잡 플래닛 리뷰분석을 통해 지원할 빅데이터 IT회사 선별

시험: 중간고사 보고서
학과: 빅데이터경영통계
학번: 20162533
이름: 장성민

INDEX

1 서론

주제: 잡 플래닛 리뷰분석을 통해 지원할 빅데이터 IT회사 선별

이번 프로젝트는 잡 플래닛에 등록되어 있는 총 3389개의 IT회사 중에서 빅데이터와 관련 있으면서 실제 현장의 업무 분위기 또한 좋은 IT회사를 선별해내는 것을 목표로 한다.

학부생들은 꿈을 향해 열심히 공부하고 있지만 안타깝게도 사회에 내가 지원하면 좋을 회사들은 어떤 회사들이 있는지, 회사들의 실제 현장 업무 분위기는 어떠한지, 어떤 장단점들이 있는지, 어떤 회사가 나에게 성향적으로 맞을지 알고 판단하기가 쉽지 않다. 이런 현실에서 나 또한 다르지 않다.

그래서 이번 프로젝트를 계기로 내가 지원을 목표로 하면 좋을 회사를 텍스트 분석을 통해서 찾아내 보고자 한다. 잡 플래닛의 리뷰를 여러 방면으로 분석하고 분석결과를 이용해 각 분석별로 기업들을 선별해낼 것이다. 그리고 선별된 기업들을 교집합하여 빅데이터와 관련 있으면서 실제 현장의 업무 분위기도 좋은 IT회사들을 최종 선별하도록 한다.

2 데이터 수집

데이터 수집 조건

1. 잡 플래닛에서 기업의 1차 산업군으로 IT/웹/통신을 선택(IT기업)

2 기업내 다양한 부서가 존재하여 같은 회사라도 부서마다 분위기가 다르기 때문에 내가 희망하고 있는 IT부서를 선택

3. 총 3389개의 IT기업 중에서 리뷰수가 적어도 100개 이상인 기업을 선택
   -> 총 297개의 기업 선정

4. 리뷰 데이터 총 12칼럼 스크랩 진행 -> 총 42641개의 데이터 수집
   - 통합 리뷰 정보: 기업명, 기업평균점수
   - 개인 리뷰 정보: 근무부서, 현재근무여부, 근무지역, 리뷰 작성년도, 기업별점점수, 기업에 대한 한줄평, 장점, 단점, 기업에게 바라는 점, 기업 추천여부

5. lmxl, selenium 모두 이용

잡 플래닛 기업 리뷰 예시

3 데이터 전처리

데이터 전처리

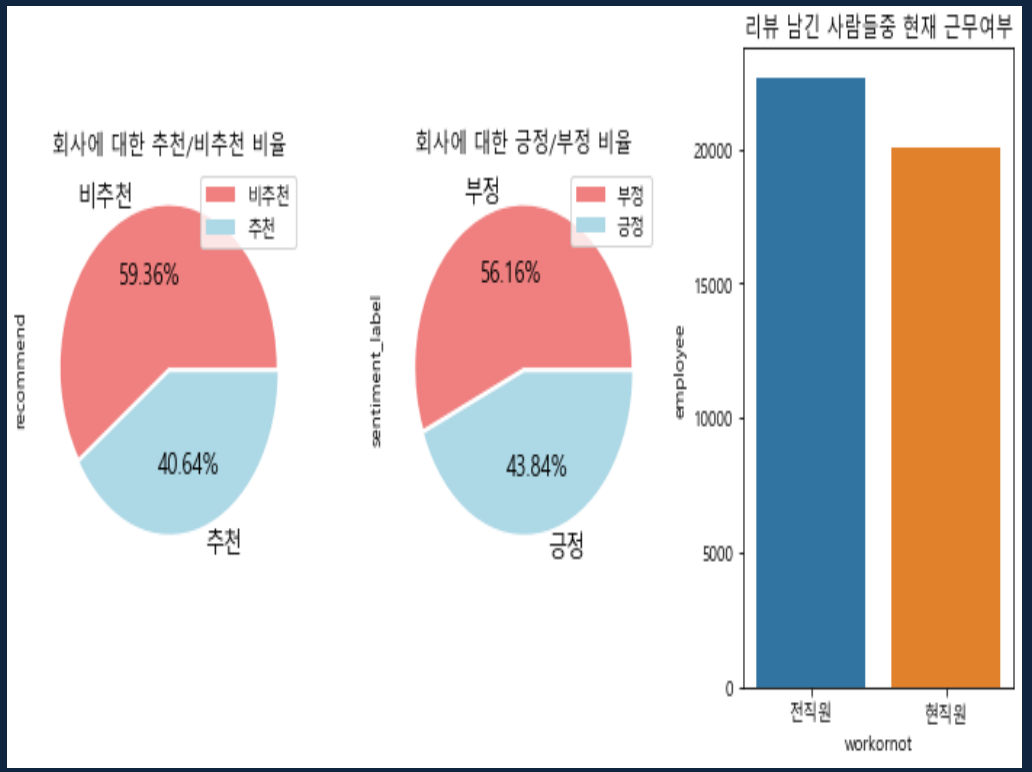
1. 데이터 수집하며 조건문, 예외문, replace, strip, split메소드를 이용하여 데이터 전처리

