* MPTI: Make self-PR Through AI

MPTI:
자기소개서 생성&한 줄 요약

KUBIG NLP 세션 대회
생성2팀

13기 기다연
12기 전지우
13기 이현규

2021.08.26

0. 진행 현황

01

데이터 크롤링

링커리어 합격 자기소개서 데이터

잡코리아 합격자기소개서 데이터

02

데이터 전처리

```
# 텍스트 정제 함수: 한글 이외의 문자는 모두 제거
def text_cleaning(text):
    hangul = re.compile('[^ ㄱ-|가-힣]+') # 한글과 띄어쓰기를 제외한 모든 글자
    result = hangul.sub('', str(text))
    return(result)
```

```
[ ] cleaned_corpus = []
    for sent in linkcareer['content']:
        cleaned_corpus.append(text_cleaning(sent))

    linkcareer['content_kor'] = cleaned_corpus
    linkcareer.head(3)
```

```
[ ] # 3) pykospacing으로 띄어쓰기 교정
    from pykospacing import Spacing
    spacing = Spacing()

    spaced_corpus = []
```

크롤링 완료된 데이터 전처리

03

GPT-3 모델 fine-tuning

GPT-3 기존 코드 공부

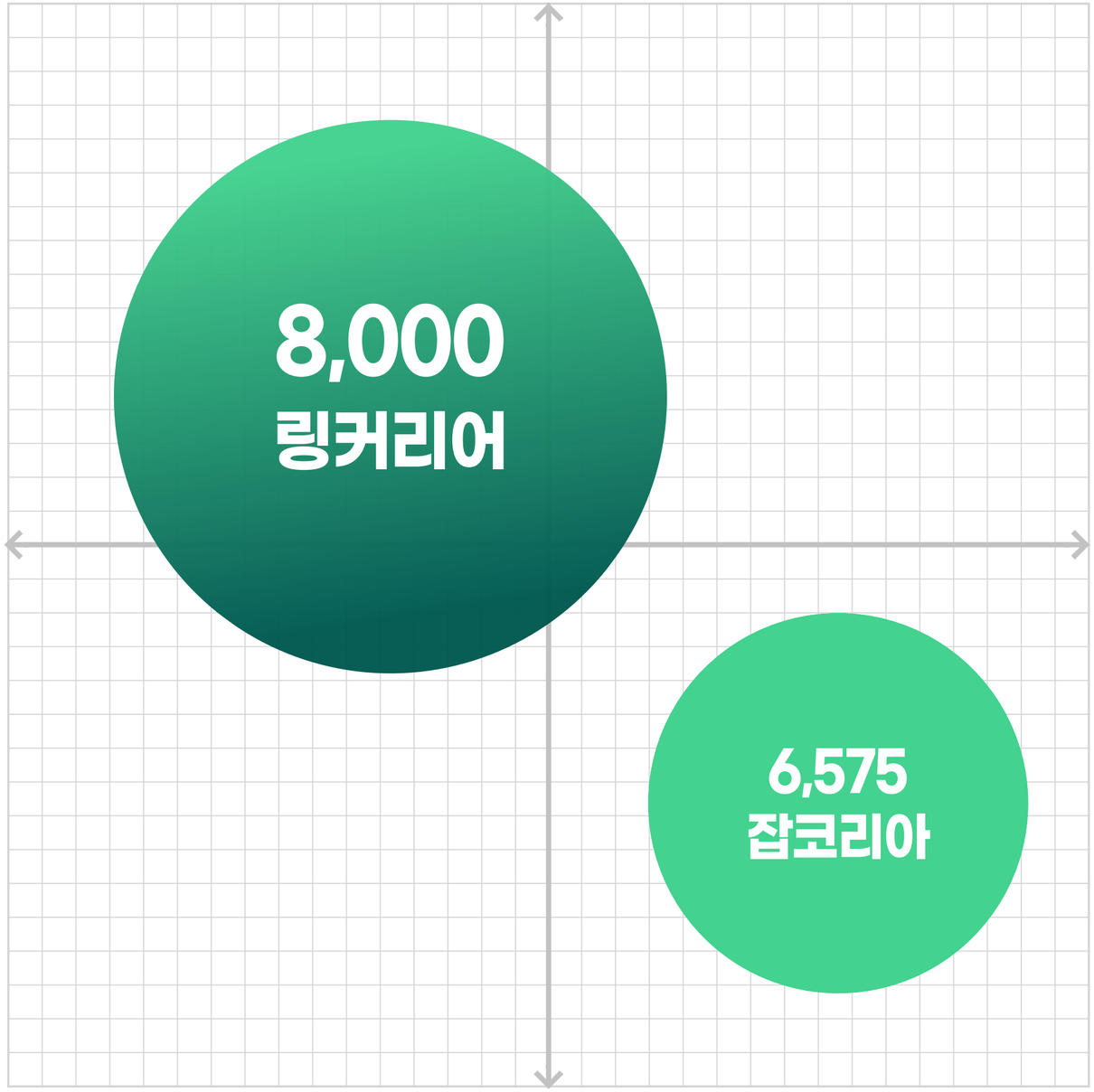
GPT-3 모델 fine-tuning

1. 데이터 크롤링

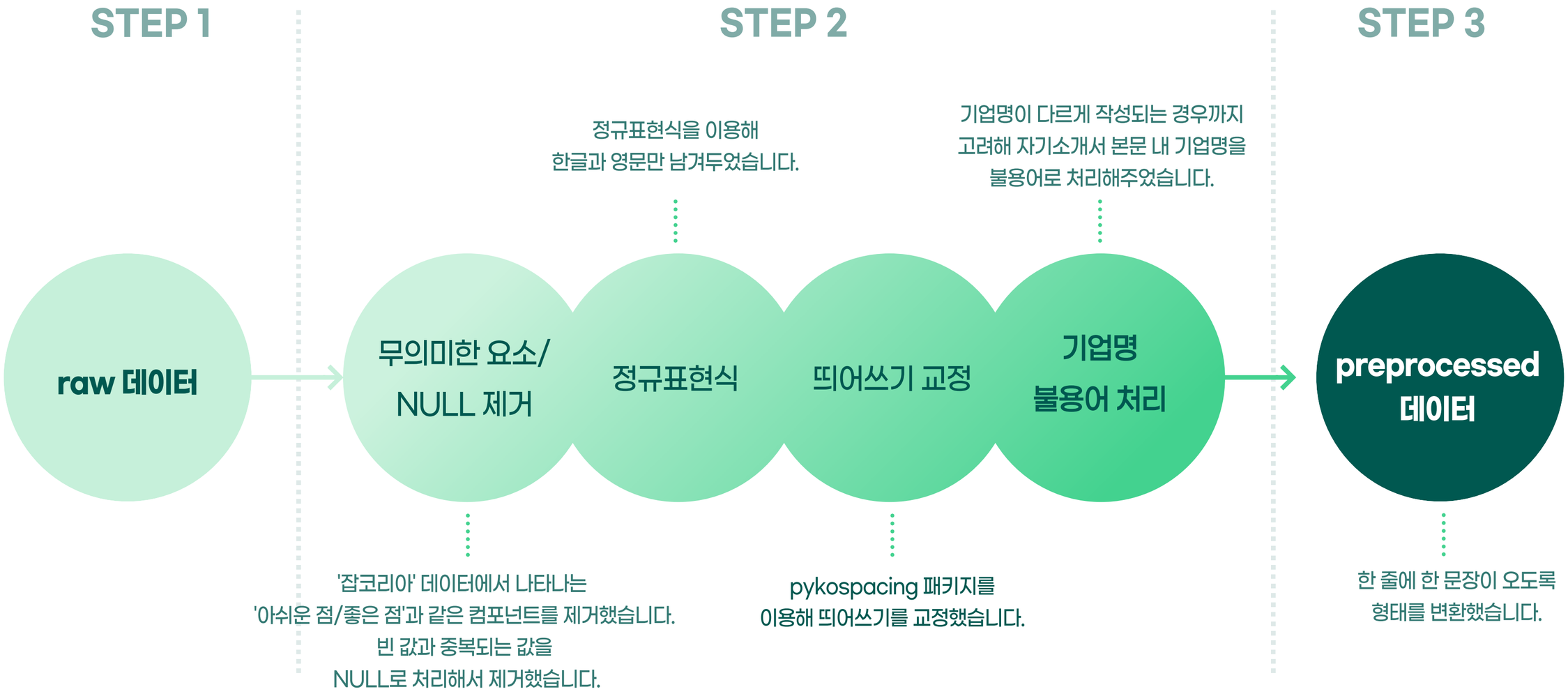
1주일간 각자 맡은 자기소개서 제공 사이트에서
데이터를 크롤링을 진행한 결과,
총 14,575건의 합격자기소개서 데이터를 얻을 수 있었습니다.

