

캡스톤 디자인 I

종합설계 프로젝트

| | |
|--------|---------|
| 프로젝트 명 | 갈 JOB 이 |
| 팀 명 | 4726 |
| 문서 제목 | 수행계획서 |

| | |
|---------|-------------|
| Version | 3.0 |
| Date | 2019-APR-12 |

| | |
|----|----------|
| 팀원 | 고현경 (조장) |
| | 이선홍 |
| | 이수민 |
| | 김혜인 |
| | 김희주 |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

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.

문서 정보 / 수정 내역

| | |
|----------|-------------------------|
| Filename | 계획서-길 JOB 이_수정.docx |
| 원안작성자 | 고현경, 김혜인, 김희주, 이선희, 이수민 |
| 수정작업자 | 고현경, 김혜인, 김희주, 이선희, 이수민 |

| 수정날짜 | 대표수정자 | Revision | 추가/수정 항목 | 내 용 |
|----------|-------------------|----------|----------|----------------------------|
| 19/03/04 | 고현경 | 1.0 | 최초 작성 | 목차 수정 및 프로젝트 개요 초안 작성 |
| 19/03/05 | 이수민 | 1.1 | 내용추가 | 소프트웨어 개발 환경 작성 |
| 19/03/06 | 김혜인 김희주 | 1.2 | 내용추가 | 개발 내용 자연어 처리 과정 초안 작성 |
| 19/03/07 | 이선희 | 1.3 | 내용추가 | 프로젝트 개요 수정 및 추진 배경 작성 |
| 19/03/08 | 김혜인 김희주 이선희 | 1.4 | 내용추가 | 개발 내용 자연어 처리 과정 수정 및 추가 작성 |
| 19/03/09 | 고현경 이수민 | 1.5 | 내용추가 | 개발 내용 웹 / 서버 과정 작성 |
| 19/03/10 | 김혜인 김희주 | 1.6 | 내용 수정 | 개발 일정 수정 |
| 19/03/12 | 고현경 | 2.0 | 내용 수정 | 목차 정리 및 상세 내역 수정 |

| | | | | |
|---|---|-------------------------|-------------|-------------|
|  | 국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | | 프로젝트 명 | 길 JOB 이 | |
| | | 팀 명 | 4726 | |
| | | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

목 차

| | |
|--------------------------------|----|
| 1. 개요 | 4 |
| 1.1. 프로젝트 개요 | 4 |
| 1.2. 추진 배경 및 필요 | 4 |
| 1.2.1. 추진배경 | 4 |
| 1.2.2. 취업 시장의 현황 및 문제점 | 5 |
| 1.2.3. 개발된 시스템과의 차별성 | 6 |
| 2. 개발 목표 및 내용 | 7 |
| 2.1. 목표 | 7 |
| 2.2. 연구/개발 내용 | 7 |
| 2.2.1 웹 / 서버 | 7 |
| 2.2.1.1 웹 | 7 |
| 2.2.1.2 서버 | 8 |
| 2.2.2 자연어 처리 | 9 |
| 2.2.2.1 데이터 수집 단계 | 9 |
| 2.2.2.2 데이터 전처리 단계 | 9 |
| 2.2.2.3 데이터 벡터화 단계 | 9 |
| 2.2.2.4 입력된 자소서 처리 결과 | 10 |
| 2.3. 개발 결과 | 11 |
| 2.3.1 시스템 기능 요구사항 | 11 |
| 2.3.2 시스템 비기능(품질) 요구사항 (우선순위별) | 12 |
| 2.3.3 시스템 구조 | 13 |
| 2.3.4 결과물 목록 및 상세 사항 | 15 |
| 2.4 기대효과 및 활용방안 | 16 |
| 3 배경 기술 | 17 |
| 3.1 기술적 요구사항 | 17 |
| 3.1.1 개발환경 | 17 |
| 3.1.2 프로젝트 결과물 확인 환경 | 17 |
| 3.2 현실적 제한 요소 및 그 해결 방안 | 18 |
| 3.2.1 소프트웨어 | 18 |
| 3.2.2 기타 | 18 |
| 4 프로젝트 팀 구성 및 역할 분담 | 20 |
| 5 프로젝트 비용 | 20 |
| 6 개발 일정 및 자원 관리 | 21 |
| 6.1 개발 일정 | 21 |
| 6.2 일정별 주요 산출물 | 22 |
| 6.3 인력자원 투입계획 | 23 |
| 6.4 비 인적자원 투입계획 | 24 |
| 7 참고 문헌 | 24 |

| | | | |
|--|-------------------------|-------------|-------------|
|  <div> <p>국민대학교</p> <p>컴퓨터공학부</p> <p>캡스톤 디자인 I</p> </div> | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

1. 개요

1.1. 프로젝트 개요

취업 시즌에 자기소개서 첨삭을 해달라는 글은 취업 관련 커뮤니티에서 흔히 볼 수 있는 글이다. 구직자들은 서로 자기소개서를 공유하고 첨삭을 주고받는다. 취업 시장은 점점 더 과열되어가고 취업의 첫 번째 관문인 서류심사를 통과하기 위해 많은 구직자들은 관련 서적, 커뮤니티, 주변 인맥을 통하여 정보를 얻고자 한다. 하지만 문제는 정보의 질이다. 서로가 서로에게 받는 ‘카더라’식의 정보는 전문적, 객관적이지 못하기 때문에 자신에게 얼마나 유용한 정보인지 가늠하기 어렵다. 구직자 자신이 지원하고자 하는 기업 군에 맞는 자기소개서는 어떻게 되는지, 작성하고자 하는 내용을 어떻게 정리할지에 대한 내용은 많은 시행착오를 겪고 나서야 비로소 완성이 된다.

