비정형 데이터 최종 발표

재직자 리뷰 기반 고 선호도 기업의 만족요인 분석

Seoul Tech
Data science
2020.12.08
김인조
이경찬

서론

연구배경

- 최근 청년 실업난이 심화되면서 사회 문제로 대두되고 있지만 더불어 중소기업 및 스타트업 또한 인력난에 시달리고 있음
- 이는 소수의 복지가 좋은 대기업에만 구직자가 몰리기 때문으로 파악됨
- 청년들이 선호하는 기업의 리뷰를 분석하여 해당 기업을 선호하는 이유를 파악하면 인력난 해소에 기여할 수 있을 것으로 파악됨

연구동기

- 청년들의 기업 선호 요인에 대해 정량적으로 분석한 연구가 많지 않음
- 직원의 입장에서 진솔하게 기업을 평가한 리뷰 데이터를 바탕으로 해당 기업을 선호/비선호 하는 이유를 분석
- 기업 평가가 높은 기업의 만족요인과 불만족 요인을 대조군과 비교분석함으로써 나타나는 차이를 기업 입장에서 활용할 수 있음

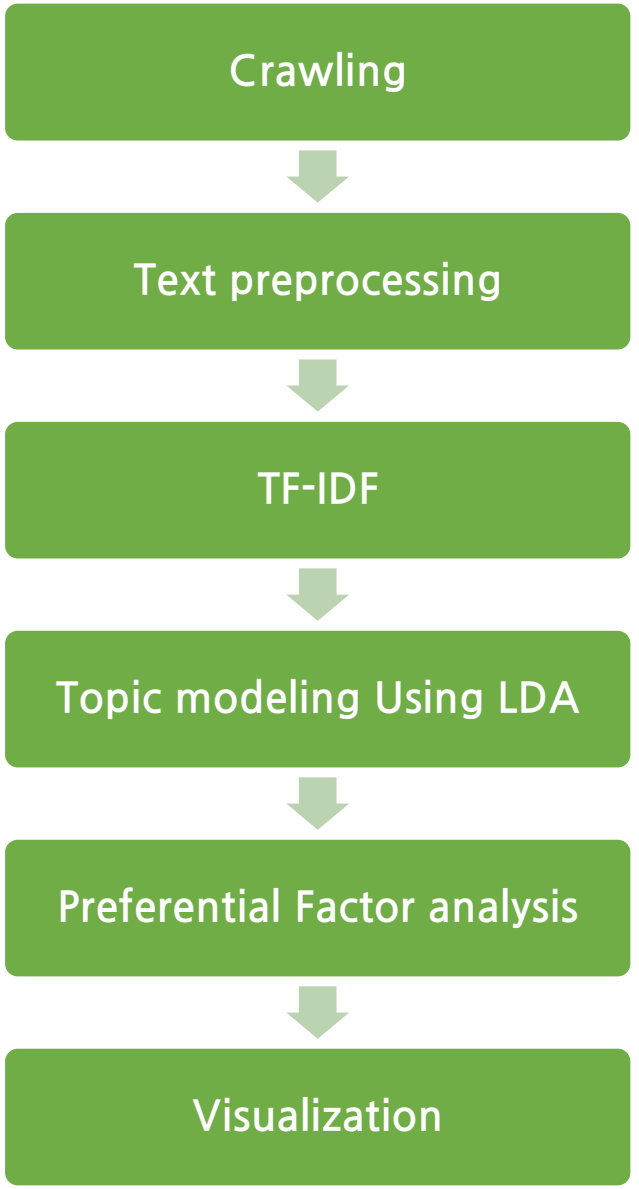
연구 목적

- 기업 정보 리뷰 사이트 '잡플래닛 ' 에서 선호도가 높은 50개의 기업과 평점 2점대 기업 중 50개 기업의 리뷰 데이터를 비교하여 선호 요인을 분석하고자 함

