과제명: 기술기반 기업의 지속성장 지원을 위한 정형-비정형 데이터 연계 기반 컨설팅 서비스 개발

경희대학교 3차년도 연구개발 목표 및 내용

주 관 기 관 : 경희대학교 산학협력단
연구책임자: 김재경 교수

3차년도 개발 개요

- **3차년도에 경희대는 기업 정보 요약 알고리즘, 기업 의사결정 알고리즘, BMC 비정형 데이터 분석 모델 등 총 세 가지 요소를 고도화함으로써, 컨설팅 서비스 제공자가 피컨설팅 기업의 강점과 약점을 분석하고 분석 결과에 최적화 된 의사 결정을 내릴 수 있도록 프로세스를 자동화하는 것을 목표로 함.**
  - **개발 개요1(기업 맞춤형 자동화 컨설팅 서비스를 위한 기업 정보 요약 알고리즘 고도화)**: 규칙 기반 시스템 도입 과정을 통해 기업에 대한 정형 및 비정형 정보를 자동으로 분석 및 요약함으로써, 2차년도 개발 과정 중의 한계점을 보완하고자 하며, 공인인증성능평가의 측정 요소 중 하나인 텍스트 요약 평가 점수를 평점 4.0 이상 획득하는 것을 목표로 함.
  - **개발 개요2(기업 의사결정 알고리즘 고도화)**: 사용 모델과 파라미터를 수정함으로써 설명력과 성능 측면에서 보다 정교한 알고리즘을 구현하여 3차년도의 성능 목표치인 정확률 85%를 초과 달성하며 동시에 일반화 성능까지 향상시키는 것을 목적으로 함.
  - **개발 개요3(BMC 비정형 데이터 분석 모델 고도화)**: 비정형 텍스트 데이터인 BMC를 구성하고 있는 각 9가지 텍스트 요인을 분석하여 자동으로 지수화하는 모델을 2차년도 과제에서 사용한 모델의 한계점을 보완하면서 고도화함으로써 자기적으로 진단할 수 있는 모델을 개발하고자 함.

개발 설계 1:

정보 요약 모델

개발 목표

- **복잡한 구조를 가지고 있는 기업 정보 데이터를 자동화 처리된 알고리즘을 통해, 피컨설팅 기업과 동종 업계에 속한 기업 간의 데이터를 빠르고 정확하게 비교·분석하여, 분석 결과를 기반으로 기업 맞춤형 자동화 컨설팅 서비스를 제공하는 것을 목표로 함.**
  - 2차년도의 경우, 딥러닝 모델(Sequence-to-sequence with Attention)을 구축하고 이를 기반으로 특정 기업 관련 뉴스 데이터를 입력하여, 시간의 흐름에 따라 특정 기업의 동향 정보를 생성하는 것을 목표로 하였음.
  - 2차년도에 사용한 뉴스 데이터는 '금융' 도메인 한 개로 한정되어 있었으며, 딥러닝 모델 학습 및 검증 과정에서 사용되어진 뉴스 데이터의 개수가 적었기 때문에, 특정 기업에 대한 뉴스에서 핵심 정보를 추출하고 표현하는데 한계점이 존재함.
  - 3차년도에는 기본적으로 기업의 기술기반 데이터(지수 정보 등) 및 비정형 데이터(뉴스, 리뷰 정도 등)를 활용해 규칙 기반(Rule-based) 기업 정보 요약 모델을 구축하는 한편, 기업 요약문을 쉽게 이해할 수 있도록 피컨설팅 기업과 동종업계에 속한 기업들을 비교·분석하는 시각화 컨텐츠를 설계하는 것을 목표로 함.
  - 규칙 기반 기업 정보 요약 모델의 경우, 피컨설팅 기업의 강점과 약점에 따라 시나리오를 구성하여 요약문을 제공하는 것이 핵심이고, 특히 약점 중에서도 가장 취약한 요인이 어떠한 것인지 쉽게 파악할 수 있도록 하며, 이에 대한 간단한 보완 가이드라인을 제시하는 것을 목표로 함.

1. 최종 텍스트 샘플(컨소시엄 피드백 반영)

1) 6대 지수에 대하여, 기업의 강점과 약점이 모두 존재하는 경우

현대제철은 제철업 분야의 대기업으로, 1953년부터 70년간 사업을 영위하고 있다. 대표자는 안동일이며, 2022년 4월 기준 종업원은 11,528명이다**(기업 개요)**.

슬라이드 10에 요약문 생성 규칙을 기술

현대제철은 동종업계 기업들의 평균 대비 R&D, 인적자원, BM 지수가 우수한 역량을 보이고 있다. 반면, 마케팅, 미래성장, 평판 지수의 경우 동종업계 기업들의 평균 대비 부족한 역량을 보이고 있다. **(동종업계 기업에 비하여 우수한 측면 및 취약한 측면을 한 번에 기술)**

특히, 마케팅 지수에서는 영업/마케팅 인력 비중, 미래성장 지수에서는 매출액증가율이 가장 취약한 것으로 확인되었으며, 그리고 평판 지수의 경우 상위 5개의 부정적 키워드는 [갑질, 서비스, 불량, 제품, 배송]이다. **(약점을 보이는 지수의 세부 요인 중에서 제일 취약한 요인을 선별 + 단 평판 지수의 경우, 상위 5개의 긍정 또는 부정적 키워드를 언급)**

따라서 현대제철은 고객응대 서비스, 인력 및 네트워크 보유 측면에서 마케팅 비용 투자를 진행할 필요가 있다. **(간단한 보완 가이드라인 제시)**

1. 최종 텍스트 샘플

2) 6대 지수에 대하여, 기업의 강점만 존재하는 경우

현대제철은 제철업 분야의 대기업으로, 1953년부터 70년간 사업을 영위하고 있다. 대표자는 안동일이며, 2022년 4월 기준 종업원은 11,528명이다**(기업 개요)**.

슬라이드 11에 요약문 생성 규칙을 기술

현대제철은 동종업계 기업들의 평균 대비 6대 지수가 모두 우수한 역량을 보이고 있다. 특히, R&D 지수에서는 최근 5개년 이내 R&D 국가실적이 동종업계 평균 대비 매우 우수한 역량을 보여주고 있다.**(전부 우수하지만, 그 중에서도 최고로 우수한 요인에 대해 언급)**

단, 마케팅 지수는 선도기업 대비 마케팅 비용 수준이 동종업계 평균 대비 크게 차이가 나지 않는 모습을 보여주고 있다.

(강점 지수 중에서 그나마 동종업계 평균과 차이가 적은 요인을 선별)

따라서, 현대제철은 각 6대 지수에 대한 강점을 유지할 수 있도록 지속적인 관심과 노력이 필요하다(**간단한 가이드라인 제시**).

