연구개발보고서

1. 연구과제명

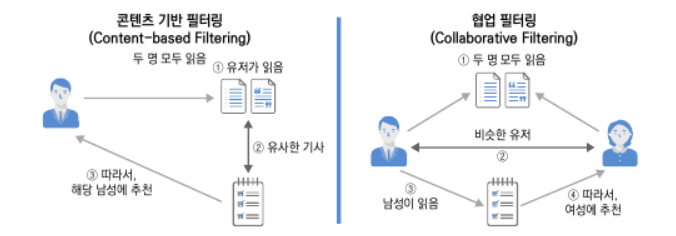
인공지능을 활용한 정부지원 R&D사업 추천 및 검색 시스템 개발에 대한 연구

1. 연구개발 과제 개요

* 연구 목표

R&D 지원사업은 기업의 성장과 도약을 뒷받침하는데 핵심적인 역할을 수행하고 있다. 실제로 세계 최고 수준의 체성분 측정기술을 보유한 ‘인바디(Inbody)’, 모바일 부동산 거래 플랫폼 ‘직방’과 같은 기업이 중기부 R&D를 통해 성장 도약하는 계기를 마련한 바 있다. 국내는 엔젤투자 환경이 발달하지 못한 탓에 창업 초기 대출에만 의존할 수밖에 없는 스타트업 자금 조달 환경을 가지고 있어 기업이 Seed Money를 확보할 수 있는 가장 좋은 방법은 R&D 정부지원사업에 참여하는 것이다. 국내 스타트업의 생존을 위해서는 R&D 지원사업 참여가 선택이 아닌 필수라는 걸 명심해야 한다. 이렇듯 기업의 R&D 지원사업 참여는 매우 중요하지만, 스타트업 또는 중소기업에 있어 공고 지원을 도와주는 서비스는 특이하게도 아직 존재하지 않는다. 이러한 필요성과 수요는 이번 연구의 목표를 뒷받침한다.

이번 연구 “인공지능을 활용한 정부지원 R&D사업 추천 및 검색 시스템 개발”에 사용할 기술인 인공지능(AI)은 학습, 문제 해결, 패턴 인식 등과 같이 주로 인간 지능과 연결된 인지 문제를 해결하는 데 주력하는 컴퓨터 공학 분야로써 4차 산업혁명의 중요한 한 부분으로 꼽히며 서비스, 비즈니스 등의 분야에 한정되지 않고 다양한 분야에 진출하고 있는 IT 기술이다. 그 중 이번 연구에 사용할 알고리즘은 바로 추천 알고리즘이다. 추천 알고리즘은 인공지능 모델로 사용자의 입력 값을 판단해 사용자가 최종적으로 보고 싶어하는 결과값을 미리 예측해 반환하는 형태의 알고리즘이다. 우리가 알고 있는 유튜브나 넷플릭스도 각 사용자의 시청 기록에 따라 다음 컨텐츠를 추천해주는 추천 알고리즘 기반 서비스다.

[]

[그림 1] 콘텐츠 기반 필터링과 협업 필터링 예제 (출처: Software carpentry)

추천 알고리즘에는 대표적으로 두가지 필터링 기법이 있는데 (필터링 기법: 추천 결과값을 도출하는 방법) 바로 “콘텐츠 기반 필터링 (Content-based Filtering)”과 “협업 필터링 (Collaborative Filtering)” 이다. 이 중 우리가 이번 연구에서 사용할 필터링 기법은 “콘텐츠 기반 필터링”인데, 채택의 이유는 인공지능 모델 개발 및 학습에 있어 다른 사용자들의 데이터 (예: 사용기록)가 필요한 협업 필터링 기법과는 다르게 콘텐츠 기반 필터링은 사용자의 입력 값 만을 분석해 유사한 문서를 찾아 필터링하는 기법이기 때문이다. 다양한 R&D 정부지원사업 공고를 머신 러닝(Machine Learning) 기법으로 인공지능 모델에 학습시켜 사용자가 입력한 키워드에 유사한 공고를 찾아내 보여준다면 사용자는 어떤 사업에 참여해야 할지, 또는 참고해야 할지, 혹은 다른 사업의 공고를 기다려야 할지 쉽고 빠르게 판단할 수 있게 될 것이고 이는 사용자 (중소기업 등) 에게 편리함을 가져다 줄 뿐만이 아니라 R&D 정부지원사업 참여도의 상승에도 영향을 미칠 것으로 생각된다.