연구 프레임 워크

Preferential Factor analysis

- 추출한 명사를 바탕으로 DTM(Document -Term Matrix) 생성
- LDA를 이용한 Topic modeling
- 추출된 토픽을 바탕으로 선호 요인 파악



Experiment
Data collection

분석 데이터

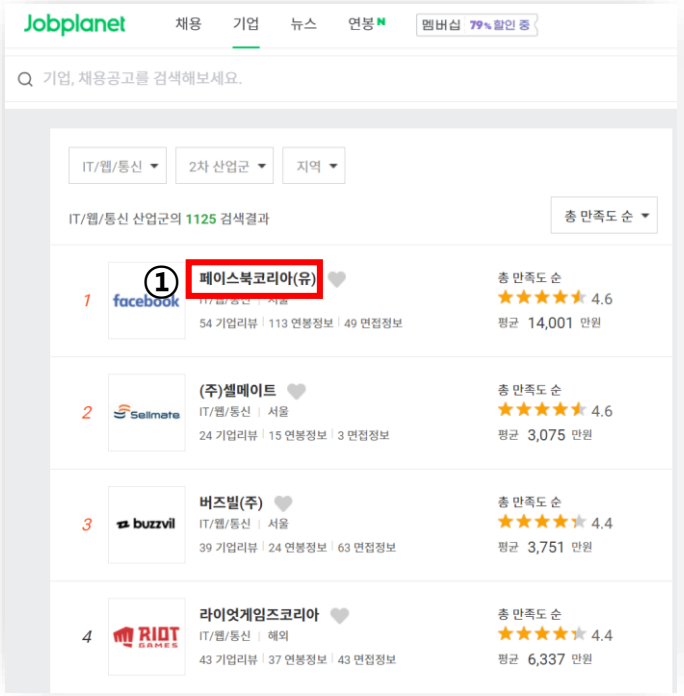
- 잡플래닛 내 IT 산업군 중 만족도 Top 50 기업 (만족도 3.6~5.0)
- 대조군으로 만족도 2점대인 기업 중 랜덤하게 50개의 기업을 샘플링
- 만족도 Top 50인 기업들을 하나의 실험군으로 2점대 기업을 대조군으로 설정

데이터 사전 처리

- 기업당 약 70-150개의 리뷰 데이터가 존재
- 삼성, SK와 같은 대기업은 700개 이상의 리뷰가 존재하여 데이터의 imbalance을 야기하고 선호 이유도 단순히 '대기업이기때문' 이므로 분석 대상에서 제외

Experiment
Data collection

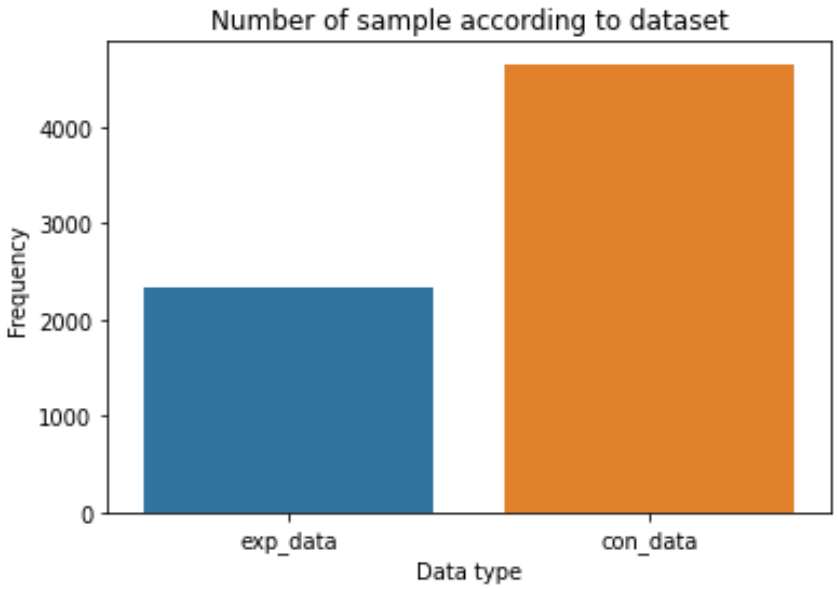
[기업리스트 페이지]