1. 최종 텍스트 샘플

3) 6대 지수에 대하여, 기업의 약점만 존재하는 경우

현대제철은 제철업 분야의 대기업으로, 1953년부터 70년간 사업을 영위하고 있다. 대표자는 안동일이며, 2022년 4월 기준 종업원은 11,528명이다**(기업 개요)**.

슬라이드 12~13에 요약문 생성 규칙을 기술

현대제철은 동종업계 기업들의 평균 대비 6대 지수가 모두 부진한 역량을 보이고 있다._ 특히, 평판 지수는 동종업계 평균 대비 매우 부진한 역량을 보이며, 관련 키워드로 [갑질, 서비스, 불량, 제품, 배송]가 있다. 또한, 마케팅 지수에서는 선도기업 대비 마케팅 비용 수준이 부진한 역량을 보이고 있고, 미래성장 지수는 미래 평판점수가 낮은 역량을 보이고 있다.**(약점 지수 WORST3와 각 지수별로 제일 취약한 요인 언급 + 단 평판 지수의 경우, 상위 5개의 부정적 키워드를 언급)**

다만, R&D 지수의 최근 5개년 이내 R&D 국가실적이 동종업계 평균과 가장 가까운 것을 확인하였다.**(그나마 적은 약점을 보이는 지수와 가장 덜 취약한 요인을 언급)**

따라서 현대제철은 고객응대 서비스, 인력 및 네트워크 보유 측면에서 마케팅 비용 투자를 진행할 필요가 있으며, 매출액 및 순이익에 악영향을 끼치는 요소를 확인하고 이에 대한 대책을 세우며, 기술성 및 사업성을 향상시키기 위한 비즈니스 모델을 구축할 필요가 있다.**(간단한 보완 가이드라인 제시)**

1. 최종 텍스트 샘플

4) 각 지수 별 보완 가이드라인 문장 형식

R&D 지수: 연구소 인력, 연구개발비 대비 실적, 그리고 최근 5개년 이내 국가 R&D 실적 등을 확인하고, 부족한 요소에 대한 보완 계획을 수립할 필요가 있음.

미래성장 지수: 매출액 및 순이익에 악영향을 끼치는 요소를 확인하고 이에 대한 대책을 세우며, 기술성 및 사업성을 향상시키기 위한 비즈니스 모델을 구축할 필요가 있음.

인적자원 지수: 경영지원 또는 전략기획 등의 인력 수준, 직원의 자기 개발 지원 정도, 종사자 1인당 매출액 수준 등의 세부 요소를 고려하여 인적자원 역량 전문화를 추진할 필요가 있음.

BM 지수: 해당 분야에 대한 시장 경쟁력, 규모, 제품 생명 주기 등을 고려하여 사업화 역량을 고도화시켜하며, 이에 따른 비즈니스 모델을 수립할 필요가 있음.

마케팅 지수: 고객응대 서비스, 인력 및 네트워크 보유 측면에서 마케팅 비용 투자를 진행할 필요가 있음.

평판 지수: 기업 내부 평판으로서 복지, 급여 및 자기 개발 지원 등의 요소를, 기업 외부 평판으로서 고객 서비스 및 대응 방식에 대한 만족도 등의 요소를 고려하여, 미비한 요소를 보완하기 위한 계획을 수립할 필요가 있음.

2. 요약문 산출 과정

1) 6대 지수에 대하여, 기업의 강점과 약점이 모두 존재하는 경우

현대제철은 동종업계 기업들의 평균 대비 R&D, 인적자원, BM 지수가 우수한 역량을 보이고 있다. 반면, 마케팅, 미래성장, 평판 지수의 경우 동종업계 기업들의 평균 대비 부족한 역량을 보이고 있다. **(동종업계 기업에 비하여 우수한 측면 및 취약한 측면을 한 번에 기술)** 특히, 마케팅 지수에서는 영업/마케팅 인력 비중, 미래성장 지수에서는 매출액증가율이 가장 취약한 것으로 확인되었으며, 그리고 평판 지수의 경우 상위 5개의 부정적 키워드는 [갑질, 서비스, 불량, 제품, 배송]이다. **(약점을 보이는 지수의 세부 요인 중에서 제일 취약한 요인을 선별 + 단 평판 지수의 경우, 상위 5개의 긍정 또는 부정적 키워드를 언급)**

- Step 1) 피컨설팅 기업의 6대 지수 값과 동종업계에 속한 전체 기업의 6대 지수 별 평균 값을 호출
- Step 2) 각 6대 지수에 대하여, 피컨설팅 기업의 값과 동종업계의 평균 값의 차이를 계산
  - ✓ 차이 = 피컨설팅 기업의 지수 값 – 동종업계 전체 기업의 지수 평균값
- Step 3) 차이가 양의 방향(+)으로 나오면 강점으로, 음의 방향(-)으로 나오면 약점 지수로 분류
- Step 4) 약점 지수에 대하여, 각 지수를 구성하는 세부요인의 차이를 계산
  - ✓ 차이 = 피컨설팅 기업의 세부요인 값 – 동종업계 전체 기업의 세부요인 평균값
- Step 5) (약점)각 지수에 대하여, 각각 세부요인의 차이가 큰 세부요인을 한 개씩 선정
  - ✓ 단, 평판 지수의 경우 빈도수 기반 상위 5개의 부정적 키워드를 출력

2. 요약문 산출 과정

2) 6대 지수에 대하여, 기업의 강점만 존재하는 경우

현대제철은 동종업계 기업들의 평균 대비 6대 지수가 모두 우수한 역량을 보이고 있다. 특히, R&D 지수에서는 최근 5개년 이내 R&D 국가실적이 동종업계 평균 대비 매우 우수한 역량을 보여주고 있다.**(전부 우수하지만, 그 중에서도 최고로 우수한 요인에 대해 언급)** 단, 마케팅 지수는 선도기업 대비 마케팅 비용 수준이 동종업계 평균 대비 크게 차이가 나지 않는 모습을 보여주고 있다. **(강점 지수 중에서 그나마 동종업계 평균과 차이가 적은 요인을 선별)**

