



국민대학교
전자정보통신대학
컴퓨터공학부


캡스톤 디자인 I

종합설계 프로젝트

프로젝트 명	길 JOB 이
팀 명	4726
문서 제목	결과보고서


Version	2.0
Date	2019-05-28

팀원	고현경 (조장)
	김혜인
	김희주
	이수민
	이선흥

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

CONFIDENTIALITY/SECURITY WARNING


이 문서에 포함되어 있는 정보는 국민대학교 전자정보통신대학 컴퓨터공학부 및 컴퓨터공학부 개설 교과목 캡스톤 디자인 I 수강 학생 중 프로젝트 "길 JOB 이"를 수행하는 팀 "4726"의 팀원들의 자산입니다. 국민대학교 컴퓨터공학부 및 팀 "4726"의 팀원들의 서면 허락없이 사용되거나, 재가공 될 수 없습니다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

문서 정보 / 수정 내역

Filename	결과보고서-길 JOB 이.doc
원안작성자	고현경, 김혜인, 김희주, 이수민, 이선희
수정작업자	고현경, 김혜인, 김희주, 이수민, 이선희

수정날짜	대표수정 자	Revision	추가/수정 항목	내 용
2019-05-22	고현경	1.0	최초 작성	프로젝트 개요 및 목표 작성
2019-05-24	이선희	1.1	내용 추가	연구/개발 내용 및 결과물 작성
2019-05-25	김혜인	1.2	내용 추가	연구/개발 내용 및 결과물 작성 수정
2019-05-26	이수민	1.3	내용 추가	기대효과 및 활용방안 작성
2019-05-27	김희주	1.4	내용 추가	매뉴얼 작성
2019-05-28	전원	2.0	검토	최종 검토

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

목 차

1	개요	5
1.1	프로젝트 개요	5
1.2	추진 배경 및 필요성	5
1.2.1	추진 배경	5
1.2.2	취업 시장의 현황 및 문제점	6
1.2.3	개발된 시스템과의 차별성	7
2	개발 내용 및 결과물	8
2.1	목표	8
2.2	연구/개발 내용 및 결과물	9
2.2.1	웹 / 서버	9
2.2.1.1	Web Application Server	9
2.2.1.2	NLP Application Server	13
2.2.2	자연어처리	13
2.2.2.1	데이터 수집	13
2.2.2.2	데이터 전처리	13
2.2.2.3	데이터 벡터화	14
2.2.2.4	데이터 학습 모델 구현	14
2.2.2.5	데이터 학습 모델 테스트 결과	15
2.2.3	시스템 기능 요구사항	17
2.2.4	시스템 비기능(품질) 요구사항	18
2.2.5	시스템 구조 및 설계도	19
2.2.6	활용/개발된 기술	21
2.2.6.1	서버	21
2.2.6.2	자연어 처리	22
2.2.7	현실적 제한 요소 및 그 해결 방안	22
2.2.8	결과물 목록	23
2.3	기대효과 및 활용방안	25
3	자기평가	26
4	참고 문헌	28
5	부록	29
5.1	사용자 매뉴얼	29
5.2	배포 가이드	30
5.2.1	flask 서버 배포 가이드	30
5.2.2	springboot 서버 배포 가이드	30
5.3	테스트 케이스	31

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

1 개요

1.1 프로젝트 개요

본 프로젝트는 현재 과열되고 있는 취업 시장 속에서 자기소개서 작성으로 어려움을 겪는 취업 준비생들을 위한 머신러닝 기반 자기소개서 분석 서비스이다.

취업준비생들은 자기소개서 작성 시 키워드를 염두에 두고 작성하는 경우가 많다. 기업, 직무별로 선호하는 키워드, 즉 인재상이 다르기 때문이다. 따라서 수집한 합격 자기소개서들과 사용자의 자기소개서를 문서 분류 모델을 통해 10 가지의 역량(글로벌역량, 능동, 도전, 성실, 소통, 인내심, 정직, 주인의식, 창의, 팀워크)을 바탕으로 분석하여 사용자가 지원하고자 하는 직무와 기업과의 적합도를 제공한다.

이때, 다양한 문서 분류 모델을 구현해본 뒤 가장 적합한 모델을 선택하여 분석 결과의 정확도를 높이는 것을 목표로 한다.

1.2 추진배경 및 필요성

1.2.1 추진배경

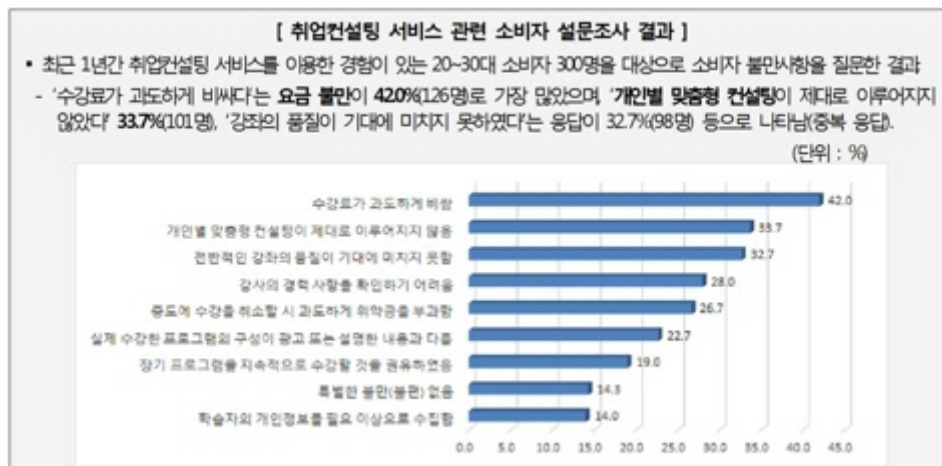


[그림 1] 에서 볼 수 있듯이, 인사담당자의 61%는 완벽한 스펙을 갖춘 지원자보다 완벽한 자기소개서를 작성한 지원자를 더 선호하는 것으로 나타났다. 이처럼 자기소개서 작성은 취업에 중요한 과정 중 하나이다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

신입사원을 대상으로 한 설문조사에 따르면 조사 대상 가운데 15.8%는 취업 준비 과정에서 비용을 들여 자기소개서 첨삭·컨설팅을 받은 경험이 있다고 밝혔다. 그러나 실제 합격에 도움이 됐다는 응답자는 이들 가운데 23.9%에 그쳤을 뿐만 아니라 컨설팅에 드는 비용이 만만치 않은 것으로 조사된다. 보통 취업 컨설팅 업체에서 한번 첨삭 시 최소 20 만원에서 80 만원까지 비용을 청구하며 지원자들이 취업이 될 때까지 첨삭에 들이는 비용은 평균 340 만원을 웃도는 것으로 드러났다. 이처럼 자기소개서 작성에 대한 고액 컨설팅 열풍이 지속되면서 첨삭 비용에 대한 부담도 점점 커지고 있다. 따라서 큰 비용을 들이지 않아도 자기소개서를 작성하는 데 도움을 줄 수 있는 서비스가 필요하다.