[개별 리뷰 Box]

	①기업명	②총평점	③직종	④재직상태	⑤작성일	⑥리뷰title	⑦장점	⑧단점	⑨경영진에 바라는점	⑩도움	⑪ Score1	Score2	Score3	Score4	Score5
1	페이스북 코리아	5	IT/인터넷	전직원	2020. 07. 31	개발 대우도 잘해주고 여러모로 근무하기 편한 분위기	개방적인 분위기에서 자유롭게 일할 수 있는 환경. 눈치보지 않을 수 있는게 제일 좋음	핵심 업무 외에 해결해야할 다른 업무들이 조금 있어서, 일에 온전하게 집중하기에는 어렵다	직원들의 요구 사항을 잘 들어주셨으면 좋겠습니다.	0	5	5	5	5	5

Experiment
EDA



실험군과 대조군간의 sample의 개수에서 차이를 보임

- 상대적으로 2점대인 기업의 수가 많아 수집된 review data의 수가 더 많음
- Sample의 수에서 약 2배정도의 차이를 보여 이들간의 균형을 맞추어 주어야함
- Con_data에서 다른 유저의 'like' 수가 5개 이하인 리뷰는 임의로 제거
- 최종적으로 실험군 2322개, 대조군 2819개 balanced data로 변환

Experiment
EDA

전,현직자의 분포 상, 대조군의 은퇴자의 리뷰가 가장 많음

- 최초 가설 : 평점이 낮은 기업의 은퇴자들이 보복성으로 안 좋은 리뷰를 많이 남길 것이다.
- 가설대로 대조군의 은퇴자들의 리뷰가 가장 많았음
- 그러나 해당 리뷰들이 대부분 평점이 낮을 것이라는 가설은 기각
- 대조군의 리뷰의 수가 상대적으로 많고, 현직자보다는 퇴직자가 리뷰를 남길 가능성이 높기 때문에 이와 같은 결과가 발생한 것으로 보임

Experiment
Text preprocessing

- 각 corpus 내에서 숫자, 기호등을 모두 무시하고 텍스트만을 추출
- Corpus별로 Tokenize
- 한국어 Lemmatizer를 제공하는 soylemma.Lemmatizer library를 이용하여 lemmatize 진행
- Okt Tag를 명사 Tagging
- Length가 1이하인 단어 제거