- Step 1) 피컨설팅 기업의 6대 지수 값과 동종업계에 속한 전체 기업의 6대 지수 별 평균 값을 호출
- Step 2) 각 6대 지수에 대하여, 피컨설팅 기업의 값과 동종업계의 평균 값의 차이를 계산
  - ✓ 차이 = 피컨설팅 기업의 지수 값 – 동종업계 전체 기업의 지수 평균값
- Step 3) 차이가 가장 큰 1개의 지수와(최고 강점 지수) 가장 적은 1개 지수 선정(강점 지수 중 그나마 동종업계와 차이가 적은 지수)
- Step 4) 선정된 2개의 지수에 대하여, 각 지수를 구성하는 세부요인의 차이를 계산
  - ✓ 차이 = 피컨설팅 기업의 세부요인 값 – 동종업계 전체 기업의 세부요인 평균값
- Step 5) 선정된 2개의 지수에 대하여, 각각 세부요인의 차이가 가장 큰/가장 작은 세부요인을 한 개씩 선정
  - ✓ 단, 평판 지수의 경우 빈도수 기반 상위 5개의 긍정적 키워드를 출력

2. 요약문 산출 과정

3) 6대 지수에 대하여, 기업의 약점만 존재하는 경우

현대제철은 동종업계 기업들의 평균 대비 6대 지수가 모두 부진한 역량을 보이고 있다._ 특히, 평판 지수는 동종업계 평균 대비 매우 부진한 역량을 보이며, 관련 키워드로 [갑질, 서비스, 불량, 제품, 배송]가 있다. 또한, 마케팅 지수에서는 선도기업 대비 마케팅 비용 수준이 부진한 역량을 보이고 있고, 미래성장 지수는 미래 평판점수가 낮은 역량을 보이고 있다.**(약점 지수 WORST3와 각 지수별로 제일 취약한 요인 언급 + 단 평판 지수의 경우, 상위 5개의 부정적 키워드를 언급)**

- Step 1) 피컨설팅 기업의 6대 지수 값과 동종업계에 속한 전체 기업의 6대 지수 별 평균 값을 호출
- Step 2) 각 6대 지수에 대하여, 피컨설팅 기업의 값과 동종업계의 평균 값의 차이를 계산
  - ✓ 차이 = 피컨설팅 기업의 지수 값 – 동종업계 전체 기업의 지수 평균값
- Step 3) 차잇값이 가장 큰 3개의 지수를 선정
- Step 4) 선정된 3개의 지수에 대하여, 각 지수를 구성하는 세부요인의 차이를 계산
  - ✓ 차이 = 피컨설팅 기업의 세부요인 값 – 동종업계 전체 기업의 세부요인 평균값
- Step 5) 선정된 3개의 지수에 대하여, 각각 세부요인의 차이가 큰 세부요인을 한 개씩 선정
  - ✓ 단, 평판 지수의 경우 빈도수 기반 상위 5개의 부정적 키워드를 출력

2. 요약문 산출 과정

3) 6대 지수에 대하여, 기업의 약점만 존재하는 경우

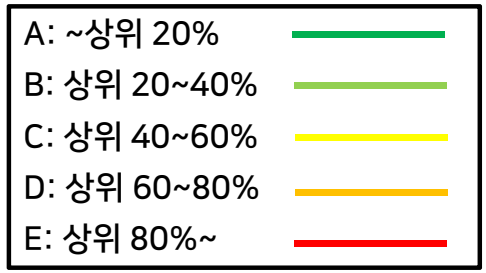
다만, R&D 지수의 최근 5개년 이내 R&D 국가실적이 동종업계 평균과 가장 가까운 것을 확인하였다.**(그나마 적은 약점을 보이는 지수와 가장 덜 취약한 요인을 언급)**

- Step 1) 6대 지수를 구성하는 세부요인에 대하여, 아래의 차이를 계산
  - ✓ 차이 = 피컨설팅 기업의 지수 세부요인 - 동종업계에 속한 전체 기업의 세부요인의 평균 값
- Step 2) 계산된 차이가 가장 작은 세부요인과 이에 대응하는 지수를 선택
  - ✓ 단, 평판 지수의 경우 빈도수 기반 상위 5개의 부정적 키워드를 출력

따라서 현대제철은 고객응대 서비스, 인력 및 네트워크 보유 측면에서 마케팅 비용 투자를 진행할 필요가 있으며, 매출액 및 순이익에 악영향을 끼치는 요소를 확인하고 이에 대한 대책을 세우며, 기술성 및 사업성을 향상시키기 위한 비즈니스 모델을 구축할 필요가 있다.**(간단한 보완 가이드라인 제시)**

- Step 1) 차이가 가장 큰 3개의 지수를 선정한 후, 내림차순으로 정렬
  - ✓ 차이 = 피컨설팅 기업의 지수 값 – 동종업계 전체 기업의 지수 평균값
- Step 2) 정렬된 순서대로, 해당 지수를 보완하기 위한 가이드라인을 제시

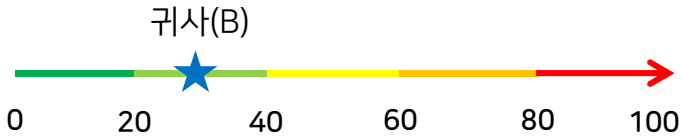
1. 피컨설팅 기업의 종합 등급·점수 및 6대 지수 정보 시각화(결과 샘플)



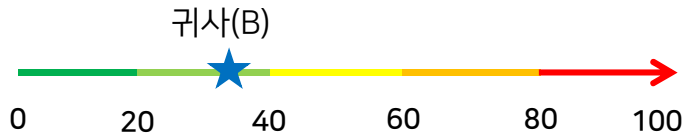
귀사의 기업 종합 점수는 70점이며, 동종업계 속한 기업 중 상위 25%에 해당합니다.
따라서, 귀사의 기업 종합 등급은 B등급입니다.

