(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0064522
(43) 공개일자 2025년05월09일

(51) 국제특허분류(Int. Cl.)
G06Q 10/10 (2023.01) *G06F 16/33* (2025.01)
G06F 40/205 (2020.01) *G06F 40/279* (2020.01)
G06F 40/30 (2020.01) *G06N 20/20* (2019.01)
G06N 3/0475 (2023.01) *G06N 3/096* (2023.01)
G06Q 10/06 (2012.01) *G06V 30/14* (2022.01)
(52) CPC특허분류
G06Q 10/1053 (2023.01)
G06F 16/3334 (2019.01)
(21) 출원번호 10-2023-0171430
(22) 출원일자 2023년11월30일
심사청구일자 2023년11월30일
(30) 우선권주장
1020230146735 2023년10월30일 대한민국(KR)

(71) 출원인
동서대학교 산학협력단
부산광역시 사상구 주례로 47(주례동, 동서대학교)
(72) 발명자
박승민
부산 사상구 주례로 93
김홍찬
서울특별시 성북구
(뒷면에 계속)
(74) 대리인
김종석

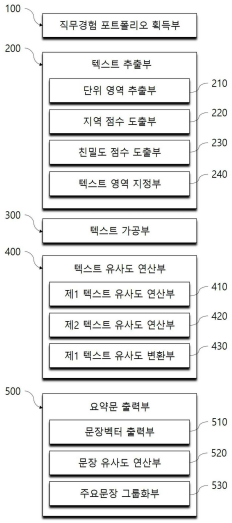
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 **인공지능 기술기반 직무경험 큐레이팅 시스템**

(57) 요 약

본 발명은 인공지능 기술기반 직무경험 큐레이팅 시스템에 관한 것으로, 보다 구체적으로 지원자 단말로부터 제공된 직무경험 포트폴리오를 획득하는 직무경험 포트폴리오 획득부, 상기 직무경험 포트폴리오로부터 텍스트를 인식한 후 추출하는 텍스트 추출부, 상기 텍스트를 형태소 단위로 토큰화(Tokenization)하고, 토큰화 텍스트를 가공하는 텍스트 가공부, 평가자 단말로부터 제공된 키워드와 상기 토큰화 텍스트 간의 텍스트 유사도를 연산하는 텍스트 유사도 연산부 및 상기 텍스트 유사도가 임계값 이상인 상기 토큰화 텍스트를 포함하는 주요문장을 문맥에 따라 배열하고, 키워드별 요약문을 출력하는 요약문 출력부를 포함하는 인공지능 기술기반 직무경험 큐레이팅 시스템에 관한 것이다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/3335 (2019.01)
G06F 16/3344 (2019.01)
G06F 40/205 (2020.01)
G06F 40/279 (2020.01)
G06F 40/30 (2020.01)
G06N 3/0475 (2023.01)
G06N 3/096 (2023.01)
G06Q 10/06398 (2013.01)
G06Q 10/10 (2023.01)
(72) 발명자
김채호
경상남도 양산시 물금읍 범어로 77
김준희
부산광역시 북구 효열로 265
박민건
부산광역시 북구 의성로 55

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193872
과제번호	2019-0-01817-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신창의인재양성
연구과제명	SW중심대학(동서대)
기 여 율	1/1
과제수행기관명	동서대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

지원자 단말로부터 제공된 직무경험 포트폴리오를 획득하는 직무경험 포트폴리오 획득부;

상기 직무경험 포트폴리오로부터 텍스트를 인식한 후 추출하는 텍스트 추출부;

상기 텍스트를 형태소 단위로 토큰화(Tokenization)하고, 토큰화 텍스트를 가공하는 텍스트 가공부;

평가자 단말로부터 제공된 키워드와 상기 토큰화 텍스트 간의 텍스트 유사도를 연산하는 텍스트 유사도 연산부; 및

상기 텍스트 유사도가 임계값 이상인 상기 토큰화 텍스트를 포함하는 주요문장을 문맥에 따라 배열하고, 키워드별 요약문을 출력하는 요약문 출력부;를 포함하는 인공지능 기술기반 직무경험 큐레이팅 시스템.

청구항 2

제1항에 있어서,

상기 텍스트 추출부는,

CRAFT 모델을 이용하여 텍스트가 비인식되는 이미지 형식의 직무경험 포트폴리오에서 하나의 숫자 또는 철자(spelling)를 포함하는 단위 영역을 추출하는 단위 영역 추출부;

상기 단위 영역의 경계가 명확해지도록 상기 단위 영역을 이루는 각 픽셀이 하나의 숫자 또는 철자(spelling)의 중심일 확률인 지역 점수(Region Score)를 도출하는 지역 점수 도출부;

상기 단위 영역의 연속성이 확인되도록 서로 다른 두 개의 단위 영역에 대한 친밀도 박스(Affinity box)를 형성하고, 상기 친밀도 박스로부터 두 개의 단위 영역의 중심일 확률인 친밀도 점수(Affinity Score)를 도출하는 친밀도 점수 도출부; 및

상기 지역 점수와 상기 친밀도 점수를 이용하여 상기 단위 영역으로부터 상기 텍스트 영역을 지정하는 텍스트 영역 지정부;를 포함하고,

기 저장된 딥러닝 모델에 상기 텍스트 영역이 입력됨으로써, 상기 텍스트 영역으로부터 텍스트를 추출하는 것을 특징으로 하는 인공지능 기술기반 직무경험 큐레이팅 시스템.

청구항 3

제1항에 있어서,

상기 텍스트 유사도 연산부는,

레벤슈타인 거리(Levenshtein distance) 기법을 이용하여 편집거리에 따른 연산횟수를 제1 텍스트 유사도로 출력하는 제1 텍스트 유사도 연산부;

시퀀스 매처(Sequence Matcher) 기법을 이용하여 상기 키워드와 상기 토큰화 텍스트를 각각 바이트(byte) 단위로 변환한 후 제2 텍스트 유사도를 출력하는 제2 텍스트 유사도 연산부; 및

상기 제1 텍스트 유사도를 바이트 단위로 변환하는 제1 텍스트 유사도 변환부;를 포함하고,

변환된 제1 텍스트 유사도와 상기 제2 텍스트 유사도를 평균하여 상기 텍스트 유사도를 도출하는 것을 특징으로 하는 인공지능 기술기반 직무경험 큐레이팅 시스템.

청구항 4

제3항에 있어서,

상기 제1 텍스트 유사도 변환부는,

하기 [수학식 2]를 이용하여 상기 제1 텍스트 유사도를 바이트 단위로 변환하는 것을 특징으로 하는 인공지능 기술기반 직무경험 큐레이팅 시스템.

[수학식 2]

$$S(A,B) = 1 - \frac{d(A,B)}{\max(len(A), len(B))}$$

여기서, S(A,B)는 바이트 단위로 변환된 제1 텍스트 유사도이고, d(A,B)는 상기 연산횟수이고, max(len(A), len(B))는 상기 키워드의 문자열 길이(len(A))와 상기 토큰화 텍스트의 문자열의 길이(len(B)) 중 가장 긴 길이를 추출하는 함수이다.

