2021-여름 KUBIG NLP 세션 2021.09.02

기소개서

는 만큼 보인다는 말은 배워가야 한다는 것을 일깨워줍니다.

통해 알게 된 게임이론은 … E 시간에 탐구해보기로 … 성 후 PPT 대회에 참가하며 … 후 경제, 경영에 대한 심도 있는 탐구를 … | 주최하는 멘토링 캠프에서 한 강의도 할 수 있었습니다. 지원 동. 관심 분.

했는지 구

학문적궁

회과 연계

하-학과에

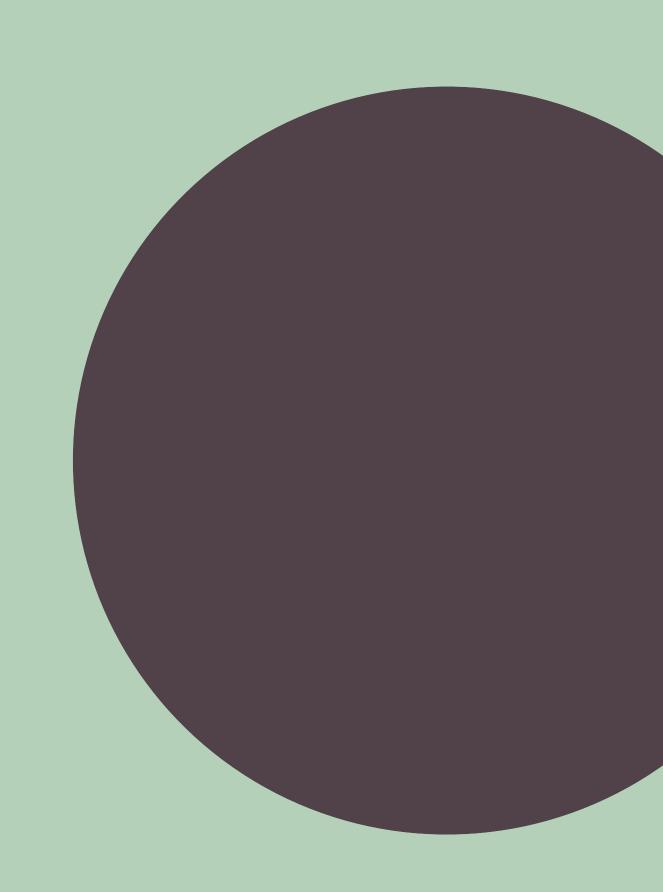
왜 서술함

!은 늘 학생이라고 생각하고 … 아하고, 하기 싫어하는 일들을 준히 실천하였습니다. 수업 시 전등 및 에어컨을 끄는 등의 너지 지킴이 활동과 ….

자기소개서생성 & 한 줄 요약

MPTI (생성2팀)

13기 기다연, 이현규, 12기 전지우



목 차

01

주제 소개

02

데이터 크롤링/

03

텍스트생성

GPT-2 fine-tuning

04

텍스트 요약 LexRank

- 아이디어 출현 배경
- 기대효과

- 링커리어 데이터 크롤링
- 잡코리아 데이터 크롤링
- 데이터 전처리 방안

- GPT란?
- fine-tuning이란?
- 하이퍼파라미터 튜닝
- 텍스트 생성

- LexRank이란?
- 텍스트 요약
- 추후 과제

주제 소개 1-1. 아이디어 출현 배경

언제까지 자기소개서로 서탈할래!

입력받은 첫 문장, 키워드를 이용해 자동으로 텍스트를 생성하고 한 줄 요약을 보여드립니다.

