자료와 정보 57458 2024년 봄학기 탐구 과제 1

제출 마감: 2024년 5월 19일(일) 23:59까지

탐구 내용: 어떤 데이터 세트에서 클러스터링 방법을 통해 몇 개의 대표값을 추출하고 그 정보를 이용하여 목적에 맞는 파생 정보(결론)를 최종 생성한다.

1. kaggle.com에서 자신의 탐구 과제에 들어맞는 데이터 세트를 고른다. (또는 다른 곳에서 얻은 데이터를 이용해도 좋습니다.)
2. 이 데이터를 활용해서 어떤 정보를 추출하고 그 정보로부터 어떤 결론에 이를 수 있는 지식(또는 파생정보, 가령, 가치가 있는 정보의 형태)을 끌어낼 수 있는지 전체 과정을 정한다.
3. 예를 들어, 우리나라 20~30세 남, 녀의 신체 수치 데이터를 선택했다면, 키와 몸무게가 유사한 그룹을 몇 개로 나눌 수 있는지 알아보고 이를 토대로 새로운 케쥬얼 상하의 세트를 맞들 때 기준이 되는 옷 치수의 표준을 만든다. 등등
4. 과정에 맞는 군집화 방법을 2가지 이상(k-means, k-medoids…. 등등 또는 이와 같은 방법의 변형 방법)을 적용하여 최종 결론을 얻고 결과의 장단점, 파생 가치 등을 논한다. 예를 들어 이때 최종 군집에서 끌어낼 정보의 유효성을 판정할 기준을 정하고, 하나의 집단의 판정 기준은 최대/최소의 범주를 전체 평균 치수의 10% 정도의 단위 이내여야 한다. 등등. 이를 토대로 파생 정보를 생성하여 그 성능을 평가한다.

제출할 내용: 다음의 내용을 자신의 학번으로 파일명으로 하는 하나의 압축 파일을 만들고 과제란에 마감 이전에 제출해야 합니다.

1. 실행 파일 및 소스 코드 (윈도우 또는 리눅스에서 실행 가능한 파일과 실행 파일을 생성 가능한 보조 파일도 함께 제출해야 합니다.)
2. 프로그램은 소스 코드, 초기 조건, 데이터 파일을 읽어 분할표를 <학번>.csv 파일에 저장한 결과 파일. 설정 조건에 따라 다수의 분할표를 만든다면 <학번>-<순번>.csv로 생성 (올바른 파일 형식이 아닐 경우 0점)
4. 분석 결과를 도출하는 조건, 소프트웨어 구조와 실험 결과를 설명하는 보고서 (A4 일반 여백, 폰트 10, 행간 130%, 5장 내외).

평가 기준 및 방법: 다음의 항목을 기준으로 과제물을 평가합니다.

1. 탐구할 내용: 30% (탐구 과정의 완성도, 제출한 결과의 신뢰도)
2. 군집화 적용: 30% (군집화의 성능 판정 정도와, 실행 시간 등. 프로그램 실행의 효율성은 각자 제출한 보고서에 기재된 프로그램 실행 방법을 기준으로 측정한다.)
3. 보고서 완성도: 40% (내용:20%, 형식:20%)

* 주의 사항
- **만일 두 사람의 과제 내용이나 데이터 세트의 구성이 일반적 수준 이상으로 같으면 두 학생의 성적 모두 낙제 처리합니다.**
- 소프트웨어 패키지나 툴을 사용하면 50%만 인정합니다. 자신이 라이브러리등의 도움 없이 클러스터링 소스 코드를 직접 작성하면 100% 인정합니다.
- 제한된 마감 시간 이후에 제출되는 어떠한 형태의 것도 평가에서 제외됩니다. 반드시 제한 시간 이전에 업로드 완료해야 합니다. 마감에 인접하여 통신장애로 인해 업로드하지 못한 것도 인정되지 않습니다. 미리미리 업로드하기 바랍니다. (수정이 필요하다면 다시 업로드하면 됩니다.)
- **과제 제출물을 제출하지 않으면 나머지 시험 점수와 관계없이 낙제(F)입니다. 단 과제 1이나 과제 2중 하나만 완료하면 됩니다. 둘다 제출하면 높은 점수를 인정합니다.**