청구항 5

제1항에 있어서,

상기 요약문 출력부는,

상기 주요문장이 기 학습된 SBERT 모델에 입력됨으로써, 문장벡터를 출력하는 문장벡터 출력부;

피어슨 유사도 기법을 이용하여 상기 문장벡터 간 문장 유사도를 연산하는 문장 유사도 연산부; 및

상기 문장 유사도가 임계값 이상인 주요문장을 그룹화하는 주요문장 그룹화부;를 포함하는 것을 특징으로 하는 인공지능 기술기반 직무경험 큐레이팅 시스템.

청구항 6

제5항에 있어서,

상기 요약문 출력부는,

키워드별 그룹화된 주요문장을 생성형 인공지능 모델에 입력함으로써, 지원자의 직무수행능력 여부와 직무적합성 여부에 대한 큐레이팅 문장을 더 출력하는 것을 특징으로 하는 인공지능 기술기반 직무경험 큐레이팅 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공지능 기술기반 직무경험 큐레이팅 시스템에 관한 것이다.

배 경 기 술

[0002] 최근 채용 프로세스에 많은 변화가 일어났다. 기업은 인재 채용에 있어 과거 학벌이나 자격증에 의존하던 때와 달리, 이제 실전에서 실력을 발휘할 수 있는 정도를 판단하는 것이 중요하다. 이를 위해 기업은 이력서 및 자기소개서와 함께 직무경험 포트폴리오를 지원자들에게 요구하고 있고, 기업의 인사담당자가 이력서, 자기소개서 및 직무경험 포트폴리오를 종합적으로 평가한 후 해당 지원자가 기업 및 직무와 적합하지를 판단하고 당락을 결정하는 추세이다.

[0003] 이력서는, 취업의 목적으로 지원자의 학력, 경력 및 스펙 등이 온라인 또는 서면으로 객관적인 사실이 기록된 서류로, 지원자의 인적사항, 학력사항, 병력사항, 자격사항, 직무관련 경력사항 및 연수사항, 가족사항 등을 포함할 수 있다. 자기소개서는, 지원자의 성격과 태도, 지원동기와 직무역량 등에 대해 주관적인 경험을 바탕으로 작성된 서류이다. 직무경험 포트폴리오는, 객관적인 사실 및 주관적인 경험이 다양한 방식으로 작성된 서류로, 텍스트 기재방식으로 한정되어 있는 이력서 및 자기소개서에 비해, 텍스트, 이미지, 영상, 사진 등 특정 기재방식에 한정되지 않고 다양한 기재방식으로 직무경험 및 지원역량이 표현될 수 있다. 예컨대, 디자이너, 프로그래머, IT 기획자, 마케터 직무 지원 시 직무경험 포트폴리오는 이전에 수행한 프로젝트 및 작품들과 프로젝트별 정량적 및 정성적 결과 등을 포함할 수 있다.

[0004] 다만, 기업의 인사담당자는 수많은 지원자의 직무경험 포트폴리오를 분석 및 평가하는데 상당한 시간을 소요하고 있다. 따라서 인사담당자가 직무경험 포트폴리오에서 중요하게 보는 내용에 중점을 두어 분석 및 요약해주는

시스템이 본 기술분야에서 절실히 필요한 실정이다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 대한민국 공개특허공보 제10-2022-0013629호

(특허문헌 0002) 대한민국 공개특허공보 제10-2006-0011297호

발명의 내용

해결하려는 과제

[0006] 본 발명은 상기와 같은 문제점을 해결하기 위한 것으로 기업의 인사담당자가 수많은 지원자의 직무경험 포트폴리오를 분석 및 평가하는데 효율적으로 업무를 수행할 수 있도록 직무경험 포트폴리오로부터 텍스트를 추출하고, 평가자 단말로부터 제공된 키워드와 텍스트 간 유사도를 연산하고, 텍스트 유사도가 임계값 이상인 주요문장을 포함한 요약문을 출력하는 인공지능 기술기반 직무경험 큐레이팅 시스템을 얻고자 하는 것을 목적으로 한다.

[0007] 본 발명이 이루고자 하는 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 본 발명의 기재로부터 당해 분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있다.

과제의 해결 수단

[0008] 상기 목적을 달성하기 위하여, 본 발명의 인공지능 기술기반 직무경험 큐레이팅 시스템은 지원자 단말로부터 제공된 직무경험 포트폴리오를 획득하는 직무경험 포트폴리오 획득부; 상기 직무경험 포트폴리오로부터 텍스트를 인식한 후 추출하는 텍스트 추출부; 상기 텍스트를 형태소 단위로 토큰화(Tokenization)하고, 토큰화 텍스트를 가공하는 텍스트 가공부; 평가자 단말로부터 제공된 키워드와 상기 토큰화 텍스트 간의 텍스트 유사도를 연산하는 텍스트 유사도 연산부; 및 상기 텍스트 유사도가 임계값 이상인 상기 토큰화 텍스트를 포함하는 주요문장을 문맥에 따라 배열하고, 키워드별 요약문을 출력하는 요약문 출력부;를 제공한다.

발명의 효과

[0009] 이상과 같이 본 발명에 의하면 지원자의 직무경험 포트폴리오로부터 요약문을 출력함으로써, 기업의 인사담당자가 수많은 지원자의 직무경험 포트폴리오를 분석 및 평가하는데 효율적으로 업무를 수행할 수 있도록 하는 효과가 있다.

[0010] 또한, 본 발명은 생성형 인공지능 모델을 이용하여 지원자의 직무경험 포트폴리오로부터 지원자의 직무수행능력 여부와 직무적합성 여부에 대한 큐레이팅 문장을 출력함으로써, 기업의 인사담당자가 지원자의 직무경험 포트폴리오를 평가하는데 객관적인 지표를 제공할 수 있는 효과가 있다.

[0011] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 상세한 설명 및 청구범위의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0012] 도 1은 본 발명의 인공지능 기술기반 직무경험 큐레이팅 시스템 구성도이다.

도 2는 본 발명의 일실시예에 따른 텍스트 추출부의 CRAFT 모델이다.

도 3은 본 발명의 일실시예에 따른 지역 점수 도출과정을 표시한 도면이다.

도 4는 본 발명의 일실시예에 따른 친밀도 점수 도출과정을 표시한 도면이다.

도 5는 본 발명의 일실시예에 따른 텍스트 가공과정을 표시한 도면이다.

도 6은 본 발명의 일실시예에 따른 텍스트 유사도 연산과정을 표시한 도면이다.

도 7은 본 발명의 일실시예에 따른 문장 유사도 연산과정을 표시한 도면이다.