2. EDA와 감성분석에 활용할 변수를 다음 조건에 맞추어 sentiment_label 생성
   조건1: score가 80이하일 경우 기업에 대한 긍정의 반응으로 1로 설정
   조건2: score가 40이하일 경우 기업에 대한 부정의 반응으로 0으로 설정
   조건3: score가 60일 경우, recommend가 추천일 경우 1,
   비추천일 경우 0으로 설정

3. 불필요한 칼럼 제거, 분석하며 에러나는 NaN값 처리

4 데이터 EDA

IT부서에 대한 만족도 EDA

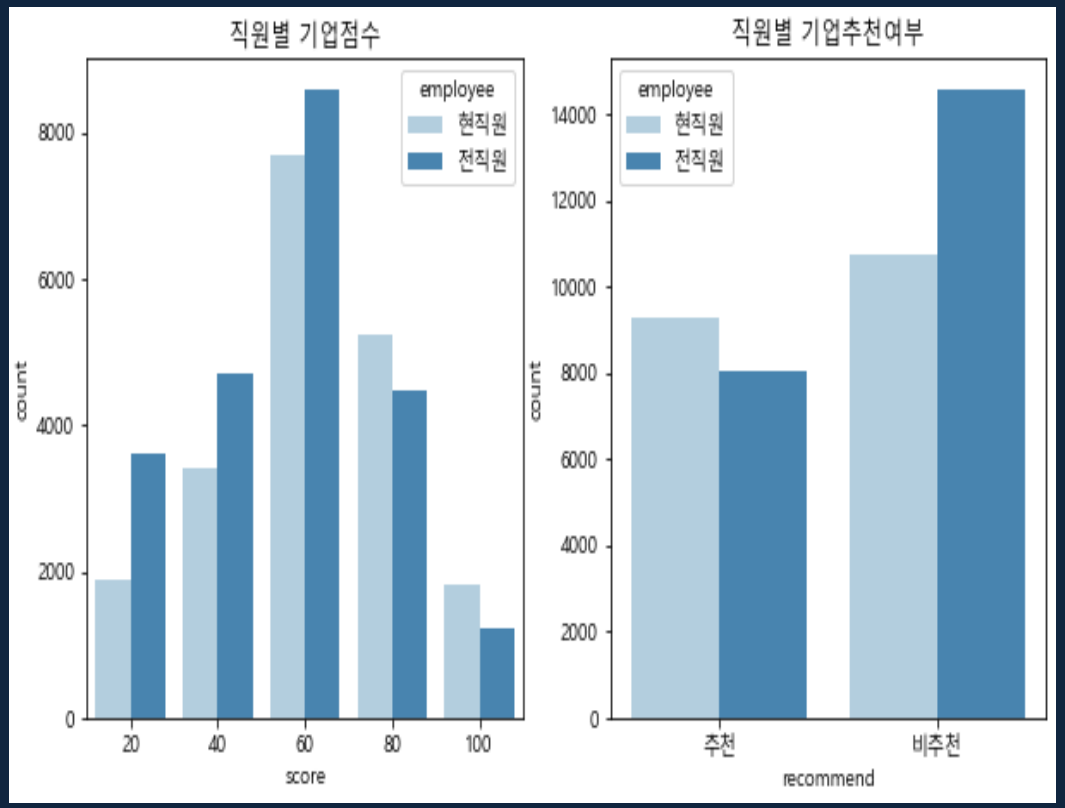


- IT부서에 대한 개인별 만족도를 시각화하면 왼쪽과 같다. 생각보다 IT부서에서 기업들에 대해 비추천, 부정의 비율이 높아서 놀랐다.

- 데이터를 수집해온 구직사이트의 특성상 퇴사 이후 새로운 기업을 탐색하면서 이전에 근무했던 곳에 대한 리뷰를 많이 달아서 그러한 것이 아닐까 추측하였다.

4 데이터 EDA

IT부서에 대한 리뷰점수 EDA



- 기업에 대한 개인별 리뷰 점수를 시각화하면 왼쪽과 같다. 현재 근무를 하고 있는 직원들은 기업 점수를 잘 주었고, 기업에 대해서도 추천을 하고 있다.