* 기업 종합 점수는 산출된 각 지수 점수에 가중치를 곱한 값을 합산하여 계산할 수 있음

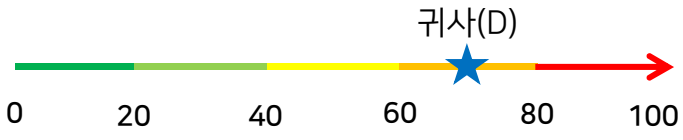
평판정보지수: 55



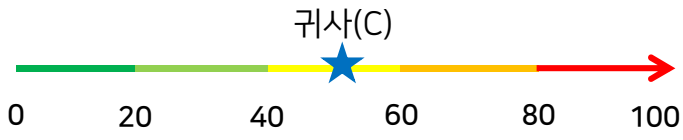
마케팅역량지수: 65



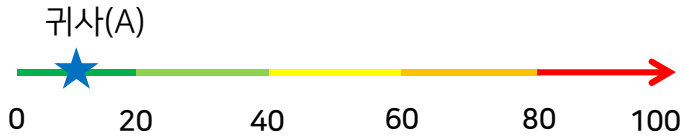
인적자원지수: 35



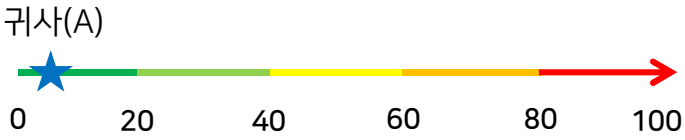
R&D역량지수: 45



BM지수: 90



미래성장지수: 95



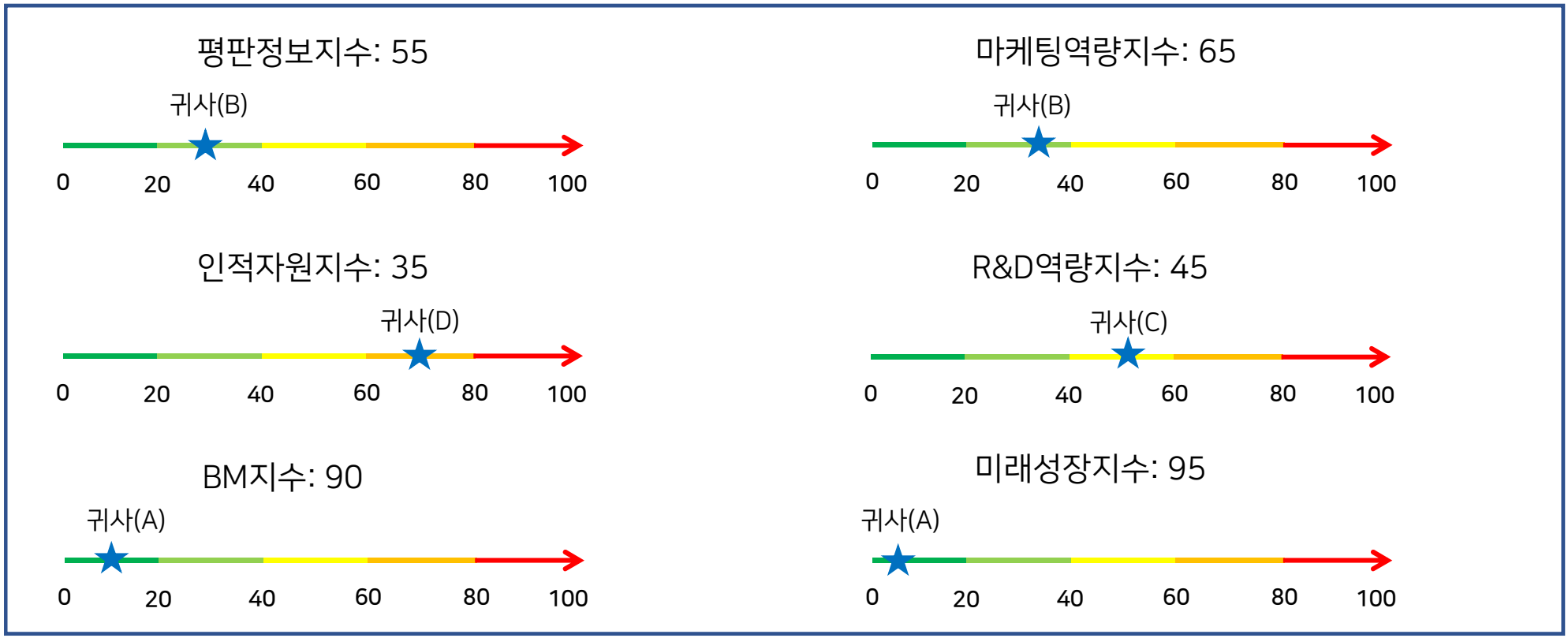
1. 피컨설팅 기업의 종합 등급 및 점수 시각화(상세 설명)

귀사의 기업 종합 점수는 70점이며, 동종업계 속한 기업 중 상위 25%에 해당합니다.
따라서, 귀사의 기업 종합 등급은 B등급입니다.
* 기업 종합 점수는 산출된 각 지수 점수에 가중치를 곱한 값을 합산하여 계산할 수 있음

A: ~상위 20%
B: 상위 20~40%
C: 상위 40~60%
D: 상위 60~80%
E: 상위 80%~

- 기업의 종합 점수는, 분석을 통해 얻은 각 지수 값에 가중치를 곱한 값을 합산하여 계산을 수행함.
  - ➢ 기업 종합점수 $Y = \sum_i Weight_i \times Score_i (i: each\ index)$
- 피컨설팅 기업과 동종업계에 속한 기업들의 종합 점수를 이용하여, 피컨설팅 기업에 대한 백분위수를 계산함.
- 따라서, 계산된 백분위수를 통해 피컨설팅 기업의 상대적인 위치를 파악할 수 있음. 즉, 해당 백분위수는 상대적인 값을 의미함.
- 기업 종합 등급은 A ~ E까지 총 5개의 등급으로 구성되어 있으며, (상대적) 백분위수 구간을 균등하게 분할하여 각 등급에 할당함.

1. 피컨설팅 기업의 6대 지수 정보 시각화(상세 설명)



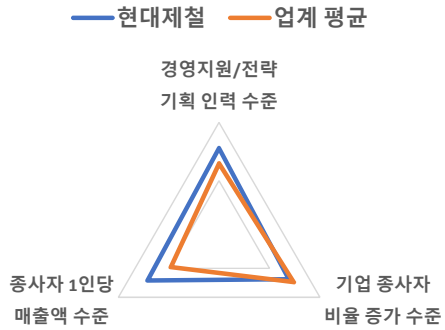
- 분석을 통해 얻은 각 지수 값(원점수)와 동종업계에 속한 기업 대비 상대적인 위치를 시각화함.
- 피컨설팅 기업과 동종업계에 속한 기업들의 지수 값을 이용하여, 피컨설팅 기업에 대한 상대적 백분위수를 계산함.
- 기업 종합 등급은 A ~ E까지 총 5개의 등급으로 구성되어 있으며, (상대적) 백분위수 구간을 균등하게 분할하여 각 등급에 할당함(종합 등급 및 점수 시각화 파트와 방법이 동일함).