1.2.2 취업 시장의 현황 및 문제점



[그림 2] 취업컨설팅 서비스 관련 소비자 설문조사 결과

취업난이 계속되는 가운데, 자기소개서의 중요성 증대, 블라인드 평가 도입 등 기업별로 채용방식이 다변화되면서 취업준비생을 대상으로 한 '취업 컨설팅 서비스'가 인기를 끌고 있다. 그러나 [그림 2] 와 같이 제공되는 서비스에 비해 수강료가 과도하게 비싸다는 불만이 많고, 개인별 맞춤 컨설팅이 제대로 이루어지지 않는다는 의견이 많다. 또한 취업 컨설팅 서비스는 주로 사용자가 자기소개서와 스펙을 제출하면 취업 전문가들이 직접 분석 및 첨삭을 하는 방식으로 진행된다. 이러한 방식에는 불필요한 부분에도 사람의 주관적인 의견이 추가되거나, 첨삭자의 역량에 따라서 결과가 달라질 가능성이 있다.

이에 따라 AI 를 이용한 자기소개서 분석 서비스가 현재 시장에 출시되었다. 사람을 통해서 컨설팅하는 것보다 비용이 저렴하고, 언제나 일정한 수준의 품질로 객관적인 자기소개서 분석을 해 줄 수 있기

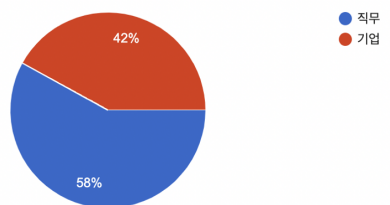
 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

때문이다.

기출시된 AI 자기소개서 서비스는 제출한 자기소개서가 해당 직무의 우수지원자와 얼마나 적합한지 비교하고 지원자의 성향과 강점을 분석한다. 하지만 직무 기준으로 보여주는 정보로는 각 기업이 요구하는 인재상에 얼마나 적합한지, 어떤 부분이 부족한지는 알 수 없다.

1.2.3 개발된 시스템과의 차별성

1. 자기소개서 분석 시 적합한 직무와 기업 중 어떤 것을 중점으로 분석받는 것을 선호하십니까?




[그림 3] 취업 준비생들을 대상으로 한 설문조사 결과

[그림 3]은 직접 취업 준비생들을 대상으로 설문조사를 실시한 결과이다. 위 결과를 통해 취업 준비생들의 관심사는 자신의 직무뿐 아니라 기업에도 초점이 맞추어져 있는 것을 알 수 있다.

또한 기업은 자기소개서를 통해 지원자의 '직무에 대한 이해도'만큼 '충성도', 즉 '안착률'을 중요하게 평가한다. 회사 차원에서는 근속기간이 길수록 신입사원 채용과 교육, 훈련에 쓴 비용(6000 만~1 억 2000 만원)을 회수하는 등 여러 방면에서 이익을 창출할 수 있기 때문이다. 그러므로 자기소개서는 직무뿐 아니라 기업에 대한 관심도와 충성도를 자기소개서에 표현하는 것도 중요하다.


이 프로젝트는 자기소개서가 지원할 직무에 맞게 작성됐는지 분석할 뿐만 아니라 기업별 합격 자기소개서 데이터들을 기반으로 자신이 작성한 자기소개서와 성향이 맞는 기업들을 추천한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2 개발 내용 및 결과물

2.1 목표

본 프로젝트는 머신러닝을 통해 자기소개서를 작성하는 취업 준비생들에게 객관적인 자기소개서 분석 서비스를 제공하는 것을 목표로 한다. 단순히 합격과 불합격을 판가름하는 것이 아니라 사용자의 강점과 약점을 분석하여 취업준비생들이 자기소개서를 작성하는 데 도움이 되는 것이 목적이다. 또한 누구나 비용, 장소의 부담 없이 자기소개서를 첨삭 받고, 자신을 객관적으로 돌아볼 수 있게 한다. 따라서 기존 AI 자기소개서 분석기보다 더 유용한 서비스를 제공할 수 있을 것이다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

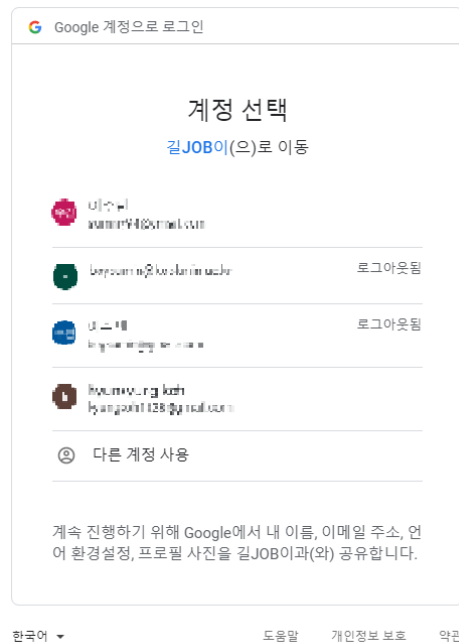
2.2 연구/개발 내용 및 결과물

2.2.1 웹 / 서버

2.2.1.1 Web Application Server

(1) 로그인

- 길 JOB 이 서비스의 로그인 기능은 Springboot, Spring Security OAuth2.0 + Google Single Sign On 을 통해 구현하였다. 사용자는 자신의 구글 계정으로 간편하게 로그인하여 회원으로 등록되고 자기소개서 분석 서비스를 이용할 수 있다.



[로그인 페이지]

(2) 자기소개서 제출 방식

- 사용자가 자기소개서 제출 시 자신이 지원하는 직무와 기업을 선택하고 문항별로 나누어 작성한 뒤 제출하기 버튼을 클릭하면 선택한 분류와 입력한 자기소개서가 POST 요청으로 전송된다. Web 서버는 데이터를 받아 객체를 생성한 뒤 Timestamp 를 추가해 DB 에 저장하고 NLP(자기소개서 분석) 서버로 Request 를 보낸 뒤 NLP 서버는 제출된 자기소개서를 분석한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

길JOB이

종 제출하기 비교하기 로그인 회원가입

제출하기

지원하는 직무 분야와 기업을 선택한 후 자기소개서를 입력하세요.