- 링커리어 데이터 8,000건
- 잡코리아 데이터 6,575건

	year	company	position	content
0	2021 상반기	현대엔지니어링	자산관리 자산설비	본인의 회사를 선택함에의 기준은 무엇이며, 왜 현대엔지니어링이 그 기준에 적합한지를 기술해 주십시오.(800~1200) - 건설기술연구원에서 지열원 히트펌프 COP을 올리기 위해 삼 주 정도 기계실에서 상주한 적이 있습니다. 시설관리자와 함께 설비를 회사란 단순히 돈을 버는 수단이 아닌 회사 성장과 함께 나 자신 자아실현의 장이 되어야 한다고 생각합니다. 현대엔지니어링은 원생지, 황강 공장 등 산업시설과 상업빌딩, 의료시설 등 일반건축에 이르기까지 넓은 사업영역을 가지고 있을 뿐만 아니라 해외에서는 개발도상국의 의료시설, 산업시설뿐만 아니라 동남아시아, 동유럽 등에 수요가 예상되며 해외에 넓은 저는 정적인 자산설비 엔지니어가 되기보단 넓은 영역에서 각각의 설비에 능통하고 최첨단 설비에 적용하여 배우며 폭넓고 높은 지식을 갖고 넓은 분야에서 성과를 올리는 동적인 자산설비 엔지니어가 되고 싶습니다. 저는 세 가지 강점이 있습니다 첫째, 자산설비 엔지니어는 자신이 담당한 설비에 대한 책임감이 있어야 한다고 생각합니다. 저는 BMTKOREA 현장실습을 할 때, 스티어링펌프개조공사중 한 챔버에 달린 두 개의 유압펌프 압력 차로 인해 오일이 누출이 발생하였고 더 큰 피해를 예방하기 둘째, 설비를 혼자 관리하는 것이 아닌 여러 인원이 관리하기 때문에 원활한 의사소통을 통해 공통적인 목표 지향이 필요합니다. 저는 인도공과대학교 연구실 인턴을 할 때 자성유체 droplet 프로젝트를 통해 현지 적응 못 하는 팀원을 위해 노력한 적이 있 셋째, 수많은 종류의 설비와 최첨단 설비에 대응하기 위해서는 입사 후에도 배움의 자세가 필요하다고 생각합니다. 건설기술연구원에서 BEMS를 위한 데이터 수집업무를 주로 하면서 부족함을 많이 느꼈고 실습이 끝나고도 BEMS 협회에서 하는 전문가양 이러한 강점을 바탕으로 자산설비 엔지니어가 되어 설비를 유지, 보수하고 또한 설비상의 문제를 해결함으로써 고객에게 쾌적한 환경을 제공할 것이고 건물 소유고객의 자산가치를 높이기 위해 노력할 것입니다. 현대엔지니어링 직무분야에 지원하게 된 이유와 선택직무에 본인이 적합하다고 판단할 수 있는 이유 및 근거를 제시해 주십시오(800~1200) - 자산설비 기술 엔지니어로서 설비 고장 분석 및 개선을 위해 데이터에 대한 통찰력과 설비 문제 파악 능력이 필요하다고 생각합니다. 다음과 같은 경험으로 필요한 역량들을 기를 수 있었습니다. [분석이 바탕 된 전력소비량 감소] 3학년 여름방학 때 건설기술연구원 건물에너지부서에서 현장실습을 하여 전력소비량 분석을 통해 전체 소비 전력량을 낮춘 경험이 있습니다. 전력 데이터 시각화 중 19년 5월부터 구조동 전력량의 기울기 증가를 확인하였습니다. 18년~19년 7월까지 월별로 로우 데이터를 통해 평균 전력량을 비교하였고 17시 이후 불필요하게 가동된 항온항습기 때문이란 것을 알았습니다. 