**피플(HR) 애널리틱스: 평가 텍스트 분석**

텍스트 분석을 하는 이유는 주어진 문서를 정량적으로 이해하는 것이다. 현재의 텍스트 분석 기술은 아직 인간의 언어를 이해하는 수준에는 한참 미치지 못한다. HR 맥락에서 대표적인 텍스트 데이터 중 하나인 평가 의견을 중심으로 텍스트 분석의 과정과 시사점을 살펴보겠다.

평가의견은 S, A, B, C 등으로 직원의 성과에 따른 등급을 부여하면서 뭘 특별히 잘해서 S등급을 주었는지, 아니면 뭘 특별히 못해서 C등급을 주었는지에 대한 평가자의 의견이다. 국내에서는 업적과 역량에 대한 평가가 종종 구분되서 이루어지는데 이 글에서는 역량에 대한 평가의견을 다루겠다. 물론, 사람이 평가 의견을 일일이 다 읽어 보고 전체적인 감상을 요약할 수도 있지만 주관이 개입될 소지가 크다.

평가 텍스트 분석은 다른 데이터 분석 작업과 마찬가지로 주어진 평가 문장에서 정량적 특성을 뽑아 요약하거나 서로 다른 집단 간 특성 차이를 비교하는 식으로 이루어진다. 다음과 같은 다양한 질문에 대한 답변을 평가 텍스트 분석으로 찾을 수 있다.

- 고/저성과자 사이에는 어떤 역량 차이가 있는가?
- 직무별로 어떤 역량 차이가 존재하는가?
- 고성과 리더와 저성과 리더의 평가 내용에 차이가 있는가?
- 피평가자 수와 평가자의 평가제도에 대한 효능감 사이에 관계가 있는가?

평가 텍스트 분석은 크게 아래 그림의 절차로 진행할 수 있다. 우선 주어진 평가 문서에서 단어나 문구(토큰 또는 키워드라고 함)를 추출한 후 추출된 토큰(Token)을 분석 목적(토큰의 빈도 랭킹, 직원 집단간 토큰 사용빈도의 차이 분석 등)에 맞게 테이블 형태로 가공하게 된다. 그 후 집단간 텍스트 특성의 차이나 토큰 사이의 관계 등을 분석하게 된다.

[그림: 평가의견 텍스트 분석 절차]


① **토큰/키워드 추출:** 평가 텍스트(문서)를 구성하는 의미있는 구성요소 (단어/문구) 추출

② **문서 변환/요약:** 메타 데이터(평가/피평가자 인사정보)와 토큰을 포함한 평가 문서를 분석 목적에 알맞은 테이블 형태로 변환/가공
③ **문서간 차이 분석:** 직원 집단별 평가 텍스트에 사용된 토큰들의 정략적 차이를 분석
④ **관계 분석:** 평가 텍스트에 사용된 단어/문구(Token)들 사이의 관계 분석

**TF(Term Frequency) – 텍스트에 사용된 단어(토큰)의 빈도**

문서를 정량적으로 이해하기 위한 출발점은 특정 문서(예, 평가의견 텍스트)를 해당 문서에서 사용된 단어(문구)와 해당 단어의 출현빈도로 정리하는 것이다. 문서에서 특정 단어가 얼마나 빈번히 사용되었는지를 측정하는 지표를 TF(Term Frequency; 용어 빈도)라고 한다. 예를 들면, 고성과자의 평가의견으로 구성된 문서와 저성과자의 평가의견으로 구성된 문서 각각에서 추출한 단어와 해당 단어의 빈도로 구성된 테이블을 만들어 서로 비교하여 볼 수 있다.

**지프의 법칙(Zipf's Law) – 주어진 맥락에서 주로 사용하는 단어들은 비슷하다.**

서로 다른 집단(예, 고성과자와 저성과)에 대한 평가의견을 두 종류의 문서로 구분하여 두 문서집합(Corpus: 말뭉치라고 한다)에 주로 사용된 단어나 문구의 빈도를 비교 분석하면 많은 경우 두 집단의 최빈단어들이 비슷하게 추출되는 것을 발견할 수 있다. 이런 결과가 나오는 것은 평가의 맥락에서 사람이 주로 사용하는 어휘들이 업무, 수행, 역량 등 제한된 단어에 집중되어 있기 때문이다. 지프의 법칙(Zipf's Law)이라는 개념이 있는데 충분히 큰 문서집단(장편 소설, 성경책)에서 추출한 단어들의 빈도 랭킹과 빈도를 곱하면 일정하게 동일한 값(상수)이 나오는 현상을 의미한다. 예를 들면, 빈도 랭킹 1위인 단어(업무)의 사용 빈도가 1,000이었다면, 2위 단어(수행)의 빈도는 5,00, 3위(역량)는 333과 같은 식이다.