2. 피컨설팅 기업의 6대 지수 정보 시각화

- 세부 구성요인 – Teams에 업로드 된 지식서비스 유형_2 파일 및 뷰테이블 정의서를 참고하여 내용을 구성
- 피컨설팅 기업의 각 지수 별 세부 구성요인을 동종업계에 속한 기업들의 평균값과 비교 가능하도록 시각화

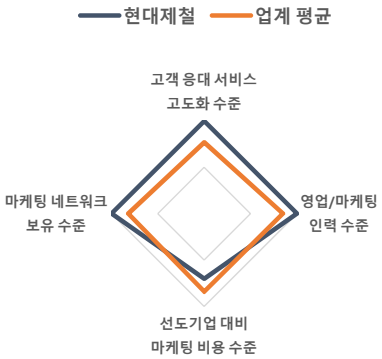
R&D 지수

인적자원 지수

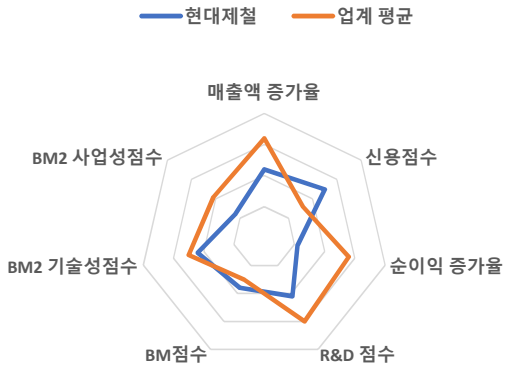


BM 지수

마케팅 지수



미래성장 지수



평판 지수

3. 키워드 워드클라우드 시각화(결과 샘플 및 상세 설명)

- 피컨설팅 기업에 대한 비정형 데이터(뉴스, 블로그 또는 카페 게시글 등)을 활용하여 키워드 워드클라우드를 생성할 수 있음.
- 키워드 클라우드를 시각화하는 과정에서 빈도수를 기반으로 시각화 될 단어의 상대적인 크기를 조정할 수 있음.
  - ➢ 노출이 많이 된 단어는 상대적으로 크게, 반대인 경우는 상대적으로 작게 표현
- 위 예시처럼 각 키워드에 대한 감성 점수가 존재할 경우, 부정적 키워드의 경우에는 빨간색으로, 긍정적 키워드의 경우는 파란색으로 표현 가능함.

- 텍스트 요약 및 시각화 컨텐츠를 구성하는 데 필요한 데이터를 정리한 표로서, IP&신용 지수를 제외한 여섯 개의 지수를 구성하는 세부 요인 및 데이터 적재 상황 등의 내용을 포함하고 있음.
  - ✓ 분석 데이터: 현재 분석 중인 데이터로서, DB 상에 해당 값이 적재되지 않은 상태를 의미
  - ✓ 설문조사 데이터: 설문조사를 기반으로 얻는 데이터로서, DB 상에 해당 값이 적재되지 않은 상태를 의미

컨텐츠	지수명	세부 구성요인	데이터 적재 상황	비고
텍스트 요약, 시각화	R&D 지수	기술인력비중 수준	O	일부 적재
		기술인력비중 수준 평균		
		전년대비 R&D 투자 증액 수준		
		전년대비 R&D 투자 증액 수준 평균		
		최근 5년 이내 국가 R&D 실적		
		최근 5년 이내 국가 R&D 실적 평균		
		4차 산업혁명 대응수준	O	기본값(0)으로 적재
		4차 산업혁명 대응수준 평균		
		R&D 지수 역량 점수	X	분석 데이터
		동종업계 R&D 지수 역량 점수 평균		

개발 프레임워크: 입력 데이터

컨텐츠	지수명	세부 구성요인	데이터 적재 상황	비고
텍스트 요약	기업 개요	기업명	O	-
		기업유형		
		대표자명		
		설립연도		일부 적재
		업종		
		기준일자		
		종업원수		
텍스트 요약, 시각화	인적자원 지수	경영지원/전략기획 인력 수준	X	설문조사 데이터
		경영지원/전략기획 인력 수준 평균		
		종사자 1인당 매출액 수준		
		종사자 1인당 매출액 수준 평균		
		기업 종사자 비율 증가 수준		
		기업 종사자 비율 증가 수준 평균		
		인적자원 지수 역량 점수		분석 데이터
		동종업계 인적자원 지수 역량 점수 평균		

개발 프레임워크: 입력 데이터

컨텐츠	지수명	세부 구성요인	데이터 적재 상황	비고
텍스트 요약, 시각화	BM 지수	고용전망	O	-
		고용전망 평균		
		체감경기 전망		
		체감경기 전망 평균		
		매출액 증가 추이		일부 적재
		매출액 증가 추이 평균		
		사업화 기반 구축 수준		-
		사업화 기반 구축 수준 평균		
		BM 지수 역량 점수	X	분석 데이터
		동종업계 BM 지수 역량 점수 평균		
	마케팅 지수	고객 응대 서비스 고도화 수준	X	설문조사 데이터
		고객 응대 서비스 고도화 수준 평균	O	일부 적재
		마케팅 네트워크 보유 수준	X	설문조사 데이터
		마케팅 네트워크 보유 수준 평균	O	일부 적재
		영업/마케팅 인력 수준	X	설문조사 데이터
		영업/마케팅 인력 수준 평균	O	일부 적재

개발 프레임워크: 입력 데이터

컨텐츠	지수명	세부 구성요인	데이터 적재 상황	비고
텍스트 요약, 시각화	마케팅 지수	선도기업 대비 마케팅 비용 수준	X	설문조사 데이터
		선도기업 대비 마케팅 비용 수준 평균	O	일부 적재
		마케팅 지수 역량 점수	X	분석 데이터
		동종업계 마케팅 지수 역량 점수 평균		
	미래성장 지수	매출액 증가율	X	분석 데이터
		매출액 증가율 평균		
		순이익 증가율		
		순이익 증가율 평균		
		신용점수		
		신용점수 평균		
		R&D 점수		
		동종업계 R&D 지수 역량 점수 평균		
		BM 점수		
		동종업계 BM 지수 역량 점수 평균		
		BM2 기술성 점수		
		BM2 기술성 점수 평균		

개발 프레임워크: 입력 데이터

컨텐츠	지수명	세부 구성요인	데이터 적재 상황	비고
텍스트 요약, 시각화	미래성장 지수	BM2 사업성 점수	X	분석 데이터
		BM2 사업성 점수 평균		
		미래성장 지수 역량 점수		
		동종업계 미래성장 지수 점수 평균		
텍스트 요약	평판 지수	긍정 키워드	X	• 분석 데이터 • 긍정 및 부정 키워드: 6대 지수 정보 및 키워드 워드클라우드 시각화 두 곳 모두 적용 가능
		부정 키워드		
		평판 지수 역량 점수		
		동종업계 평판 지수 점수 평균		
시각화		긍정 백분율		
		긍정 백분율 평균		
		중립 백분율		
		중립 백분율 평균		
		부정 백분율		
		부정 백분율 평균		
		평판 지수 역량 점수		
		동종업계 평판 지수 점수 평균		

정보 요약 모델
함수 정의서

- 슬라이드 6~13에 설계한 요약문 형식을 'khu_get_summarize.py' 파이썬 파일 형태로 구현하였으며, 요약문을 생성하고 출력하는 프로세스는 아래와 같음.