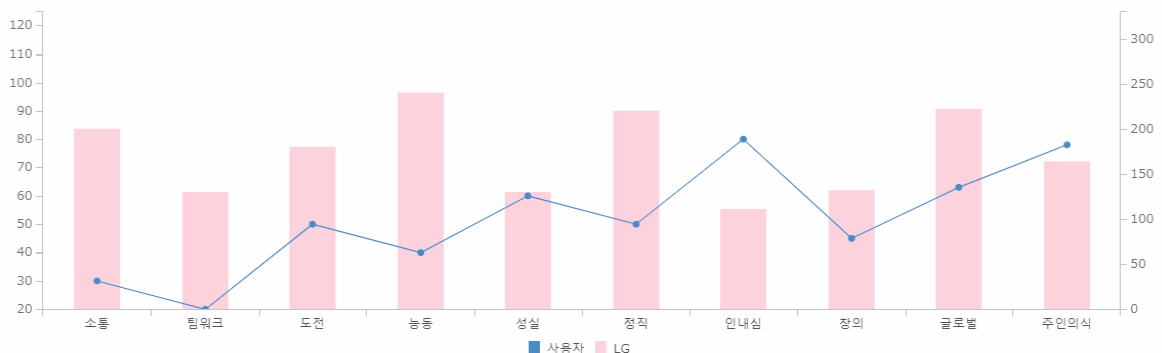
직무 기업

1번 문항

[제출하기 페이지]

(3) 자기소개서 분석 결과 매핑

- 결과페이지의 자기소개서 분석 결과 그래프를 생성할 때 결과 데이터를 매핑하는 방식은 스프링 내부의 구현체인 ModelAndView 형식과 thymeleaf 페이지방식으로 구현하였다. NLP 서버에서 분석 결과를 받으면 데이터를 결과페이지의 thymeleaf 파라미터 값에 매핑한 뒤 생성된 결과페이지를 보여준다. 그래프는 billboard.js 와 toast ui chart 를 수정하여 사용했다.

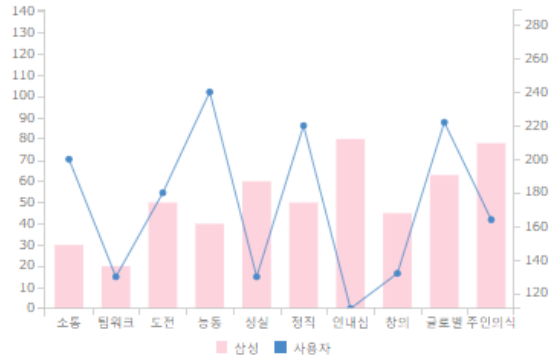
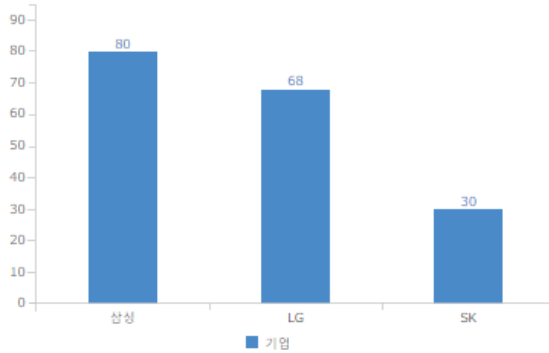


[분석결과 1] 선택 기업 핵심역량 그래프

사용자가 자기소개서 제출 시 선택한 기업과 사용자의 자기소개서가 핵심 역량 별로 얼마나 일치하는지 보여주는 그래프이다. 분홍색 Bar Chart 는 기업의 역량 수치를 보여주고 파란색 Line Chart 는 사용자의 핵심 역량 수치를 보여준다. 제출 페이지에서 선택한 기업의 핵심 역량 데이터를 데이터베이스에서 조회하고 사용자의 자기소개서 분석 결과 데이터를 그래프의 파라미터에 매핑하여 보여준다.

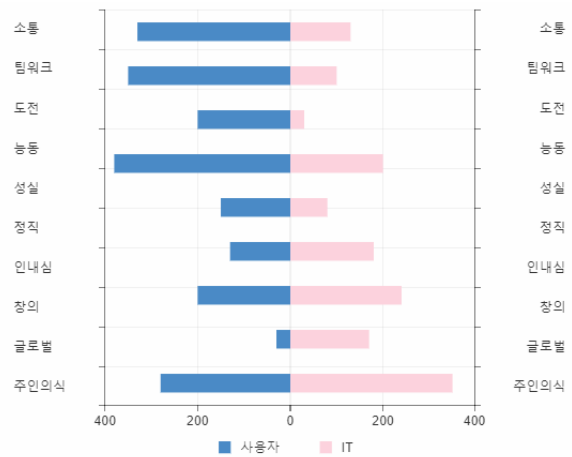
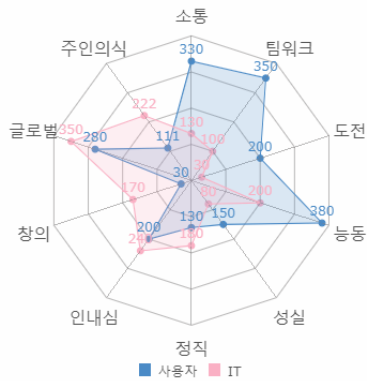
 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

[기업별 적합도 그래프]



[분석결과 2] 추천기업 순위 그래프(좌), 1 순위 기업 핵심역량 그래프(우)

사용자의 자기소개서 분석 결과, 기업 합격 자기소개서와 유사한 정도를 적합도로 나타내어 수치가 가장 높은 기업 3 개를 선정하여 보여준다. 오른쪽 그래프는 가장 적합도가 높은 기업의 핵심역량 수치를 보여주는 그래프이다



[분석결과 3] 선택 직무 핵심역량 그래프

좌측에는 사용자의 자기소개서와 사용자가 선택한 직무의 합격 자기소개서를 분석한 결과를 방사형 차트로 나타낸다. 방사형 차트에 핵심역량 수치를 그래프로 2 개의 그래프로 나타내어 비교 분석이 가능하도록 한다. 우측에는 사용자의 역량 수치가 우수한 순서로 보여주고 선택한 직무와 비교가 가능하도록 한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.2.1.2 NLP Application Server

(1) 자기소개서 분석

- 길 JOB 이의 제출 페이지에서 전송된 데이터를 분석하여 결과 데이터를 JSON 형식으로 반환해주는 API 를 제공하는 서버를 python 기반의 flask 를 사용해 구현하였다. NLP 서버 또한 AWS EC2 를 통해 구축하였다.
- Web Application Server 와의 통신은 REST API 방식으로 구현하였다.