도 8은 본 발명의 일실시예에 따른 요약문 출력과정을 표시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0013] 본 명세서에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0014] 다르게 정의되지 않는 한 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0015] 이하, 본 발명에 따른 실시예를 첨부한 도면을 참조하여 상세히 설명하기로 한다. 도 1은 본 발명의 인공지능 기술기반 직무경험 큐레이팅 시스템 구성도이다. 도 2는 본 발명의 일실시예에 따른 텍스트 추출부(200)의 CRAFT 모델이다. 도 3은 본 발명의 일실시예에 따른 지역 점수 도출과정을 표시한 도면이다. 도 4는 본 발명의 일실시예에 따른 친밀도 점수 도출과정을 표시한 도면이다. 도 5는 본 발명의 일실시예에 따른 텍스트 가공과정을 표시한 도면이다. 도 6은 본 발명의 일실시예에 따른 텍스트 유사도 연산과정을 표시한 도면이다. 도 7은 본 발명의 일실시예에 따른 문장 유사도 연산과정을 표시한 도면이다. 도 8은 본 발명의 일실시예에 따른 요약문 출력과정을 표시한 도면이다.

[0016] 우선 도 1을 보면, 본 발명은 인공지능 기술기반 직무경험 큐레이팅 시스템은 지원자 단말(10)로부터 제공된 직무경험 포트폴리오를 획득하는 직무경험 포트폴리오 획득부(100), 상기 직무경험 포트폴리오로부터 텍스트를 인식한 후 추출하는 텍스트 추출부(200), 상기 텍스트를 형태소 단위로 토큰화(Tokenization)하고, 토큰화 텍스트를 가공하는 텍스트 가공부(300), 평가자 단말(20)로부터 제공된 키워드와 토큰화 텍스트 간의 텍스트 유사도를 연산하는 텍스트 유사도 연산부(400) 및 상기 텍스트 유사도가 임계값 이상인 상기 토큰화 텍스트를 포함하는 주요문장을 문맥에 따라 배열하고, 키워드별 요약문을 출력하는 요약문 출력부(500)를 포함한다.

[0017] 상기 지원자 단말(10)은 기업에 지원하고자 직무경험 포트폴리오를 제출한 지원자의 단말이다. 평가자 단말(20)은 기업에 지원한 수많은 지원자 중 기업의 해당 직무에 적합한 지원자를 직무경험 포트폴리오를 통해서 찾는 인사담당자의 단말이다.

[0018] 보다 구체적으로, 상기 직무경험 포트폴리오는 객관적인 사실 및 주관적인 경험이 다양한 방식으로 작성된 서류로, 텍스트 기재방식으로 한정되어 있는 이력서 및 자기소개서에 비해, 텍스트, 이미지, 영상, 사진 등 특정 기재방식에 한정되지 않고 다양한 기재방식으로 직무경험 및 지원역량이 표현될 수 있다.

[0019] 도 2의 일실시예를 보면, 상기 직무경험 포트폴리오 획득부(100)는 pdf 파일형식의 직무경험 포트폴리오를 획득하는 것이 가장 바람직하다. 이때, 상기 직무경험 포트폴리오는 한글 파일, 파워포인트(ppt) 파일 및 워드(word) 파일이 pdf 파일로 형식 변환된 것이라면, 상기 텍스트 추출부(200)는 상기 직무경험 포트폴리오로부터 용이하게 텍스트를 인식 및 추출할 수 있다. 예컨대, 상기 텍스트 추출부(200)는 상기 직무경험 포트폴리오에 '나는 학교에 간다' 라는 하나의 문장이 있다면, '나는', '학교에', '간다' 로 각각 텍스트를 추출할 수 있다. 즉, 본 발명에서 언급하는 텍스트는 어절 단위이다.

[0020] 반면에, 상기 직무경험 포트폴리오는 인쇄물을 스캔한 스캔본 등 이미지 파일이 pdf 파일로 형식 변환된 것이라면, 상기 텍스트 추출부(200)는 곧바로 상기 직무경험 포트폴리오로부터 텍스트를 인식 및 추출할 수 없다.

[0021] 이러한 문제를 해결하고자 상기 텍스트 추출부(200)는 CRAFT 모델을 이용하여 텍스트가 비인식되는 이미지 형식의 직무경험 포트폴리오에서 하나의 숫자 또는 철자(spelling)를 포함하는 단위 영역을 추출하는 단위 영역 추출부(210) 및 상기 단위 영역의 경계가 명확해지도록 상기 단위 영역을 이루는 각 픽셀이 하나의 숫자 또는 철자(spelling)의 중심일 확률인 지역 점수(Region Score)를 도출하는 지역 점수 도출부(220)를 포함할 수 있다.

[0022] 도 3의 (a)의 일실시예를 보면, 우선 상기 지역 점수 도출부(220)는 이미지 형식의 직무경험 포트폴리오의 크기

(H*W)의 절반 크기(H/2*W/2)의 2차원 등방성 가우시안 맵(Isotropic gaussian map)을 생성할 수 있다. 예컨대, 직무경험 포트폴리오의 크기가 500*500이라면, 2차원 등방성 가우시안 맵의 크기는 250*250으로 생성될 수 있다.

[0023] 도 3의 (b)의 일실시예를 보면, 상기 지역 점수 도출부(220)는 이미지 형식의 직무경험 포트폴리오에서 추출된 단위 영역과 대응되는 대응영역을 상기 2차원 등방성 가우시안 맵에 지정할 수 있다. 여기서, 상기 단위 영역은 '3'이라는 하나의 숫자를 포함하는 영역일 수 있고, '나'라는 음절을 포함하는 영역일 수 있고, 's'라는 스펠링을 포함하는 영역일 수 있다. 즉, 상기 단위 영역은 숫자 또는 철자를 포함하되, 경계가 확정되지 못한 상태이다. 상기 단위 영역은 정사각형, 마름모, 사다리꼴, 직사각형 등 4개의 꼭지점을 이은 도형의 형상으로 형성될 수 있다. 숫자 또는 철자의 중심일 확률은 상기 단위 영역의 중심일 확률이고, 상기 단위 영역의 중심은 마주보는 두 꼭지점을 연결한 두 대각선이 만나는 지점일 수 있다.

[0024] 그리고 상기 지역 점수 도출부(220)는 상기 단위 영역과 대응되는 대응영역에 하나의 숫자 또는 철자의 중심일 확률인 지역 점수를 도출할 수 있다. 상기 지역 점수는 0 내지 1 범위 이내일 수 있다. 본 발명의 일실시예에 따르면, 각 픽셀이 하나의 숫자 또는 철자의 중심일 확률이 높을수록, 즉, 1에 가까울수록 빨간색으로 표시될 수 있고, 각 픽셀이 하나의 숫자 또는 철자의 중심일 확률이 낮을수록 즉, 0에 가까울수록 남색으로 표시될 수 있다.

[0025] 도 3의 (c)의 일실시예를 보면, 상기 단위 영역 추출부(210)는 '1', '8', '0', '0', 'm'에 대한 각각의 단위 영역을 추출한다. 상기 지역 점수 도출부(220)는 각각의 단위 영역과 대응되는 상기 2차원 등방성 가우시안 맵에서의 대응영역에 상기 지역 점수를 도출할 수 있다.

[0026] 따라서, 상기 텍스트 추출부(200)는 상기 단위 영역 추출부(210) 및 지역 점수 산출부(220)를 통해서 이미지 형식의 직무경험 포트폴리오로부터 숫자 또는 철자를 포함하되, 경계가 명확해진 단위 영역을 추출할 수 있다.