이러한 내용들을 종합하여 본 프로젝트에서는 텍스트 마이닝을 통해 합격 자기소개서를 분석하고 학습된 데이터를 바탕으로 사용자의 자기소개서의 성향을 분석하는 웹 애플리케이션을 만드는 프로젝트이다.

1.2. 추진 배경 및 필요

1.2.1. 추진배경

한 연구 결과에 따르면 내년엔 우리나라 대졸자 비율이 절반을 넘고 2035 년엔 70%까지 올라설 것이라고 한다. 하지만 채용 인원이 줄어들어 취업준비생들은 고용 시장에서 바늘구멍 같은 취업의 문을 뚫기가 쉽지 않다. 또한 잡코리아 조사에 따르면 신입사원들이 재직 중인 회사에 최종 합격할 때까지 써낸 자기소개서가 14.4 건으로 평균 10 건을 훌쩍 넘는다. 특히 응답자의 7.2%는 무려 50 건 이상 작성했다고 밝혔다. 이처럼 자기소개서는 점점 더 취업에 중요한 비중을 차지하고 있다.



[그림 1] 인사담당자의 61%는 완벽한 스펙을 갖춘 지원자보다 완벽한 자기소개서를 작성한 지원자를 더 선호하는 것으로 나타났다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

또한, 구직자들에게 자기소개서를 작성하는 데 있어서 가장 큰 도움이 된 요소는 '취업포털·구직커뮤니티의 취업 자료'라는 응답이 51.4%로 가장 많았고, '취업 선배나 지인들의 합격 자기소개서'(47.9%)와 '취업 사교육'(22.3%) 등의 순으로 조사되었다.

이 밖에 조사 대상 신입사원들 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 들어는 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.

본 프로젝트에서는 현재 과도화된 고액 컨설팅 시스템을 개선하고 구직자들의 자기소개서 작성에 도움을 줄 수 있는 자기소개서 종합적 분석 시스템을 개발하는 것을 목표로 한다. 텍스트마인딩을 기반으로 한 자기소개서 분석 시스템은 더욱 객관적인 데이터 통계를 통해 취업준비생에게 신뢰도 있는 분석 결과를 보여준다. 이러한 시스템은 현재 불필요하게 고액인 컨설팅 비용을 완화해 취업을 준비하는 데 있어서 실질적인 도움을 얻을 수 있게 될 것으로 기대된다.

1.2.2. 취업 시장의 현황 및 문제점



[그림 2] 취업컨설팅 서비스 관련 소비자 설문조사 결과

취업난이 계속되는 가운데, 블라인드 평가 도입 등 기업별로 채용방식이 다변화되면서 취업준비생들 대상으로 한 '취업컨설팅 서비스'가 인기를 끌고 있다. 그러나 그림[2]와 같이 제공되는 서비스에 비해 수강료가 과도하게 비싸다는 불만이 많고, 개인별 맞춤형 컨설팅이 제대로 이루어지지 않는다는 의견이 많다. 또한 취업 컨설팅 서비스는 주로 사용자가 자기소개서나 스펙을 제출하면 취업 전문가들이 직접 분석

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

및 검색을 하는 방식으로 진행된다. 이러한 방식에는 불필요한 부분에도 사람의 주관적인 의견이 추가될 가능성이 있다.

AI 를 이용한 자기소개서 분석 서비스가 그 대안으로 출시되었다. 사람을 통해서 컨설팅하는 것보다 비용이 저렴하고 언제나 일정한 수준의 품질로 객관적인 자기소개서 분석을 해 줄 수 있기 때문이다.


기출시된 AI 자기소개서 분석기는 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

취업 준비생들 대상으로 하는 설문 조사들은 대부분 직무보다 기업에 초점을 맞춰 이루어진다. 어떤 직무를 선호하는지를 물어보는 것이 아니라 어떤 기업을 선호하는지를 물어보는 식이다. 취업 준비생들도 직무에 관심을 두는 것만큼 어떤 기업에 합격할 가능성이 있을지에 대해 궁금해한다.

또한 기업은 자기소개서를 통해 지원자의 '직무에 대한 이해도'만큼 '충성도'를 중요하게 평가한다. 회사 차원에서는 근속기간이 길수록 신입사원 채용과 교육, 훈련에 쓴 비용을 회수하는 등 이득이기 때문이다. 그러므로 자기소개서는 직무뿐만이 아니라 기업에 대한 관심도와 충성도를 자기소개서에 표현하는 것도 중요하다.