원문

“연차제도가 굉장히 좋고 상사의 눈치가 없음. 그리고 무엇보다 수평적인 분위기“

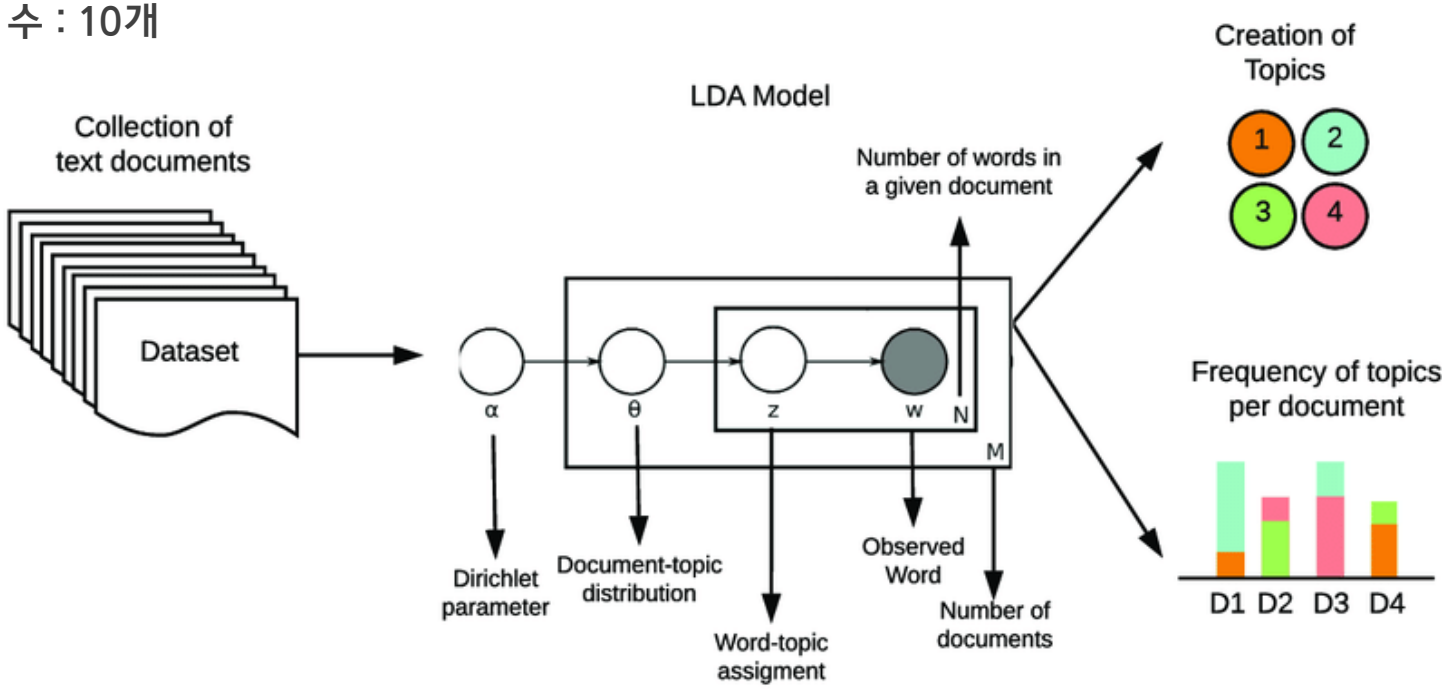
전처리 결과

[연차, 제도, 상사, 눈치, 수평, 분위기]