2.2.2 자연어처리

2.2.2.1 데이터 수집

데이터 수집의 첫 번째 단계는 자기소개서의 성향에 대한 데이터 수집이다. 자기소개서를 분석하기 위해 소통, 팀워크, 인내심 등과 같이 사용자의 특성을 정의할 수 있는 10 가지 역량을 선정해 해당하는 어휘들을 포함하는 문장 데이터를 뉴스 기사, 웹 문서를 통해 수집하였다. 이 역량들은 구인 구직 사이트에서 기업의 인사 담당자들에게 자기소개서 핵심 키워드를 조사한 것을 바탕으로 결정한 것이다. 두 번째 단계는 자기소개서 수집이다. 사용자의 자기소개서와 비교할 데이터를 수집하기 위해 온라인으로 공개된 기업 합격 자기소개서를 수집했다. 웹 크롤러는 beautiful soup 과 selenium 을 사용하여 구현하였다.

2.2.2.2 데이터 전처리

데이터 전처리 단계는 수집된 데이터들을 분석할 수 있도록 정제하는 단계이다. 데이터 수집 단계에서 수집된 키워드를 포함한 원문 상태의 데이터들은 불필요한 어미, 조사 등을 포함하고 있어 분석에 영향을 미치기 때문에 꼭 필요하다. 전처리는 한글 토큰화 과정에 가장 많이 활용되는 Konlpy 라이브러리를 활용하였다. 수집한 원문 데이터는 첫 번째로 토큰화(tokenization)를 진행하여 데이터들을 품사별로 분리한다. 토큰화된 데이터에서 분석에 필요한 품사(명사, 동사) 위주로 추출하여 표현 방법이 다른 단어들을 통합시키는 정규화, 의미를 담고 있는 부분을 원형으로 바꿔주는 어근화 단계를 진행하였다. 그 후 분석에 큰 의미가 없는 불용어를 제거하여 데이터를 분석하기 위한 전처리 단계를 완료했다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.2.2.3 데이터 벡터화


데이터 벡터화는 수집한 텍스트 형태의 데이터를 분석에 알맞게 수치로 변경하는 단계이다. 이 프로젝트에서는 TF-IDF 방식을 선택해 단순히 단어 출현 빈도에 따라 중요도를 결정하는 것이 아니라 문서 내 단어 중요도를 의미하도록 했다.

Scikit-learn 의 서브 패키지 feature_extraction 에서 제공하는 TfidfVectorizer 를 이용하여 벡터화를 완료하였다. 이때, 벡터화 단계에서 파라미터를 어떻게 설정하냐에 따라 모델 구현 시 예측 점수 차이가 크게 났다. 많은 테스트 끝에 *n-gram_range 파라미터를 (1,3)으로 넘겨 한 단어에서 세 단어까지로 *window size 를 지정하고, *max_df = 0.95, *min_df = 0 으로 설정하는 것이 우리의 목적에 최적의 결과를 낸다고 판단했다.

- n-gram : 연속된 단어열은 하나의 단어보다 표현할 수 있는 정보가 뚜렷하다. 세 개 이상 연속된 단어를 n-gram 이라고 하는데, 표현력이 좋기 때문에 문서 분류의 base model 로 이용된다.
- window : 타겟 단어 앞뒤에 있는 단어들의 범위이다.
- max_df : 단어장에 포함되기 위한 최대 빈도. max_df = 0.95 는 문서의 95% 이상에 나타나는 단어를 무시한다는 뜻이다.
- min_df : 단어장에 포함되기 위한 최소 빈도. min_df = 0.01 은 문서의 1% 미만으로 나타나는 단어를 무시한다는 뜻이다. 우리는 min_df를 0으로 설정함으로써 어떤 용어도 무시하지 않았다.

2.2.2.4 데이터 학습 모델 구현

문서 분류(document classification) 모델을 구현하기 위해 지도 학습을 통해 기계 학습을 진행하였다. 문서분류는 라벨링 한 데이터를 학습시킨 후, 클래스를 예측한다. 현재 지도학습에는 다양한 알고리즘이 존재하는데, 대표적인 문서 분류 모델로는 SVM, Naive Bayes, Logistic Regression 등이 있다. 본 프로젝트에서는 각 알고리즘을 활용하여 여러 모델을 구현한 뒤에 예측점수를 비교했다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.2.2.5 데이터 학습 모델 테스트 결과

2.2.2.4 에 언급한 문서 분류 모델들을 구현 한 뒤 예측 점수를 계산한 결과이다.

	SGDClassifier	Naive Bayes	LinearSVC & OneVsRestClassifier	Logistic Regression & OneVsRestClassifier
train	0.9845416506288879	0.9836476737977874	0.9958405244664076	0.9788549646754988
test	0.9155007688623358	0.8971993244094885	0.9409866646499786	0.9309032241800902


LinearSVC 모델의 성능이 가장 적합하다고 판단하였다. 모델의 종류를 확정 한 후, 파라미터값과 classifier 를 변경해가며 모델을 테스트했다.

1) 직무

입력 데이터의 각 클래스에 대해 직무별, 기업별로 구한 확률값으로 사용자의 자기소개서와 비교해야 하기 때문에 predict_proba 함수가 있는 모델만을 선택해 테스트했다. 앞서 테스트한 LinearSVC 모델은 decision_function 만 존재할 뿐, predict_proba 함수는 없었기 때문에 Classifier 를 변경한 모델과 SVC 모델을 테스트했다.

여러 모델을 테스트한 뒤, LinearSVC(ver.CalibratedClassifierCV, cv=10)모델로 프로젝트를 진행하기로 결정했다.

	SVC(gamma='auto')	LinearSVC(ver.CalibratedClassifierCV, cv=10)	LinearSVC(ver.CalibratedClassifierCV, cv=5)
train	0.8952557146227293	0.9956542792932617	0.9953314543264754
test	0.8639239708588571	0.9383145529254582	0.9361214046232574


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2) 기업

기업 모델 또한 SVC 알고리즘으로 진행하였다. 순수한 자기소개서만을 가지고 예측할 수 있도록 데이터에서 문항이나 기업 이름 등을 제거한 후 모델을 만들었다. 의미가 없지만 자주 등장해서 예측에 영향을 끼칠 수 있기 때문이다.

학습 결과 SVC(kernel='Linear', gamma='scale')의 예측 수치가 높게 나와 기업 분류 모델로 선택했다.