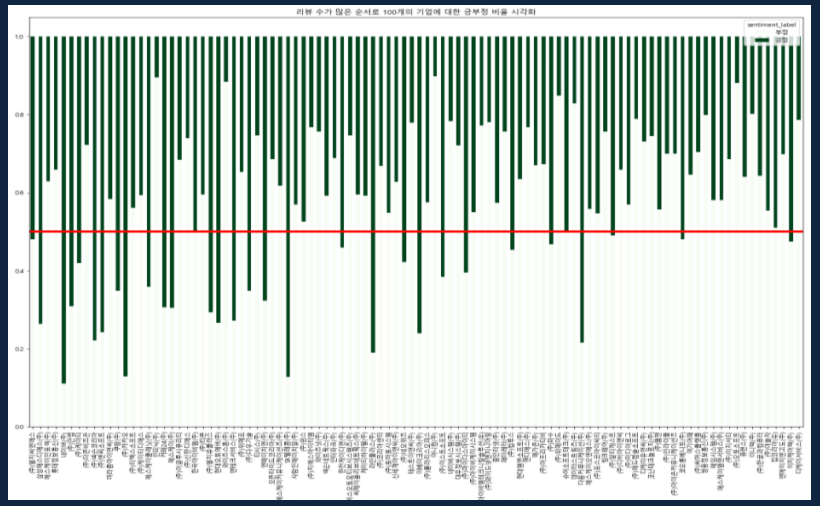
- 추측하였던 대로 퇴사이후 새로운 일자리를 찾는 사람들이 이전에 근무했던 곳에 대해서 낮은 점수를 주고 비추천을 하는 것을 확인할 수 있었다.

- 점수 시각화로 회사에 대한 긍부정에 있어서 전직원들보다 회사에 남아있는 현직원일수록 긍정의 비율이 더 높음을 알 수 있었다.

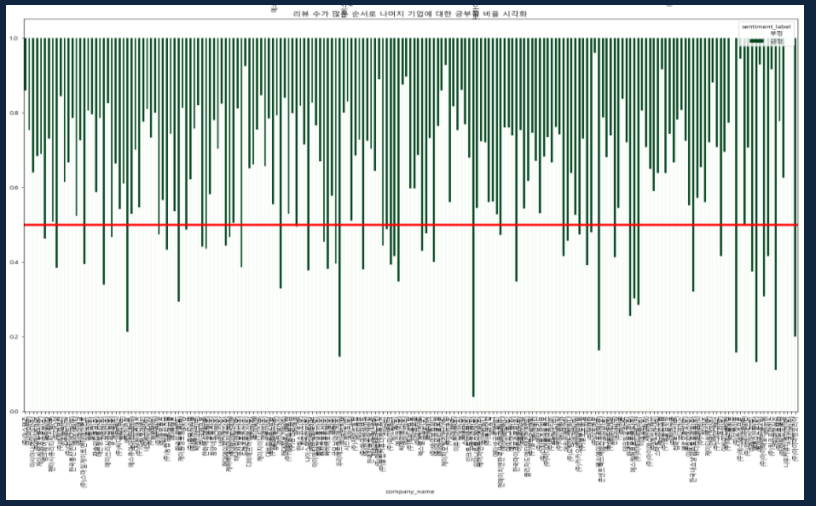
4 데이터 EDA

IT부서에 대한 긍부정 EDA

기업 규모순으로 100개의 기업에 대한 긍부정 비율



기업 규모순으로 100개 이후 나머지 기업에 대한 긍부정 비율



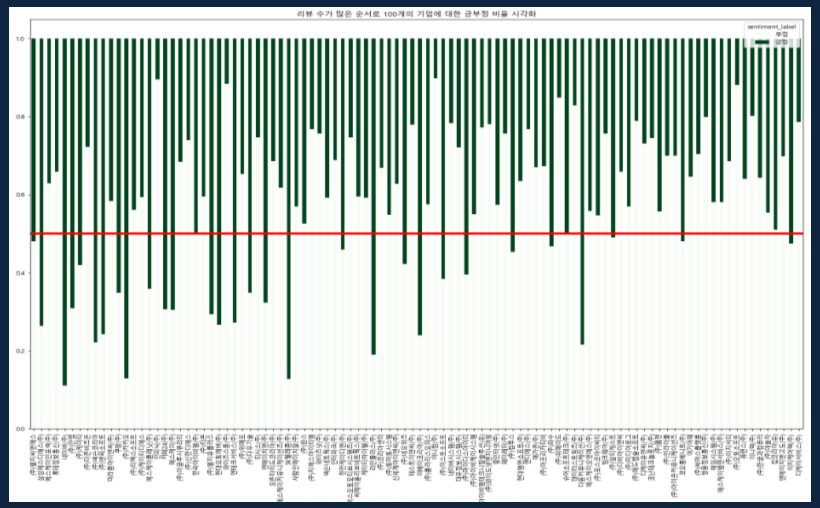
- 리뷰수가 많은 IT 대기업일수록 긍정의 비율이 높지않을까 하여 리뷰 개수를 세서 기업 규모순으로 기업별 긍정의 비율을 시각화하여 보았다.