그에 맞춰 이 프로젝트는 자기소개서가 지원할 직무에 맞게 작성됐는지 분석할 뿐만 아니라 기업별 합격 자기소개서 데이터들을 기반으로 자신이 작성한 자기소개서와 성향이 맞는 기업들을 추천해준다. 또한 누구나 비용, 장소의 부담 없이 자기소개서를 검색 받고, 자신을 객관적으로 돌아볼 수 있게 한다. 따라서 기존 AI 자기소개서 분석기보다 더 유용한 서비스를 제공할 수 있을 것이다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

2. 개발 목표 및 내용

2.1. 목표

본 프로젝트는 취업을 목적으로 하는 사람들에게 더욱 전문적인 자기소개서 분석 서비스를 제공하는 것을 목표로 한다. 자기소개서 분석을 하는 과정에 텍스트마이닝을 활용하여 기존 검색 서비스보다 객관적이고 분석적인 데이터를 만드는 것을 목표로 한다. 또한 분석 결과를 다양한 차트로 보여주어 사용자 정보를 확인하는데 용이하도록 한다.

2.2. 연구/개발 내용

2.2.1 웹 / 서버


<길 JOB 이>는 웹 페이지에서 클라이언트가 자기소개서를 제출하고 자기소개서에 대한 분석 결과를 확인할 수 있는 서비스를 제공한다. 웹 서비스로 제공하여 누구든지 쉽게 이용할 수 있고 실시간으로 분석 결과를 확인할 수 있다. 웹 페이지는 java 를 사용하는 스프링부트 기반으로 개발하고 AI 분석기는 파이썬 어플리케이션으로 파이썬 서버를 따로 구축하여 두 서버를 연동시키는 방법으로 서비스 시스템 구조를 구축한다.

2.2.1.1 Web Application Server

서버 개발 환경은 스프링 프레임워크 기반의 스프링부트를 사용하여 간편한 환경설정과 의존성 관리 통해 개발 기간을 단축시킨다.

웹 페이지 뷰는 데스크탑, 랩탑, 모바일 등 다양한 플랫폼에서도 사용할 수 있도록 반응형 웹으로 개발하여 어떤 사용자라도 문제없이 사용할 수 있게 한다.

- 로그인
 - Google Login API 를 사용하여 Google Cloud Platform 에서 API 인증을 받은 후 로그인/회원 가입을 구현하여 사용자는 OAuth 인증방식의 회원가입을 진행할 수 있다. 이때 사용자가 회원 가입 진행 중 입력한 사용자 정보는 Maria DB 에 저장되고 비밀번호는 보안을 위해 해시함수 (PBKDF2)를 통해 처리 된 후 저장된다.
- 자기소개서 입력
 - 클라이언트는 자기소개서를 제출하기 전 미리 분류된 직무 분류 중 자신이 지원하는 직무 분야와 지원하는 기업을 선택하고 자기소개서를 텍스트로 제출한다. 제출된 데이터는 웹 서버에서 Json 형식으로 자연어 처리 서버로 전송된다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

● 분석 / 결과 페이지

- 자연어 처리 서버에서 자기소개서를 분석한 결과 데이터를 DB 에 저장하여 기록을 조회하거나 비교할 수 있게 만든다. 결과 데이터를 웹 서버로 전송하면 controller 가 받아 분석 결과에 대한 정보를 view 에게 넘긴다. view 는 받은 데이터를 이용하여 웹 페이지를 생성하여 분석 결과를 시각적으로 보여준다.
- 결과 페이지에 나타날 차트는 총 세가지로 1) 적합 기업 그래프 2) 직무 적합도 그래프 3) 사용자 선택 기업 적합도 그래프로 나타난다.

[1] 적합 기업 그래프 : 사용자가 입력한 자기소개서가 어떤 기업에 적합한지 상위 3 개의 기업을 보여준다.

[2] 직무 적합도 그래프 : 직무 분야별로 몇 가지의 키워드를 통해 나타난 직무 적합도를 사용자가 입력한 자기소개서에 대하여 분석하여 나타난 적합도 점수와 비교하여 그래프를 통해 나타낸다.

[3] 사용자 선택 기업 적합도 : 사용자가 자기소개서를 분석한 결과가 사용자가 분석 받기 원하는 기업의 결과간의 카테고리 별 적합도를 점수로 나타낸다.

- 결과 페이지는 차트뿐만 아니라 사용자의 성향에 적절한 텍스트도 표현이 된다. 이는 전문가의 도움을 받아 작성 할 예정이다.

2.2.1.2 자연어처리 Application Server

클라이언트가 선택하는 변수 값과 자기소개서 데이터를 Json 형식의 데이터로 자연어 처리 Application Server 에 전달한다. 데이터를 전달 할 때는 자연어처리 Application 을 AWS Lambda 에 배포하여 실시간으로 데이터를 처리한 후 처리한 데이터를 DynamoDB 에 업데이트 하고 AWS EC2 에 구축 된 Web Server 에서 데이터를 받아와 View 와 연동하게 된다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

2.2.2 자연어 처리

합격한 자기소개서를 종합하여 분석하고 특징을 분석해내려면 자연어처리가 반드시 필요하다. 또한 우리나라 구직자들의 자기소개서를 토대로 분석하므로 모든 자연어처리 작업은 한글로 진행된다. 자연어처리의 과정은 크게 데이터 수집단계, 데이터 전처리 단계, 데이터 벡터화 단계, 입력된 분석 처리 총 4 단계로 나뉜다.