[0027] 다음으로, 상기 텍스트 추출부(200)는 상기 단위 영역의 연속성이 확인되도록 서로 다른 두 개의 단위 영역에 대한 친밀도 박스(Affinity box)를 형성하고, 상기 친밀도 박스로부터 두 개의 단위 영역의 중심일 확률인 친밀도 점수(Affinity Score)를 도출하는 친밀도 점수 도출부(230)를 더 포함할 수 있다. 상기 친밀도 점수를 통해서 상기 단위 영역의 연속성을 확인함으로써, 두 음절 및 두 철자가 동일한 텍스트를 형성하는지, 형성하지 않는지를 확인할 수 있다.

[0028] 도 4의 (a)의 일실시예를 보면, 상기 친밀도 점수 도출부(230)는 상기 지역 점수를 통해서 경계가 명확해진 서로 다른 두 개의 단위 영역을 대상으로 한다. 상기 친밀도 점수 도출부(230)는 각 단위 영역의 마주보는 두 꼭지점을 연결하여 두 대각선을 형성할 수 있다. 그러면, 4개의 삼각형이 형성되는데, 이때 상기 친밀도 점수 산출부(230)는 상하로 마주보는 두 삼각형의 중심을 도출할 수 있다. 즉, 상기 친밀도 점수 산출부(230)는 하나의 단위 영역에 대한 상하로 마주보는 두 삼각형의 중심과, 또 다른 하나의 단위 영역에 대한 상하로 마주보는 두 삼각형의 중심인 총 4개의 중심점을 이어 친밀도 박스(Affinity box)를 형성할 수 있다. 즉, 상기 친밀도 박스는 빨간색의 선분으로 이루어진 박스이다.

[0029] 도 4의 (b) 및 (c)의 일실시예를 보면, 상기 친밀도 점수 도출부(230)는 상기 친밀도 박스와 대응되는 대응영역을 상기 2차원 등방성 가우시안 맵에 지정할 수 있다. 상기 친밀도 박스는 상기 단위 영역 '1'과 '8' 사이에 지정될 수 있고, 상기 단위 영역 '8'과 '0' 사이에 지정될 수 있고, 상기 단위 영역 '0'과 '0' 사이에 지정될 수 있고, 상기 단위 영역 '0'과 'm' 사이에 지정될 수 있다. 상기 친밀도 박스는 정사각형, 마름모, 사다리꼴, 직사각형 등 4개의 꼭지점을 이은 도형의 형상으로 형성될 수 있다. 서로 다른 두 개의 단위 영역의 중심은 마주보는 두 꼭지점을 연결한 두 대각선이 만나는 지점일 수 있다.

[0030] 그리고 상기 친밀도 점수 도출부(230)는 상기 대응영역에 서로 다른 두 개의 단위 영역의 중심일 확률인 친밀도 점수를 도출할 수 있다. 상기 친밀도 점수는 0 내지 1 범위 이내일 수 있다. 본 발명의 일실시예에 따르면, 친밀도 박스 내 각 픽셀이 서로 다른 두 개의 단위 영역의 중심일 확률이 높을수록, 즉, 1에 가까울수록 빨간색으로 표시될 수 있고, 각 픽셀이 두 개의 단위 영역의 중심일 확률이 낮을수록 즉, 0에 가까울수록 남색으로 표시될 수 있다.

[0031] 다음으로, 상기 텍스트 추출부(200)는 상기 지역 점수와 상기 친밀도 점수를 이용하여 상기 단위 영역으로부터 상기 텍스트 영역을 지정하는 텍스트 영역 지정부(240)를 더 포함하고, 기 저장된 딥러닝 모델에 상기 텍스트 영역이 입력됨으로써, 상기 텍스트 영역으로부터 텍스트를 추출하는 것을 특징으로 한다.

[0032] 상기 텍스트 영역 지정부(240)는 상기 지역 점수 도출부(220) 및 상기 친밀도 점수 도출부(230)로부터 도출된

상기 지역 점수와 상기 친밀도 점수를 이용하여 하나 이상의 단위 영역인 '1', '8', '0', '0', 'm'을 상기 텍스트 영역인 '1800m'을 지정할 수 있다. 그리고 상기 텍스트 영역 지정부(240)는 상기 텍스트 영역을 각각 잘라내어 별도의 이미지 파일로 저장할 수 있다.

[0033] 그리고, 본 발명에서 언급하는 기 저장된 딥러닝 모델은 OCR(Optical Character Recognition) 모델일 수 있다. 기 저장된 딥러닝 모델에 이미지 형식의 상기 텍스트 영역인 '1800m'가 입력됨으로써, 어절 단위의 텍스트를 최종적으로 추출할 수 있는 것이다.

[0034] 따라서 본 발명은 상기 텍스트 추출부(200) 내 단위 영역 추출부(210), 지역 점수 도출부(220), 친밀도 점수 도출부(230) 및 텍스트 영역 지정부(240)를 포함함으로써, 이미지 형식의 직무경험 포트폴리오에서 최종적으로 어절 단위의 텍스트를 추출할 수 있는 현저한 효과가 있다.

[0035] 다음으로, 상기 텍스트 가공부(300)는 어절 단위의 상기 텍스트를 형태소 단위로 토큰화(Tokenization)하고, 토큰화 텍스트를 가공한다. 여기서, 형태소는 뜻을 가진 가장 작은 말의 단위로, '복숭아'라는 텍스트는 '복숭아'가 형태소이고, '책가방'이라는 텍스트는 '책'과 '가방'이 각각 형태소이다. 토큰화는 텍스트를 토큰의 단위로 분할하는 행위를 말한다. 상기 텍스트 가공부(300)는 상기 텍스트를 형태소 단위로 토큰화(Tokenization)함으로써, 상기 토큰화 텍스트에 대한 명사, 형용사, 동사, 관사, 전치사, 조사 및 접속사를 포함하는 품사 정보를 얻을 수 있다.

[0036] 도 5의 일실시예를 보면, 상기 텍스트 가공부(300)는 3가지의 가공과정을 수행할 수 있다. 우선, 상기 텍스트 가공부(300)는 토큰화 텍스트가 기 설정된 표준형식으로 변환되는 정규화(Normalization)를 수행할 수 있다. 예컨대, 상기 텍스트 가공부(300)는 같은 의미의 영문의 토큰화 텍스트이지만 표기가 다른 토큰화 텍스트들을 하나의 토큰화 텍스트로 정규화할 수 있다. 'korea', 'kr'인 경우 'kr'로 정규화할 수 있다. 상기 텍스트 가공부(300)는 대문자와 소문자로 기재된 영문의 토큰화 텍스트를 소문자로 기재된 영문의 토큰화 텍스트로 정규화할 수 있다. 문장의 가장 선단이 있는 'I'인 경우 'i'로 정규화할 수 있다. 상기 텍스트 가공부(300)는 국문의 토큰화 텍스트인 경우 기 저장된 국문 사전(Dictionary)에 등재된 단어만을 추출하는 정규화를 수행할 수 있다. 한국어 사전에서 '진쟈'는 등재되지 않은 단어임으로 삭제될 수 있고, 한국어 사전에서 '진짜'는 등재된 단어임으로 추출될 수 있다.