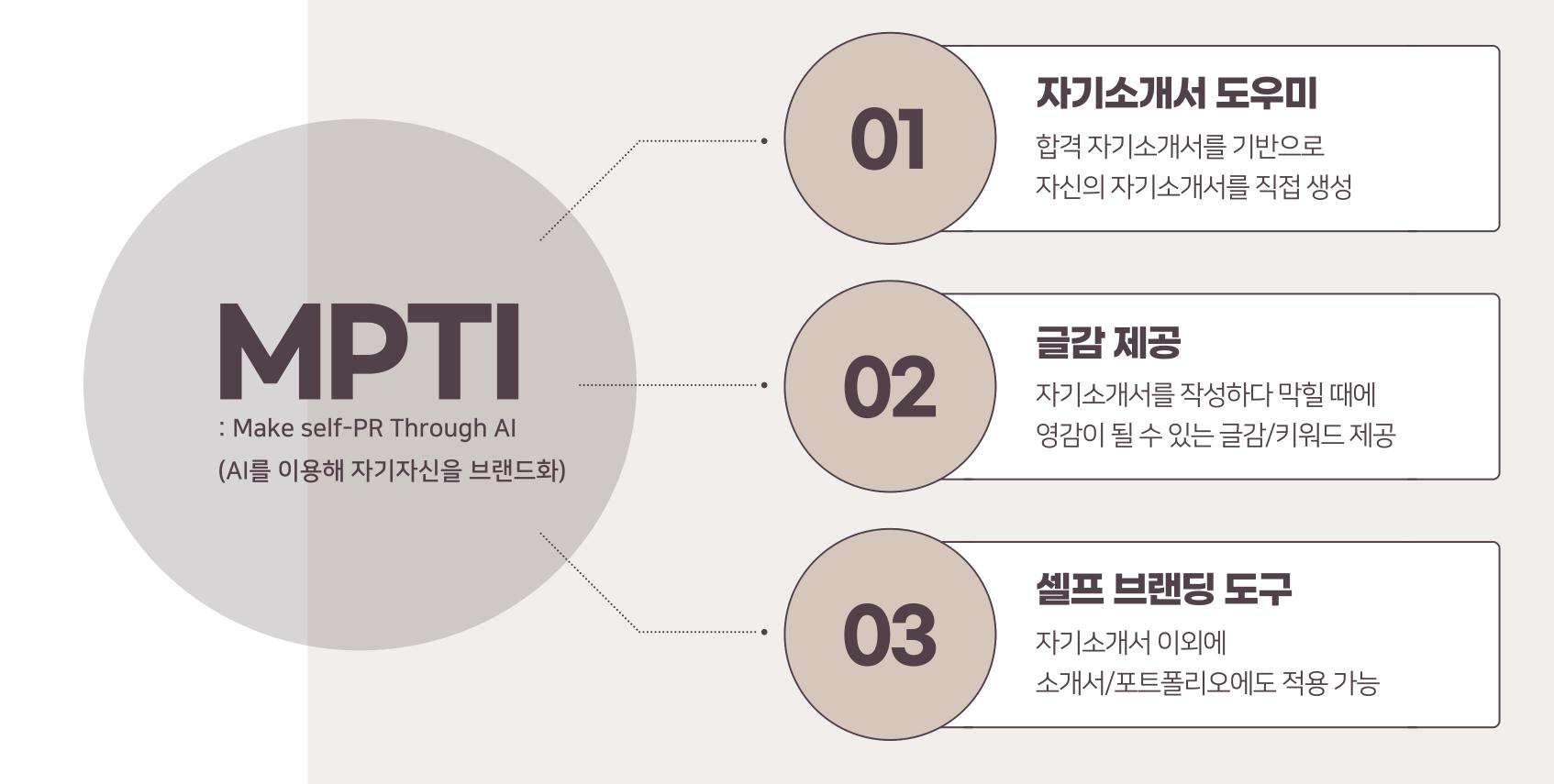
대학 입시부터 학회 지원, 취업까지, 자기소개서는 자신을 표현할 수 있는 가장 효 과적인 수단으로 떠오르고 있습니다. 때문에 기업체나 대학에서는 자기소개서 내 용을 더욱 더 중요한 평가기준으로 두고 있습니다.

합격을 부르는 자기소개서를 단번에 작성할 수 있다면 가장 베스트겠지만, 모두가 그런 자기소개서를 작성하기는 어려우므로 AI의 힘을 빌려봅시다!



저희는 합격자기소개서 데이터를 기반으로 키워드나 첫 줄 입력시 자기소개서를 자동 생성하고, 각 문항별로 한 줄로 요약해주는 모델을 만들고자 합니다.