2.2.2.1 데이터 수집 단계

첫 번째 데이터 수집은 다시 하위 2 단계로 나뉘게 된다. 하위 1 단계는 자기소개서 분석 결과 항목에 대한 데이터 수집이다. 자기소개서 분석의 결과는 10 가지 항목에 따라 방사형 그래프로 나타내며 각 항목은 소통(의사소통 능력), 협력(팀워크 능력), 인내심 등과 같이 사용자의 특성을 정의할 수 있는 용어들로 구분된다. 이 용어들은 구인 구직 사이트에서 기업의 인사 담당자들에게 자기소개서 핵심 키워드를 조사한 것을 바탕으로 결정하였다. 각 항목의 해당하는 어휘들을 포함하는 문장 데이터를 뉴스기사, 웹 문서, 소셜 네트워크를 통해 수집하여 그 항목으로 구분된 어휘들의 데이터를 수집한다. 하위 2 단계는 자기소개서에 대한 데이터 수집이다. 사용자의 자기소개서와 비교할 데이터를 수집하기 위해 온라인으로 공개된 기업 합격 자기소개서들을 모은다.

2.2.2.2 데이터 전처리 단계

데이터 전처리 단계에서는 수집된 방대한 데이터들을 분석할 수 있도록 정제하는 단계이다. 데이터 수집 단계에서 수집된 데이터들은 키워드를 포함한 원문 상태의 데이터이며, 이 데이터들은 불필요한 어미, 조사 등을 포함하고 있어 분석에 영향을 미치게 된다. 따라서 데이터를 분석하기 전 정제를 해야하고 정제 단계는 크게 4 단계로 구분된다. 각 단계는 1) 데이터 토큰화, 2) 데이터 정규화, 3) 데이터 어근화, 4) 불용어 제거 순으로 진행된다.

수집한 원문데이터는 첫 번째로 토큰화(tokenization)를 진행하여 데이터들을 품사별로 분리하게 된다. 이 과정에서 한글 토큰화 과정에 가장 많이 활용되는 Konlpy 라이브러리를 활용하여 진행한다. 이러한 라이브러리를 활용하게 되면 “안녕하세요, 반갑습니다” 같은 문장은 안녕(일반명사, NNG) 하(동사 파생 접미사, XSV)세요(종결어미, EF), (침표, SP) 반갑(형용사, VA)습니다(종결어미, EF) .(마침표, SP)와 같이 분석된다. 이와 같은 토큰화된 데이터에서 분석에 필요한 품사(단어(명사), 동사) 위주로 추출하여 표현방법이 다른 단어들을 통합시키는 정규화, 의미를 담고 있는 부분을 원형으로 바꿔주는 어근화 단계를 진행한다. 데이터를 정제하고 최종적으로 분석에 큰 의미가 없는 불용어를 제거하고 나면 데이터를 분석하기 위한 전처리 단계가 완료된다. 이 단어들 중 명사만을 뽑아 그래프에 보이게 될 키워드와 중요하게 연관된 단어 순으로 나열한다. 중요하게 연관된 단어 순으로 나열하는 것은 다음

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

데이터 벡터화 단계에서 TF-IDF 를 활용한다.

2.2.2.3 데이터 벡터화 단계

중요하게 연관된 정도는 단순히 빈도수로 정의하긴 어려우므로 텍스트마이닝에서 사용되는 TF-IDF 를 사용하여 나열하게 된다. TF-IDF 는 출현빈도를 사용하여 어떤 단어가 문서 내에서 얼마나 중요한지 나타내는 수치이다. 이 수치가 높을수록 단어는 문서를 대표하는 성격을 띠게 된다고 볼 수 있다. TF-IDF 의 TF 는 Term Frequency 로 단어의 문서 내에 출현한 횟수를 의미한다. 그리고 IDF 는 Inverse Document Frequency 로 그 단어가 출현한 문서의 숫자의 역수(inverse)를 의미한다. TF 는 단어가 문서 내에서 출현한 횟수이다. 따라서 그 숫자가 클수록 문서 내에서 중요한 단어일 확률이 높다. 하지만, 'the'와 같은 단어도 TF 값이 매우 클 것이다. 하지만 'the'가 중요한 경우는 거의 없으므로 이때 IDF 가 필요하다. DF 는 그 단어가 출현한 문서의 숫자를 의미 하므로, 그 값이 클수록 'the'와 같이 일반적으로 많이 쓰이는 단어일 가능성이 높다. 따라서 IDF 를 구해 TF 에 곱해줌으로써, 'the'와 같은 단어들에 대한 페널티를 준다.

최종적으로 우리가 얻는 숫자는, 다른 문서들에서는 잘 나타나지 않지만, 특정 문서에서만 잘 나타난 경우에 횟수가 높아지기 때문에, 특정 문서에서 얼마나 중요한 역할을 차지하는지 나타내는 수치가 될 수 있다. 따라서 TF-IDF 가중치를 사용하여 방사형 그래프로 나타낼 때 필요한 항목별 특징 벡터를 만든다. 그 후, 기업별, 직무별 기준으로 2.2.2.1 단계에서 수집한 합격 자기소개서 자료를 두 번 구분하고 토큰화한 뒤 각각 특징 벡터를 만든다. 예를 들어 기업별 기준으로는 삼성, LG, 롯데, 현대, SK, CJ 등이 있고, 직무별 기준으로는 개발, 영업, 마케팅, 인사 등의 특징 벡터가 있다. 앞에서 구한 항목별 특징 벡터와 기업별, 직무별 벡터들의 코사인 유사도를 이용해 지원자의 자기소개서와 비교할 최종 기업별, 직무별 벡터를 만든다.