[0037] 그리고 상기 텍스트 가공부(300)는 정규화된 토큰화 텍스트 중에서 관사, 전치사, 조사 및 접속사인 토큰화 텍스트를 불용어로 제거할 수 있다. 이는, 관사, 전치사, 조사 및 접속사는 문장을 형성하기 위한 보조적인 수단으로써 의미를 파악할 수 없기 때문에 향후 평가자 단말(20)로부터 제공된 키워드와 비교할 필요가 없기 때문에 연산량을 줄이기 위함이다. 다음으로, 상기 텍스트 가공부(300)는 불용어가 제거된 토큰화 텍스트를 각각 분리할 수 있다. 즉, 상기 텍스트 가공부(300)는 명사, 형용사 및 동사 중 하나의 품사 정보를 갖는 정규화된 토큰화 텍스트를 각각 출력할 수 있다.

[0038] 도 6의 일실시예를 보면, 상기 텍스트 유사도 연산부(400)는 레벤슈타인 거리(Levenshtein distance) 기법을 이용하여 편집거리에 따른 연산횟수를 제1 텍스트 유사도로 출력하는 제1 텍스트 유사도 연산부(410), 시퀀스 매처(Sequence Matcher) 기법을 이용하여 상기 키워드와 상기 토큰화 텍스트를 각각 바이트(byte) 단위로 변환한 후 제2 텍스트 유사도를 출력하는 제2 텍스트 유사도 연산부(420) 및 상기 제1 텍스트 유사도를 바이트 단위로 변환하는 제1 텍스트 유사도 변환부(430)를 포함하고, 변환된 제1 텍스트 유사도와 상기 제2 텍스트 유사도를 평균하여 상기 텍스트 유사도를 도출하는 것을 특징으로 한다. 상기 제1 텍스트 유사도는 0 또는 양의 정수이다.

[0039] 본 발명에서 언급하는 레벤슈타인 거리 기법은 문자열이 얼마나 비슷한지를 나타내는 알고리즘으로, 상기 토큰화 텍스트를 상기 평가자 단말(20)로부터 제공된 키워드와 동일하도록 기 설정된 편집과정에 따라 편집했을 때, 몇 번의 문자열 조작이 필요한지 카운트(Count) 함으로써 편집거리에 따른 연산횟수를 출력할 수 있다. 이때, 편집과정은 삽입, 삭제, 대체 등의 연산을 수행하는 것을 말한다. 예컨대, 상기 평가자 단말(20)로부터 제공된 키워드가 '위키미디아'이고, 상기 토큰화 텍스트가 '위키매니아'라면, '위키매니아'의 '매'를 '미'로 대체하는 횟수 1회, '위키매니아'의 '니'를 '디'로 대체하는 횟수 1회의 조작이 필요하다. 즉, 상기 편집거리에 따른 연산횟수인 2회가 상기 제1 텍스트 유사도일 수 있다.

[0040] 본 발명에서 언급하는 시퀀스 매처 기법은 상기 평가자 단말(20)로부터 제공된 키워드를 바이트(byte) 단위로 변환하고, 상기 토큰화 텍스트를 바이트(byte) 단위로 변환한 후 각 자리를 매칭하여 비교함으로써 0과 1 사이의 값을 갖는 상기 제2 텍스트 유사도를 출력할 수 있다. 예컨대, 상기 평가자 단말(20)로부터 제공된 키워드가

'위키미디아' 로 5자리이고, 상기 토큰화 텍스트가 '위키매니아' 로 5자리일 수 있다. 상기 제2 텍스트 유사도 연산부(420)는, 하기 [수학식 1]로 제2 텍스트 유사도를 출력할 수 있다.

수학식 1

[0041] $$D_{ro} = \frac{2K_m}{|S_1| + |S_2|}$$

[0042] 여기서, D_{ro}는 상기 제2 텍스트 유사도이고, K_m은 일치하는 자릿수이고, S_1은 상기 키워드의 자릿수이고, S_2은 상기 토큰화 텍스트의 자릿수이다. 즉, 상기 제2 텍스트 유사도 연산부(420)는 (2×3)/(5+5)로 0.6을 상기 제2 텍스트 유사도로 출력할 수 있다.

[0043] 다음으로, 상기 제1 텍스트 유사도 변환부(430)는 하기 [수학식 2]를 이용하여 상기 제1 텍스트 유사도를 바이트 단위로 변환하는 것을 특징으로 한다.

수학식 2

[0044] $$S(A,B) = 1 - \frac{d(A,B)}{\max(len(A), len(B))}$$

[0045] 여기서, S(A,B)는 바이트 단위로 변환된 제1 텍스트 유사도이고, d(A,B)는 상기 제1 텍스트 유사도 연산부(410)로부터 출력된 제1 텍스트 유사도이고, max(len(A), len(B))는 상기 키워드의 길이(len(A))와 상기 토큰화 텍스트의 길이(len(B)) 중 가장 긴 길이를 추출하는 함수이다.

[0046] 예컨대, 상기 평가자 단말(20)로부터 제공된 키워드가 5자리이고, 상기 토큰화 텍스트가 5자리이고, 상기 제1 텍스트 유사도가 3회라면, 상기 제1 텍스트 유사도 변환부(430)는 상기 [수학식 2]를 통해서 상기 제1 텍스트 유사도를 0.4으로 변환할 수 있다. 따라서 변환된 제1 텍스트 유사도가 0.4이고 상기 제2 텍스트 유사도가 0.6이라면, 상기 텍스트 유사도 연산부(400)는 이를 평균하여 평가자 단말(20)로부터 제공된 키워드와 상기 토큰화 텍스트 간 텍스트 유사도를 0.5로 도출할 수 있다. 따라서 본 발명은 두 기법으로 유사도를 연산한 후 평균하여 상기 텍스트 유사도를 출력함으로써, 키워드와 토큰화 텍스트 간 보다 정확한 비교가 가능한 현저한 효과가 있다.

[0047] 또한, 상기 텍스트 유사도 연산부(400)는 평가자 단말(20)로부터 제공된 키워드가 5개이고, 상기 토큰화 텍스트가 15개라면 키워드별 토큰화 텍스트를 비교하기 위해서 총 75번을 반복하여 텍스트 유사도를 연산할 수 있다.

[0048] 도 7의 일실시예를 보면, 상기 요약문 출력부(500)는, 상기 주요문장이 기 학습된 SBERT 모델에 입력됨으로써, 문장벡터를 출력하는 문장벡터 출력부(510), 피어슨 유사도 기법을 이용하여 상기 문장벡터 간 문장 유사도를 연산하는 문장 유사도 연산부(520) 및 상기 문장 유사도가 임계값 이상인 주요문장을 그룹화 하는 주요문장 그룹화부(530)를 포함하는 것을 특징으로 한다.