* 주요 내용

연구 프로세스는 개념적으로 크게 4단계로 구분하여 진행하였으며, 각 단계를 통하여 정부지원사업에 참여해보고자 하는 사용자의 아이디어 (키워드)를 분석하여 축적된 R&D 정부지원사업 공고 데이터와 비교해 유사한 공고를 찾아 사용자에게 해당 공고의 링크를 제공할 수 있는 시스템 구축을 제안한다. 머신 러닝 기반의 학습 방법 예시와 자연어 처리 등의 데이터 가공법에 대한 사전 조사를 토대로 빠르고 가벼운 인공지능 모델 구현을 통해 사용자가 쉽고 간편하게 사업 추천을 받을 수 있는 방향으로 알고리즘을 구성하였다.

- 본 연구의 배경 및 필요성

- 연구 목적 및 범위

- R&D 지원사업 데이터 분석

- 인공지능 (머신러닝) 알고리즘 채택

- 데이터 가공 및 인공지능 모델 학습 (머신러닝 기법)

- 시범 운용을 위한 샘플 UI 구현

- 연구 적용 결과 확인

- 구현에 따른 효과 분석 및 향후 방향 제시

**1. 연구개발 목적 및 범위**

**2. 연구개발 이론적 고찰**

**3. 연구개발 수행**

**4. 연구개발 결과 분석**

[그림 2] 연구의 흐름도

인공지능(AI)은 사용자에게 편리하고 효율적인 서비스를 제공하고자 개발이 되고 있지만 언제나 그렇듯 항상 기술적인 문제를 마주한다. 인공지능 모델 개발을 위해 필요한 데이터를 보유하고 있다 하더라도 데이터의 양과 질, 목적에 알맞은 알고리즘을 채택해야 하며, 모델 학습 방법도 천차만별이라 효율적인 성능을 위해서는 충분한 사전조사와 지식 습득이 필요하다. 또한 중복된 데이터가 존재할 시 학습에 부정적인 영향을 주기 때문에 데이터를 가져올 사이트는 다양하지만 그 중 하나로 추려야 했다. 다행히도 각 사이트마다 공유되어 있는 공고가 비슷하기 때문에 사이트를 하나로 추려도 데이터의 양과 내용이 달라지진 않았다.

텍스트이(가) 표시된 사진

자동 생성된 설명

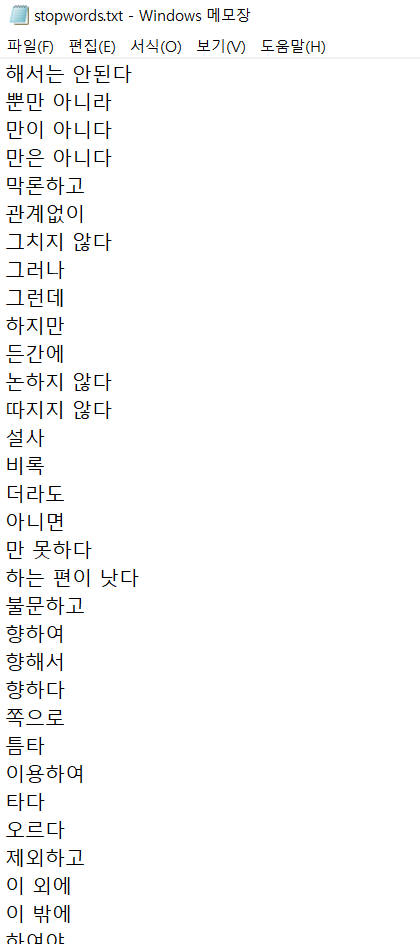
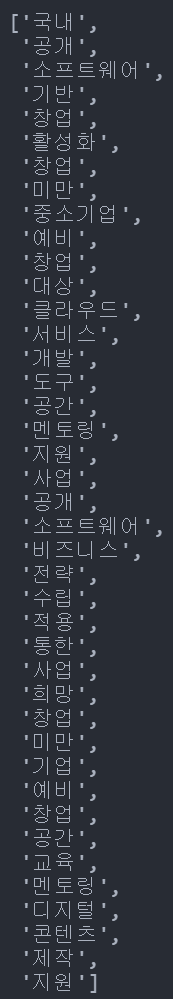
[그림 3] bizinfo 카테고리 선택과 다운로드 방법