기업의 6대 지수, 지수 세부 구성 요인, 동종업계 평균 정보를 포함하는 데이터프레임

ElasticSearch server

2. 동종업계 역량평가 데이터

테이블명		STDAD_DATE	INDUST_CODE_2	TOTCR_SC	TOT_IP_SC	TOT_BM_SC	TOT_RI_SC	TOT_MR_SC	TOT_HR_SC	TOT_RD_SC	TOT_FI_SC
KHU_AVG_SC		기준년월	동종업계분류코드	신용지수역량점수	IP지수	BM지수	평판정보지수	마케팅역량지수	인적자원지수	R&D역량지수	미래성장지수
		202205	58	64.45	54.51	34.54	64	51		65	46
		202205	64	54	6	54	6	15	15.5	64.8	54

1. 기업 역량평가 데이터

테이블명	BIZ_NO	INDUST_CODE	STDAD_DATE	INDUST_CODE_2	TOT_CR_SC	TOT_IP_SC	TOT_BM_SC	TOT_RI_SC	TOT_MR_SC	TOT_HR_SC	TOT_RD_SC	TOT_FI_SC
KHU_CMP_SC	사업자번호	기업산업분류코드	기준년월	동종업계분류코드	신용지수역량점수	IP지수	BM지수	평판정보지수	마케팅역량지수	인적자원지수	R&D역량지수	미래성장지수
	1234567890	58221	202205	58	46		78		45	46	15	
	6456465465	58461	202205	58	64	46	15	51	51	61	16	15

현재 엘라스틱서치에 존재하는 모든 기업들의 정보를 데이터프레임 형식으로 로컬에 저장

khu_get_summarize.py

요약문 생성 함수 실행 (print_summarize_sentence())

요약문 생성 및 출력

현대제철은 제철업 분야의 대기업으로, 1953년부터 70년간 사업을 영위하고 있다. 대표자는 안동일이며, 2022년 4월 기준 종업원은 11,528명이다.

현대제철은 동종업계 기업들의 평균 대비 R&D, 인적자원, BM 지수가 우수한 역량을 보이고 있다. 반면, 마케팅, 미래성장, 평판 지수의 경우 동종업계 기업들의 평균 대비 부족한 역량을 보이고 있다.

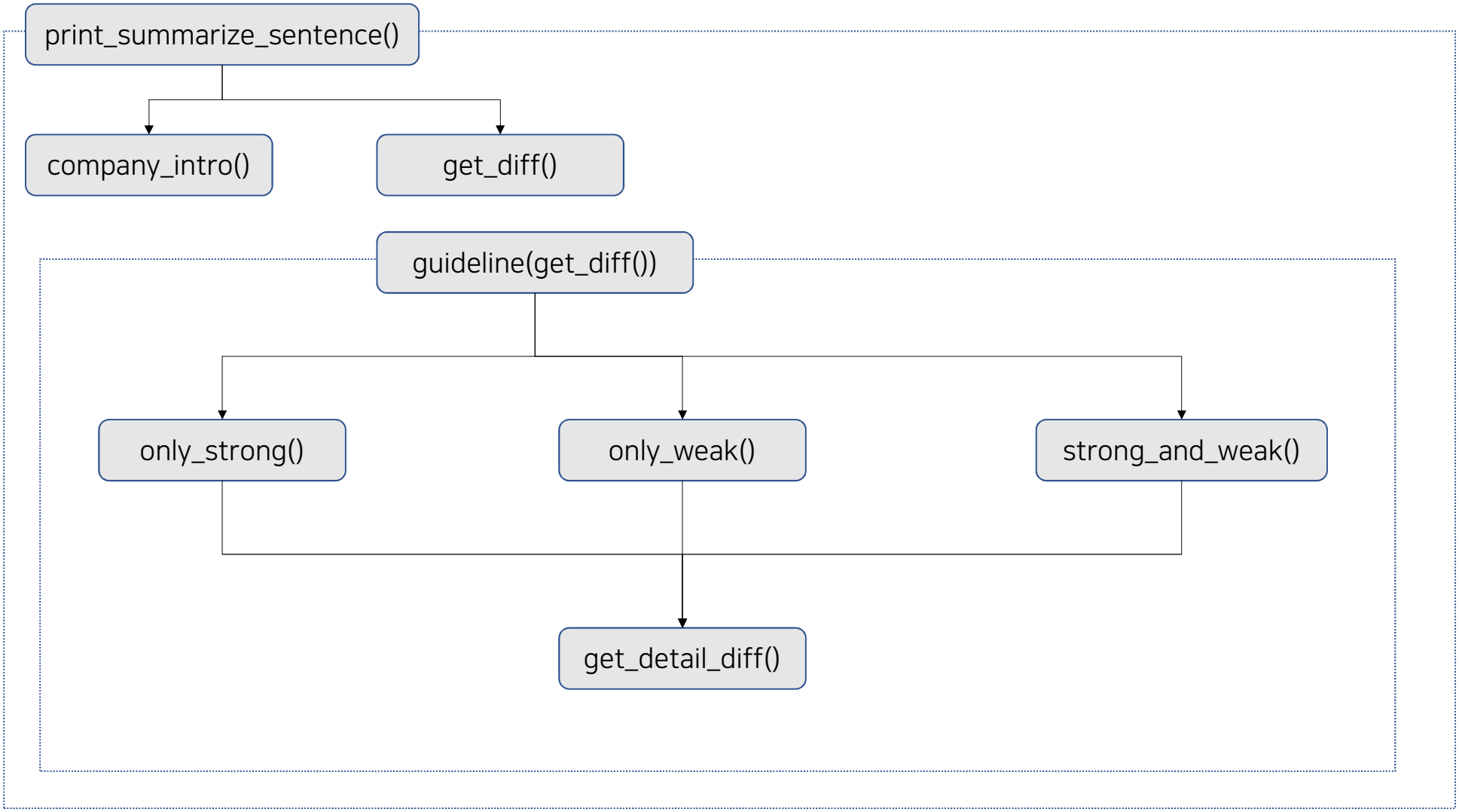
특히, 마케팅 지수에서는 영업/마케팅 인력 비중, 미래성장 지수에서는 매출액증가율이 가장 취약한 것으로 확인되었으며, 그리고 평판 지수의 경우 상위 5개의 부정적 키워드는 [갑질, 서비스, 불량, 제품, 배송]이다.