[0049] 본 발명에서 언급하는 SBERT 모델은 BERT의 문장 임베딩(Embedding) 성능을 우수하게 개선시킨 모델로, 한국어에 최적화되도록 KorNLI와 KorSTS로 사전학습될 수 있다. 피어슨 유사도 기법은 두 문장벡터 간 피어슨 상관계수를 연산하고, 피어슨 상관계수가 높을수록 두 주요문장은 서로 유사한 것이고, 피어슨 상관계수가 낮을수록 두 주요문장은 서로 상이한 문장이라는 것이다. 여기서, 유사하다는 것은 문맥(context)이 유사하다는 의미이다.

[0050] 다음으로, 상기 주요문장 그룹화부(530)는 키워드별로 문장 유사도가 임계값 이상인 주요문장을 그룹화함으로써, 키워드별 유사도가 높은 주요문장을 그룹화할 수 있다. 이는, 상기 요약문 출력부(500)는 키워드별 요약문을 용이하게 출력할 수 있도록 하기 위함이다.

[0051] 다음으로, 상기 요약문 출력부(500)는, 상기 주요문장 그룹화부(530)로부터 키워드별로 그룹화된 주요문장을 생성형 인공지능 모델에 입력함으로써, 키워드별 그룹화된 주요문장 전체가 문맥에 따라 배열된 키워드별 요약문

을 출력할 수 있다. 또는, 상기 요약문 출력부(500)는 상기 주요문장 그룹화부(530)로부터 키워드별로 그룹화된 주요문장을 생성형 인공지능 모델에 입력함으로써, 키워드별 그룹화된 주요문장이 문맥에 따라 배열될 뿐만 아니라, 이를 기 설정된 요약기준에 따라 요약된 키워드별 요약문을 출력할 수 있다.

[0052] 도 8의 일실시예를 보면, 상기 요약문 출력부(500)는, 상기 주요문장 그룹화부(530)로부터 키워드별 그룹화된 주요문장을 생성형 인공지능 모델에 입력함으로써, 지원자의 직무수행능력 여부와 직무적합성 여부에 대한 큐레이팅 문장을 더 출력하는 것을 특징으로 한다.

[0053] 본 발명에서 언급하는 생성형 인공지능 모델은 KoGPT일 수 있다. KoGPT은 한국어를 사전적, 문맥적으로 이해하고 사용자의 의도에 맞춘 문장을 생성해 제공하는 파인-튜닝(Fine-tuning)된 생성형 인공지능 모델이다. 즉, 상기 요약문 출력부(500)는 KoGPT를 이용하여 키워드별 요약문에 지원자가 실무에 투입되었을 때 직무를 수행할 수 있는 능력이 있는지, 없는지, 높은지 낮은지 등에 대한 키워드별 큐레이팅 문장을 출력할 수 있고, 지원자가 기업의 해당 직무에 적합한지, 적합하지 않은지 등에 대한 키워드별 큐레이팅 문장을 출력할 수 있다.

[0054] 따라서 본 발명은 요약문 출력부(500)를 포함함으로써, 기업의 인사담당자가 수많은 지원자의 직무경험 포트폴리오를 분석 및 평가하는데 키워드별 요약문을 참고함으로써, 신속하게 업무를 수행할 수 있도록 하는 현저한 효과가 있다.

[0055] 실시예들은 하드웨어, 소프트웨어, 펌웨어, 미들웨어, 마이크로코드, 하드웨어 기술 언어, 또는 이들의 임의의 조합에 의해 구현될 수 있다. 소프트웨어, 펌웨어, 미들웨어 또는 마이크로코드로 구현되는 경우, 필요한 작업을 수행하는 프로그램 코드 또는 코드 세그먼트들은 컴퓨터 판독 가능 저장 매체에 저장되고 하나 이상의 프로세서에 의해 실행될 수 있다.

[0056] 그리고 본 명세서에 설명된 주제의 양태들은 컴퓨터에 의해 실행되는 프로그램 모듈 또는 컴포넌트와 같은 컴퓨터 실행 가능 명령어들의 일반적인 맥락에서 설명될 수 있다. 일반적으로, 프로그램 모듈 또는 컴포넌트들은 특정 작업을 수행하거나 특정 데이터 형식을 구현하는 루틴, 프로그램, 객체, 데이터 구조를 포함한다. 본 명세서에 설명된 주제의 양태들은 통신 네트워크를 통해 링크되는 원격 처리 디바이스들에 의해 작업들이 수행되는 분산 컴퓨팅 환경들에서 실시될 수도 있다. 분산 컴퓨팅 환경에서, 프로그램 모듈들은 메모리 저장 디바이스들을 포함하는 로컬 및 원격 컴퓨터 저장 매체에 둘 다에 위치할 수 있다.

[0057] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 으로 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0058] 그러므로 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