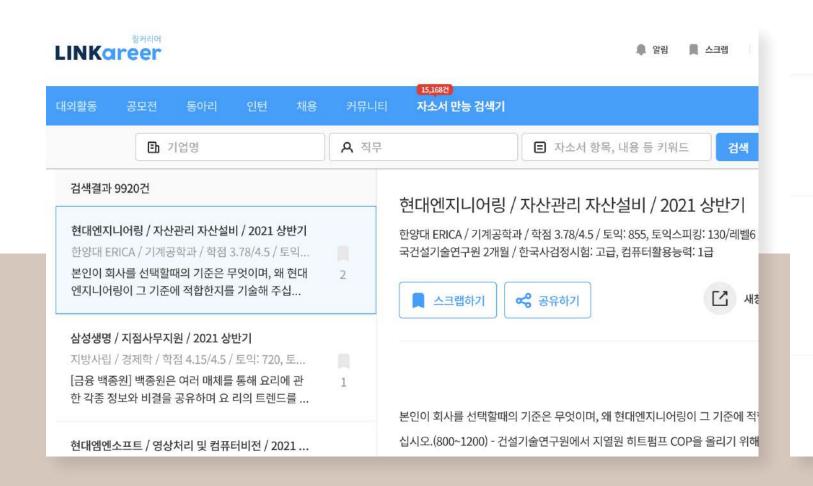
1-2. 기대효과

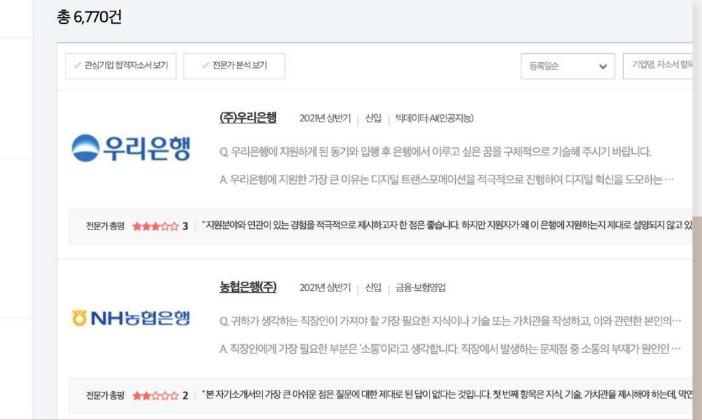




데이터 크롤링/전처리

2-1. 데이터 크롤링





링거리어 (8,078건)

잡코리아 (6,575건)

크롤링이 가능한 합격자기소개서 제공 사이트 중 '링커리어'와 '잡코리아'에서 데이터를 크롤링 했습니다. 2명의 팀원은 링커리어을 절반씩 나누어서, 1명의 팀원은 잡코리아에서 크롤링 작업을 진행했습니다. 각 자기소개서마다 **합격연도(2021 상반기), 기업명, 직무명, 자소서 본문**을 크롤링해 csv 파일로 저장했습니다.

2-2. 데이터 전처리

'아쉬운 점/좋은 점'과 같은 컴포넌트를 제거했습니다.

빈 값과 중복되는 값을

NULL로 처리해서 제거했습니다.



이용해 띄어쓰기를 교정했습니다.

