* TF-IDF 기반 이직 대상 기업 추천 시스템 설계 및 구현
* TF-IDF
* 문서 내에서 특정 단어가 갖는 중요도를 나타내는 가중치
* 문서 내에서 특정 단어가 출현하는 빈도와 특정 단어를 갖고 있는 문서가 전체 문서에서 차지하는 비율의 역수에 로그를 취한 값을 곱하여 구한 값
* 텍스트 마이닝 연구에 많이 쓰이는 방법 중 한 가지
* 추천 시스템
* 사용자간 취향의 유사도를 이용하는 협업 필터링 기법 Collaborative Filtering
* 추천 대상인 아이템의 유사도를 이용하는 내용 기반 필터링 Content based Filtering
  + 내용 기반 필터링 Content based Filtering
    - 정보 검색 기술에 바탕을 둔 시스템
    - 아이템의 콘텐츠를 분석하여 아이템-아이템 혹은 아이템-사용자 선호도 간 유사성을 분석하여 이를 토대로 고객에게 아이템을 추천해주는 방식
  + 협업 필터링 Collaborative Filtering
    - 사용자별 아이템에 대한 선호도를 기초로 사용자간 유사도를 계산하여 아이템을 추천하는 방식
* 본 논문
* 이직 대상 기업을 추천하는 방법을 제안
* 이직이라는 행위 = 소비자가 온라인 쇼핑몰에서 상품을 구매하는 행위와 큰 차이가 있음
* 이직은 소비자가 물건을 구매하거나 영화를 보는 것과 같이 빈번하게 일어나는 행위가 아님
* 이러한 특성으로 협업 필터링 기반의 추천 방식을 적용하는 것이 제한됨
* 이에 본 연구에서는 기업에 대한 리뷰를 수집하여 각 기업별 리뷰 문서를 정의하고 그 문서의 내용을 분석하여 추천하는 방법을 활용
* TF-IDF 기반 이직 대상 기업 추천 시스템 연구 방법
* 리뷰 수집을 위해 웹 크롤링 Web Crawling
* 기업 별로 50개의 리뷰를 한 개의 value로 MariaDB에 저장
* 서버에는 MariaDB에서 리뷰를 불러와 불용어 제거, 토큰화, 품사 정보 입력 등의 전처리를 실행
* 전처리 된 리뷰는 형태소 단위로 나뉜 상태, 그 중에 명사만을 추출
* 문서 별 단어들의 TF-IDF를 오름차순으로 정렬하여 Keyword 추출
  + 유사도를 측정하는 기준
* 클라이언트가 리뷰를 입력하면 앞과 동일한 전처리 과정을 거침
* 전처리 이후 명사를 추출하고 이 명사들과 기업 별 Keyword를 비교하여 유사도를 계산

텍스트, 도표, 스크린샷, 평면도이(가) 표시된 사진

자동 생성된 설명

* 리뷰 데이터 수집 및 전처리
* 웹 크롤링 방식으로 데이터를 수집
  + 웹 크롤링 Web Crawling
    - 브라우저 상의 HTML 태그들을 바탕으로 필요한 부분을 자동으로 추출하는 것
    - Beautifulsoup ( 파이썬의 라이브러리 ) / Selenium ( 테스트 도구 ) 을 활용
* 수집 대상은 코스피, 코스닥에 상장된 기업 중 각 시장 별 시가총액 상위 50대 기업 중에서 리뷰의 개수를 50개 이상 갖는 59개 기업에 대한 리뷰
* 본 연구는 기업 별 리뷰 모음을 한 개의 문서로 정의
* 수집된 전체 데이터는 문서들의 모음이고 corpus라고 정의
* 각 기업 별 최신 순으로 50갸의 리뷰를 수집, 실험에 사용된 리뷰의 개수는 2950개
* 데이터 전처리
* 볼용어 제거 : 의미 있는 정보를 갖고 있지 않고 문법을 지키기 위해 사용하는 단어
* 토큰화 : 문장을 형태소로 나누는 작업
  + 형태소 : 의미를 갖는 가장 작은 언어 단위
* 품사 정보 입력 :
* 전처리 후
* TF-IDF를 계산하고 그 중에서 명사만 추출하여 Keyword로 사용
* Corpus와 문서에서 높은 출현 빈도를 갖고 있으나 문서의 내용을 설명하는 데는 역할을 하지 않음
* TF-IDF 계산
* 한 단어가 문서 내에서 갖고 있는 중요성의 정도를 나타내는 지표
* TF Term Frequency
  + 특정 단어가 전체 문서에서 차지하는 빈도
* IDF Inverse Document Frequency
  + 특정 단어가 등장하는 문서의 개수를 전체 문서의 개수로 나눈 값에 로그를 취한 값
* TF-IDF는 TF와 IDF의 곱의 결과 값으로 한 문서 안에서 단어들이 갖는 중요도를 나타냄
* 한 문서에서 특정 단어의 TF-IDF 값이 클수록 그 단어는 다른 문서에는 덜 등장하고 현재 문서에서 높은 빈도로 등장한다고 직관적으로 판단
* 본 논문에서는 기업 별로 리뷰를 수집하고 분석
* 이 때 기업 별 리뷰 모음을 한 개의 문서로 정의

3.2까지의 정리