	SVC(gamma='auto')	LinearSVC(ver.CalibratedClassifier CV, cv=10)	LinearSVC(ver.CalibratedClassifier CV, cv=5)	SVC(kernel = 'Linear', gamma = 'scale')
train	0.33144535347178516	1.0	1.0	1.0
test	0.34071550255536626	0.6712095400340715	0.6707836456558773	0.7001703577512777

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.2.3 시스템 기능 요구사항

대분류	소분류	기능	완료여부
Web Service	로그인	사용자는 구글 계정을 통해 간편하게 로그인하여 서비스를 이용할 수 있다.	완료
	자기소개서 입력	분석을 원하는 기업과 직무를 선택한 후 사용자는 자기소개서를 입력한다. ->사용자는 자기소개서를 문항별로 질문과 답변을 각각 입력한다.	변경 완료
	분석 결과 확인	사용자는 기업별, 직무별 그래프를 통해 자기소개서 분석 결과를 확인할 수 있다. -> 사용자는 결과 페이지에서 자기소개서 분석 결과를 확인할 수 있다. 결과 페이지에서는 4 가지의 그래프를 통해 직무별, 기업별, 추천 기업별, 문항별 분석 결과를 확인할 수 있다.	변경 완료
	이전 결과 비교	자기소개서를 분석했던 결과를 선택하여 비교페이지에서 비교 결과를 확인할 수 있다.	완료
NLP	자기소개서의 핵심역량 분석	핵심 역량 분류 모델을 통하여 사용자의 자기소개서와 합격 자기소개서를 분석한다.	완료
	사용자 자기소개서와 기업/직무 적합도 비교	분석된 사용자 자기소개서의 핵심역량과 기업/직무별 핵심역량 수치를 비교한다. -> 합격 자기소개서로 학습시킨 기업 분류 모델을 통하여 사용자의 자기소개서를 분석한다.	변경 완료

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.2.4 시스템 비기능(품질) 요구사항

1. 데이터 분석의 정확성 : 모델 구현이 완료된 후, 테스트 데이터 셋을 이용하여 예측 정확도를 측정한다. 테스트 데이터 셋의 예측 점수가 80 점 이하일 경우 신뢰성이 낮다고 판단한다.

[달성]

2. 처리 속도

- 1) 웹페이지 로딩 : 웹 페이지 로딩 속도를 평균 로딩 속도인 5.30 초 이내가 되도록 한다.

[달성]

- 2) 자기소개서 분석 : 자기소개서를 제출한 후 결과 페이지 로딩이 20 초 이내가 되도록 한다. [달성]

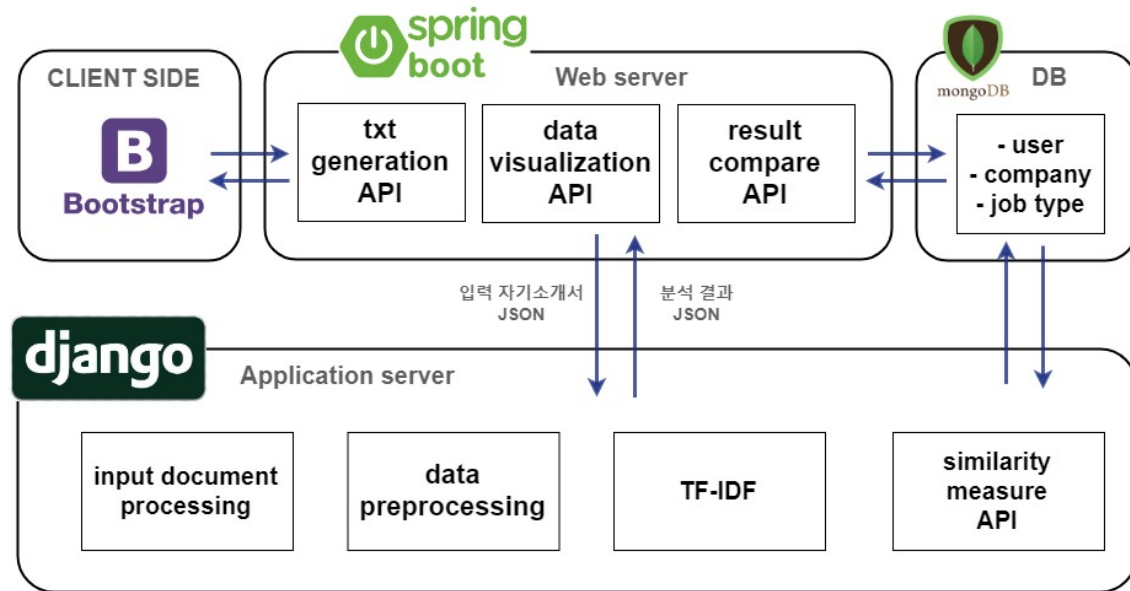
3. 사용자 편리성

- 1) 사용자가 직관적으로 웹페이지를 이용할 수 있도록 UI/UX 측면에서 웹 페이지를 디자인한다. [달성]

- 2) 사용자가 시스템에 제출한 자기소개서가 온라인에 공개되지 않게 설정할 수 있도록 한다. 사용자가 비공개로 설정하면 자기소개서 분석 내용은 공개되지 않는다. [달성]

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

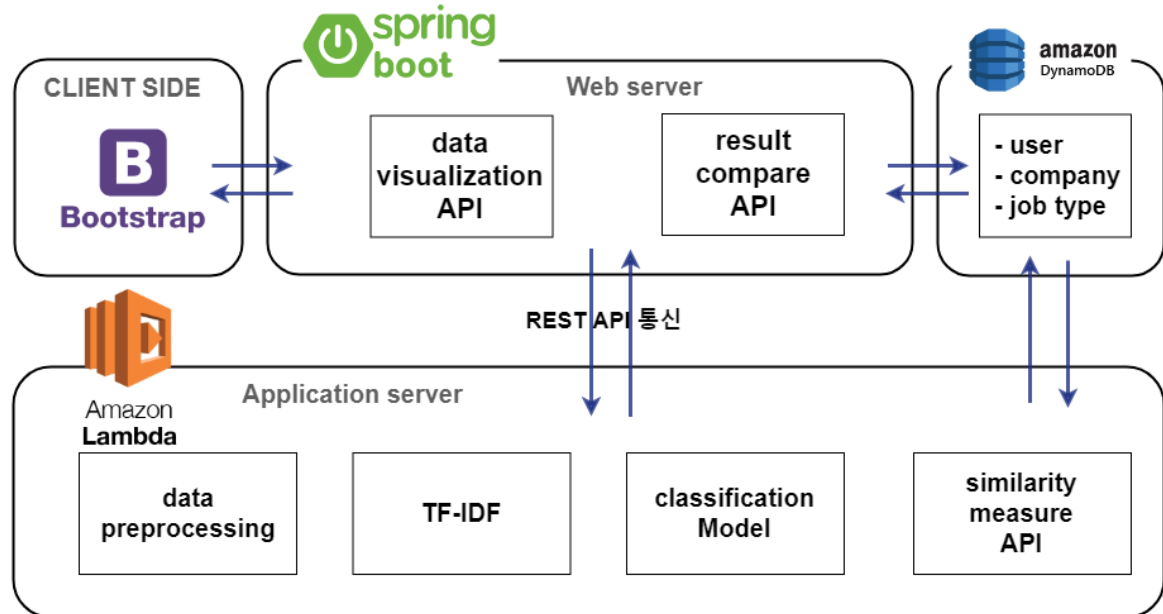
2.2.5 시스템 구조 및 설계도



[수행계획서에서 제시한 시스템 구조도]

최초 계획 작성 시 웹 서버 구축은 개발 환경 설정과 구현이 편한 스프링부트를 사용하고 NLP 서버는 자연어처리 python 모델을 server 할 수 있는 python 기반의 django 서버로 구현할 예정이었다. nosql 기반의 데이터베이스를 사용하면 간단한 설정과 쿼리를 활용할 수 있게 되어 mongodb 를 사용할 예정이었다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28



[최종 시스템 구조도]

길 JOB 이는 웹 서버와 자기소개서 분석을 처리하는 어플리케이션 서버를 통해 서비스를 제공한다. 사용자가 접속하는 웹 페이지의 뷰는 디자인 템플릿 Bootstrap 을 사용하여 디자인하였다. Bootstrap 을 사용하여 반응형 웹으로 구현되어 있어 서로 다른 환경에서 서비스를 사용하는 다양한 사용자에게 대응할 수 있도록 구현하였다.