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명

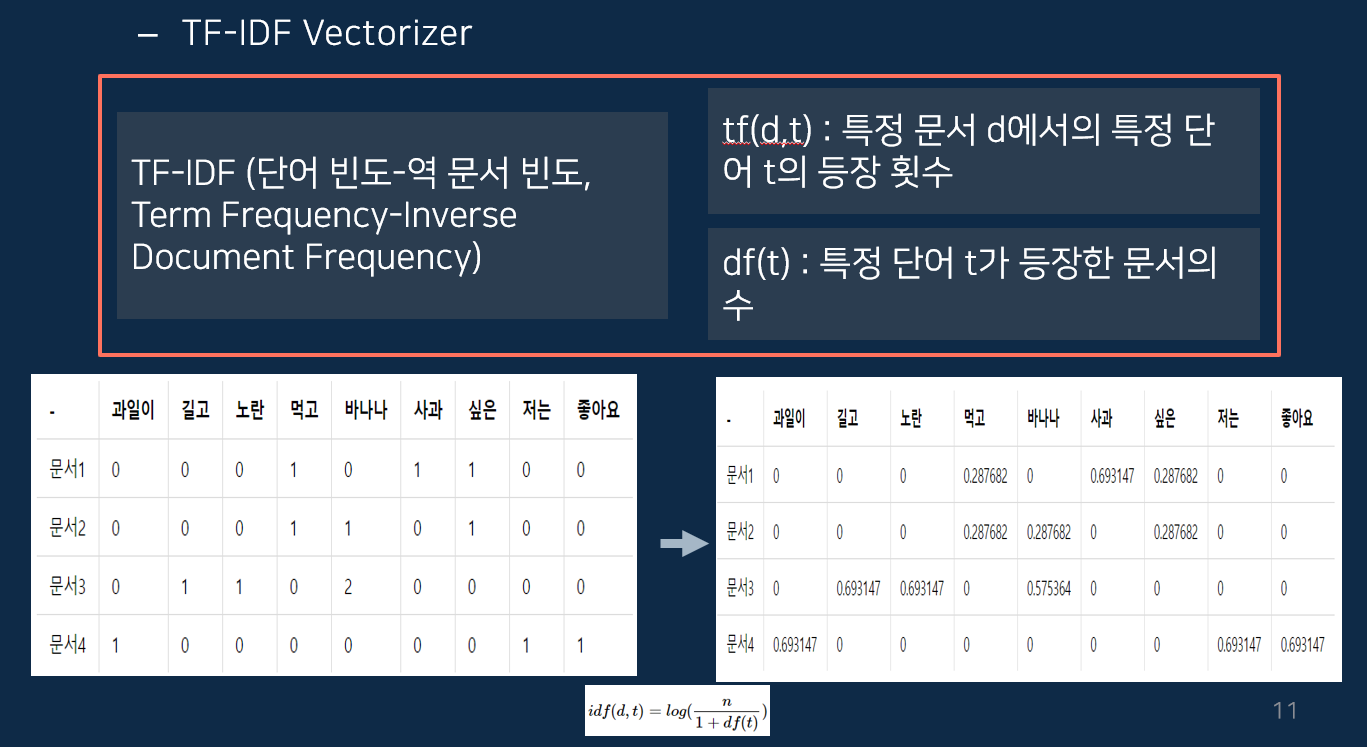
[그림 4] bizinfo 2020년~2021년 R&D정부지원사업 공고 (6642건)

모델 학습에 사용할 데이터는 기업마당 (bizinfo)에서 쉽게 구할 수 있었다. 엑셀(xlsx) 형식으로 다운로드도 제공하고 필요 시 무료로 API를 연동해 데이터를 가져오기에도 편했기 때문에 3가지 후보 사이트 중 기업마당을 선택해 약 1년간의 R&D정부지원사업 데이터를 다운로드하였다. 기술 분야, 소관부처, 신청시작일, 신청 마감일을 포함해 6642건의 데이터를 얻을 수 있었으며 인공지능 모델의 성능 개선을 위해 진행되고 있는 공고뿐만 아니라 마감된 공고까지도 다운로드 하였다. 연구 시작 단계에서는 데이터의 크롤링 (업무 자동화)을 고민하였었는데 기업마당에서 원하는 기간, 카테고리를 선택해 다운로드 할 수 있어 편리했고 업무 자동화 단계를 건너 뛸 수 있어 시간이 많이 절약되었다.