Experiment
Preference factor analysis

Topic modeling using LDA

- LDA를 이용하여 Topic modeling 진행
- Topic의 수 : 10개
- Topic 당 단어의 수 : 10개



Experiment
Preference factor analysis

Experimental Data

```
pros topics
(0, '0.042*"제공" + 0.029*"회사" + 0.028*"분위기" + 0.027*"수" + 0.021*"업무" + 0.019*"점심" + 0.016*"저녁" + 0.015*"아침" + 0.014*"직원" + 0.013*"개인"')
(1, '0.038*"분위기" + 0.020*"사람" + 0.020*"복지" + 0.018*"연차" + 0.017*"자기" + 0.014*"사용" + 0.014*"회사" + 0.012*"자율" + 0.011*"수" + 0.011*"성장"')
(2, '0.031*"시장" + 0.026*"기업" + 0.023*"중고차" + 0.018*"자체" + 0.017*"성장" + 0.016*"위치" + 0.016*"업체" + 0.013*"시세" + 0.012*"계속" + 0.012*"근무"')
(3, '0.057*"눈치" + 0.038*"연차" + 0.035*"휴가" + 0.030*"퇴근" + 0.025*"분위기" + 0.021*"업무" + 0.020*"안보" + 0.018*"출근" + 0.018*"복지" + 0.017*"수평"')
(4, '0.033*"복지" + 0.032*"직원" + 0.023*"사내" + 0.021*"분위기" + 0.018*"회사" + 0.015*"지원" + 0.014*"야근" + 0.014*"문화" + 0.013*"개발" + 0.012*"성장"')
(5, '0.047*"지원" + 0.021*"연차" + 0.018*"도서" + 0.013*"구입" + 0.013*"장점" + 0.011*"시스템" + 0.010*"사용" + 0.010*"대표" + 0.010*"자유" + 0.009*"공부"')
(6, '0.056*"복지" + 0.030*"회사" + 0.025*"야근" + 0.022*"장점" + 0.018*"연봉" + 0.016*"연차" + 0.016*"분위기" + 0.016*"수준" + 0.014*"제공" + 0.014*"자유"')
(7, '0.038*"회사" + 0.029*"기업" + 0.029*"게임" + 0.019*"문화" + 0.018*"사람" + 0.016*"업계" + 0.015*"안정" + 0.015*"매출" + 0.013*"복지" + 0.013*"업무"')
(8, '0.051*"회사" + 0.047*"분위기" + 0.034*"사람" + 0.034*"업무" + 0.020*"기회" + 0.018*"수평" + 0.018*"복지" + 0.016*"개인" + 0.014*"환경" + 0.012*"본인"')
(9, '0.051*"문화" + 0.045*"기업" + 0.019*"분위기" + 0.016*"수평" + 0.016*"외국" + 0.016*"조직" + 0.016*"회사" + 0.015*"글로벌" + 0.014*"복지" + 0.012*"직원"')
--------------------------------------------------------------------------------------------------------------------------------
cons topics
(0, '0.019*"연봉" + 0.018*"대한" + 0.016*"복지" + 0.015*"신입" + 0.015*"업무" + 0.015*"경력" + 0.014*"체계" + 0.013*"교육" + 0.013*"직원" + 0.013*"회사"')
(1, '0.028*"회사" + 0.020*"직원" + 0.013*"느낌" + 0.013*"조직" + 0.012*"체계" + 0.012*"부분" + 0.012*"경우" + 0.012*"성장" + 0.011*"대해" + 0.011*"문제"')
(2, '0.043*"사람" + 0.024*"분위기" + 0.015*"회사" + 0.014*"업무" + 0.013*"조직" + 0.013*"야근" + 0.012*"직원" + 0.012*"보고" + 0.010*"다소" + 0.009*"자기"')
(3, '0.047*"업무" + 0.030*"회사" + 0.013*"사람" + 0.012*"야근" + 0.012*"경우" + 0.012*"때문" + 0.012*"개인" + 0.011*"자기" + 0.011*"경향" + 0.011*"서비스"')
(4, '0.025*"직원" + 0.021*"사업" + 0.017*"개발" + 0.017*"경영" + 0.017*"문화" + 0.012*"대한" + 0.011*"진의" + 0.010*"수" + 0.010*"이상" + 0.009*"보수"')
(5, '0.048*"연봉" + 0.025*"승진" + 0.022*"부서" + 0.016*"사람" + 0.012*"사업" + 0.012*"정치" + 0.012*"기회" + 0.012*"회사" + 0.011*"복지" + 0.011*"분위기"')
(6, '0.017*"관리" + 0.014*"개발" + 0.011*"부분" + 0.010*"고객" + 0.010*"이사" + 0.010*"인사" + 0.010*"인수" + 0.010*"환경" + 0.009*"차별" + 0.009*"본사"')
(7, '0.089*"업무" + 0.025*"사람" + 0.023*"단점" + 0.019*"체계" + 0.018*"강도" + 0.014*"회사" + 0.014*"생각" + 0.014*"평가" + 0.013*"조금" + 0.012*"장점"')
(8, '0.023*"사람" + 0.018*"퇴사" + 0.017*"야근" + 0.016*"회사" + 0.016*"개발" + 0.014*"변화" + 0.013*"정도" + 0.012*"개발자" + 0.012*"직원" + 0.011*"업무"')
(9, '0.034*"회사" + 0.023*"사람" + 0.021*"기업" + 0.014*"때문" + 0.014*"문화" + 0.013*"분위기" + 0.011*"야근" + 0.011*"생각" + 0.011*"단점" + 0.009*"외국"')
```