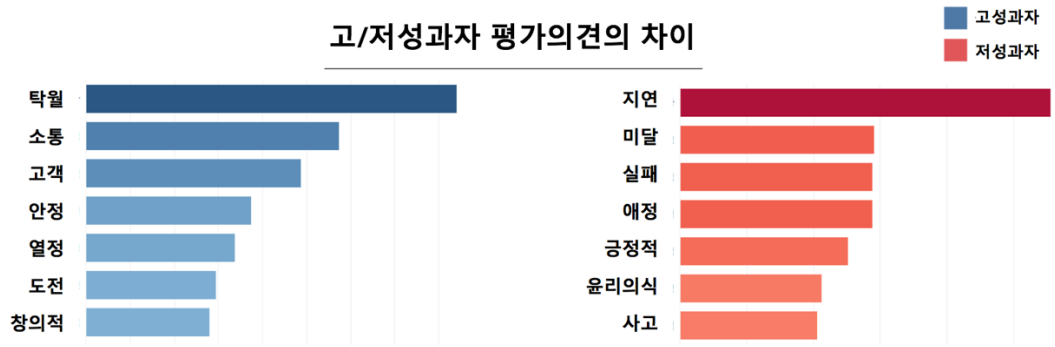
**IDF(Inverse Document Frequency) – 문서의 특성을 잘 나타내는 단어 찾기**

그렇다면, 텍스트 분석에서 문서 간 차이를 효과적으로 드러내는 방법은 무엇일까? 서로 다른 문서 간 특성 차이를 정량적으로 이해하는데 요긴한 지표로 IDF(Inverse Document Frequency)가 있다. IDF는 해당 토큰(단어, 문구)이 다른 문서들에서는 잘 사용되지 않은 경우 증가하게 된다. 특정 문서(고성과자 평가의견)에 사용된 특정 토큰(예, 적극적 의견개진)의 IDF값이 크다는 것은 그 토큰(적극적 의견개진)이 다른 문서들에서는 거의 사용되지 않아서 해당 문서(고성과자 평가의견)의 특징을 잘 대변해 준다고 생각할 수 있다. (수학적으로는 특정 단어의 IDF 값은 총문서의 갯수를 해당 단어가 포함된 문서의 갯수로 나눈 값에 대한 자연로그 값이다. 예를 들어 총 문서의 갯수가 두개였고 특정 단어*(*예, 업무*)*가 두개의 문서에서 모두 사용되었다면 IDF값은 0(ln1)이 되며, 특정 단어(예, 역발상)가 하나의 문서에서만 사용되었다면 IDF는 대략 0.7(ln2)이 된다.)

TF값에 IDF값을 곱한 것을 TF-IDF라고 하는데 개념적으로는 토큰의 빈도(TF)에 토큰의 희소성(IDF)을 곱한 값이라고 이해하면 된다. 아래 그림은 고성과자, 저성과자에 대한 평가의견에 사용된 단어들을

TF-IDF 값으로 순위를 매겨 비교한 예이다. 서로 다른 문서집단 간의 차이가 잘 들어나는 것을 확인할 수 있다.

[그림: TF-IDF로 고/저 성과자의 평가의견에 나타난 역량 차이 발견]

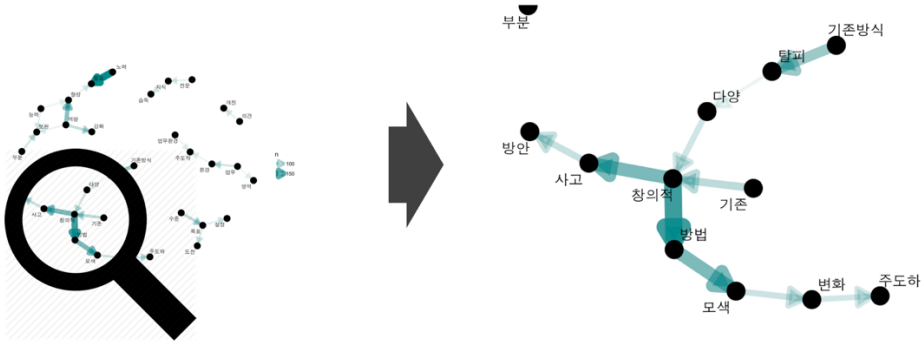



**관계 분석 – 문서를 구성하는 단어들 간의 관계 발견**

개별 토큰(단어나 문구)의 절대(TF) 혹은 상대적(TF-IDF) 빈도로 문서, 혹은 문서 간 특성을 이해하는 방법 이외에 하나의 문서집단(코퍼스) 안에 사용된 토큰들 간의 관계를 분석, 시각화하여 주어진 문서의 특성을 이해하는 방법도 있다. 텍스트에 사용된 토큰들 사이의 관계를 분석하는 대표적인 두가지 방법을 살펴보자.

