빅데이터의이해 HW2

- 워드클라우드(Word Cloud)에 대하여-

201614010 영어영문학과(휴먼ICT) 박도준

오늘날 우리가 살고 있는 현대사회에서는 하루에도 수많은 양의 데이터가 수집되어 축적되고 있습니다. 글로벌 시장 조사 업체인 IDC의 보고서에 따르면 2016년 기준 하루에 생산되는 데이터 량은 약 440억 기가바이트에 달한다고 합니다. 이 중 90%는 글, 사진, 영상과 같은 비 구조적인 형태의 비정형 데이터로 구성되어 있습니다. 텍스트 마이닝은 비정형 데이터 마이닝의 한 종류로서 토픽 추출, 문서 요약, 텍스트 유사도 분석 등 다양한 분석 방법이 있습니다. 텍스트 마이닝의 여러 분석 기법 중 텍스트에 출현한 단어 빈도 수를 시각적인 이미지로 표현하는 기법인 워드클라우드에 대해 살펴보겠습니다.

워드클라우드(word cloud)는 방대한 양의 텍스트로부터 문서의 중심 키워드 및 핵심 개념을 시각적 정보로 제공하여 글에 대한 빠른 직관을 얻도록 도와줍니다. 텍스트를 워드클라우드 이미지로 표현하기 위해서는 몇 가지 중요한 전 처리 과정을 거치게 됩니다. 우선 텍스트에서 특수문자나 불용어 등을 제거하는 텍스트 정제 작업이 우선됩니다. 이 작업은 좋은 결과물을 얻기 위해 수행되는 중요한 과정 중 하나입니다. 다음으로 정제된 텍스트 데이터를 특정 토큰 단위로 나누는 토크나이징(Tokenizing) 단계를 거칩니다. 이 때 한국어 텍스트의 경우 한국어의 언어적 특성에서 발생하는 한 가지 난해한 문제에 부딪힙니다. 영어의 경우 띄어쓰기를 기준으로 각각의 단어가 분리되어 있기에 어절 단위로 문자를 분리할 경우 문장 내에서 단어를 쉽게 구분해 낼 수 있습니다. 반면 한국어의 경우 조사, 높임말, 맺음말 등 여러 의미를 가지는 형태소들이 함께 결합되어 하나의 어절을 구성합니다. 그래서 영어로 된 텍스트처럼 어절 단위로 단어를 구분할 수 없고, 문장의 단어들의 품사 정보를 태깅(Tagging)하는 형태소를 분석 작업을 거쳐야 합니다. 한국어 텍스트를 위한 형태소 분석기로는 한나눔, 꼬꼬마, komoran, mecab, okt가 있으며, 파이썬에서는 konlpy, R에서는 konlp 라이브러리를 통해 쉽게 접근하여 사용할 수 있습니다. 이렇게 텍스트 전 처리 작업이 완료되었다면 출현 단어의 빈도 수를 계산하여 이미지로 출력하는 워드클라우드 분석을 수행하여 이미지를 결과물을 얻을 수 있고, 다양한 옵션 정보를 추가하여 출력 이미지, 글자색, 글꼴 등의 변화를 줄 수 있습니다.

워드클라우드 이외에도 다양한 텍스트마이닝 기법이 존재합니다. 저는 하루가 다르게 발전하고 있는 데이터분석의 발전 토대는 분명 빅데이터에 기반한다고 생각합니다. 빅데이터에 대한 높은 이해는 앞으로 빠르게 변화해 나가는 현대사회 속에서 수동적인 소비자가 아닌, 이 사회가 필요로 하는 능력을 갖춘 능동적인 생산자가 되는 좋은 밑거름이 될 것이라고 생각합니다.