미래기술마당 개선 프로젝트

공공빅데이터 일경험 수련생

분석 결과 보고서

03: 자연어 처리 기반 특허 기술 분류

기 관 명 과학기술일자리진흥원
수행기간 2023년 01월 ~ 2023년 02월
수 련 생 김인수 황양하

행정안전부

NIA 한국지능정보사회진흥원

CSLEE
㈜씨에스리

목차

S-BERT기반 수요기술 매칭 서비스

01

프로젝트 개요

전체프로젝트 개요

미래기술마당 **서비스 개선**을 위한 프로젝트 진행

프로젝트 1:
S-BERT 기반 수요기술 매칭 서비스

수요자가 신청한 수요기술의 내용을 기반으로 유사한 특허를 찾아내어 수요자에게 유사 특허 리스트를 제공

프로젝트 2:
BI 기반 통계 서비스

미래기술마당 관리자 통계 화면의 문제점을 개선하고, 사용자를 위한 검색어 통계 시각화 서비스를 제공

프로젝트 3:
자연어 처리 기반 특허 기술 분류

자연어 처리 기술을 활용하여 특허 데이터를 미래유망기술(6T) 분류 기준으로 분류하는 모델을 제작

프로젝트 진행 과정


9월
10월
11월
12월
1월
2월
사업화유망기술 DB정제
SBERT 기반 수요기술 매칭 서비스
BI기반 통계 서비스
모델 개선 & 자연어 처리 기반 특허 분류

02

분석 배경

미래유망신기술이란?

목적 및 근거

PROBLEM

내가 찾은 **특허**,  **미래유망신기술** 분류별로 나누어 볼 순 없을까?

앞서 진행했던 프로젝트인 수요기술 매칭 서비스의 결과를 사용자에게 효과적으로 보여주기 위해
자연어 처리 기술을 이용하여 특허를 미래유망신기술 분류별로 나누어 보고자 함.

미래유망신기술이란?

미래유망신기술(6T)이란 인류의 미래를 주도할 첨단 산업기술로서 주목받고 있는 6가지의 기술을 말한다.

정보기술
IT

생명공학기술
BT

나노기술
NT

환경공학기술
ET

우주항공기술
ST

문화콘텐츠기술
CT

수요기술 신청은 6T 분류 기준으로 신청하지만
추천 특허들은 6T기준으로 분류 되어 있지 않아
사용자들이 6T기준으로 특허를 볼 수 없는 불편함이 있음.

특허를 **6T 기준으로 분류**하여 사용자에게 제공하고자 함.

03

모델 프로세스

프로젝트 개요

모델 프로세스

최종 모델

DATA

프로젝트 활용 데이터

데이터	건수
NTIS 사업 과제 정보	451,070건
KIPRIS 특허 서지정보	1,588,185건

TOOL

프로젝트에서 사용한 툴

jupyter

PostgreSQL

TensorFlow

Keras

미래유망기술 분류 **프로세스**

특허를 6T 기준으로 분류하여 기술 수요자들의 편의성을 향상시키고자 함.

NTIS에 등록된
국가 사업 과제 정보를
학습데이터로 활용

학습데이터를 활용해
여러 분류 모델의 성능을
비교하여 최적의 모델 선정

최종적으로
선정된 모델을 활용하여
특허를 6T 기준으로
분류하고 시각화

PROCESS 01: 데이터 전처리

NTIS 국가 과제 정보 데이터의 '초록'과 Target 데이터인 '6T분류' 데이터 정제

```
df.abstractfullteaser.iloc[0]
```