따라서 현대제철은 고객응대 서비스, 인력 및 네트워크 보유 측면에서 마케팅 비용 투자를 진행할 필요가 있다.

- ‘khu_get_summarize.py’ 파일을 구성하는 함수 목록은 아래와 같음.

  - print_summarize_sentence(): 피컨설팅 기업에 대한 전체 요약문을 출력 -> *main function*
  - guideline(): 피컨설팅 기업의 보완점을 문장 형식으로 출력
  - company_intro(): 피컨설팅 기업의 개요를 출력
  - get_diff(): 6대 지수에 대하여, 피컨설팅 기업의 값과 동종업계 평균 값의 차이를 계산
  - get_detail_diff(): 특정 지수에 대하여, 세부 구성 요인값의 차이를 계산(피컨설팅 기업의 값 – 동종업계 평균 값)
  - only_strong(): 피컨설팅 기업의 전체 지수 값이, 동종업계 평균보다 높은 경우에 해당하는 요약문 생성&출력
  - only_weak(): 피컨설팅 기업의 전체 지수 값이, 동종업계 평균보다 낮은 경우에 해당하는 요약문 생성&출력
  - strong_and_weak(): 전체 지수 값 중 일부 지수는 동종업계 평균보다 낮고, 나머지 지수는 동종업계 평균보다 높은 경우에 해당하는 요약문 생성& 출력
  - 단, 특정 지수가 결측값인 경우, 해당 지수에 대한 문장 구성은 생략함.

- ‘khu_get_summarize.py’ 파일을 구성하는 함수 간 호출 관계도는 아래와 같음.



정보 요약 모델 함수 정의서: company_intro()

함수명	입력	출력 및 반환 형식	설명	비고
company_intro	1. 사업자 번호 2. 데이터프레임	• 출력: 기업 개요에 해당하는 요약문 **(슬라이드 6~8의 첫 번째 문단)** • 반환 형식: 없음(None)	• 모든 정보가 있는 경우와, 일부 정보만 있는 경우를 구분함 • 기업 개요를 구성하는 정보가 하나도 없는 경우, 다른 사업자 번호를 입력받게 함	입력 데이터프레임은 각 기업 당 6대 지수 값, 세부 요인 값 및 동종업계 평균 값이 들어간 정보를 의미함. **(세윤에서 제공한 khu_get_data_all() 함수의 반환 데이터프레임)**

✓ def company_intro(company_number: Union[int, float], df: pd.DataFrame)
- company_number: 사업자 번호
- df: 데이터프레임

정보 요약 모델 함수 정의서: get_diff()

함수명	입력	출력 및 반환 형식	설명	비고
get_diff	1. 사업자 번호 2. 딕셔너리 3. 데이터프레임	• 출력: 없음(None) • 반환 형식: dict -> 6대 지수에 대하여, 피컨설팅 기업과 동종업계 평균 값의 차이를 계산한 딕셔너리 **(계산 방식은 슬라이드 10의 step 2 참고)** -> Key: 지수 이름, value: 차이 ✓ 예) {'R&D역량지수': np.nan, '인적자원지수': 20, …}	• 아래에 해당하는 경우, 반환 딕셔너리의 value를 **결측값 (np.nan)**으로 정의함. ✓ 피컨설팅 기업의 지수값 또는 해당 지수에 대한 동종업계 평균 값 중, 어느 하나라도 결측값인 경우	• 입력 데이터프레임 형식은 슬라이드 27의 내용과 동일 • 입력 딕셔너리는 (지수 이름 – 세부 구성 요인) 쌍으로 이루어졌으며, 정의 내용은 아래 'detail_dict' 를 참고

✓ def get_diff(company_number: Union[int, float], detail_dict, df: pd.DataFrame) -> dict
- company_number: 사업자 번호
- detail_dict: (지수명-세부요인 리스트)의 쌍으로 구성된 딕셔너리
- df: 데이터프레임(기업 당 6대 지수, 지수 세부 구성 요인, 동종업계 평균값을 포함)

✓ detail_dict = {"R&D역량지수": ['기술인력비중평가점수', '전년대비R&D투자증액수준평가점수', '국가R&D실적평가점수', '4차산업혁명대응수준평가점수'],
"인적자원지수": ['경영지원/전략기획인력수준평가점수', '종사자1인당매출액수준평가점수', '기업종사자비율증가수준평가점수'],
"BM지수": ['고용전망평가점수','체감경기전망평가점수', '매출액증가추이평가점수', '사업화기반구축수준평가점수'],
"미래성장지수": ['매출액증가율', '순이익증가율', '신용점수', 'R&D점수', 'BM점수', 'BM2기술성점수', 'BM2사업성점수'],
"평판정보지수": ['긍정백분율', '중립백분율', '부정백분율'],
"마케팅역량지수": ['고객응대서비스고도화수준평가점수','마케팅네트워크보유역량평가점수', '영업·마케팅인력비중평가점수','선도기업대비마케팅비용비중평가점수']}

정보 요약 모델 함수 정의서: get_detail_diff()

함수명	입력	출력 및 반환 형식	설명	비고
get_detail_diff	1. 사업자 번호 2. 딕셔너리 3. 데이터프레임 4. 지수명	• 출력: 없음(None) • 반환 형식: dict -> 입력받은 지수에 대하여, 지수를 구성하는 각 세부요인의 차이를 계산한 딕셔너리 **(슬라이드 10의 step 4 참고)** -> Key: 세부 요인명, value: 차이	• 아래에 해당하는 경우, 반환 딕셔너리의 value를 **결측값 (np.nan)**으로 정의함. ✓ 피컨설팅 기업의 세부요인값 또는 해당 세부요인에 대한 동종업계 평균값 중, 어느 하나라도 결측값인 경우	• 입력 데이터프레임 형식은 슬라이드 27의 내용과 동일 • 입력 딕셔너리는 (지수 이름 – 세부 구성 요인) 쌍으로 이루어짐. **(슬라이드 28의 detail_dict)**

✓ def get_detail_diff(company_number: Union[int, float], detail_dict: dict, factor: str, df: pd.DataFrame) -> dict
- company_number: 사업자 번호
- detail_dict: (지수명-세부요인 리스트)의 쌍으로 구성된 딕셔너리
- factor: 지수명
- df: 데이터프레임(기업 당 6대 지수, 지수 세부 구성 요인, 동종업계 평균값을 포함)

정보 요약 모델 함수 정의서: only_strong()