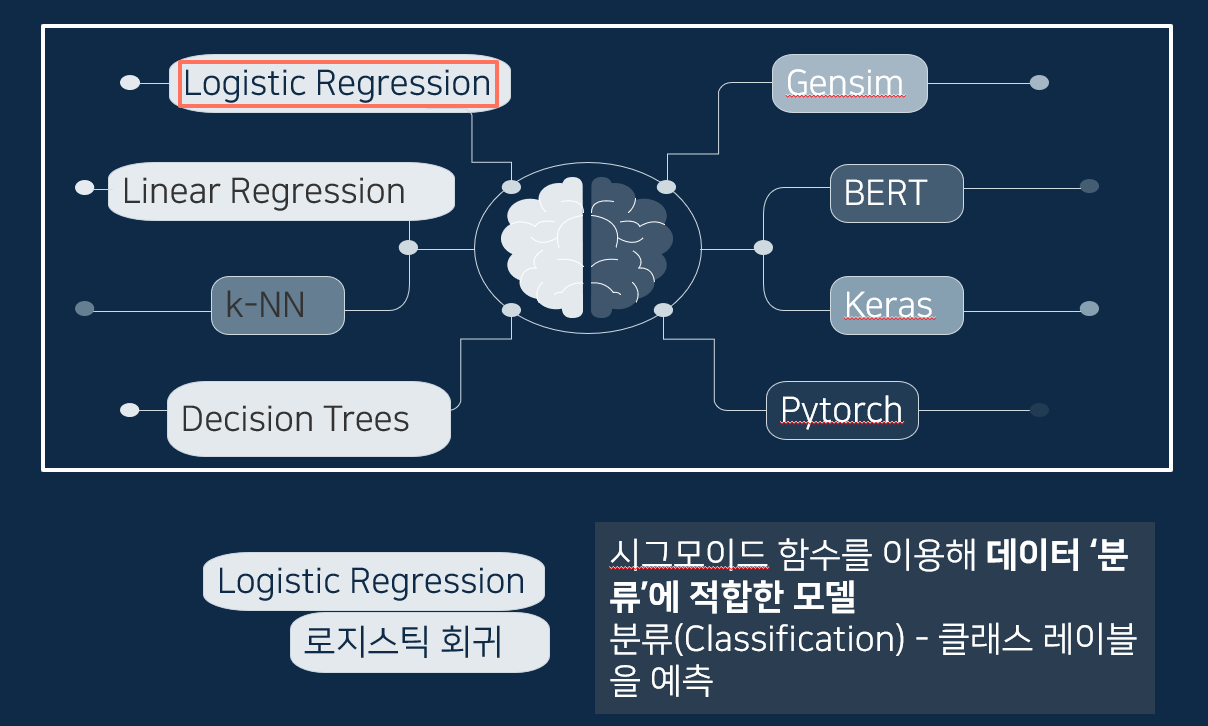
[그림 5] 불용어(Stopwords)를 사용한 데이터 가공 (전처리)

데이터 수집 후 첫번째로 진행한 작업은 자연어 전처리 작업이다. 키워드를 통해 문서 유사도를 측정하는 기법의 알고리즘을 채택했기에 R&D 정부지원사업 공고 데이터에 명사를 제외한 나머지의 부사, 동사 등의 단어들은 필요가 없었다. 가장 기초적인 전처리 기법인 불용어 제거 후의 단어 Tokenizing (토큰화)을 사용해 불용어를 우선 제거하고, 내용에서 필요한 단어들만 뽑아냈다. 불용어 사전 (txt파일)은 많이 사용되고 있는 <https://www.ranks.nl/stopwords/korean> 의 불용어 사전을 사용했다. 단어 토큰화 라이브러리 중 대표적으로 사용되는 “Okt” 와 “Kkma”를 한 번씩 적용해 테스트하여 육안으로도 성능이 좋아 보이는 “Okt”를 채택했다. [그림 5]와 같이 ‘소프트웨어’, ‘서비스’, ‘클라우드’ 등의 핵심 명사를 성공적으로 추출해낼 수 있었다.