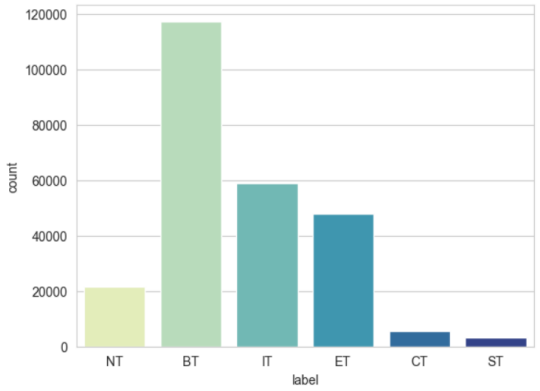
```
'▶ 전면적/선택적 나노표면 구현을 위한 나노 금형 인서트 제작기술 \n\n\n- 나노 금형 인서트 설계: 나노 제작공정의 한계를 고려한 금형설계\n\n\n- 나노
금형 인서트 제작: (1) 양극산화 알루미늄 공정과 롤로이드 리소그래피 공정을 응용한 나노구조 템플릿 제작, (2) 니켈 나노전주도금을 이용한 고내구성 나노
금형 인서트 제작\n\n\n- 나노 금형 인서트 평가: 제작된 나노 금형 인서트에 대한 기계적 특성 평가\n\n\n▶ 나노표면 대량생산을 위한 성형기술\n\n\n- 나
노표면 성형공정모사: 컴퓨터 모사기술을 통한 나노표면 공정변수 도출 및 공정 최적화\n\n\n- 나노표면 성형공정기술: 나노사출/압축성형, 나노엠보싱공정을
기반으로 나노표면의 제작 및 대량생산 가능성 확인\n\n\n- 나노표면 성형공정평가: 제작된 나노표면의 성형성 평가 및 화학적/물리적 특성 평가\n\n\n▶ 세
포외기질 단백질의 미세패턴 전사기술\n\n\n- 단백질 미세패턴 표면 설계: 미세 제작공정의 한계를 고려한 연성 스탬프 및 스텐실 설계\n\n\n- 단백질 미세패
턴 표면 제작: (1) 마이크로 콘택트 프린팅 공정과 (2) 스텐실 공정을 응용한 단백질 미세패턴 표면의 제작\n\n\n- 단백질 미세패턴 표면 평가: 세포외기질
단백질 미세패턴이 전사된 표면의 화학적/물리적 특성 평가\n\n\n▶ 나노표면과 단백질 미세패턴 표면을 이용한 단일줄기세포 거동제어\n\n\n- 세포외기질 단
백질 미세패턴에 의한 단일줄기세포 거동제어 연구 (화학적 요인)\n\n\n- 전면적/선택적 나노표면에 의한 단일줄기세포 거동제어 연구 (물리적 요인) \n\n\n-
선택적 나노표면과 세포외기질 단백질 미세패턴의 복합적 영향에 의한 단일줄기세포 거동제어 연구 (물리/화학적 요인)\n\n\n▶ 단일줄기세포 거동제어용 선택
적 나노표면 세포배양조 대량생산 기술 확보\n\n\n- 세포배양조 설계: 거동제어 연구를 통해 도출된 결과를 바탕으로 세포배양조를 설계  \n\n\n- 세포배양
조 제작: (1) 성형공정 기술을 고려한 세포배양조 금형설계, (2) 컴퓨터 모사기술을 통한  공정변수 도출 및 공정 최적화, (3) 성형공정 최적화를 통한 대량
생산 기반 기술 확보\n\n\n- 세포배양조 평가: 제작된 세포배양조의 화학적/물리적 특성평가'
```

- 괄호와 괄호 안에 있는 문자 제거
- 모든 영문자는 소문자로 변경
- 특수 문자와 한 칸 이상의 공백 제거
- HTML 태그 제거
- 숫자 넘버링(로마자, 영어) 제거
- 중복 데이터와 빈 열 제거

```
BT(생명공학기술)                                  180580
위의 미래유망신기술(6T) 103개 세분류에 속하지 않는 기타 연구      88205
IT(정보기술)                                      77866
ET(환경기술)                                      62157
NT(나노기술)                                      32578
CT(문화기술)                                       6930
ST(우주항공기술)                                   5041
```

- Label 데이터의 특수문자 제거
- 괄호와 괄호 안 문자 제거
- Label 중 기타에 해당하는 분류는 성능과 관련 없는 데이터로 간주해 훈련과 평가 데이터에서 제외

PROCESS 01: 데이터 전처리



Train 데이터의 Target 분포를 확인해 본 결과,
데이터의 불균형이 매우 심한 것을 확인 할 수 있었음.

불균형 데이터 레이블의 균형을 맞춰주기 위해
Text Data Augmentation을 진행함.

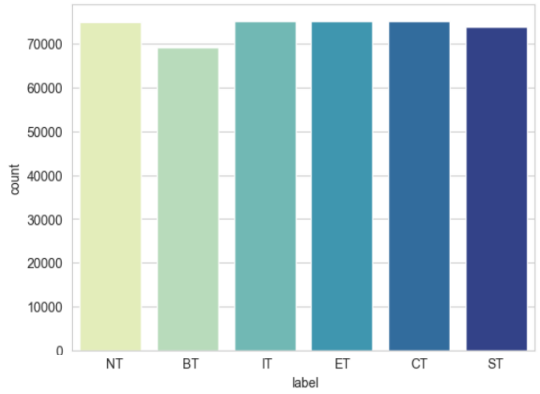
Synonym Replacement
특정 단어를 유의어로 교체하는 방식

Random Deletion
임의의 단어를 삭제하는 방식

Random Insertion
임의의 단어를 삽입하는 방식

Random Swap
문장 내 임의의 단어의 위치를 바꾸는 방식

대표적인 4가지 텍스트 데이터 증강 기법 중
적용이 권장되는 방법은 RD와 RS.
둘 중 데이터와 문맥의 손실이 적다고 판단되는
Random Swap 방법을 사용하여 데이터를 증강시킴.
증강 후 가장 많은 데이터를 기준으로 샘플링 함.