[0059] 100.. 직무경험 포트폴리오 획득부

200.. 텍스트 추출부

210.. 단위 영역 추출부

220.. 지역 점수 도출부

230.. 친밀도 점수 도출부

240.. 텍스트 영역 지정부

300.. 텍스트 가공부

400.. 텍스트 유사도 연산부

410.. 제1 텍스트 유사도 연산부

420.. 제2 텍스트 유사도 연산부

430.. 제1 텍스트 유사도 변환부

500.. 요약문 출력부

510.. 문장벡터 출력부

520.. 문장 유사도 연산부

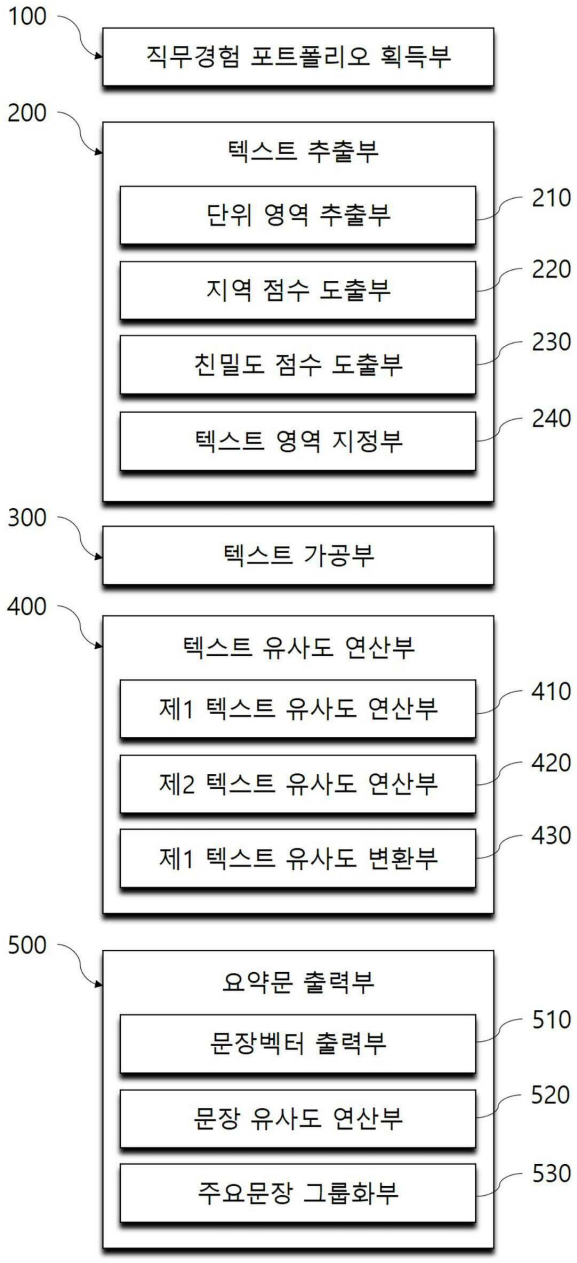
530.. 주요문장 그룹화부

10.. 지원자 단말

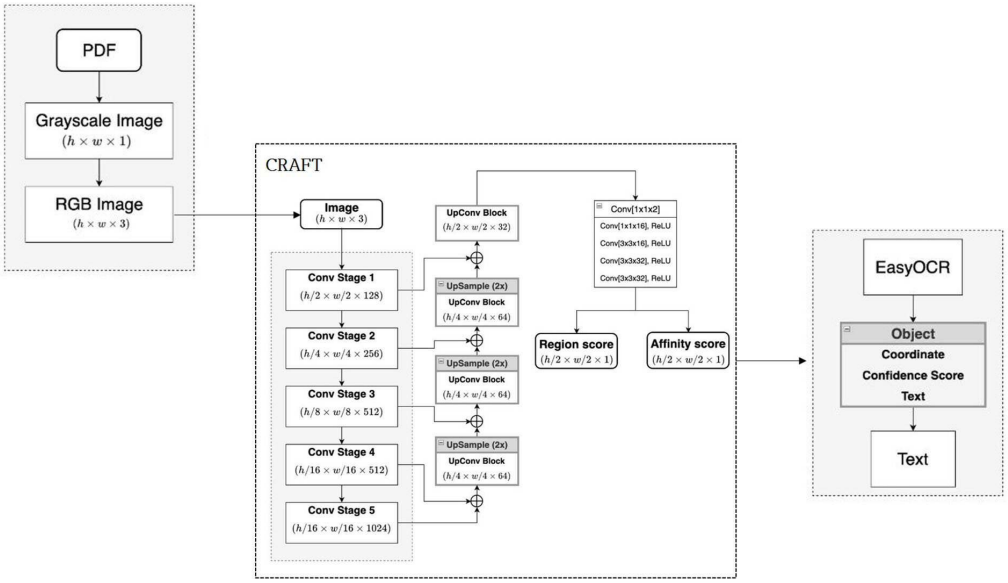
20.. 평가자 단말

도면

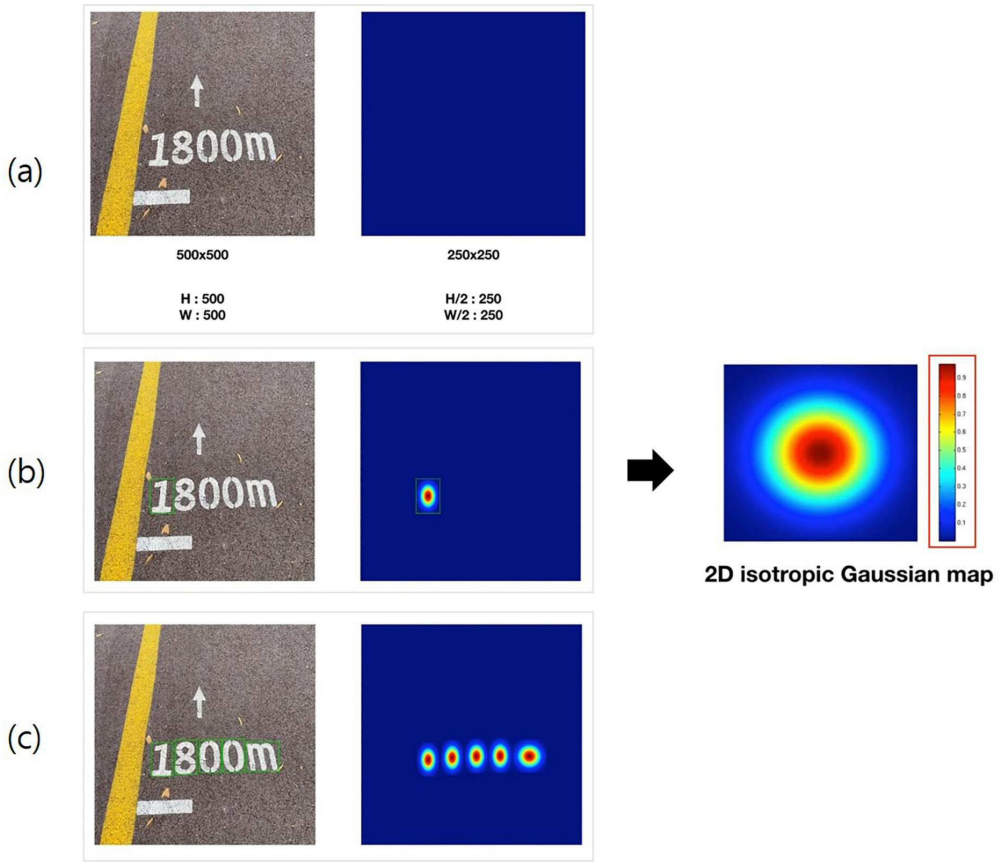
도면1



도면2

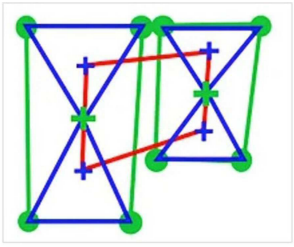

PDF
Grayscale Image
$(h \times w \times 1)$
RGB Image
$(h \times w \times 3)$
CRAFT
Image
$(h \times w \times 3)$
Conv Stage 1
$(h/2 \times w/2 \times 128)$
Conv Stage 2
$(h/4 \times w/4 \times 256)$
Conv Stage 3
$(h/8 \times w/8 \times 512)$
Conv Stage 4
$(h/16 \times w/16 \times 512)$
Conv Stage 5
$(h/16 \times w/16 \times 1024)$
UpConv Block
$(h/2 \times w/2 \times 32)$
UpSample (2x)
UpConv Block
$(h/4 \times w/4 \times 64)$
UpSample (2x)
UpConv Block
$(h/4 \times w/4 \times 64)$
UpSample (2x)
UpConv Block
$(h/4 \times w/4 \times 64)$
Conv[1x1x2]
Conv{1x1x16}, ReLU
Conv{3x3x16}, ReLU
Conv{3x3x32}, ReLU
Conv{3x3x32}, ReLU
Region score
$(h/2 \times w/2 \times 1)$
Affinity score
$(h/2 \times w/2 \times 1)$
EasyOCR
Object
Coordinate
Confidence Score
Text
Text