기업명이 다르게 작성되는 경우까지

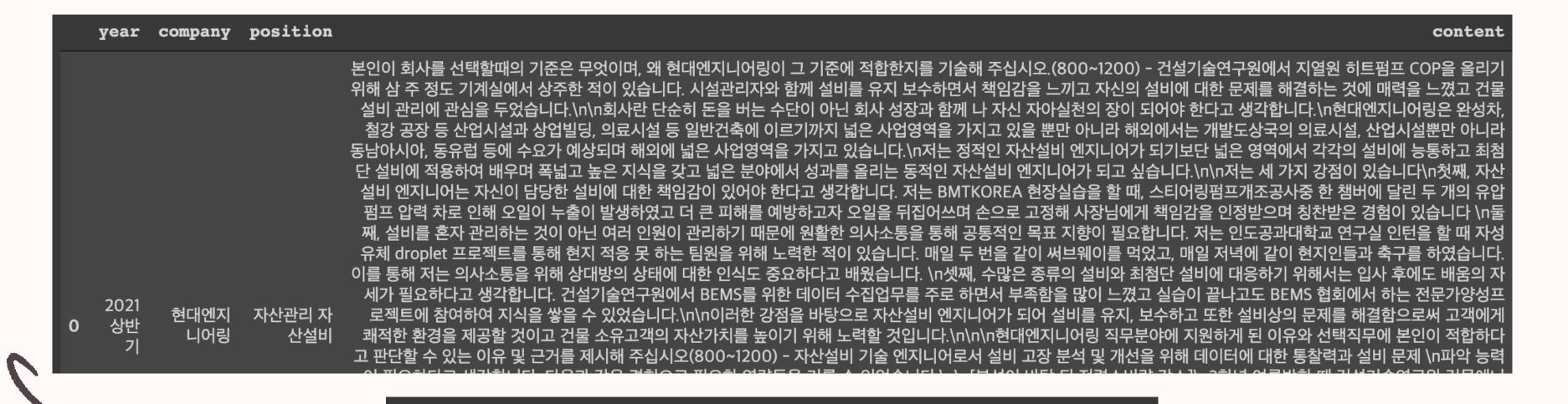
예) '글자수', '단락', '이내', '이상', '최소', '최대',

'입력가능', '공백', '포함', '작성요령', '소제목',

'신명조', '진하게', '글자크기', '글자', 'BYTE',

'byte', 'Byte', 'Bytes'

2-2. 데이터 전처리



) 본인이 회사를 선택할때의 기준은 무엇이며 왜 그 기준에 적합한지를 기술해 주십시오 ...

content

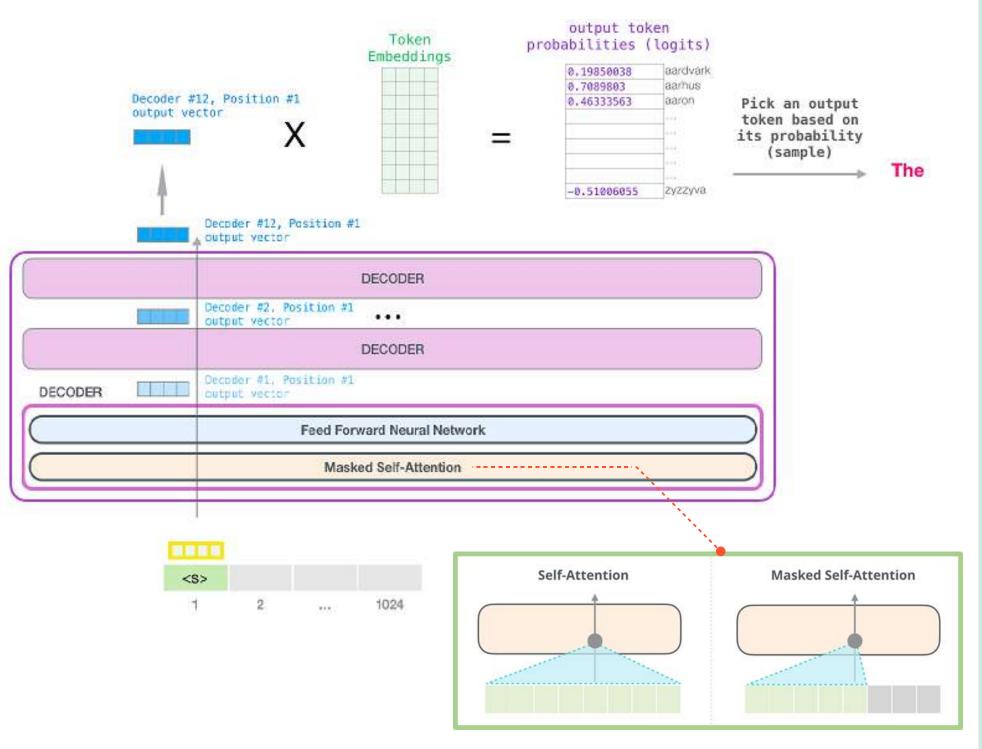
1 금융 백종원 백종원은 여러 매체를 통해 요리에 관한 각종 정보와 비결을 공유하며 요...

기존 데이터 중 NULL 값 제거, 한글과 영문만 남기고, 띄어쓰기 교정, 기업명을 불용어로 처리해 자기소개서 본문에서 제거했습니다. 최종 데이터는 한 줄에 하나의 자기소개서가 나오는 리스트 형태로 변환했습니다. 총 데이터의 행 수는 14,653개입니다.