계획서 작성 후 연구 과정에서 txt generation API 기능은 구현이 어려울 뿐만 아니라 명확한 결과물을 얻을 수 없을 것으로 예상되어 제외하였다.

웹 서비스는 AWS EC2 를 통해 배포하기 때문에 AWS DynamoDB 를 활용하면 데이터 관리가 용이할 것으로 판단되어 mongoDB 에서 DynamoDB 를 사용하도록 변경하게 되었다.

NLP Application Server 는 최초 계획 시 Django 를 사용하여 구현할 예정이었으나 서버 관련 연구를 진행 중 AWS Lambda 를 사용하면 서버 리스로 간단하게 분석 API 를 제공할 수 있을 거라 판단되어 계획을 변경하였다. 하지만 AWS Lambda 서비스가 제공하는 모델은 최대 250mb 까지만 가능하여 Flask 로 변경하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

NLP Application 에서는 최초 수행계획에서 데이터를 TF-IDF 로 벡터화 과정을 거친 결과를 벡터 간의 비교를 통해 문서를 분석할 예정이었다. 하지만 TF-IDF 값의 단순한 비교는 자기소개서에 대한 분석 결과에 대한 정확도를 측정하기 어렵다고 판단하여 딥러닝 알고리즘을 추가하기로 하였다. Classification Model 알고리즘을 추가함으로써 데이터를 보다 분석적으로 학습하고 예측하는 모델을 만들 수 있었다. 이 학습 모델을 통하여 입력된 텍스트가 어떠한 클래스에 속하는지 분류를 할 수 있게 되었다.

2.2.6 활용/개발된 기술

2.2.6.1 서버

1. Front-end

디자인의 시간 단축을 위하여 bootstrap templates 을 이용하여 view 를 작성하였고, Thymeleaf 와 JQuery Ajax 를 이용하여 server 와 client 를 연동하였다.

결과 페이지에서 보이는 차트의 경우 billboard.js 와 toast ui chart 를 이용해 line chart, bar chart, pie chart 를 제공하였다.

제출페이지에서 select-box 를 이용해 직무와 기업 분류를 선택할 수 있도록 하였고, 문항별로 자기소개서를 제출할 때 text-area 를 사용하였다. form 태그를 이용하여 Data 를 post request 로 전송하였다. post request 를 처리하는 controller 는 model 객체에 data 를 저장한 후 결과 페이지를 생성하여 사용자에게 보여준다.


2. Back-end

Spring boot 프레임 워크를 기반으로 Aws 와 spring embeded tomcat 을 이용하여 Web Server 환경을 구축하였고, Flask 를 활용하여 NLP server 를 구축하였다.

NLP server 와 Web server 는 모두 AWS EC2 를 이용하여 배포하였다.

데이터베이스는 AWS 에서 제공하는 DynamoDB 를 사용하여 데이터베이스 설계와 데이터 관리를 용이하게 하였다.

NLP Server 와 Web Serve 간의 통신은 REST API 를 이용하여 연동을 완성하였다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.2.6.2 자연어 처리

1. Data Crawling

beautiful soup 과 selenium 을 이용하여 데이터 수집을 진행하였다. beautiful soup 은 사용자의 행동을 특정해서 데이터를 가져올 때 빠르게 처리할 수 있다는 장점이 있고 selenium 은 브라우저를 직접 동작시키기 때문에 그보다는 느리지만 사용자의 행동을 동적으로 처리할 수 있다.

2. Machine Learning

Python 기반의 Scikit-Learn 라이브러리를 통해 데이터 학습을 시켜 문서 분류 모델을 구현하였다. 문서 분류 알고리즘은 SVC 를 사용하였고, 다중 클래스 분류는 one-vs-rest 방식을 사용해 각 클래스가 얻은 조건부 확률값을 더해서 이 값이 가장 큰 클래스를 선택하도록 했다. 이때, Pipeline 함수로 전처리와 벡터화, 문서 분류 알고리즘을 한꺼번에 묶은 후 예측을 진행했다. Pipeline 은 기계 학습 알고리즘을 적용하기 전에 전처리와 알고리즘이 수행하는 부분을 한 번에 연결해서 구현할 수 있도록 한다.

2.2.7 현실적 제한 요소 및 그 해결 방안

1) 기업 직무 개수

현재 길 JOB 이의 취업준비생들이 가장 많이 지원하는 기업과 직무 각 5 개만을 분석한다. 현재 구할 수 있는 자기소개서의 양이 한정돼있기 때문에 이렇게 정한 것인데, 이 문제는 기업과 연계하여 자기소개서를 수집하거나 적은 양의 데이터로 좋은 성능을 낼 수 있도록 모델을 개선하면 해결할 수 있을 것이다.


2) 딥러닝 클라우드 실행 속도

개인 데스크탑보다 우수한 성능으로 자기소개서 분석모델을 학습시키기 위해 학교에서 제공하는 딥러닝 프라이빗 클라우드를 이용했다. 클라우드를 이용한 뒤 개인 컴퓨터를 꺼도 실행이 되어 편리한 부분도 있었지만 속도 면에서 개인 컴퓨터보다 더 떨어졌다. 따라서 다양한 모델을 만들고 테스트하는 데에는 시간적인 면에서 불편했다. 학교에서 제공한 딥러닝 클라우드보다 더 나은 성능을 가진 구글 클라우드를 사용하면 해결될 것으로 보인다.