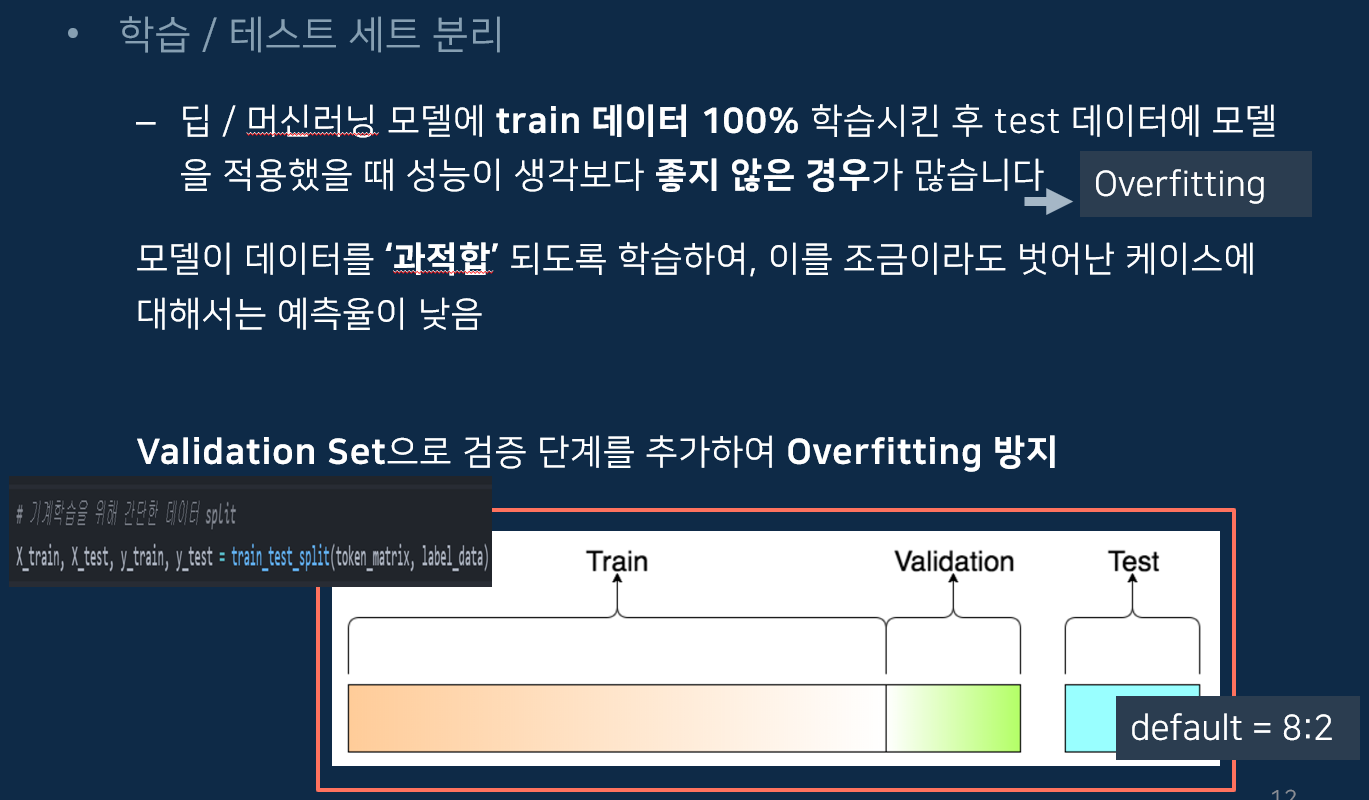
[그림 6] TF-IDF Vectorizer 알고리즘

데이터의 내용에서 명사 위주의 단어를 추출해 낸 후 진행한 작업은 문서 유사도 측정 작업이다. 단어의 유사 점수를 측정하는 라이브러리에는 대표적으로 TF-IDF Vectorizer와 Counter Vectorizer가 있는데, 이번 연구에서는 단어의 등장 횟수와 연관도에 따라 각 단어의 유사 점수를 측정해주는 TF-IDF Vectorizer를 사용했다. 이렇게 각 단어들의 유사도를 측정해 점수를 맥이고, 최종적으로 문서 간의 유사도를 검사하기 위해 데이터셋에 컬럼을 추가해 각 점수를 저장했다. R&D 정부지원사업 공고의 분야가 다양했기 때문에 TF-IDF Vectorizer 점수 또한 넓은 분포도를 보였다.



[그림 7] 인공지능 모델 채택 – Logistic Regression

앞서 진행한 데이터 정제 및 전처리를 통해 얻은 데이터셋을 확보한 후에는 인공지능 모델 학습을 진행해야 한다. 다만 무료로 사용할 수 있게 공개되어 있는 인공지능 모델은 한정적이고 목적이나 사용법에 따라 성능이 다 다르기 때문에 한 번씩 테스트를 거쳐 성능을 측정해보아야 이번 연구에 사용할 최적의 모델을 골라낼 수 있다. 약 6000건의 데이터는 딥러닝 기법의 학습을 진행하기엔 적은 양의 데이터라 머신 러닝 (Machine Learning) 기법을 사용하는 것으로 선택 폭을 줄인 후 scikit-learn 라이브러리의 Logistic Regression(로지스틱 회귀), Linear Regression(선형 회귀), Decision Trees(결정 트리) 알고리즘들을 후보군으로 추려 성능을 비교해보는 것으로 결정했다.