03

GPT-2 fine-tuning

3-1. GPT(Generative Pretrained Transformer)란?



GPT란?

: 트랜스포머(Transformer) 체계에서 디코더(decoder)만 사용해 데이터의 분포를 모델링하는 생성 모델

Step

- 1. 각 단어의 순서들을 고려한 임베딩 행렬을 입력 벡터로 넣습니다.
- 2. 각 디코더셀의 self-attention 과정을 거친 뒤 신경망 레이어를 통과합니다.

디코더 블록을 거친 최종 결과물은 입력값에 대한 최종 셀프 어텐션값입니다.

- 3. 출력값을 우리가 가진 임베딩 벡터와 곱해줍니다.
- 그 결과값은 각 단어가 다음 단어로 등장할 확률값으로 출력됩니다.
- 4. 이 중 가장 확률값이 높은 것을 출력하며 이 값은 곧 다음 입력값이 됩니다.

X Masked Self-Attention

- 디코더 과정은 타깃 단어 이후의 단어를 보지 않고 예측해야 하므로 타깃 단어 뒤의 단어들이 Self Attention에 영향을 주지 않도록 마스킹 처리를 합니다.
- → 따라서 GPT 모델은 해당 단어 이후에 있는 맥락정보들을 이용할 수 없으나 다음 단어의 예측 능력은 뛰어나므로 <u>문장 생성</u>에 강점을 지닙니다.

(GPT-1, GPT-2, GPT-3의 차이는 바로 사이즈!

→ 기본적인 모델 구조나 학습 원리는 모두 같으나 모델과 데이터의 크기가 커지고 성능이 높아지고 있습니다.)

3-2. fine-tuning이란?



Pre-Training

- 특정한 목적 태스크가 없는 대량의 자연어 데이터셋(Unlabeled corpora)을 언어 모델에 학습시킵니다.
- ko-GPT 모델은 국내 주요 커머스 리뷰, 모두의 말뭉치, 위키피디아 등의 텍스트 데이터 70GB(127억 개 토큰)를 학습했습니다.

Fine-Tuning

- 모델의 파라미터를 미세하게(정교하게) 조정하는 과정
- Pre-training을 마친 모델에 태스크에 맞는데이터셋(Labeled corpora)을 추가로 학습시켜 파라미터를 업데이트합니다.
- 모델의 웨이트를 다운스트림하는 태스크에 대해 미세 조정을 실시합니다.

3-3. 하이퍼마라미터 튜닝

1. batch_size

- 메모리의 한계로 한 번의 epoch에 모두 넣지 않는데,
- 이 때 batch_size로 데이터를 나눠서 iteration마다 넣었습니다.

2. optimizer

- 일부의 경우, 일반화 효과를 덜 받게 되는 Adam 대신 AdamW 방식을 사용
- 가중치 업데이트 식에 weight decay 텀을 추가
- learning rate schedule: 매 가중치 업데이트마다 learning rate를 일정 비율 감소시킴

-

최종 하이퍼파라미터 조합

- batch_size = 2
- epochs = 2
- learning_rate = 5e-4
- warmup_steps = 1e2
- sample_every = 1,000
- epsilon = 1e-8

3. learning_rate

- Local minimum에 효율적으로 도달할 수 있도록 하는 값으로, SGD에서는 기울기에 learning_rate를 곱해 다음 지점을 결정
- 0.1에서 0.0005 사이에서 0.0005를 선택

4. warmup_steps

- Adaptive learning rate 방식에서 학습 초기에 최적이 아닌 local optima에 수렴하게 되는 문제의 해결책
- warmup_steps가 10, learning_rate가 0.01인 경우,
- 처음 10 step 동안 0.001, 0.002, 0.003 등으로 조금씩 learning_rate를 변화시킴

3-4. 텍스트 생성 예시

사전 학습된 GPT-2에 합격자기소개서 데이터를 입력해 fine-tuning 시켜줬습니다.

 coverletter_generator('저는 데이터 분석 능력에 능합니다. ')

 Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.

 tensor([[42000, 14156, 16642, 15626, 19082, 8008, 3066, 14015, 2016, 3]])