데이터 증강 전에 비해 Target 값의 불균형이
많이 해소된 것을 볼 수 있음.
모델 학습 시에는 **증강 전과 후 데이터**에
모두 학습하여 그 성능을 비교 함.

PROCESS 02: 모델 학습 - LSTM

데이터의 형식을 LSTM의 학습에 적합하도록 수정해 줌.

문장 샘플들을 신경망 모델의 Input이 될 수 있도록 문장을 **토큰화** 한 후, **정수 시퀀스**로 만들어 주어야 함.

Mecab에 KoPatElectra의 사용자 단어 사전을 추가하여 만든 토크나이저로 문장을 토큰화시켜 전처리함.

고리서열 재설계를 통한
단백질 접힘 제어에 관한 연구

→

불용어 제외

단어 길이가 2 미만인 단어 제외

명사, 어근, 외국어만 추출

→

'고리서열' '재설계'
'단백질' '접힘' '제어' '연구'

PROCESS 02: 모델 학습 - LSTM

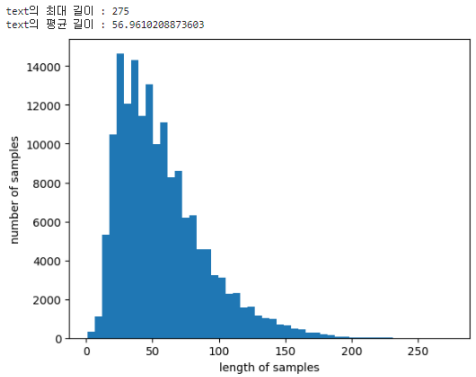
```
단어 집합(vocabulary)의 크기 : 345120
등장 빈도가 2번 이하인 희귀 단어의 수: 213442
단어 집합에서 희귀 단어의 비율: 61.845734816875286
전체 등장 빈도에서 희귀 단어 등장 빈도 비율: 2.804399131304
단어 집합의 크기 : 131679
```

빈도수가 낮은 단어는 훈련 데이터에서 **제외**하기 위해 등장 빈도수가 2회 이하인 희귀 단어들의 분포를 확인.

희귀 단어들은 전체 데이터에서 2.84%를 차지하고 있음
모델 훈련 과정에서 중요하지 않다고 판단하여 제외.

희귀 단어 제외로 인해 값이 없어진 샘플들은 삭제.

Keras의 tokenizer를 통해
텍스트 시퀀스를
정수 시퀀스로 변환




```
전체 샘플 중 길이가 150 이하인 샘플의 비율: 98.37
```

정수 시퀀스로 변환된 샘플들의
서로 다른 길이를 동일하게 맞춰 줌.
전체 데이터 길이의 분포를 파악한 후,
전체 훈련 데이터의 98%를 반영하는 150으로
모든 데이터의 **길이를 통일**

PROCESS 02: 모델 학습 - LSTM

다중분류 학습에 적합하게 LSTM모델의 하이퍼 파라미터를 조정함

```
embedding_dim = 256
hidden_units = 256
num_classes = 6

model = Sequential()
model.add(Embedding(vocab_size, embedding_dim, input_length=150))
model.add(LSTM(hidden_units, activation = 'relu', return_sequences = False))
model.add(Dense(num_classes, activation='softmax'))

es = EarlyStopping(monitor='val_loss', mode='auto', verbose=1, patience=7)
mc = ModelCheckpoint('bm_ntis_lstm0222.h5', monitor='val_acc', mode='max', verbose=1, save_best_only=True)

model.compile(loss='categorical_crossentropy', optimizer=tf.keras.optimizers.RMSprop(learning_rate=0.0001), metrics=['acc'])
history = model.fit(X_tr, y_tr, batch_size=1000, epochs=20, callbacks=[es, mc],  validation_data=(X_val, y_val)) #validation_split=0.2
```