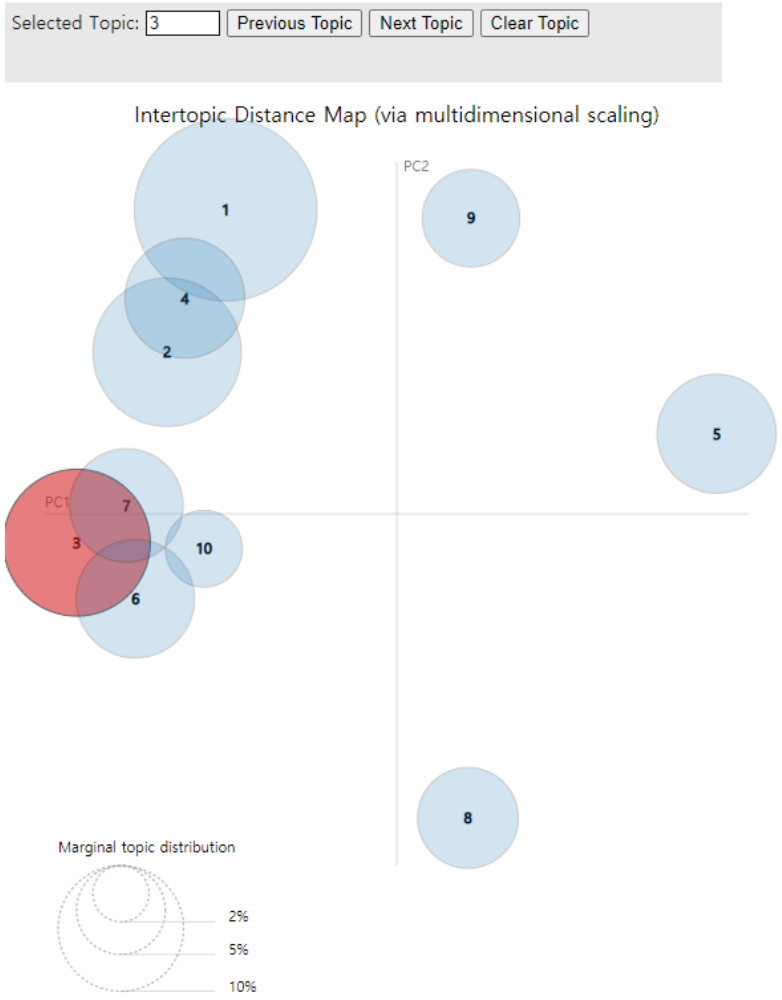
Experiment
Preference factor analysis

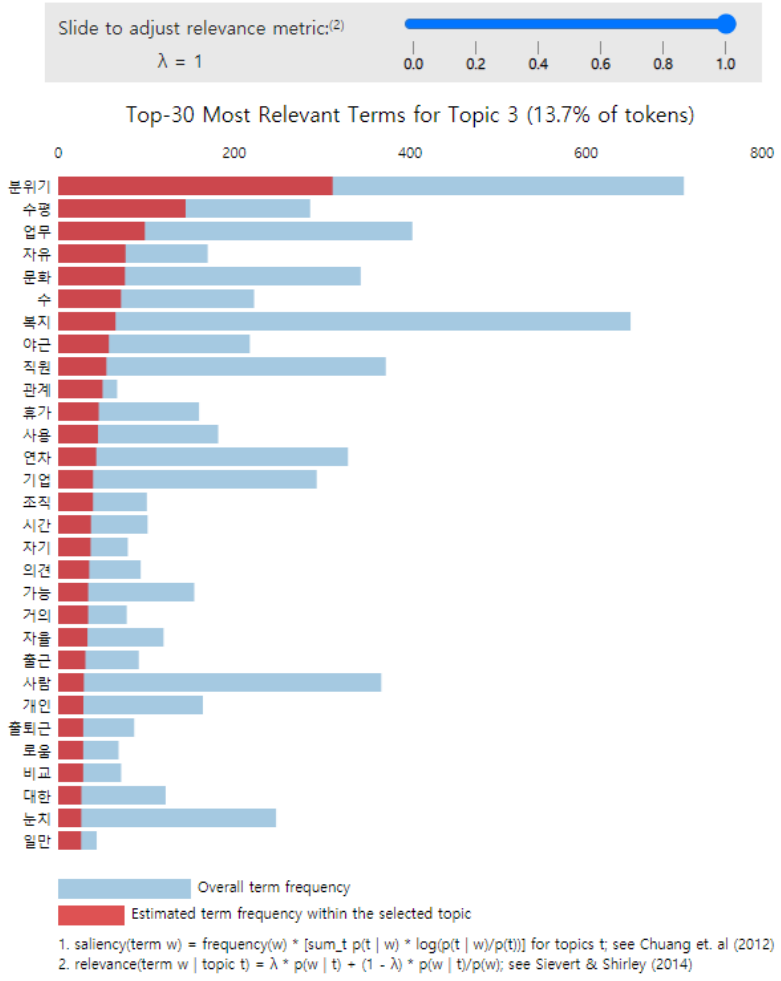
Control Data

```
pros topics
(0, '0.076*"분위기" + 0.063*"직원" + 0.037*"회사" + 0.019*"수평" + 0.016*"복지" + 0.016*"기업" + 0.015*"문화" + 0.014*"관계" + 0.014*"사원" + 0.014*"끼리"')
(1, '0.085*"장점" + 0.069*"월급" + 0.029*"밀리" + 0.023*"정도" + 0.021*"회사" + 0.015*"사람" + 0.014*"프로젝트" + 0.013*"급여" + 0.012*"생각" + 0.012*"하나"')
(2, '0.062*"업무" + 0.030*"경험" + 0.026*"경력" + 0.018*"회사" + 0.013*"생각" + 0.013*"시간" + 0.012*"입사" + 0.012*"신입" + 0.012*"능력" + 0.011*"개인"')
(3, '0.070*"회사" + 0.019*"안정" + 0.015*"복지" + 0.014*"업계" + 0.013*"기업" + 0.012*"매출" + 0.012*"대표" + 0.011*"구조" + 0.010*"개발" + 0.010*"여자"')
(4, '0.042*"위치" + 0.034*"건물" + 0.033*"회사" + 0.025*"카페" + 0.023*"직원" + 0.021*"장점" + 0.015*"식당" + 0.015*"이용" + 0.013*"복지" + 0.012*"구내식당"')
(5, '0.048*"분위기" + 0.045*"사람" + 0.044*"업무" + 0.028*"회사" + 0.014*"수" + 0.013*"강요" + 0.011*"직원" + 0.011*"회식" + 0.011*"복리" + 0.010*"후생"')
(6, '0.066*"연차" + 0.052*"사용" + 0.036*"자유" + 0.033*"분위기" + 0.024*"야근" + 0.024*"근무" + 0.021*"복장" + 0.020*"가능" + 0.018*"로움" + 0.018*"복지"')
(7, '0.073*"제공" + 0.048*"점심" + 0.037*"저녁" + 0.037*"야근" + 0.034*"지원" + 0.031*"커피" + 0.023*"연봉" + 0.023*"식대" + 0.022*"카페테리아" + 0.020*"간식"')
(8, '0.024*"경험" + 0.022*"사업" + 0.022*"업무" + 0.021*"기업" + 0.019*"대기업" + 0.016*"연봉" + 0.015*"분야" + 0.015*"기회" + 0.015*"업체" + 0.014*"회사"')
(9, '0.058*"눈치" + 0.038*"연차" + 0.031*"가능" + 0.029*"아침" + 0.026*"칼퇴" + 0.024*"안보" + 0.023*"퇴근" + 0.022*"시간" + 0.015*"부서" + 0.015*"휴가"')