 0: 저는 데이터 분석 능력에 능합니다. 대학교 년 때 ㅇㅇ의 데이터베이스를 만들고 구축하며 기초 지식을 쌓았고 또한 이를

 1: 저는 데이터 분석 능력에 능합니다. 특히 통계 프로그램을 통해서 수치화된 그래프들을 보며 데이터를 모으고 분석하는 일

 2: 저는 데이터 분석 능력에 능합니다. 전공수업을 통해 데이터분석 및 설계 수업을 들었고 데이터 분석에 필요한 기본요소가

저는 데이터 분석 능력에 능합니다.

특히 통계 프로그램을 통해서 수치화된 그래프들을 보며 데이터를 모으고 분석하는 일을 하고 싶습니다. 또한 여러 분야의 데이터분석프로그램들과 인사이트를 공유하고 협력해 나가는 것을 즐기고 있습니다. 이러한 저의 강점은 제가 하는 일에 자부심을 가질 수 있다는 믿음을 가지게 해주었습니다. 지원동기와 입사 후 포부미래를 선도하는 데이터 분석 전문가가 되고 싶습니다.

제가 한 분야 최고가 되기 위해서는 학부에서의 프로그래밍 실력 이외에도 다양한 분야의 지식을 함양하는 것과 그것을 활용할 줄 아는 소통 능력이 필요하다고 생각합니다. 이를 위해서 공학인증을 통해 경영 경제 관련 자격증을 취득하였고 경영경제분야에서 it분야 빅데이터 활용능력 의사소통능력을 키우는 것을 목표로 하고 있습니다. 그리고 이를 위해서 컴퓨터공학도로서 네트워크 관련 기본 지식을 쌓기 위해 정보처리기사와 전산구조해석사 자격증을 취득하였습니다. 특히 정보보안기사를 준비하면서 데이터 분석 및 분석능력뿐만 아니라 보안 문제를 해결하는 능력을 키웠습니다. 뿐만 아니라 네트워크 과목에서 배운 내용들과 실제 네트워크상에서 사용되는 프로토콜들의 사용법을 익히는 것과 동시에 정보보안기사와 전산구조해석사를 준비하면서 언어적인 능력도 더욱 향상시킬 수 있었습니다. 따라서 입사 후에는 다양한 분야의 지식과 실무 경험을 쌓아 제가 원하는 분야에 전문적인 전문가로 성장하고자 합니다. 특히 장기적으로는 it 분야의 전문적인 지식을 지속적으로 학습하고 그것을 바탕으로 차세대 인프라 사업을 이끌어나가는 데 이바지할 것입니다.

3-4. 텍스트 생성 예시 (이어서)

```
coverletter_generator('저의 강점은 소통입니다.')

Setting `pad_token_id` to `eos_token_id`:3 for open-end generation.
tensor([[42000, 16972, 38964, 8078, 20312, 13992, 2016, 3]])
0: 저의 강점은 소통입니다. 제 장점은 항상 상대방의 이야기를 잘 들어주고 공감해주려는 태도입니다. 그래서 다른 사람들에.
1: 저의 강점은 소통입니다. 소통능력은 다양한 사람들과의 소통을 통해 좋은 방향으로 나아갈 수 있다고 생각합니다. 저는 더
2: 저의 강점은 소통입니다. 대학교 학년 때 교환학생을 다녀왔습니다. 이 경험으로 저는 다양한 사람들을 만날 수 있었고 그
```

저의 강점은 소통입니다.

소통능력은 다양한 사람들과의 소통을 통해 좋은 방향으로 나아갈 수 있다고 생각합니다. 저는 대학생활 중 많은 사람을 만나며 소통 능력을 길렀습니다. 이러한 경험을 바탕으로 일원이 되어 보다 좋은 서비스를 제공하고 싶습니다. 이를 위해 먼저 사람들과 원활히 소통할 방법을 찾았습니다. 우선 매주 월요일에 모여서 학교 축제를 홍보하는 동아리 활동을 했습니다.

특히 년 동안 꾸준히 활동해온 동아리는 다양한 사람들 그리고 서로 간의 정보를 공유하는 창구가 되었습니다. 또한 동아리 부회장으로서도 소통하며 많은 사람들과 친해졌습니다. 이를 활용하여 더욱 활성화된 동아리가 될 수 있도록 노력하겠습니다.