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.2.8 결과물 목록

대분류	소분류	기능	진행 현황
서버	메인 페이지	길 JOB 이 서비스의 메인 페이지	완료
	로그인 페이지	가입 완료 후 인증된 계정을 통해 로그인을 진행하는 페이지	완료
	제출 페이지	클라이언트가 지원하는 직무, 기업을 선택하고 자기소개서를 작성하여 제출하는 페이지	완료
	로딩 페이지	입력한 자기소개서의 분석 진행 정도를 보여주는 페이지	완료
	결과 페이지	클라이언트의 자기소개서가 선택한 직무, 기업에 얼마나 적합한지를 보여주는 적합도 그래프와 성향이 맞는 기업을 추천해 보여주는 페이지	완료
	비교 페이지	자기소개서 분석한 결과 중 같은 직무, 기업을 선택한 기록 2 개를 선택하는 페이지	완료
	비교 결과 페이지	비교 페이지에서 선택한 기록 간의 비교 결과를 비교 차트를 통해 보여준다.	완료
	웹 서버 구축	AWS EC2 를 통해 웹 서버 구축	완료
	회원가입 기능	Spring boot, Spring security, OAuth 를 통해 회원 시스템을 구축하여 클라이언트가 name, email, pw 를 입력하여 회원 가입할 수 있다.	중단 (SSO 로 대체)
	SSO 인증 기능	구글 계정을 가진 사용자가 간단한 인증을 통해 가입 및 로그인하도록 만든다.	완료
	페이지 이동 기능	페이지 간 버튼을 통해 이동하는 기능	완료

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

	로그인 기능	계정정보를 입력하여 로그인하거나 인증된 구글 계정을 통해 로그인 가능	완료
	자기소개서 제출기능	제출 페이지에서 선택한 직무, 기업 정보와 입력한 자기소개서 텍스트 데이터를 제출하면 자연어처리 서버로 전송	완료
자연어처리	데이터 수집	웹크롤러를 이용한 데이터 수집	완료
	데이터 전처리	원문 데이터로부터 분석을 위한 형태로 전처리	완료
	데이터 벡터화	사이킷런 라이브러리를 이용한 데이터 벡터화	완료
	사용자 데이터 분석	학습된 데이터와 입력된 사용자 데이터를 분석	완료
	프로그램 신뢰성 테스트	테스트 데이터셋을 통하여 분석 결과의 신뢰성 체크	완료


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

2.3 기대효과 및 활용방안

구직자들에게 자기소개를 작성하는 데 있어서 가장 큰 도움이 된 요소는 '취업포털·구직커뮤니티의 취업 자료', '취업 선배나 지인들의 합격 자기소개서', '취업 사교육' 등의 순이었다. 취업 사교육이나 특강 같은 경우에는 서울이나 수도권에 거주하지 않으면 참여하기 어렵고, 취업 선배나 지인들의 합격 자기소개서는 주변 지인이 없다면 얻기 힘든 자료다. 본 프로젝트의 결과물이 사용자가 자기소개를 작성할 때 외부 요인에 영향받지 않고 자신의 역량을 최대한 펼칠 수 있도록 도움을 줄 것을 기대한다.

또한 누구나 부담 없이 자기소개를 분석 받고 그 과정을 통해 사용자 자신의 강점이나 약점을 파악할 수 있는 프로젝트가 될 것이다. 특히 일반적인 컨설팅은 자신이 작성한 자기소개를 멘토에게 공개해야 하므로, 그러한 사항이 불편했던 사용자는 직접적인 대면 없이 머신러닝을 통해 분석하는 본 프로젝트가 많은 도움이 될 것이라고 기대된다.

자연어처리를 통해 자기소개를 분석하는 본 프로젝트는 다양한 방면으로 활용, 발전될 수 있다. 기업의 인사담당자가 지원자의 자기소개를 직접 읽어보지 않고 미리 설정한 키워드를 바탕으로 역량을 평가할 수 있는 서비스로 발전할 수 있으며, 더 나아가 기업관계자가 사용자들이 작성한 자기소개서 중 기업 성향에 맞는 자기소개를 골라 열람할 수 있도록 하여 하나의 리크루팅 시스템을 형성할 수 있다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28


3 자기평가

프로젝트 주제를 선정하면서 제일 중요하게 생각한 것은 실생활에서 사용 가능한 서비스를 제공하는 것과 쉽게 접해보지 못한 머신러닝 기술을 사용해 보는 것이었다. 따라서 현재 팀원들에게 가장 현실적으로 다가오는 취업 문제에 도움이 될 수 있는 자기소개서 분석을 프로젝트 주제로 선정했다. 더불어, 자료조사를 하다 알게 된 100 만명에 달하는 취업 준비생들과 그중 취업을 하기 위해 고액을 사용하여 컨설팅을 받는다는 사회문제 역시 길 JOB 이 프로젝트를 통해 해결하고자 했다.

웹서버는 기업에서 현재 많이 사용하는 것으로 알려진 Spring boot 을 활용했다. 웹서버 구현에는 큰 어려움이 없었으나, 자연어 처리 서버를 구현하는데 많은 변경사항이 있었다. 초반에는 AWS 에서 제공하는 Lambda API 를 활용하려고 했지만 용량이 250mb 로 제한되어 우리 프로젝트에서 사용하기에는 적합하지 않았다. 결국 flask 로 서버를 변경해 자연어 처리 서버를 구현하게 되었다. 서버를 구현하더라도 머신러닝에 대한 배경지식을 쌓고 구현을 시작하였더라면 시스템 구조를 여러 번 변경하는 일이 없지 않았을까 하는 아쉬움이 남는다. 다음에 프로젝트를 진행한다면 이러한 경험을 바탕으로 프로젝트에 대한 전반적 이해를 한 뒤 진행을 해야겠다는 교훈을 얻었다.


자연어 처리라는 다소 생소한 프로젝트를 진행하면서 많은 점을 배울 수 있었다. 인공지능 분야에서 한국어 자연어처리가 어려운 이유로 띄어쓰기 문법이 어려운 점, 구어와 문어의 차이가 큰 점, 청자와 화자의 관계에 따른 높임법이 다양한 점, 주어, 목적어의 빈번한 생략이 발생한다는 점을 꼽는다. 따라서 한국어 자연어처리는 현재도 연구가 진행 중이므로 자료도 적고, 우리가 원하는 비슷한 유형의 자료가 있어도 다 영어라서 적용하기가 쉽지 않았다. 이러한 이유로 프로젝트 구현을 위한 스터디를 진행하는데도 어려움이 있었다. 프로젝트 계획이 변경되는 경우가 빈번하여 시간을 효율적으로 사용하지 못하였다. 추후에 프로젝트를 진행할 때에는 자료조사와 더 체계적인 스터디를 계획하여 프로젝트를 진행해야겠다고 느꼈다.

자연어 처리는 주로 scikit-learn 과 konlpy 라이브러리를 활용하여 자기소개서 분석에 맞는 결과를 얻기 위해 여러 번 테스트 과정을 거쳤다. 결과적으로 성공적인 모델을 개발했지만, 딥러닝 클라우드를 이용했음에도 한 모델을 학습시키는 데 최소 3 시간 이상이 걸리는 등 시간이 너무 많이 걸린다는 문제점이 있었다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28


개발하면서 처음 예상과는 다르게 자연어 처리 부분뿐만 아니라 서버에서 돌릴 코드를 작성하는데에도 신경 써야 했다. 기본적인 변수명부터, 리턴타입까지 세세하게 주의를 기울였음에도 서로 얘기를 나뉘가며 많은 수정이 필요했다.