함수명	입력	출력 및 반환 형식	설명	비고
only_strong	1. 사업자 번호 2. 딕셔너리 3. 데이터프레임	• 출력: 요약문(**슬라이드 7의 두 번째 및 세 번째 문단 내용 참고**) • 반환 형식: str -> 요약문은 print 함수를 통해 출력하고, 세부요인명을 반환 ✓ 예) 'R&D 역량지수' ✓ 해당 반환 값은 guideline() 함수에서 보완점을 출력하기 위한 용도로 사용	• 피컨설팅 기업의 6대 지수 값 전체가 동종 업계 평균 값보다 클 경우에 해당하는 요약문을 출력 • 반환 값은, 동종업계 평균 값과 가장 차이가(+ 방향) 적은 지수에 대한 이름 ✓ 보완 가이드라인 문장 생성에 이용됨. • 결측값을 가지는 지수는 요약문 출력 대상에서 제외	• 입력 데이터프레임 형식은 슬라이드 27의 내용과 동일 • 입력 딕셔너리 형식의 예시는 아래 'main_factor_2' 변수와 정의 방식과 같음 ✓ **결측값인 지수 정보를 제외한 상태**

✓ def only_strong(company_number: Union[int, float], main_factor_2: dict, df: pd.DataFrame) -> str
- company_number: 사업자 번호
- main_factor_2: 딕셔너리
- df: 데이터프레임

✓ example) main_factor_2 = {'R&D역량지수': 40, '인적자원지수': 20, '마케팅역량지수': 75}
- main_factor_2는 get_diff(슬라이드 28) 함수의 반환 값에서 결측값에 해당하는 지수 정보를 제외한 딕셔너리
- 즉, 6대 지수의 정보 중 일부 지수의 정보만 포함이 될 수 있음

정보 요약 모델 함수 정의서: only_weak()

함수명	입력	출력 및 반환 형식	설명	비고
only_weak	1. 사업자 번호 2. 딕셔너리 3. 데이터프레임	• 출력: 요약문(**슬라이드 8의 두 번째 및 세 번째 문단 내용 참고**) • 반환 형식: str -> 요약문은 print 함수를 통해 출력하고, 세부요인명을 반환 ✓ 예) 'R&D 역량지수' ✓ 해당 반환 값은 guideline() 함수에서 보완점을 출력하기 위한 용도로 사용	• 피컨설팅 기업의 6대 지수 값 전체가 동종업계 평균 값보다 작을 경우에 해당하는 요약문을 출력 • 반환 값은, 동종업계 평균 값과 차이가(- 방향) 큰 3개의 지수에 대한 이름 ✓ 보완 가이드라인 문장 생성에 이용됨. • 결측값을 가지는 지수는 요약문 출력 대상에서 제외	• 입력 데이터프레임 형식은 슬라이드 27의 내용과 동일 • 입력 딕셔너리 형식은 슬라이드 30의 'main_factor_2' 변수 형태와 동일 ✓ **결측값인 지수 정보를 제외한 상태**

✓ def only_weak(company_number: Union[int, float], main_factor_2: dict, df: pd.DataFrame) -> str
- company_number: 사업자 번호
- main_factor_2: 딕셔너리
- df: 데이터프레임

정보 요약 모델 함수 정의서: strong_and_weak()

함수명	입력	출력 및 반환 형식	설명	비고
strong_and_weak	1. 사업자 번호 2. 딕셔너리 3. 데이터프레임	• 출력: 요약문(**슬라이드 6의 두 번째 및 세 번째 문단 내용 참고**) • 반환 형식: list -> 요약문은 print 함수를 통해 출력하고, 지수 이름 리스트를 반환 ✓ 예) ['BM지수', '마케팅역량지수'] ✓ 리스트에 포함된 요소는 동종업계 평균 값보다 낮은 지수들로 구성 ✓ 해당 반환 값은 guideline() 함수에서 보완점을 출력하기 위한 용도로 사용	• 피컨설팅 기업의 6대 지수 값이 동종업계 평균에 비해 큰 경우와 작은 경우가 모두 존재할 때 요약문 출력 • 반환 값은, 동종업계 평균 값보다 값이 작은 지수(들)의 이름 리스트 ✓ 보완 가이드라인 문장 생성에 이용됨. • 결측값을 가지는 지수는 요약문 출력 대상에서 제외	• 입력 데이터프레임 형식은 슬라이드 27의 내용과 동일 • 입력 딕셔너리 형식은 슬라이드 30의 'main_factor_2' 변수 형태와 동일 ✓ **결측값인 지수 정보를 제외한 상태**

✓ def strong_and_weak(company_number: Union[int, float], main_factor_2: dict, df: pd.DataFrame) -> list
- company_number: 사업자 번호
- main_factor_2: 딕셔너리
- df: 데이터프레임

정보 요약 모델 함수 정의서: guideline()

함수명	입력	출력 및 반환 형식	설명	비고
guideline	1. 사업자 번호 2. 가이드라인 딕셔너리 3. 지수 딕셔너리 4. 데이터프레임	• 출력: 요약문**(슬라이드 6~8의 마지막 문단 내용 참고)** • 반환 형식: 없음(None)	• 피컨설팅 기업에 대한 보완 가이드라인을 생성 및 출력 • 결측값을 가지는 지수는 요약문 출력 대상에서 제외	• 입력 데이터프레임 형식은 슬라이드 27의 내용과 동일 • 가이드라인 딕셔너리는 (지수 이름: 가이드라인) 쌍으로 이루어짐. ✓ 슬라이드 9 참고 • 지수 딕셔너리 형식은 슬라이드 30의 'main_factor_2' 변수 형태와 동일 ✓ 결측값인 지수 정보를 제외한 상태

✓ def guideline(company_number: Union[int, float], guide_dict: dict, main_factor_2: dict, df: pd.DataFrame)
- guide_dict: 가이드라인 딕셔너리 ('지수 이름': 가이드라인) -> 파이썬 파일 내부에 상수로 정의함
- main_factor_2: 지수 딕셔너리 ('지수 이름': 지수 값)
- df: 데이터프레임
- company_number: 사업자 번호

정보 요약 모델 함수 정의서: print_summarize_sentence() -> **main function**

함수명	입력	출력 및 반환 형식	설명	비고
print_summarize_sentence	1. 사업자 번호 2. 데이터프레임	• 출력: 피컨설팅 기업에 대한 전체 요약문(슬라이드 6, 7, 8 中 하나의 요약문 출력) • 반환 형식: 없음(None)	• 피컨설팅 기업 개요, 기술기반 정보 및 보완 가이드라인 내용을 출력 • 결측값을 가지는 지수는 요약문 출력 대상에서 제외	• 입력 데이터프레임 형식은 슬라이드 27의 내용과 동일

✓ def print_summarize_sentence(company_number: Union[int, float], df: pd.DataFrame)
- company_number: 사업자 번호
- df: 데이터프레임