마지막으로 다양한 사람들을 만났습니다. 년 때 한 친구가 저희 동아리를 소개하지 않았습니다.

이 사실을 알고 있었던 저는 동아리 홍보를 하기 위해 직접 연락해 도움을 요청했습니다.

동아리원들의 연락처를 알려주고 함께 이야기 나눌 수 있는 자리를 마련해주었습니다. 결국 두 친구와 잘 어울리면서 동아리를 알릴 수 있었습니다. 입사 후에도 고객과 마주하는 영업 사원으로서 열린 마음으로 동료들과 협력하고 고객의 소리를 듣고 문제점을 찾아내고 개선점을 만들어가는 신입사원이 되겠습니다.

3-4. 텍스트 생성 예시 (이어서)



3-4. 텍스트 생성 예시 (이어서)

이제 여러분이 원하시는 문장으로 돌려보세요!

Google Colab 링크: https://colab.research.google.com/drive/1D5DCA-ulr_J_h6H12ZgcotHRfmlzS0BA?usp=sharing

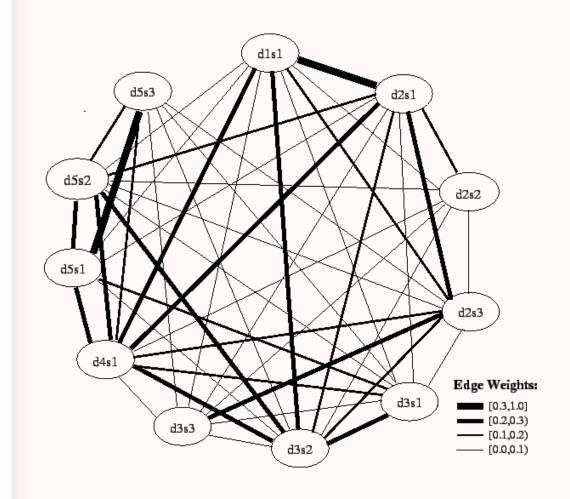
LexRank이란?

LexRank

Google의 검색엔진에 적용되었던 PageRank라는 알고리즘을 문서 요약에 적용한 방법
→ 문서 내 문장과 문장의 관계를 그래프로 나타내기 위하여 TF-IDF Cosine Similarity를 사용하여
문장 간의 관계를 그래프로 표현하는 방법입니다.

- 1 Graph-based Extractive Summarization → 문서 내 문장을 하나의 노드로 해 그래프 표현
- 2 Centrality-based Sentence Salience로 문서에서 중요 문장을 추출

 → 클러스터에서 가장 중심이 되는 문장을 식별하는 작업과 동일함,
 즉, 다른 많은 문장들과 유사한 문장이 문서 클러스터의 중심(central or salient)에 위치한다고 가정
- 3 각 문장 → STM(Sentence Term Matrix) 형태로 표현 각 문장의 단어 → TF-IDF로 표현 두 문장간 유사도 →코사인 유사도(idf-modified cosine-similarity)로 정의