활성화 함수	*Sigmoid*
Loss function	*Categorical crossentropy*
Optimizer	*RMSprop*
Batch size	*1000*

PROCESS 02: 모델 학습 - LSTM

	precision	recall	f1-score	support
0	0.86	0.92	0.89	36116
1	0.56	0.54	0.55	1386
2	0.79	0.78	0.78	12432
3	0.82	0.85	0.83	15572
4	0.69	0.54	0.61	6516
5	0.75	0.61	0.67	1008
accuracy			0.81	73030
macro avg	0.75	0.71	0.72	73030
weighted avg	0.80	0.81	0.80	73030

데이터 불균형 해소를 위해 **데이터를 증강**하여 학습시킨
모델 성능을 평가해 본 결과
Accuracy는 0.81, Macro F1 Score는 0.72으로
전체적으로 준수한 성능을 보임.

	precision	recall	f1-score	support
0	0.89	0.92	0.90	23546
1	0.00	0.00	0.00	1098
2	0.48	0.64	0.55	9578
3	0.66	0.79	0.72	11766
4	0.00	0.00	0.00	4333
5	0.00	0.00	0.00	654
accuracy			0.73	50975
macro avg	0.34	0.39	0.36	50975
weighted avg	0.65	0.73	0.69	50975

데이터를 **증강하지 않고** 학습을 진행한
모델 성능을 평가해 본 결과
Accuracy는 0.73, Macro F1 Score는 0.36으로
특정 클래스의 결과만 잘 측정하는 모델임.

PROCESS 02: 모델 학습 - Fasttext

sixtechnologylarge	abstractfullteaser
NT(나노기술)	▶ 전면적/선택적 나노표면 구현을 위한 나노 금형 인서트 제작기술 \n\n\n- 나...
NT(나노기술)	본 연구에서는 연구책임자가 기개발한 그래핀 제조법들을 이용하여 초고도 소수성과 우수...
BT(생명공학기술)	1. 두경부암 세포주 및 환자에서 PAK의 기능 및 역할 규명\n\n\n\n -...
BT(생명공학기술)	본 연구 과제의 최종 목표는 기능성 나노 프로브와 마이크로 패터닝 기술을 이용하여 ...

데이터의 형식을 FastText의
지도 학습에 적합하도록 수정해 줌.

label	sentence
__label__NT	전면적선택적 나노표면 구현을 위한 나노 금형 인서트 제작기술 나노 금형 인서트 설...
__label__NT	본 연구에서는 연구책임자가 기개발한 그래핀 제조법들을 이용하여 초고도 소수성과 우수...
__label__BT	두경부암 세포주 및 환자에서 pak의 기능 및 역할 규명 두경부암 세포주에서 pa...
__label__BT	본 연구 과제의 최종 목표는 기능성 나노 프로브와 마이크로 패터닝 기술을 이용하여 ...

FastText의 레이블은 __label__ 로 시작함.
데이터의 클래스 앞에 __label__ 을 붙여주어
학습이 가능한 형식으로 만들어 줌.

PROCESS 02: 모델 학습 - Fasttext

```
model = fasttext.train_supervised(input='train_dataset.txt',
                                  autotuneValidationFile='ntis_test.txt',
                                  autotuneMetric="f1:__label__CT")
```

Train 데이터로 지도 학습을 수행함.
Auto Tuner를 이용하여
하이퍼 파라미터를 자동 조정 해주었음.
선행분석에서 F1-Score가 가장 낮았던 CT의 F1-Score가
가장 높을 때의 모델을 선정하는 것으로 설정함.

	precision	recall	f1-score	support
BT	0.94	0.92	0.93	23594
CT	0.54	0.60	0.57	1127
ET	0.80	0.78	0.79	9622
IT	0.80	0.84	0.82	11820
NT	0.71	0.73	0.72	4379
ST	0.70	0.66	0.68	661
accuracy			0.85	51203
macro avg	0.75	0.76	0.75	51203
weighted avg	0.85	0.85	0.85	51203

데이터 불균형 해소를 위해 **데이터를 증강**하여 학습시킨
모델 성능을 평가해 본 결과
Accuracy는 0.85, Macro F1 Score는 0.75으로
전체적으로 준수한 성능을 보임.