2.2.2.4 입력된 자소서 처리 결과

데이터 벡터화 단계(2.2.2.3)까지 완료되면 사용자의 자기소개서 분석을 위한 모든 준비가 끝난 상태다. 사용자가 자기소개서를 웹사이트에 입력하면 먼저 2.2.2.2 의 전처리과정을 거친 후 벡터화를 시킨다. 완성된 벡터로 항목별 벡터와 코사인유사도를 측정한 후 그 값을 JSON 형식으로 서버에 전달한다.

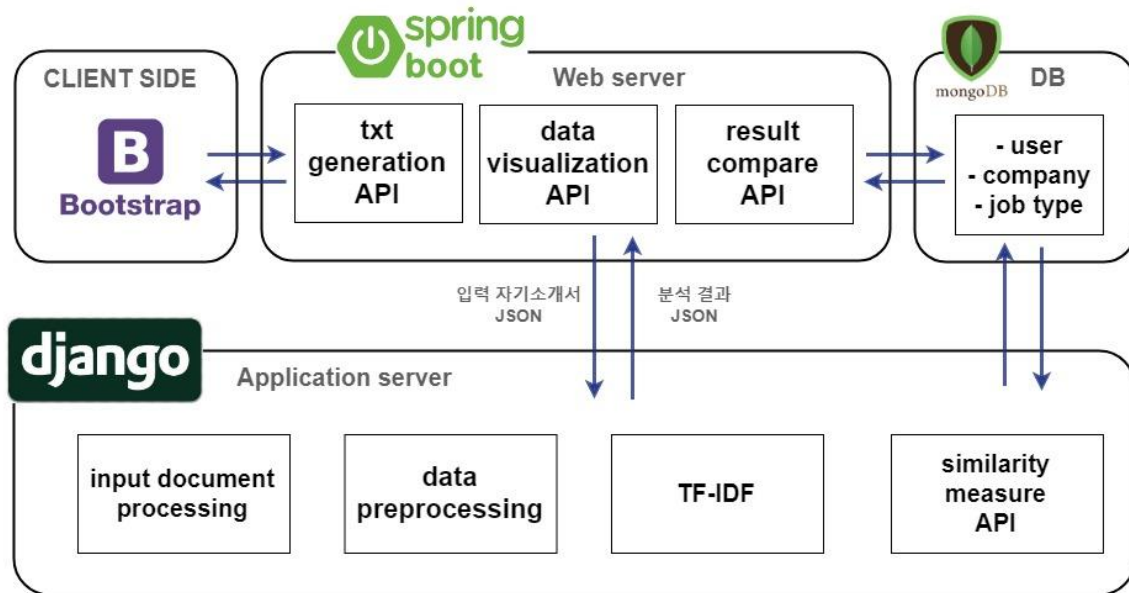
| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

2.3.2 시스템 비기능(품질) 요구사항 (우선순위순)

1. 데이터 분석의 정확성 (신뢰성) : 시스템 개발이 완료되면, 테스트 데이터 셋(합격 자기소개서 등)을 이용하여 구현한 알고리즘이 얼마나 정확히 예측하는지 정확도를 측정한다. 테스트 분석 결과의 항목별 점수 평균을 내어 합격 자기소개서로 만든 벡터의 항목별 점수와 비교한다. 테스트 데이터 셋의 점수가 합격 자기소개서의 점수의 80% 이하일 경우 신뢰성이 낮다고 판단한다.
2. 처리 속도 :
 - 1) 웹페이지 : 웹페이지 로딩 속도를 크롬 평균 로딩속도인 5.30 초 이내로 처리할 수 있어야 한다.
 - 2) 자기소개서 분석 : 자기소개서 분석에서 중요한 것은 데이터 분석 결과의 정확성이다.
 - 3) 따라서 자기소개서 분석의 속도를 우선순위로 두고 사용자가 기다리기에 지루하지 않을 정도의 속도인 20 초를 최대로 둔다.
3. 사용자 편리성: 사용자가 시스템에 제출한 자기소개서가 온라인에 공개되지 않게 설정할 수 있도록 한다. 사용자가 비공개로 설정하면 자기소개서 분석 내용은 공개되지 않는다.

| | | | |
|--|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

2.3.3 시스템 구조



- Client
 - 클라이언트에게 제공되는 웹 페이지는 웹 프론트 프레임워크인 부트스트랩 기반으로 작성되었다.
 - 클라이언트는 웹 브라우저를 이용해 길 JOB 이 웹 페이지로 접속하여 회원가입, 로그인 과정을 거쳐 서비스를 이용하게 된다. _
 - 자기소개서 제출 단계에서 클라이언트는 자신의 자기소개서를 입력한 후 자신의 직무 분야와 지원하는 기업을 선택한다.
 - 제출된 자기소개서의 분석이 끝나면 클라이언트는 시각화된 자신의 자기소개서 분석결과를 확인할 수 있다. 분석 결과는 직무에 대한 적합도를 10 가지 키워드로 나눈 그래프와 기업의 성향에 얼마나 적합한지에 대한 점수를 비교한 그래프 그리고 적합 기업 추천 그래프로 확인할 수 있다.
- Web server _
 - 웹 서버는 웹 애플리케이션 개발 프레임워크인 스프링부트 기반으로 구현한다.
 - Data visualization API: Application server 에서 받은 자기소개서 분석 결과를 시각화된 페이지로 생성하여 클라이언트에게 보여준다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |


- Result compare API : 클라이언트가 이전에 분석했었던 기록 중에서 2 가지를 선택하여 분석결과를 비교하여 보여주는 기능을 제공한다.
- Txt generation API : 자기소개서를 분석한 결과에 따라서 몇 가지 성향을 보일 수 있는데 이에 따른 설명 텍스트를 생성하여 결과 페이지에 제공한다.

● Application server

- 자기소개서 분석 애플리케이션은 Python 을 기반으로 작성할 것이므로 애플리케이션 서버는 파이썬 웹 프레임워크인 Django 를 사용하여 개발한다.
- 웹 서버와의 데이터 통신은 JSON 형식의 데이터 타입을 주고 받고 통신은 웹 소켓을 이용하여 구현한다
- Input document processing: Web server 에서 입력받은 자기소개서를 json 형식으로 전달 받아 사용자의 feature vector 를 생성한 후 분석결과를 Web server 에 다시 전달한다.
- Data preprocessing: 기업, 직무별 vector 를 만들기 위해 분석에 필요한 데이터만 추출하기 위해서 데이터 전처리 한다. -
- TF-IDF: 한 문서내에서 중요한 키워드 순으로 계산하고, 전체 데이터 내 출현빈도를 사용하여 단어의 중요도를 판단한다. -
- Similiarity measure API: 사용자가 입력한 자기소개서의 벡터와 미리 학습된 기업별, 직무별 벡터를 코사인 유사도를 통해 유사도를 측정한다.


● FLOW

사용자에게 자기소개서를 입력 받아 불용어를 제거한 뒤 명사를 추출하여 전처리한다. 전처리한 자기소개서를 tf-idf 를 이용해 벡터화하고 미리 만들어놓은 키워드, 직무, 기업별 벡터와 유사도 비교 후 결과를 json 으로 전달한다. 웹서버로 전달 된 json 형식을 db 에 저장 하고 선택된 직무, 기업 분류로 미리 만들어진 벡터를 db 에서 가져와서 결과페이지를 생성하여 사용자에게 보여준다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

2.3.4 결과물 목록 및 상세 사항

| 대분류 | 소분류 | 기능 | 형식 | 비고 |
|--------|-------------------|---|-------|----|
| 자연어 처리 | 데이터 수집 모듈 | 자기소개서 분석을 위한 데이터를 수집한다. | 모듈 | |
| | 데이터 정제 모듈 | 원문 데이터를 분석에 활용할 수 있도록 정제한다. | 모듈 | |
| | 데이터 벡터화 모듈 | 정제한 데이터를 벡터화 한다. | 모듈 | |
| | 유사도 비교 모듈 | 기존에 벡터화한 데이터와 사용자 입력 데이터 벡터를 비교한다. | 모듈 | |
| 서버 | Sign-In & Sign-up | 회원가입을 통한 유저 DB 를 구축한다. | DB | |
| | 자기소개서 입력 | 분석 할 자기소개서를 입력한다. | | |
| | 분석 결과 출력 | 사용자가 선택한 기업의 키워드와 사용자의 자기소개서의 키워드 간의 유사도 수치를 시각화한다. | Chart | |
| | | 사용자가 선택한 직무의 키워드와 사용자의 자기소개서의 키워드 간의 유사도 수치를 시각화한다. | Chart | |
| | | 사용자의 자기소개서와 유사도가 높은 순으로 기업을 나열한다. | Chart | |
| | | 사용자의 분석 결과에 따른 코멘트를 보여준다. | Text | |
| | 분석결과 비교 | 사용자가 분석했던 자소서들의 비교 결과를 차트로 표현한다. | Chart | |
| 통신 | 웹 서버 | 자연어처리 어플리케이션에서 전달받은 유사도를 차트를 통해 시각화 한다. | API | |
| | | 자연어처리 어플리케이션에서 전달받은 유사도를 기반으로 사용자에게 제공할 text 를 생성한다 | API | |
| | | 사용자가 분석했던 자기소개서들의 결과를 비교한다. | API | |
| | 통신 서버 | 자연어 처리 어플리케이션에서 만든 유사도를 웹서버에 전달한다. | API | |
| | | 웹서버에서 입력받은 자기소개서를 json 형식으로 자연어처리 어플리케이션 보내준다. | API | |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

2.4 기대효과 및 활용방안

구직자들에게 자기소개를 작성하는 데 있어서 가장 큰 도움이 된 요소에는 '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 것으로 기대된다.

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 기계학습을 통해 분석하는 본 프로젝트가 많은 도움이 될 것이라고 기대된다.