제어 시 전력 이득을 예측하여 관리자에게 중요성을 일깨웠고 올바른 스케줄제어가 되어 결국 3달 후 전체 소비 전력량 21% 감소해 일조하였습니다. [설비 부품 교체를 통한 COP 향상] 설비 효율 유지를 위한 BEMS 모니터링 중 기계실 지열원 히트펌프의 효율이 낮아진 것을 발견하였습니다. 원인 분석을 위해 관련 논문 4편을 통해 유량, 입 출구 온도 차 등 주요 파라미터를 확인하여 18년도 데이터와 비교 분석하였습니다. 하지만 오차 범위의 정확한 인지가 어려워 문제점에 대한 방향성을 못 잡았고 이를 매뉴얼을 통해 이상적인 공급온도 13도 사수와 기계실의 설비를 직접 체크하였고 압축기가 노후화되어 냉매가 누설되고 있었던 것을 확인했습니다. Leak 이외에도 장비 수명을 고려하여 압축기를 교체하였고, 19년도 11월에 지열 COP을 4.5까지 올려 효율의 향상을 이루어 냈습니다. 이를 통해 데이터 분석 결과 활용의 중요성을 몸소 깨닫았고 설비의 관리에 있어서 미세한 의심 증상을 파악해 고장을 예방하는 필요성 터득하였습니다. 자산설비 엔지니어가 되어 데이터 파악능력을 바탕으로 설비에 측적된 자료를 활용해 설비에 일어날 문제를 예측하여 미리 대응할 것이고 또 새로운 문제에 대한 통찰력을 바탕으로 설비정지로 인한 피해가 없도록 대응할 것입니다.
1	2021 상반기	삼성생명	지점사무지원	[금융 백종원] 백종원은 여러 매체를 통해 요리에 관한 각종 정보와 비결을 공유하며 요 리의 트렌드를 이끌고 있습니다. 금융은 요리보다 더 정보의 비대칭으로 말미암은 손실이 큰 분야입니다. 저는 평소에 비전공자인 주위 사람들에 게 다양한 금융상품의 유쾌한 소통을 통한 친화력은 제 큰 장점입니다. 상인이 된 이후로 방학, 휴학기간을 이용하여 국내외의 다양한 서비스직을 경험했습니다.또한 군 로 장학생으로 근무할 당시 지원과 학생들의 소통을 주도 하였습니다. 사 실을 기반으로 서로의 상황을 이해, 사람을 잘 믿고 거절을 못 하는 습관이 업무현장에서 단점으로 작용한 적 이 있었습니다. 하나 두 차례의 금융기관 인턴을 진행하며 업무에서만큼 은 맹목적인 신뢰가 단점이 될 수 있음을 알게 되었습니다. 이후 동료나 고객의 말을 정리해서 되물어 일차적 [기업금융 전문가] 회사와 나, 나와 고객이 함께 성장하는 상생의 가치를 믿습니다. 삼성생명 은 최고의 브랜드 파워로 남다른 고객기반을 보유하고 있습니다. 그 때문 에 많은 현장경험으로 나를 한층 더 발전시킬 수 있을 것이라 생각해 지원 을 결심했습니
2	2021 상반기	삼성전자	메모리사업부 설비기	삼성전자를 지원한 이유와 입사 후 회사에서 이루고 싶은 꿈을 기술하십시오. (700자) [반도체 공정의 관심을 키울 수 있었던 마이크로 구조체 제작] 대학교 3학년 때, 전공 수업의 과정으로 '마이크로 구조체'를 제작할 기회가 있었습니다. 소형 마그네틱 스위치를 특정 자석의 세기 이상에서만 작동하도록 하는 것이 목표였습니다. 포토마스크를 디자인한 후 포토리소그래피 공정부터 식각 공정까지 진행