[그림 8] 인공지능 모델 예측 성능 테스트

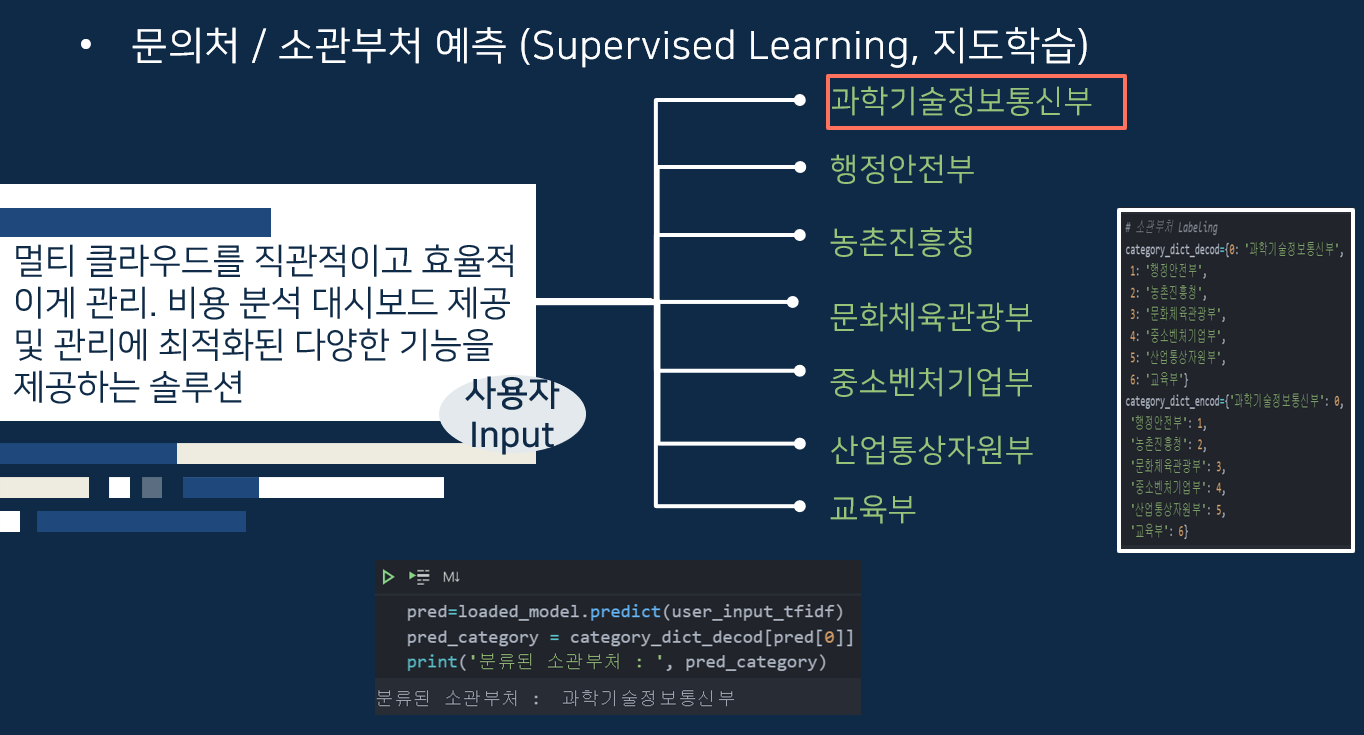
인공지능 모델을 학습하고 테스트할 때는 보통 데이터를 일정 양으로 분리해서 진행한다. 평균적으로는 8 (학습) : 2 (테스트) 의 비율로 Train/Validation 을 진행한다. 모든 데이터를 사용해 인공지능 모델 학습을 진행하면, 테스트해 볼 데이터를 제공하지 못해 성능이 떨어지게 되는 것이 데이터셋 분리의 이유다. 따라서 우리도 이번 연구에서 8:2의 비율을 지켜 각 알고리즘에 대해 예측 성능 테스트를 진행하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

[그림 9] 인공지능 모델 채택 및 파라미터 튜닝 – Logistic Regression

각 알고리즘 성능 테스트에서 Logistic Regression (로직스틱 회귀) 알고리즘의 예측율이 약 70%로 세 가지의 머신 러닝 알고리즘 중 1위를 차지했다. 최종적으로 로지스틱 회귀 알고리즘을 채택한 후 추가적인 성능 개선을 위해 하이퍼 파라미터 튜닝을 진행했다. 하이퍼 파라미터 튜닝은 scikit-learn 라이브러리에서 제공한 파라미터의 값을 데이터에 적합하게 수정하여 모델의 예측을 좁혀 성능을 개선시키는 기법이다. 우리가 준비한 데이터셋과 같은 경우에는 Max\_iter (학습 횟수), solver(세부 알고리즘 선택)의 값을 수정한 경우에 최적의 성능을 보였다.