PROCESS 02: 모델 학습 - Fasttext

```
model = fasttext.train_supervised(input='train_dataset.txt',
                                  autotuneValidationFile='ntis_test.txt',
                                  autotuneMetric="f1:__label__CT")
```

Train 데이터로 지도 학습을 수행함.
Auto Tuner를 이용하여
하이퍼 파라미터를 자동 조정 해주었음.
선행분석에서 F1-Score가 가장 낮았던 CT의 F1-Score가
가장 높을 때의 모델을 선정하는 것으로 설정함.

	precision	recall	f1-score	support
BT	0.95	0.96	0.96	36116
CT	0.81	0.57	0.67	1386
ET	0.86	0.85	0.85	12432
IT	0.85	0.90	0.87	15572
NT	0.88	0.81	0.84	6516
ST	0.87	0.68	0.77	1008
accuracy			0.90	73030
macro avg	0.87	0.80	0.83	73030
weighted avg	0.90	0.90	0.90	73030

데이터를 **증강하지 않고** 학습을 진행한
모델 성능을 평가해 본 결과
Accuracy는 0.90, Macro F1 Score는 0.83으로
증강한 데이터에서 학습했을 때보다 좋은 성능을 보임.

PROCESS 03: 성능 평가

	LSTM		**Fasttext**	
	증강 전	증강 후	**증강 전**	증강 후
Accuracy	0.73	0.81	**0.90**	0.85
Marco F1-Score	0.36	0.72	**0.83**	0.75

최종적으로 학습시킨 모든 모델의 성능을 평가해본 결과,
Accuracy와 Marco F1- Score에서 모두 가장 우수한 성능을 보인
증강하기 전 데이터로 학습한 **Fasttext모델**을 최종 모델로 선정함.

선택된 최종 모델을 특허 서지정보에 적용하여 **6T분류를 예측**해 본 결과
아래와 같이 **객관적**이고 **신뢰성** 있는 분류가 가능함을 확인 할 수 있었음.

특허1

이물질 제거 가능한 펌프를 구비하는 수처리 장치
이물질 제거가 가능한 오 • 폐수를 정화하기 위한 수처리 장치에 관한 특허

해당 특허의 IPC 분류	C02 물, 폐수, 하수 또는 오니의 처리 F04 액체용 용적형 기계 B01 파쇄, 분쇄 또는 미분쇄

→ 예측한 6T 분류 **ET**

IPC분류와 특허 초록의 내용을 고려했을 때
수질 오염 처리 및 재이용기술을 포함하는
ET에 잘 분류된 것을 확인할 수 있음.

특허2

스마트 기기를 이용한 근거리 무선 통신 기반 커피 프린팅 서비스 시스템
애플리케이션을 통해 서버에 프린팅 이미지를 송신하는 시스템에 관한 특허

해당 특허의 IPC 분류	G06 전산 A23 식품 또는 식료품 H04 전기통신기술

→ 예측한 6T 분류 **IT**

IPC분류와 특허 초록의 내용을 고려했을 때
전기통신기술을 포함하는
IT에 잘 분류된 것을 확인할 수 있음.

04

프로젝트 결과

프로젝트 결과

프로젝트 의의

특허명	six
차량 도어 충돌방지 장치	ET
복수의 전동기를 제어하기 위한 제어장치	ET
양변기용 쾌변유도형 시트	ET
콜레스테릭 액정을 포함하는 액정 캡슐 및 이의 제조방법	IT
차량에 구비된 차량 제어 장치 및 그의 제어 방법	IT
스마트 기기를 이용한 근거리 무선 통신 기반 커피 프린팅 서비스 시스템	IT
이물질 제거 가능한 펌프를 구비하는 수처리장치	ET
고이송 편면형 절삭 인서트 및 이를 장착한 절삭 공구	ST
송풍장치	ET
멸치 선별기	ET

기관에 기 존재했던 특허 데이터 중,
등록/공개된 특허 데이터를 모두 6T 기준으로
분류하여 **새로운 데이터**를 만듦

첫번째 프로젝트였던
'유사 특허 추천 서비스'의 과정에 적용하여
특허 추천 결과를 6T 기준으로
시각화하여 서비스함.

의의 1:
객관적 기준 제공

특허를 객관적인 증거 기반으로 분류하여
국가 연구개발 사업 성과 활용의 근거로 활용.
또한 미래유망기술 발굴 및 사업화에 도움.

의의 2:
성과 조사의 기준

관련 데이터의 연계를 통해
국가 연구개발 사업 진행 현황에 대한
효율적이고 체계적인 연구성과 조사 기준 제공

부록 참고문헌

딥 러닝을 이용한 자연어 처리 입문. (2022.10.25). URL: https://wikidocs.net/book/2155

Fasttext 공식 문서. (2023. 01. 20). URL: https://fasttext.cc/

감사합니다