2. 데이터 전처리


STEP 1
STEP 2
STEP 3
raw 데이터
무의미한 요소/
NULL 제거
정규표현식
띄어쓰기 교정
기업명
불용어 처리
preprocessed
데이터
정규표현식을 이용해
한글과 영문만 남겨두었습니다.
기업명이 다르게 작성되는 경우까지
고려해 자기소개서 본문 내 기업명을
불용어로 처리해주었습니다.
'잡코리아' 데이터에서 나타나는
'아쉬운 점/좋은 점'과 같은 컴포넌트를 제거했습니다.
빈 값과 중복되는 값을
NULL로 처리해서 제거했습니다.
pykospacing 패키지를
이용해 띄어쓰기를 교정했습니다.
한 줄에 한 문장이 오도록
형태를 변환했습니다.

2. 데이터 전처리

1) raw 데이터

1. '현대엔지니어링'과 같은 기업명 있음
2. ₩n과 같은 문자 기호 있음
3. (800~1200) 같은 문자 기호/숫자 있음

	year	company	position	content
0	2021 상반기	현대엔지니어링	자산관리 자산설비	본인이 회사를 선택할때의 기준은 무엇이며, 왜 현대엔지니어링이 그 기준에 적합한지를 기술해 주십시오.(800~1200) - 건설기술연구원에서 지열원 히트펌프 COP을 올리기 위해 삼 주 정도 기계실에서 상주한 적이 있습니다. 시설관리자와 함께 설비를 유지 보수하면서 책임감을 느끼고 자신의 설비에 대한 문제를 해결하는 것에 매력을 느꼈고 건물 설비 관리에 관심을 두었습니다.\n\n회사란 단순히 돈을 버는 수단이 아닌 회사 성장과 함께 나 자신 자아실천의 장이 되어야 한다고 생각합니다.\n현대엔지니어링은 완성차, 철강 공장 등 산업시설과 상업빌딩, 의료시설 등 일반건축에 이르기까지 넓은 사업영역을 가지고 있을 뿐만 아니라 해외에서는 개발도상국의 의료시설, 산업시설뿐만 아니라 동남아시아, 동유럽 등에 수요가 예상되며 해외에 넓은 사업영역을 가지고 있습니다.\n저는 정적인 자산설비 엔지니어가 되기보단 넓은 영역에서 각각의 설비에 능통하고 최첨단 설비에 적용하여 배우며 폭넓고 높은 지식을 갖고 넓은 분야에서 성과를 올리는 동적인 자산설비 엔지니어가 되고 싶습니다.\n\n저는 세 가지 강점이 있습니다\n첫째, 자산설비 엔지니어는 자신이 담당한 설비에 대한 책임감이 있어야 한다고 생각합니다. 저는 BMTKOREA 현장실습을 할 때, 스티어링펌프개조공사중 한 챔버에 달린 두 개의 유압펌프 압력 차로 인해 오일이 누출이 발생하였고 더 큰 피해를 예방하고자 오일을 뒤집어쓰며 손으로 고정해 사장님에게 책임감을 인정받으며 칭찬받은 경험이 있습니다 \n둘째, 설비를 혼자 관리하는 것이 아닌 여러 인원이 관리하기 때문에 원활한 의사소통을 통해 공통적인 목표 지향이 필요합니다. 저는 인도공과대학교 연구실 인턴을 할 때 자성유체 droplet 프로젝트를 통해 현지 적응 못 하는 팀원을 위해 노력한 적이 있습니다. 매일 두 번을 같이 써브웨이를 먹었고, 매일 저녁에 같이 현지인들과 축구를 하였습니다. 이를 통해 저는 의사소통을 위해 상대방의 상태에 대한 인식도 중요하다고 배웠습니다. \n셋째, 수많은 종류의 설비와 최첨단 설비에 대응하기 위해서는 입사 후에도 배움의 자세가 필요하다고 생각합니다. 건설기술연구원에서 BEMS를 위한 데이터 수집업무를 주로 하면서 부족함을 많이 느꼈고 실습이 끝나고도 BEMS 협회에서 하는 전문가양성프로젝트에 참여하여 지식을 쌓을 수 있었습니다.\n\n이러한 강점을 바탕으로 자산설비 엔지니어가 되어 설비를 유지, 보수하고 또한 설비상의 문제를 해결함으로써 고객에게 쾌적한 환경을 제공할 것이고 건물 소유고객의 자산가치를 높이기 위해 노력할 것입니다.\n\n\n현대엔지니어링 직무분야에 지원하게 된 이유와 선택직무에 본인이 적합하다고 판단할 수 있는 이유 및 근거를 제시해 주십시오(800~1200) - 자산설비 기술 엔지니어로서 설비 고장 분석 및 개선을 위해 데이터에 대한 통찰력과 설비 문제 \n파악 능력

