9880 mmhg pulse 64 bpm supine bp 70palpable mmhg pulse 84bpm standing rr 12 afebrile appeared fatiguedcn unremarkablemotor sensory exam unremarkablecoord slowed otherwise unremarkable movementsreflexes 22 symmetric throughout 4 extremity plantar response flexor bilaterallythe rest neurologic general physical exam unremarkablelab na 121 meql k 42 meql cl 90 meql co2 20meql bun 12mgdl cr 10mgdl glucose 99mgdl esr 30mmhr cbc wnl nl wbc differential urinalysis sg 1016 otherwise wnl tsh 28 iuml ft4 09ngdl urine osmolality 246 mosmkg low urine na 35 meqlcourse patient initially hydrated iv normal saline orthostatic hypotension resolved returned within 2448hrs laboratory study revealed aldosterone serum2ngdl low 30 minute cortrosyn stimulation test pre 69ugdl borderline low post 185ugdl normal stimulation rise prolactin 155ngml baseline given fsh lh within normal limit male testosterone 33ngdl wnl sinus xr series done history headache showed abnormal sellar region enlarged sella tursica destruction posterior clinoids also abnormal calcification seen middle sellar region left maxillary sinus opacity airfluid level seen goldman visual field testing unremarkable brain ct mri revealed suprasellar mass consistent pituitary adenoma treated fludrocortisone 005 mg bid within 24hrs despite discontinuation iv fluid remained hemodynamically stable free symptom orthostatic hypotension presumed pituitary adenoma continues managed fludrocortisone writing 11997 though developed dementia felt secondary cerebrovascular disease stroketia

→

```
(0, 258)	0.06595154696392722
(0, 777)	0.07472498255963922
(0, 345)	0.07806531578134777
(0, 241)	0.08105077066440206
(0, 194)	0.09194272855188064
(0, 872)	0.0648063402938391
(0, 372)	0.07651999038587738
(0, 831)	0.05290078020141891
(0, 728)	0.09426182853960348
(0, 359)	0.06458404764036384
(0, 105)	0.08853567546916061
(0, 545)	0.056949213176627614
(0, 915)	0.08198089230868141
(0, 190)	0.07151053262086858
(0, 533)	0.06017325470310224
(0, 566)	0.0792972274833733
(0, 210)	0.07438234728703298
(0, 121)	0.08557243883872438
(0, 889)	0.09194272855188064
(0, 351)	0.08671191708231989
(0, 499)	0.06595154696392722
(0, 495)	0.04171662090827811
(0, 547)	0.097774123454979
(0, 782)	0.11203818306469654
(0, 57)	0.045984586466009794
```

- TF-IDF는 텍스트를 토크나이징한 후, 단어의 중요도를 평가하여 문서 내에서의 중요도를 계산하고 이를 벡터로 변환합니다.
- 예시: (0, 258) 0.06596는 첫 번째 문서의 258번째 단어의 중요도가 0.06596 (범위: 0 ~ 1)이라는 뜻입니다.

Step 2: 머신러닝 **(Stacking Algorithm)**

[Stacking Algorithm 모델 생성**]**

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import StackingClassifier
from sklearn.naive_bayes import MultinomialNB
from xgboost import XGBClassifier

base_models = [
    ('lr', LogisticRegression(max_iter = 2000)),
    ('nb', MultinomialNB()),
    ('xgb', XGBClassifier(use_label_encoder=False, objective='multi:softmax', num_class = 6, eval_metric='mlogloss')) # multi:softmax bc we are classifying into one of 6
]

meta_model = LogisticRegression()

stacking_clf = StackingClassifier(estimators=base_models, final_estimator=meta_model)
```

- 텍스트 데이터는 단어의 순서, 의미, 문맥 등을 고려해야 하므로 여러 모델의 강점을 결합하여 예측 정확성을 높일 수 있는 스태킹 알고리즘을 사용합니다.
- 로지스틱 회귀는 텍스트 데이터의 선형적 관계를 잘 포착하고, XGBoost는 복잡한 텍스트 데이터의 비선형적 관계를 잘 설명하며, Naive Bayes Classifier는 텍스트 데이터를 간단하고 효과적으로 분류할 수 있습니다.
- 로지스틱 회귀는 비교적 단순한 모델이기 때문에 훈련 데이터에 과적합될 위험이 적어 meta model로 사용합니다.

[하이퍼파라미터 튜닝]

```python
from sklearn.model_selection import GridSearchCV

param_grid = {
    'lr__C': [0.1, 1, 10], # regularization prevent overfitting
    'xgb__learning_rate': [0.01, 0.1, 0.2], # step sizes for updating the model weights (bigger step will learn faster but less deatailed)
    'xgb__max_depth': [3, 6, 9], # increasing makes the model more complex and capable of learning more detailed patterns
    'xgb__n_estimators': [100, 200, 300] # number of trees in the ensemble (each tree is built to correct the errors made by the previous trees)
}

grid_search = GridSearchCV(estimator=stacking_clf, param_grid=param_grid, cv=3, scoring='f1_weighted', verbose=2, n_jobs=-1)

grid_search.fit(X_train_v, y_train['medical_specialty'])

best_params = grid_search.best_params_
print("best parameters:", grid_search.best_params_)
```

- 스태킹 알고리즘에 사용된 로지스틱 회귀와 XGBoost의 파라미터들을 여러 값으로 설정:
  - lr__C: 규제 강도를 조절하여 과적합을 방지.
  - xgb__learning_rate: 모델 가중치를 업데이트하는 스텝 크기; 큰 값은 빠르게 학습하지만 덜 세밀합니다.
  - xgb__max_depth: 모델의 복잡성을 증가시켜 더 세밀한 패턴을 학습할 수 있습니다.
  - xgb__n_estimators: 앙상블의 트리 수; 각 트리는 이전 트리의 오류를 수정합니다.
- GridSearchCV를 통해 모든 조합을 테스트하여 가장 높은 f1_weighted 점수가 나온 파라미터 조합을 저장합니다.

[예측]

```python
best_model = grid_search.best_estimator_

y_pred = best_model.predict(X_test_v)
```

- f1_weighted 점수가 가장 높게 나온 파라미터 조합을 사용하여 테스트 데이터 셋의 결과값을 예측합니다.

[점수 **(Mean F-1 Score)**]

- Mean F-1 score의 값이 0.85로 이는 모델이 높은 정밀도 (모델이 True로 분류한 사례 중 85%가 실제로 True임)와 높은 재현율 (실제 True로 분류한 사례 중 85%를 모델이 예측함)을 구현합니다.
- 모델이 실제로 긍정인 사례를 잘 찾아내며, False Positive (오탐)와 True Positive를 놓치는 경우 (누락)가 적다는 것을 의미합니다.