텍스트 마이닝

60161589 권선구

1. 텍스트 마이닝이란?

2. 텍스트 마이닝의 절차

3. 자연어처리의 주요 개념

4. 텍스트마이닝을 활용한 사례

5. 참고문헌

**텍스트 마이닝이란?**

텍스트 마이닝이란 HTML등 전자문서등 텍스트로 구성된 자료에서 유용한 정보를 추출하고 분석하는 기법으로 텍스트는 가장 일반적인 비정형 데이터로 텍스트를 통계와 언어학, 컴퓨터 공학을 이용하여 목적에 맞는 유의미한 정보를 찾고 뽑아내는 것을 텍스트 마이닝이라고 한다.

**텍스트 마이닝의 절차**

텍스트 마이닝은 정보수집->정보처리->정보추출->정보분석으로 절차가 진행되는데 추출을 하는 과정이 전체 과정의 70~80%정도에 해당될 정도로 추출하는 과정이 오래 걸리는 기술이다. 그 이유로는 텍스트는 비정형 데이터라는 것이 문제인데 여기서 정보를 추출하기 위해서는 언어학에 대해서 알아야 진행이 가능하다. 주요 언어학의 개념으로 의미를 가지는 최소단위인 형태소와 문법적 규칙, 품사 등을 알아야하므로 추출하는 과정이 오래 걸리는 기술이다.

**자연어처리의 주요 개념**

텍스트를 처리하는데 있어서 주요개념으로는 형태소 분석, 문자열 분석, 핵심어구 추출, 동시출현 단어, 토픽 모델링, 감정분석등이 있다.

형태소 분석은 의미를 가지는 최소 단위인 형태소나 단어를 통해서 텍스트를 분석하는 방식이다. 이 방식은 가장 보편적인 방식으로 뽑히며 문법적인 규칙이나 품사등을 사용하여 형태소를 분석하는 방식이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

형태소 분석기의 일종인 komoran을 이용한 형태소 분석 예시

문자열 분석의 경우에는 한국어의 글자의 개수를 사용하여 해당 문자열 다음으로 어떤 글자가 나올 지 확률을 통해 예측하는 방식이다.

핵심 어구 추출은 정보를 의미있는 단위로 묶는 청킹을 사용하여 핵심 단어를 추출하는 방식이다.

동시출현 단어의 경우에는 일정한 문맥에서 두 단어가 동시에 등장하는 빈도를 구하여 통계적으로 분석하여 의미있는 단어 쌍을 구하는 방식이다.

토픽 모델링의 경우에는 텍스트에서 구조를 파악하여 카테고리화하는 방식이다.

감정분석의 경우에는 텍스트에서 긍정, 중립, 부정등 감정을 정량화하는 방식이다.

그 밖에는 TF-IDF라는 역문서 빈도 가중치 모델을 사용하여 문서 내부에서 어떤 단어가 중요한지 파악하는 방식이 존재하는데 이 방식은 뉴스 기사에서 자주 언급되는 단어를 파악하는데 유용한 방식으로 TF는 해당 문서가 얼마나 자주 등장하는지를 의미하는 것이고 IDF의 경우에는 전체 문서에서의 희귀도를 의미하는 것으로 이 모델은 TF와 IDF를 곱하여 해당 문서에서 자주 등장하면서도 전체 문서에서는 희귀성이 높은, 가치가 큰 핵심적인 단어를 찾는 방식이다. 이 모델은 다른 마이닝 기법과 사용되는 것으로 예를 들어 감정분석과 같이 사용되면 감정에 영향을 주는 주요 단어를 확인할 수 있다.

테이블이(가) 표시된 사진

자동 생성된 설명

TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석에서의 분석 결과

**텍스트 마이닝을 활용한 사례**

2008년 미국 대통령 선거와 2010년 총선, 2012년 대선 결과를 트위터 게시물을 텍스트 마이닝을 활용해 선거 예측에 활용되었으며 미국에서는 검색엔진에 특정 질병이나 증상에 대한 검색이 급증하는 경우 해당 지역으로 백신이나 의약품을 전달하기도 했다. 국내에서는 뉴스의 감정을 분석하여 뉴스를 통해서 기업에 투자하는 모델을 만들었으며 텍스트 마이닝을 활용하여 SNS의 게시글에서 개인 정보의 포함여부를 판단하기도 했다.



금통위 의사록 워드 클라우드

참고문헌

김수현, 이영준, 신진영, 박기영, 경제분석을 위한 마이닝, 한국은행, 2019

원종호, 이한별, 문혜정, 손원, 텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류, 2017

박종영, 서충원, TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석, 2015

김도훈, 텍스트는 어떻게 분석되는가?, 통계청 통계교육원, 2017, <http://sti.kostat.go.kr/window/2017b/html/2017_win_3.html>

재정정보 분야 텍스트 마이닝 활용 방안 연구, 재정정보원, 2019