2. 데이터 전처리

2) NULL값/빈 값 처리

```
# 1) check NULL value
# NULl값이 없더라도 빈 값 유무 확인하기 위해 모든 빈 값을 NULL로 변환하고, 다시 NULL값이 있는지 확인
linkcareer.replace("", float("NaN"), inplace=True)
print(linkcareer.isnull().values.any())
```

3) 정규표현식으로 한글+영문만 남겨두기

```
# 2) 정규표현식 (영어/숫자 기호/문자 제거)
# 텍스트 정제 함수: 한글과 영문 이외의 문자는 모두 제거
def text_cleaning(text):
    hangulenglish = re.compile('[^ \.\!a-zA-Z\u3131-\u3163\uac00-\ud7a3]+') # 한글과 영문 제외한 모든 글자
    result = hangulenglish.sub('', str(text)) # 해당 글자 공백으로 대체
    return(result)
```

4) pykospacing으로 띄어쓰기 교정

```
# 3) pykospacing으로 띄어쓰기 교정
from pykospacing import Spacing
spacing = Spacing()

spaced_corpus = []
for sent in linkcareer['content_kor']:
    spaced_corpus.append(spacing(sent))

linkcareer['content_spaced'] = spaced_corpus
linkcareer.head(3)
```

2. 데이터 전처리

5) 기업명 불용어 처리 [(주)/(재)/㈜ 등 포함]

```
# 4) 기업명 불용어 처리
# (주)가 제외된 기업명 stopwords_joo에 저장
stopwords_joo = []
for i in range(len(linkcareer)):
  company_name = linkcareer['company'][i]
  joo = re.compile('[()+[주]+[)]') # ^[()+[주]+[)]와  [()+[주]+[)]$ 모두 포함
  result = joo.sub('', str(company_name)) # 해당 글자 공백으로 대체
  stopwords_joo.append(result)

stopwords_joo
```

```
# (재)가 제외된 기업명 stopwords_jae에 저장
stopwords_jae = []
for i in range(len(linkcareer)):
  company_name = linkcareer['company'][i]
  joo = re.compile('[()+[재]+[)]') # ^[()+[재]+[)]와  [()+[재]+[)]$ 모두 포함
  result = joo.sub('', str(company_name)) # 해당 글자 공백으로 대체
  stopwords_jae.append(result)

stopwords_jae
```

```
# stopwords_joo, stopwords_jae, 기존 기업명 df 합친뒤 중복 제거
stopwords_orig = list(np.array(linkcareer['company'].tolist()))
stopwords = stopwords_joo + stopwords_jae + stopwords_orig
stopwords_set = set(stopwords) # 집합set으로 변환
stopwords_list = list(stopwords_set) # list로 변환
print(len(stopwords_list)) #총 747개의 기업명 확보
```

6) 최종 데이터 형태로 변환

1. '현대엔지니어링이'과 같은 기업명이 포함된 토큰 없음
2. ₩n과 같은 문자 기호 없음
3. (800~1200) 같은 문자 기호/숫자 없음
4. 한 줄에 한 문장씩 들어가도록 변환됨

```
# 5-1) 최종 데이터 형태로 변환 (1줄에 1문장이 들어가도록) - 구두점 있는 ver
answer = []
for element in stopwords_corpus:
  answer += element
stopwords_sent = ' '.join(answer)
final_list = sent_tokenize(stopwords_sent)

final_data = []
for k in range(len(final_list)):
  final_tokens = re.sub(' \.', '.', final_list[k])
  final_data.append(final_tokens)
final_data[:3]
```

```
['본인이 회사를 선택할때의 기준은 무엇이며 왜 그 기준에 적합한지를 기술해 주십시오.',
 '건설기술연구원에서 지열원 히트펌프 COP을 올리기 위해 삼 주 정도 기계실에서 상주한 적이 있습니다.',
 '시설관리자와 함께 설비를 유지 보수하면서 책임감을 느끼고 자신의 설비에 대한 문제를 해결하는 것에 매력을 느꼈고 건물 설비 관리에 관심을 두었습니다.
```