4-2. 텍스트 요약 예시

```
# 생성된 자기소개서 불러오기
coverletter = '''
저는 데이터 분석 능력에 능합니다.
특히 통계 프로그램을 통해서 수치화된 그래프들을 보며 데이터를 모으고 분석하는 일을 하고 싶습니다.
또한 여러 분야의 데이터분석프로그램들과 인사이트를 공유하고 협력해 나가는 것을 즐기고 있습니다.
이러한 저의 강점은 제가 하는 일에 자부심을 가질 수 있다는 믿음을 가지게 해주었습니다.
지원동기와 입사 후 포부미래를 선도하는 의 데이터 분석 전문가가 되고 싶습니다.
제가 한 분야 최고가 되기 위해서는 학부에서의 프로그래밍 실력 이외에도 다양한 분야의 지식을 함양하는 것과 그것을 활용할 줄 아는 소통 능력이 필요하다고 생각합니다.
이를 위해서 공학인증을 통해 경영 경제 관련 자격증을 취득하였고 경영경제분야에서 it분야 빅데이터 활용능력 의사소통능력을 키우는 것을 목표로 하고 있습니다.
그리고 이를 위해서 컴퓨터공학도로서 네트워크 관련 기본 지식을 쌓기 위해 정보처리기사와 전산구조해석사 자격증을 취득하였습니다.
특히 정보보안기사를 준비하면서 데이터 분석 및 분석능력뿐만 아니라 보안 문제를 해결하는 능력을 키웠습니다.
뿐만 아니라 네트워크 과목에서 배운 내용들과 실제 네트워크상에서 사용되는 프로토콜들의 사용법을 익히는 것과 동시에 정보보안기사와 전산구조해석사를 준비하면서 언어적인 능력도 더욱 향상시킬 수 있었습니다.
따라서 입사 후에는 다양한 분야의 지식과 실무 경험을 쌓아 제가 원하는 분야에 전문적인 전문가로 성장하고자 합니다.
특히 장기적으로는 it 분야의 전문적인 지식을 지속적으로 학습하고 그것을 바탕으로 차세대 인프라 사업을 이끌어나가는 데 이바지할 것입니다.
LexRank.summarize(coverletter)
num_line = 0.1 # 1줄만 나오도록 설정
summaries = LexRank.probe(num_line)
for summary in summaries:
   print(summary)
제가 한 분야 최고가 되기 위해서는 학부에서의 프로그래밍 실력 이외에도 다양한 분야의 지식을 함양하는 것과 그것을 활용할 줄 아는 소통 능력이 필요하다고 생각합니다
```

LexRank **알고리즘**을 이용해 위에서 생성한 자기소개서를 한 줄로 요약했습니다. 클러스터에서 가장 중심이 되는 문장을 문서에서 중요 문장으로 간주해 해당 문장으로 요약함을 알 수 있습니다.

이 데이터 전처리 구체화

"garbage in, garbage out" 이라는 말처럼 양과 질의 데이터가 입력되어야 GPT-2 fine-tuning이 제대로 이뤄질 것 같습니다.

팀원들끼리 여러차례 논의해 확정한 데이터 전처리 과정은 모두 포함을 시켰지만, **시간적인 제약 때문에 데이터 전처리 과정을 더욱 구체화하지는 못했습니다.**

- 기업명을 자기소개서 본문에서 제거한 것과 같이 직무명 제거
- 000 등으로 마스킹 처리된 개인정보 부분 (작성자 이름, 동아리 이름 등) 하나하나 처리
- 맞춤법 교정

02 fine-tuning 과정

개인 노트북의 GPU memory와 리소스가 부족해 fine-tuning을 기존에 계획했던 대로 '한 행에 한 문장'씩 시키지 못했습니다. fine-tuning 과정에서 계속 'CUDA Out of Memory' 에러를 경험하며 데이터의 입력 형태를 바꾸기로 했습니다. 따라서 저희는 한 행에 한 개의 자기소개 서'가 오도록 변경해 fine-tuning을 진행했습니다. 또한, GPU 성능을 고려해 epoch을 3회 이상 돌리지 못한 아쉬움이 남습니다.

생성된 자기소개서를 보면 특정 직무와 관련되게 내용이 나오는 경향을 보실 수 있습니다. 각 자기소개서별로 <직무>를 라벨링한 뒤, 해당 라벨까지 학습해 직무 키워드에 기반한 자기소개서 답변을 생성할 수도 있을 것 같습니다.

2021-여름 KUBIG NLP 세션 2021.09.02

기소개서

는 만큼 보인다는 말은 배워가야 한다는 것을 일깨워줍니다.

통해 알게 된 게임이론은 … E 시간에 탐구해보기로 … 성 후 PPT 대회에 참가하며 … 후 경제, 경영에 대한 심도 있는 탐구를 … 시 주최하는 멘토링 캠프에서 한 강의도 할 수 있었습니다.

!은 늘 학생이라고 생각하고 … 아하고, 하기 싫어하는 일들을 준히 실천하였습니다. 수업 시 전등 및 에어컨을 끄는 등의 너지 지킴이 활동과 …. 지원 동. 관심 분니 했는지 구 학문적 궁. 학과 연계 학·학과에 해 서울함

감사합니다

피드백 환영합니다!