[그림 10] 소관부처 예측 기능 추가

연구 및 개발을 진행하다 새로운 아이디어가 떠올라 기능을 추가하기로 결정했다. 사용자에게 추천 공고의 링크들을 제공하는 기능만 사용할 수 있는 것이 아니라, 사용자가 궁금증이나 문의할 사항이 있으면 쉽고 간편하게 어떤 부처에 문의를 해야 하는지 알 수 있게 소관부처 예측 기능을 추가하기로 했다. 해당 기능 추가는 단순했다. 데이터셋에는 ‘소관부처’ 컬럼이 존재해서 이미 7개의 소관부처가 각 공고마다 나뉘어져 있었다. 이를 각 숫자로 라벨링(0~6) 하여 한 번 더 인공지능 모델 학습을 시킨 후 테스트를 진행해본 결과 공고 링크를 예측했을 때보다 더 높은 예측율을 보였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

[그림 11] 사용자에게 입력 받은 데이터 정제

인공지능 모델의 학습 및 테스트가 끝난 후 마지막으로 사용자의 입력 데이터를 처리하는 기능을 개발하였다. 사용자가 키워드가 아닌 문장을 입력해도 정상적으로 예측 처리가 될 수 있게 사용자의 입력에도 공고 내용을 정제 및 전처리 했던 것처럼 자동 수정 기능을 추가해주었다. 따라서 사용자가 R&D 정부지원사업과 연관이 없는 키워드나 문장을 입력한다 해도 프로그램이 알아서 이를 적합하게 수정해준다. 사용자의 입력을 받은 후 자연어 전처리를 통해 핵심 키워드를 추출해 인공지능 모델을 통해 유사한 공고를 검색해 가장 유사한 TOP 10개의 지원 가능 공고를 보여준다. 여기에 추가로 지난 공고를 참고할 수 있게 이미 마감이 된 공고에서도 유사한 10개의 공고를 보여준다.

텍스트이(가) 표시된 사진

자동 생성된 설명

[그림 12] 사용자의 입력에 따른 공고 추천 결과

1. 연구개발수행 내용 및 결과

* 연구개발 수행 기간

수행기간 : 2021년 3월부터 6월까지 약 3개월

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 3월 | | | | 4월 | | | | 5월 | | | | | 6월 | | | |
|  | 1주 | 2주 | 3주 | 4주 | 5주 | 6주 | 7주 | 8주 | 9주 | 10주 | 11주 | 12주 | 13주 | 14주 | 15주 | 16주 | 17주 |
| 인공지능 AI 개발을 위한 필요 지식 학습 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| R&D 정부지원사업 데이터 분포와 공고 사이트 조사 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 모델 학습 및 테스트를 위한 라이브러리와 프로그램 설치 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 필요 데이터 수집 (R&D 정부지원사업 공고 데이터) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 인공지능 알고리즘 성능 테스트 및 선택 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 인공지능 모델 학습 및 개발 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 시연을 위한 샘플 UI/UX 개발 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 시연 및 테스트 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

[표 1] 과제 수행 WBS

* 연구개발 수행 인력

이정호 ([jhlee@didim365.com](mailto:jhlee@didim365.com)) 디딤365㈜ 기업부설 연구소 연구원

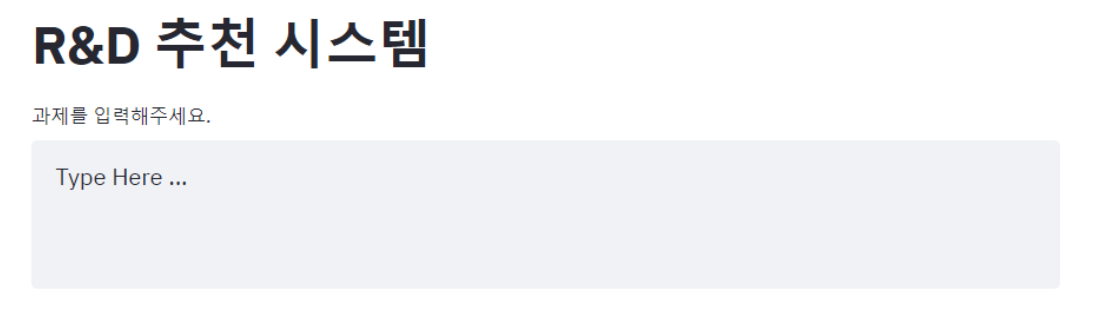
이남영 ([jkll@didim365.com](mailto:jkll@didim365.com)) 디딤365㈜ 기업부설 연구소 연구원