3. GPT-3 fine-tuning

```
======== Epoch 1 / 1 ========
Training Start
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
  Batch 1,000  of  38,652. Loss: 0.14883312582969666.   Elapsed: 0:09:10.
0: 잠금 그 때문에 업무 수행에 어려움을 겪었지만 포기하지 않고 끊임없이 노력하였습니다.
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
  Batch 2,000  of  38,652. Loss: 0.05966154485940933.   Elapsed: 0:18:19.
0: 가정에서 그 결과 학년 후배들에게 도움이 될 수 있었고 그로 인하여 동기부여를 받을 수도 있었습니다.
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
  Batch 3,000  of  38,652. Loss: 0.08254604041576385.   Elapsed: 0:27:29.
0: 잔 이를 통해 조직에 대한 이해 그리고 협력의 중요성을 배울 수 있었습니다.
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
  Batch 4,000  of  38,652. Loss: 0.10620874911546707.   Elapsed: 0:36:38.
0: 죠 이후 저는 팀장으로서 책임감을 느끼고 맡은 역할에 성실한 자세를 취했습니다.
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
  Batch 5,000  of  38,652. Loss: 0.059424713253974915.   Elapsed: 0:45:47.
0: [unused250] 하지만 제가 할 수 있는 역할들은 한계가 있었고 결국에는 동아리 활동의 일정에 차질이 생기기도 했습니다.
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
  Batch 6,000  of  38,652. Loss: 0.10559176653623581.   Elapsed: 0:54:56.
0: 번거로 학점. 에 대하여 상세히 기술해 주십시오 최소 자 최대 글자수작성안개를 걷어내다 년 월부터 개월간 공학과 학생들 중
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
  Batch 7,000  of  38,652. Loss: 0.13449738919734955.   Elapsed: 1:04:07.
0: 뤟 이런 경험을 바탕으로 저는 새로운 방식을 익히는 데 도움이 되는 되고 싶습니다.
Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
```

1. huggingface 모델 불러오기

Huggingface에서 제공하는 tokenizer / GPT-3 pretrained model 불러오기 (eos, bos, pad와 같은 special token 추가)

2. 합격자기소개서로 pytorch 데이터셋 생성

크롤링/전처리된 합격자기소개서 데이터에 BERTtokenizer 적용한 뒤, pytorch 데이터셋 생성

3. 하이퍼파라미터 설정

- AdamW 옵티마이저로 사용
- learning rate scheduler 사용 (학습이 경과함에 따라 lr 변동)
- 기타 하이퍼파라미터 임의로 설정 (추후 튜닝 진행)

4. 모델 fine-tuning

train, validation loss와 elapsed time 기록

4. 추후 일정


100
80
60
40
20
0
01
데이터 수집/전처리
1. 텍스트 생성 분야 아이디어 확정 ☑
2. 데이터 크롤링 ☑
3. 데이터 전처리 ☑
4. 공유 드라이브/팀 Github 생성 ☑
02
KoGPT-2 모델
1. 기존 코드 공부, 쉐도잉 ☑
2. KoGPT-2 기반 fine-tuning ☑
3. 하이퍼파라미터 튜닝
03
텍스트 요약 모델
1. TextRank/Attention 기존 코드 공부
2. 모델 적용
3. 하이퍼파라미터 튜닝
04
최종 발제
목표:
키워드/특정 문장 입력시
자기소개서 문장 생성
* 합격할 수 있을 정도로 잘 쓰기 *
10
15
20
50
35
45
60
50
90
1주차 현황
2주차 현황

THANK YOU

감사합니다.

질문/피드백 부탁드립니다!

2021.08.26