--------------------------------------------------------------------------------------------------------------------------------------
cons topics
(0, '0.032*"업무" + 0.027*"야근" + 0.019*"직원" + 0.016*"회사" + 0.016*"사람" + 0.015*"퇴사" + 0.011*"체계" + 0.009*"시간" + 0.009*"복지" + 0.009*"강요"')
(1, '0.067*"업무" + 0.022*"회사" + 0.022*"직원" + 0.019*"사람" + 0.012*"회의" + 0.012*"반복" + 0.011*"임원" + 0.011*"개발" + 0.009*"단순" + 0.009*"생각"')
(2, '0.040*"영업" + 0.039*"복지" + 0.015*"연봉" + 0.015*"직원" + 0.012*"야근" + 0.011*"사장" + 0.011*"압박" + 0.009*"사람" + 0.008*"실적" + 0.007*"모든"')
(3, '0.071*"회사" + 0.023*"사람" + 0.016*"분위기" + 0.013*"부서" + 0.011*"직원" + 0.010*"월급" + 0.009*"자기" + 0.008*"사원" + 0.008*"발전" + 0.008*"정치"')
(4, '0.071*"연봉" + 0.049*"사람" + 0.022*"직원" + 0.019*"인사" + 0.018*"복지" + 0.018*"회사" + 0.017*"매우" + 0.014*"평가" + 0.011*"승진" + 0.010*"업무"')
(5, '0.088*"야근" + 0.032*"수당" + 0.030*"퇴근" + 0.024*"주말" + 0.021*"눈치" + 0.019*"시간" + 0.019*"출근" + 0.019*"연차" + 0.016*"근무" + 0.015*"연봉"')
(6, '0.023*"연차" + 0.015*"사업" + 0.013*"결재" + 0.013*"휴가" + 0.013*"사용" + 0.013*"지인" + 0.012*"개발" + 0.010*"투자" + 0.009*"회사" + 0.008*"프로세스"')
(7, '0.026*"회사" + 0.018*"사람" + 0.017*"장점" + 0.017*"단점" + 0.015*"기업" + 0.013*"책임" + 0.010*"부서" + 0.010*"생각" + 0.009*"수도" + 0.008*"부재"')
(8, '0.023*"직원" + 0.019*"프로젝트" + 0.019*"사업" + 0.017*"사람" + 0.015*"인력" + 0.015*"업무" + 0.013*"대한" + 0.012*"문제" + 0.012*"개발" + 0.011*"관리자"')
(9, '0.046*"직원" + 0.028*"회사" + 0.022*"경영" + 0.012*"사내" + 0.011*"진의" + 0.011*"대표" + 0.008*"제품" + 0.008*"신경" + 0.007*"생각" + 0.007*"중소기업"')
```