우리는 프로젝트를 진행하며 약 2 만건의 합격 자기소개서를 기반으로 학습된 분석 모델로 사용자의 자기소개서를 분석해주는 길 JOB 이 서비스를 개발하였다. 길 JOB 이 서비스는 자기소개서의 기업별, 직무별, 문항별 핵심역량을 분석하여 알려준다. 사용자는 길 JOB 이가 제시하는 객관적인 데이터를 활용하여 자신의 자기소개서가 어떤 역량을 가졌는지 확인할 수 있고 미리 학습된 분류별 결과와 비교하여 자기소개서를 작성하는 데 도움 받을 수 있다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

4 참고 문헌

번호	종류	제목	출처	발행년 도	저자	기타
1	웹페이지	Natural Language Processing with PyTorch	https://kh-kim.gitbook.io/natural-language-processing-with-pytorch/00-cover-4/06-vectorization			
2	웹페이지	자연어 (NLP) 처리 기초 정리	http://hero4earth.com/blog/learning/2018/01/17/NLP_Basics_01/			
3	웹페이지	한국어 형태소 분석기 - 강승식교수연구실	http://nlp.kookmin.ac.kr/			
4	서적	(텐서플로와 머신러닝으로 시작하는) 자연어 처리		2019	전창욱, 최태균, 조중현	
5	웹페이지	스프링부트 EC2 배포하기	https://jojoldu.tistory.com/263			
6	웹페이지	Flask 와 EC2 를 이용해서 Dynamo DB 와 연동하는 API Gateway 구현	http://hochulshin.com/aws-ec2-			

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

			flask-dynamodb-angularjs/			
7	웹페이지	Spring boot and OAuth2 + SSO 로그인 기능 튜토리얼	https://spring.io/guides/tutorials/spring-boot-oauth2/			

5 부록

5.1 사용자 매뉴얼

자기소개서 분석

길 JOB 이 홈페이지 <http://ec2-52-79-90-93.ap-northeast-2.compute.amazonaws.com:8080/> 으로 접속한다.

상단 바의 로그인 버튼을 클릭하여 구글 계정을 통해 로그인한다.

상단 바의 제출하기 버튼을 클릭하여 제출 페이지로 이동한다.

지원하는 직무와 기업을 선택한 후 항목 수에 맞게 항목을 추가한 후 자기소개서를 입력한다.

제출하기 버튼을 눌러 분석을 시작한다.

자기소개서 분석이 완료되면 결과 페이지로 넘어가게 된다.

결과 페이지에서 4 가지의 그래프를 통해 자신이 작성한 자기소개서의 역량을 확인한다.


- 테스트용 자기소개서는 Git 에 업로드되어 있습니다.

분석 결과 비교하기

로그인 후 상단메뉴의 비교하기 버튼을 눌러 비교하기 페이지로 이동한다.

비교하고 싶은 자기소개서의 직무와 기업 분류를 선택한 후 분석 기록을 선택한 후 비교하기 버튼을 클릭한다.

비교 결과 페이지에서 변경된 수치를 확인한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

5.2 배포 가이드

5.2.1 flask 서버 배포 가이드

putty 를 다운로드한다.

putty 를 실행시킨 후 Configuration 에서

- 1) Session > Host Name : 13.209.217.136 입력한다.
- 2) Connection > SSH > Auth 의 Private key file for authentication : flask 를 입력한다.
- 3) Open 을 누른다.
 1. login : Ubuntu
 2. source hyunenv/bin/activate
 3. git clone https://github.com/kyungkoh/flask_deploy.git
 4. cd flask_deploy
 5. python app.py


주소창에 13.209.217.136 를 입력한다.

5.2.2 springboot 서버 배포 가이드

5.2.1 을 진행 한다.


putty 를 실행시킨 후 Configuration 에서

- 1) Session > Host Name : 52.79.90.93 입력한다.
 - 2) Connection > SSH > Auth 의 Private key file for authentication : springboot 를 입력한다.
 - 3) Open 을 누른다.
 1. login : ec2-user
 2. \$ cd app/git
 3. \$ java -jar server-0.0.1-SNAPSHOT.jar &
- 서버가 시작되는 것을 확인 후 <http://52.79.90.93:8080> 으로 접속한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

5.3 테스트 케이스

대분류	소분류	기능	테스트 방법	기대 결과	테스트 결과
서버	SSO 인증 기능	구글 계정을 통해 길 JOB 이 로 로그인한다.	상단 메뉴바에서 로그인 버튼을 누르면 구글 로그인 페이지로 이동한다. 로그인을 진행하는 동시에 회원가입이 완료된다.	로그인에 성공하여 access 권한을 얻는다.	성공
	자기소개서 제출 기능	직무, 기업 분류 정보와 작성한 자기소개서를 분석 서버로 제출한다.	상단 메뉴바에서 제출하기를 클릭한다. 1) 제출 페이지에서 직무, 기업 분류를 선택박스에서 선택한다 2) 문항을 추가하여 자기소개서를 입력하고 제출하기 버튼을 눌러 제출한다.	자연어처리 서버로 데이터가 전송되어 분석이 진행된다.	성공
	결과 페이지	제출된 자기소개서를 분석한 결과를 그래프로 나타내준다.	제출하기 과정을 진행한다.	분석된 데이터가 그래프의 파라미터값으로 매핑되어 그래프가 정상적으로 출력된다.	성공
	비교 페이지,	자기소개서 분석 기록을	비교하기 페이지에서 데이터베이스에 저장된 사용자가 이전에 진행한 분석기록을 선택하여 비교하기 버튼을 누른다.	데이터베이스에서 조회된 데이터로 생성된 그래프들이 비교	성공

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	결과보고서		
	프로젝트 명	길 JOB 이	
	팀 명	4726	
	Confidential Restricted	Version 2.0	2019-MAY-28

	비교 결과페이 지	비교하여 보여준다.		결과 페이지에 출력된다.	
자연어처 리	사용자 데이터 분석	사용자의 자기소개서 를 분석해준다	제공되는 API 를 통해 입력 데이터를 전송한다.	자기소개서를 분석한 핵심 역량 수치를 반환한다.	성공
	프로그램 신뢰성 테스트	구현한 모델의 예측 성능을 테스트한다	수집 데이터의 2/3 를 학습시켜 모델을 만든 뒤 나머지 1/3 로 신뢰성 테스트를 진행한다.	테스트 데이터로 예측한 결과가 80% 이상 일치해야 한다.	핵심역량 모델 90% 일치/ 기업 모델 70% 일치