자연어처리를 통해 자기소개를 분석하는 본 프로젝트는 다양한 방면으로 활용, 발전될 수 있다. 첫 번째로 사용자들이 기업 성향에 맞는 자기소개를 작성했다면 기업 관계자가 그러한 자기소개를 열람할 수 있도록 하여 하나의 리크루팅 시스템을 형성할 수 있다. 두 번째로는 취업용 자기소개서 다음으로 많이 작성되는 것이 대입 자기소개서임을 고려하여, 학교별 성향에 맞게 자기소개서 분석하는 것으로도 활용할 수 있다.

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

3 배경 기술

3.1 기술적 요구사항

3.1.1 개발환경

- 운영체제 : Windows, Mac OS, Linux
- 컴파일 : JDK
- 개발 언어 : Python, Java, JavaScript
- Front-end : Bootstrap
 - 쉽고 빠르고 다양한 기능들을 제공한다. 또한 다양한 디자인과 동적인 효과를 사용할 수 있다. 반응형 웹으로 구현이 되어있기 때문에 다양한 환경에서 사용이 가능하다.
 - 높은 퀄리티가 보장되어 비용절감 효과가 나타난다.
- Back-end : Spring boot
 - Tomcat 이 내장되어 있기 때문에 웹프로젝트를 배포하는 시간이 독립적인 Tomcat 으로 배포하는 시간보다 절반이상 단축된다.
 - 기존 Spring Framework 를 개발하는데 있어 필요한 의존성과 XML 구성등의 일반적인 공통 시나리오를 자동으로 구성해주고 개발자는 즉시 애플리케이션 로직을 개발 할 수 있도록 하여 비용절감에 효과적이다.
- 라이브러리
 - beautiful soup : 자동화된 방식으로 데이터를 수집 할 수 있는 웹 크롤러를 쉽게 만들 수 있다.
 - KoNLpy : 다양한 형태소 분석 라이브러리를 파이썬에서 쉽게 사용할 수 있도록 모아 놓은 패키지이다. 이 패키지는 한글 데이터 형태소 분석 및 정제하기 위해 사용된다.
 - pytorch : 데이터 기계학습에 활용되는 용도로 사용된다.
 -

3.1.2 프로젝트 결과물 확인 환경

웹 브라우저를 이용 가능한 모든 환경

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

3.2 현실적 제한 요소 및 그 해결 방안

3.2.1 소프트웨어

- 웹/서버
 - 대규모 데이터를 처리하는 데 있어서 스프링 내장 서블릿 컨테이너를 사용 시 쓰레드 생성 비용 등의 문제로 속도가 느려질 수 있다. 이를 방지하기 nginx 웹 서버를 사용하여 대규모 데이터 처리에도 속도가 저하되지 않게 만든다.
 - 부트스트랩을 이용하여 구글 크롬을 기준으로 웹 페이지를 작성했을 시 IE, Firefox 등 다른 웹 브라우저에서 뷰가 제대로 나오지 않아 서비스를 제대로 이용할 수 없는 문제가 발생할 수 있다. 뷰의 문제는 html 에서 meta 태그를 수정하는 등 코드상에서 각 브라우저마다 호환성 관련 속성을 추가하는 것으로 해결할 수 있을 것으로 보인다.
- 자연어처리
 - 자기소개서 분석을 위해 텍스트 데이터를 학습시키는 데 있어서 물리적인 이유로 속도가 느릴 수 있다. 따라서 개인 소유의 노트북이나 데스크탑을 사용하지 않고 클라우드 플랫폼을 이용한다.
 - 부족한 데이터 양을 보완하여 정확한 분석이 가능하도록 하기 위해 기존 알고리즘인 TF-IDF 와 문장의 공기관계를 같이 활용한다.

3.2.2 기타

waterfall 방식으로 개발하게 되면 개발 진행 중 지속적으로 발생하는 변화에 대응하기 어렵고 상위 업무가 완료될 때까지 하위 업무를 진행할 수 없기 때문에 비효율적인 방식이라고 생각된다. 그렇기 때문에 Agile 모델을 이용하여 계속해서 프로토타입을 만들어 내며 그때 그때 필요한 요구를 더하고 수정하여 개발해 나간다.

| | | | |
|--|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

4 프로젝트 팀 구성 및 역할 분담

| 이름 | 역할 | |
|-----|--|---|
| 이수민 | <ul style="list-style-type: none"> - Spring API 개발 - DB 설계 및 DB Query 시스템 개발 | |
| 고현경 | <ul style="list-style-type: none"> - Web Interface 개발 - Web Server 와 자연어 처리 Application 간 연동 | |
| 이선희 | - 데이터 전처리 | <ul style="list-style-type: none"> - 학습 데이터 설계 - 수집 데이터 벡터화 - 사용자 데이터 처리 |
| 김혜인 | - 데이터 수집 | |
| 김희주 | - 데이터 수집 | |

5 프로젝트 비용

| 항목 | 예상치 (MD) |
|------------------|----------|
| 아이디어 구상 | 3 |
| 관련 정보 수집 및 연구 | 20 |
| 자연어처리 어플리케이션 개발 | 80 |
| 웹 서비스 UI/UX 구성 | 20 |
| 웹 서버 구축 | 40 |
| 어플리케이션과 웹 서비스 연동 | 20 |
| 프로젝트 테스트 및 유지보수 | 4 |
| 프로젝트 관련 문서작업 | 3 |
| 합 | 190 |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