김신구 ([ksg@didim365.com](mailto:ksg@didim365.com)) 디딤365㈜ 기업부설 연구소 연구원

황미영 (popoo@didim365.com) 디딤365㈜ 기업부설 연구소 연구원

* 연구개발 주요 성과

본 연구는 중소기업이 쉽게 R&D 정부지원사업에 참여할 수 있도록 키워드를 통한 공고 검색 및 추천을 해주는 인공지능 모델의 개발을 중점으로 진행된 연구이다. R&D 지원사업 참여도의 향상을 도모할 수 있도록 구성되었다.



[그림 14] R&D 정부지원사업 추천 시스템 사용자 입력

테이블이(가) 표시된 사진

자동 생성된 설명

[그림 14] R&D 정부지원사업 추천 출력

R&D 정부지원사업 추천 시스템의 입력 란에 관심사, 문장 또는 키워드를 입력하면 실시간으로 프로그램이 동작해 사용자의 입력 값을 정제해 로컬에 저장되어 있는 인공지능 모델에 전송해 데이터셋과 사용자의 입력을 비교해 문서 유사도를 계산한다. 문서 유사도 측정이 완료되면 각 데이터 row에 점수가 저장되고 높은 순으로 정렬된다. 그 후 마감일이 지나지 않은 공고 10개와, 전체 공고의 10개를 UI로 전송해 사용자에게 보여준다. 링크 클릭 시 해당 공고로 이동할 수 있다.

해당 프로세스는 Intel i5-10400, 16GB RAM, GTX1650의 환경에서 동작하며 사용자가 결과를 보기까지 평균 3초의 시간이 소요된다. 고사양의 클라우드 컴퓨팅을 사용하면 평균 1초의 속도로 단축시킬 수도 있을 것이다. 키워드를 R&D 정부지원사업 추천 시스템이 아닌 구글이나 네이버에 검색했을 때의 결과보다 훨씬 만족스러운 추천 결과를 얻을 수 있었다.

* 연구의 한계 및 향후 연구방향

첫째, 인공지능 모델 학습에 사용할 데이터의 양이 약 6000건 정도로 적은 편에 속해 예측 성능 개선이 쉽지 않았다. 특히 진행 중인 공고는 마감된 공고보다 더 적어 공고를 추천해 준다고 하더라도 1~3개의 공고를 제외하고는 키워드와 전혀 다른 추천 결과값이 나오기도 했다.

둘째, 공고를 추천해주는 것만이 아닌, 지원 서류를 작성할 때 유리한 단어 선택 및 아이디어 표출을 도울 서비스가 구현되면 더 많은 사용자에게 인기있을 것이라고 생각한다. 공고를 추천해준다 하더라도 대부분의 사업 참여자들은 제대로 된 서류를 작성하지 못해 떨어지는 추세라고 한다.

그럼에도 불구하고 웬만한 키워드에 대해서는 만족스러운 R&D 정부지원사업 공고를 추천하는 등 유의미한 영향을 주는 것으로 나타났다. 이는 불필요한 검색 및 유선상의 문의를 해소하였고 중소기업의 아이디어 사업화를 돕는 기술로 자리 잡았음을 뜻한다. 연구 결과의 정확성을 확보하기 위해서는 더 많은 공고 데이터를 수집하는 등의 추후 연구를 진행하기 위한 노력을 기울일 필요가 있다.



1. 연구 증빙 자료

* 연구 예상 비용

TODO: 경영지원실에서 등록

- 전담부서의 조직·직원 현황 및 연구요원의 자격을 증명하는 서류

- 연구요원 등의 급여지급 명세서

- 연구개발업무에 사용하는 견본품·부품·원재료·시약류 구입 명세서

- 기타 연구･인력개발비 관련 명세

1. 참고문헌
2. 딥 러닝을 이용한 자연어 처리 입문

<https://wikidocs.net/book/2155>

1. News-Simpler 1nsidewill Github

https://github.com/1nsidewill/News-Simpler