Experiment
Visualization

Topic : 수평적인 분위기





Result

장/단점	토픽	키워드
장점	성장 가능성	경력, 자기계발, 자격증, 취득
	업계 대비 준수한 연봉	월급, 꼬박꼬박, 금융, 수가, 매출
	휴가를 눈치 안 보고 사용 가능	명절, 탄력, 휴무, 휴가
	직원들과 원만한 관계	관계, 나이, 느낌, 열정, 유대, 적임
	복지	복리, 후생, 지하철역, 의사결정, 절약
	수평적인 분위기	분위기, 수평, 자유, 문화, 복지
	식사 및 간식 제공	제공, 점심, 식대, 간식, 음료, 식권
	자율적으로 일할 수 있는 환경	자유, 복장, 명절, 선물, 휴무, 탄력
단점	낮은 연봉	연봉, 급여, 수준, 별로, 비합리적
	인사 고과 평가	영업, 압박, 실적, 성과
	차별	여자, 단점, 급여, 편이, 수준
	사내 정치	비합리적, 신입, 진급, 사람, 직원
	높은 업무 강도	업무, 야근, 퇴사, 강요, 눈치
	좋지 않은 내부 분위기	경영진, 퇴사, 눈치, 정치가
	잦은 야근 및 주말 출근 강요	야근, 주말, 눈치, 특근, 단점
	임원들에 대한 눈치	임원, 대표, 결정, 분위기, 아부, 정치가

Thank you for listening