6 개발 일정 및 자원 관리


6.1 개발 일정

| 항목 | 세부내용 | 1 월 | 2 월 | 3 월 | 4 월 | 5 월 | 비고 |
|--------|--------------------------|-----|-----|-----|-----|-----|----|
| 요구사항분석 | 요구 분석 | | | | | | |
| | 아이디어 구상 | | | | | | |
| 관련분야연구 | 자연어처리 및 텍스트마이닝 연구 | | | | | | |
| | 관련 시스템 분석 | | | | | | |
| 설계 | 시스템 설계 | | | | | | |
| 구현 | 데이터 크롤링 | | | | | | |
| | 웹 UI 및 서버 구현 | | | | | | |
| | 자연어처리 모듈 구현 | | | | | | |
| | 프로토 타입 구현 | | | | | | |
| | 자연어처리 어플리케이션과 웹 서비스 간 연동 | | | | | | |
| 테스트 | 자연어처리 어플리케이션 신뢰성 테스트 | | | | | | |
| | 자연어처리 어플리케이션과 서버간 통신 테스트 | | | | | | |
| 최종 발표 | 발표 준비 및 발표 | | | | | | |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |


6.2 일정별 주요 산출물

| 마일스톤 | 개요 | 시작일 | 종료일 |
|--------|--|------------|------------|
| 계획서 발표 | 프로젝트 아이디어 선정 프로젝트 아키텍처 구상 관련 정보 수집 및 연구 개발 환경 완성 (개발 프레임워크 설치 및 환경 설정 완료, 필요 물품 정산) 산출물 : 1. 프로젝트 수행 계획서 2. 프로젝트 수행 발표 자료(PPT) 3. 지원비 정산내역서, 지원비 신청서 | ~ | 2019-03-15 |
| 설계 완료 | 시스템 설계 완료 산출물 : 1. 시스템 설계 사양서 | 2019-03-16 | 2019-03-20 |
| 중간 보고 | 프로토타입 구현 완료 산출물 : 1. 프로젝트 중간 보고서 2. 프로젝트 진도 점검표 3. 프로토타입 구현 소스 코드 4. 프로토타입 어플리케이션 | 2019-03-21 | 2019-04-19 |
| 구현 완료 | 시스템 구현 완료 산출물: 1. 구현 소스코드 2. AI 자기소개서 분석 웹 서비스 | 2019-04-20 | 2019-05-14 |
| 테스트 | 시스템 통합 테스트 산출물: 1. 자기소개서 테스트 결과 | 2019-05-15 | 2019-05-21 |
| 최종 보고서 | 최종 보고 산출물: 1. 최종 보고서 | 2019-05-22 | 2019-05-28 |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

6.3 인력자원 투입계획

| 이름 | 개발항목 | 시작일 | 종료일 | 총개발일 (MD) |
|-----------------|------------------------|------------|------------|--------------|
| 전원 | 요구 분석 / 아이디어 구상 | 2019-01-01 | 2019-02-25 | 50 |
| 고현경 | 웹 인터페이스 디자인 구현 | 2019-03-04 | 2019-03-11 | 6 |
| 고현경 | 웹, 어플리케이션 서버 시스템 구조 설계 | 2019-03-12 | 2019-03-17 | 5 |
| 이수민 | 웹, 어플리케이션 서버 구현 | 2019-03-18 | 2019-04-29 | 35 |
| 이수민 | 서버 테스트 | 2019-04-30 | 2019-04-31 | 2 |
| 김희주 / 김혜인 | 학습 데이터 수집 | 2019-02-20 | 2019-03-15 | 20 |
| 이선홍 | 데이터 정제 : 형태소 분석 | 2019-02-24 | 2019-03-15 | 7 |
| 김희주 / 김혜인 / 이선홍 | 데이터 벡터화 : 워드 임베딩 | 2019-03-16 | 2019-04-14 | 30 |
| 김희주 / 김혜인 / 이선홍 | 벡터화된 데이터 분류 구현 | 2019-04-15 | 2019-05-01 | 20 |
| 전원 | AI 자기소개서 분석 서비스 구현 | 2019-05-01 | 2019-05-14 | 40 |
| 전원 | 최종 테스트 | 2019-05-15 | 2019-05-21 | 10 |
| 전원 | 최종 발표 준비 | 2019-05-22 | 2019-05-28 | 10 |

| | | | |
|---|-------------------------|-------------|-------------|
|  국민대학교 컴퓨터공학부 캡스톤 디자인 I | 계획서 | | |
| | 프로젝트 명 | 길 JOB 이 | |
| | 팀 명 | 4726 | |
| | Confidential Restricted | Version 3.0 | 2019-MAR-16 |

6.4 비 인적자원 투입계획

| 항목 | Provider | 시작일 | 종료일 | Required Options |
|-----------------------|----------------------------|------------|------------|------------------|
| 서버용 PC 1 대 | AWS | 2019-03-08 | 2019-05-20 | |
| Google Cloud Platform | Google | 2019-04-01 | 2019-05-20 | |
| 개발용 노트북 5 대 | Apple, LG, samsung, Lenovo | 2019-03-04 | 2019-05-28 | |

7 참고 문헌

| 번호 | 종류 | 제목 | 출처 | 발행년도 | 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 | | | |