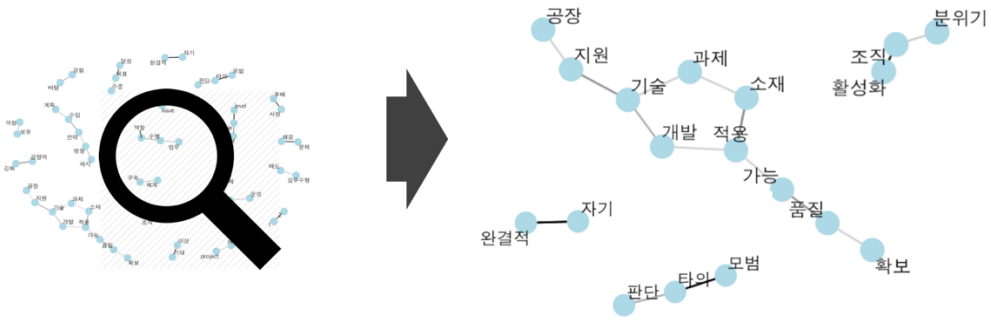
우선, 고성과자에 대한 개별 평가의견들을 별도의 독립된 문서로 보고 각각의 평가의견들에서 연속하여 함께 사용된 단어들의 쌍(예, 창의적 – 사고, 기존방식 – 탈피 등)을 추출하여 단어간의 연관성(Co-occurrence) 분석을 할 수 있다. 아래 그림은 고성과자 평가의견에 사용된 단어들 사이의 연관성을 분석한 그림인데, 그래프(Graph)로 관계를 시각화하는 경우 분석의 결과가 보다 직관적으로 잘 들어나게 된다.

[그림: 토큰 사이의 방향성(directionality)를 고려한 Co-occurrence(연관성) 분석 시각화]



다른 방법으로는 단어(토큰)들 간의 통계적 상관 관계를 분석하는 방법이 있다. 마찬가지로 고성과자에 대한 개별 평가의견들을 별도의 독립된 문서로 간주하여 두 개의 단어가 한 문서에서 함께 사용되는 경향성이 높은 경우 높은 통계값을 부여하는 방식이다. 아래 그림은 단어 간 상관관계를 계산한 결과를 그래프(Graph)로 시각화한 예이다.

[그림: 토큰 사이의 상관관계(phi coefficient 사용)를 계산하여 토큰 사이의 상관 관계를 시각화]




**텍스트 분석의 한계 및 활용방안**

텍스트 분석의 장점은 사람이 일일이 읽기 어려운 분량의 비정형화된 텍스트를 정량적으로 요약했다는 것이고 단점은 언어를 언어로 요약했기에 여전히 사람의 주관과 해석의 여지가 크다는 것이다.

이번에는 본 단락 초기에 언급했던 가설 중 하나인 "피평가자 수가 많을수록 평가자의 평가제도에 대한 효능감이 떨어질 것이다."를 텍스트에서 추출한 순수한 정량적 지표를 사용해서 분석해보자. 위 가설을 검증하기 위해 서베이를 실시하여 평가자들에게 평가제도가 구성원의 역량 향상에 실질적 도움이 되는지에 대한 설문을 실시한 후 평가자 별로 피평가자 수와 평가의견의 길이(평균 글자수)를 계산하여 간단하게 위 가설을 검증할 수 있다. 아래 그림은 평가자의 평가제도에 대한 효능감을 묻는 설문에 도움이 된다고 느끼는 정도가 클 수록 평가의견의 길이는 길어지고 피평가자수는 줄어드는 것을 보여준다.

[그림: 평가자가 느끼는 평가제도에 대한 효능감과 평가 길이, 피평가자수의 관계]


텍스트 분석 기술을 사용하여 평가 텍스트에서 시사점을 발견하는 것에서 한 발 더 나아가 평가시스템을 "Intelligent"하게 개선하는 일에도 활용할 수 있다. 예를 들면 특정 단어가 쓰인 맥락(이전에 사용된 단어들)에 따라 다음에 사용하면 좋을 단어를 예측/추천하는 기능을 생각할 수 있다. 구글이나 네이버 검색 창의 자동완성(autocomplete) 기능을 생각하면 좋겠다. 좋은 평가 텍스트를 학습해서 평가자가 평가의견을 입력할 때 성과개선과 코칭의 관점에서 사용하면 좋을 바람직한 단어들을 기계가 자동으로 추천하게 할 수 있다. 물론, 추천받은 단어나 문구를 사용하고 안하고는 평가자의 선택이다.