(1) 텍스트 마이닝, 데이터 마이닝의 차이를 알아보기

공통점 : 패턴을 찾는다.

차이점 : 텍스트 마이닝(비정형 데이터), 데이터 마이닝(정형 데이터)

(2) 텍스트 분석 프로세스의 6단계를 NCS 모듈 교재에서 찾아서 적어보자.

요구사항 분석 - 텍스트 수집 - 텍스트 저장 및 전처리 - 텍스트 분석 - 텍스트 분석 서비스 제공 - 산출물 관리 및 공유

(3) 텍스트 군집은 무엇인지 찾아서 간단하게 적어보자.

- 특성을 분석하여 그 내용 혹은 형태가 유사한 텍스트를 군집하기는 기술

분류의 군집의 차이 : 카테고리가 정해져 있는 것과 아닌 것.

(4) 형태소의 정의에 대해 찾아서 적어보자.

- 의미가 있는 최소 단위, 더 이상 분리가 불가능한 가장 작은 단위.

(5) 카테고리 구축 절차에 대해 찾아 적어보자.

(1) 카테고리 키워드 구축을 위한 입력 텍스트를 준비

(2) 입력 텍스트에 대하여 자연어 처리를 한다.

(3) 추출된 형태소에 대하여 가중치 분석을 통한 핵심 키워드를 추출한다.

(4) 추출된 키워드를 분류하기 분류 체계를 구성한다.

(5) 추출된 키워드에 대하여 키워드 분류 엔진을 적용하여, 각 카테고리에 대한 키워드를 구축한다.

(6) 카테고리별로 구축된 키워드에 대하여 검증 과정을 거쳐, 카테고리의 특성을 반영하지 못하거나 의미적 모호성이 있는 키워드는 제거한다.

(6) 데이터 유형의 3가지를 적어보자.

1. 정형 데이터(structured data) - RDB, 엑셀

2. 비정형 - 동영상, 텍스트 데이터

3. 반정형 - XML, HTML

(7) 텍스트 분류 방법에 대해 알아보자.

- knn

- 의사결정트리

- SVM

- 나이브 베이즈

(8) nltk 패키지에 대해 아는 대로 적어보자.

- 자연어 처리 문서 분석용 파이썬 패키지

- 50개 이상의 말뭉치(Corpus)와 어휘가 있다.

- 말뭉치 토큰 생성, 형태소 분석, 품사 태깅을 자동으로 해준다.

(9) konlpy 패키지에 대해 아는 대로 적어보자.

- 박은정.

- python

- 한국어 정보 처리

- 한글 엔진

꼬꼬마, 한나눔, 트워터, MeCab(리눅스), 코모란