도면3

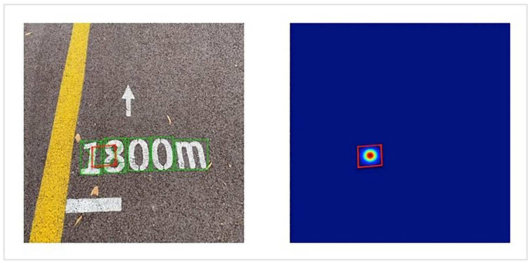

(a)
1800m
500x500
H : 500
W : 500
250x250
H/2 : 250
W/2 : 250
(b)
1800m
2D isotropic Gaussian map
(c)
1800m

도면4

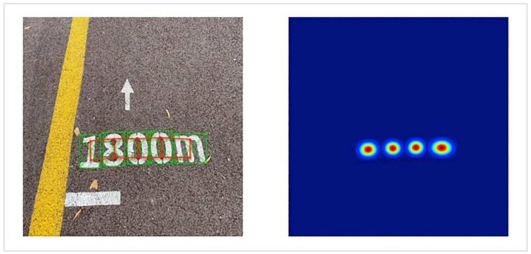
(a)



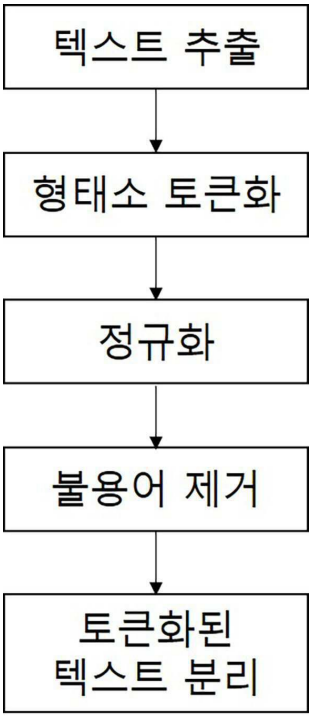
(b)



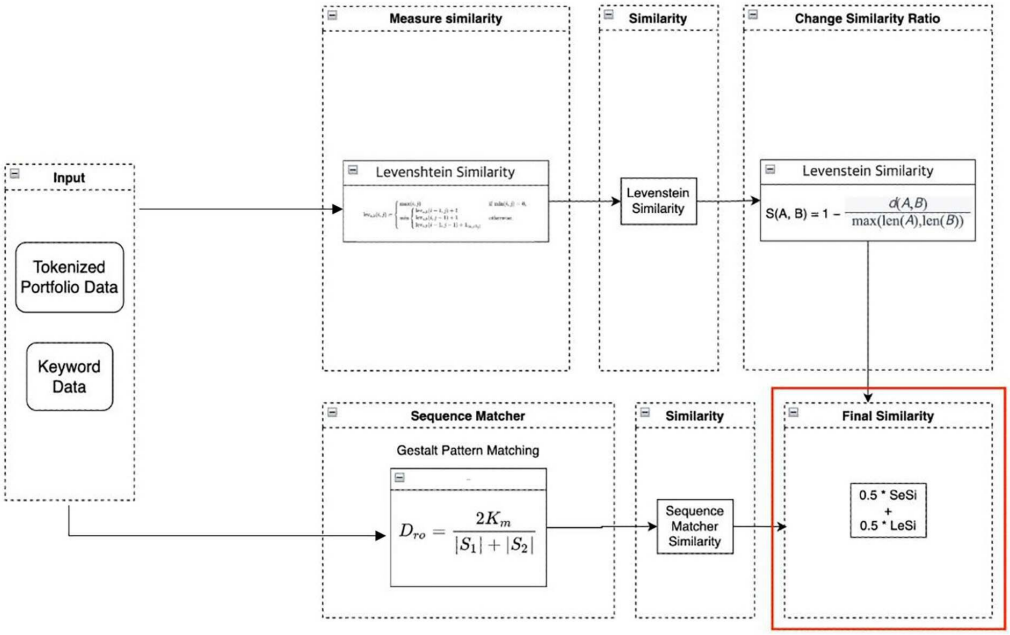
(c)



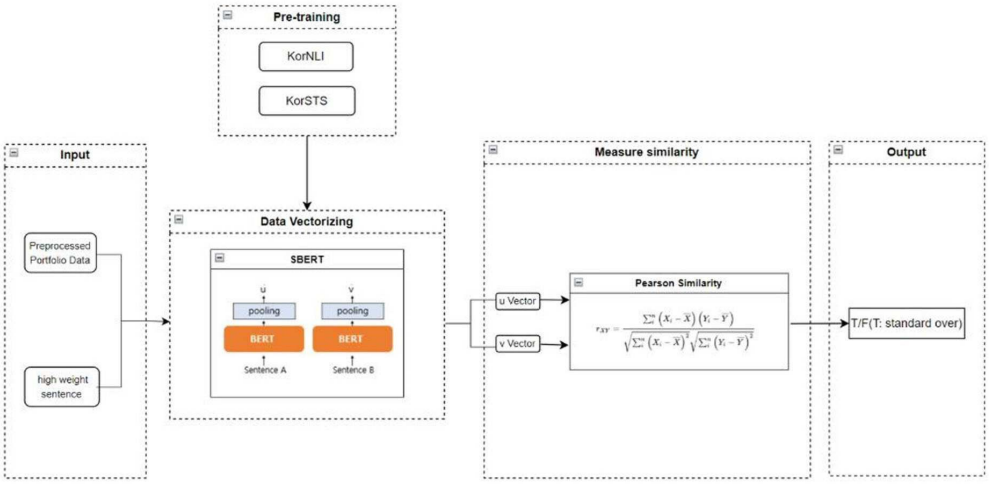
도면5



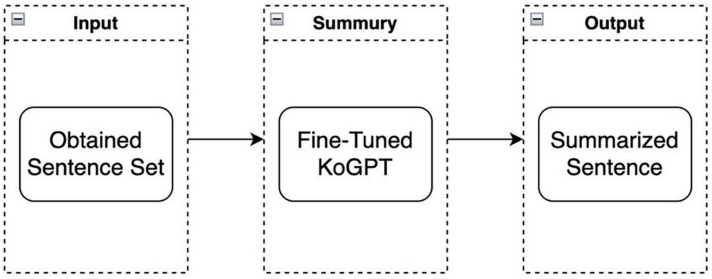
도면6


Input
Tokenized Portfolio Data
Keyword Data
Measure similarity
Levenshtein Similarity
Similarity
Levenstein Similarity
Change Similarity Ratio
Levenstein Similarity
S(A, B) = 1 − d(A,B) / max(len(A),len(B))
Sequence Matcher
Gestalt Pattern Matching
$D_{ro} = \frac{2K_m}{|S_1| + |S_2|}$
Similarity
Sequence Matcher Similarity
Final Similarity
0.5 * SeSi + 0.5 * LeSi

도면7


Pre-training
KorNLI
KorSTS
Input
Preprocessed Portfolio Data
high weight sentence
Data Vectorizing
SBERT
u
v
pooling
pooling
BERT
BERT
Sentence A
Sentence B
Measure similarity
u Vector
v Vector
Pearson Similarity
Output
T/F(T: standard over)

도면8


Input
Obtained Sentence Set
Summury
Fine-Tuned KoGPT
Output
Summarized Sentence