- 앞 부분 대기업들의 긍정 비율이 높긴 하지만 다른 회사들도 모두 살펴보았을 때 대기업이 아니어도 긍정의 비율이 높은 회사들이 많았다. IT회사 규모가 클수록 IT부서의 긍정의 비율이 높을 가능성이 있지만 회사의 규모가 IT부서에 있어서 지배적인 요인은 아님을 알 수 있었다.

4 데이터 EDA

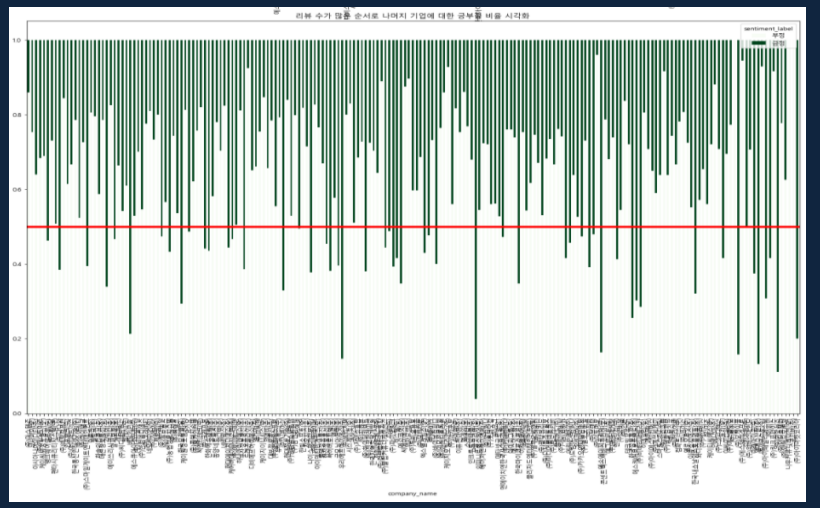
EDA를 통한 기업 선별

기업 규모순으로 100개의 기업에 대한 긍부정 비율



0.5

기업 규모순으로 100개 이후 나머지 기업에 대한 긍부정 비율



- 위의 그래프에서 긍정의 비율이 0.5보다 높은 기업들을 선별하였다.
  선별된 기업들을 변수 part4_visualization_positive_companies에 저장

- EDA 결과 선별된 기업의 수: 83

5 단어 빈도와 단어구름

WordCloud

- IT회사의 `IT부서에 대한 장단점`은 무엇이 있는지 WordCloud로 살펴본다.

1. 장점/단점 TDM만들기
- 비교1: wiki vs stanze
- 비교2: 토큰 추출 함수 ~ 명사만, 동사만, 명사와동사
- 비교3: max_features 100 vs 150 vs 200 vs 250

2. 추가한 것1: drop_one_word 함수
- 의미 없는 한글자 단어들을 제외시키는 함수 구현

3. 추가한 것2: search_review함수
- 잡 플래닛에서는 유료로 WordCloud를 제공중이다.
WordCloud에 나오는 단어를 클릭하면 그와 관련된
리뷰들이 나오도록 해 놓았기에 그러한 기능을 구현하여 보았다.
이 함수로 인해 IT부서의 장단점을 더 잘 파악할 수 있었다.

→

비교결과 사용한
파라미터

wiki,
CountVectozier
extract_noun(명사만),
max_features 150,
두글자 이상 토큰만

5 단어 빈도와 단어구름

WordCloud

장점 단어구름

단점 단어구름

- 장단점별 단어구름을 만든 결과 장점 단점별로 긍정적이거나 부정적인 단어들이 잘 분류되었다. 하지만 그 중에는 '회사', '야근', '분위기' 등과 같이 겹치는 단어들도 있음을 볼 수 있다. 같은 단어를 두고 다르게 말할 수 있음을 확인할 수 있었고 구현한 함수로 검색하여 어떤 점에서 장점이 될 수 있고 단점이 될 수 있는지 확인하였다.

- 장점 단어구름에 나온 단어들을 이용하여 기업별로 장점 칼럼에서 몇 개씩 나왔는지를 세고, 장점단어의 수가 전체 평균보다 높은 기업들 선별. 변수 part5_WordCloud_positive_companies에 저장

- 단어구름 결과 선별된 기업의 수: 141

6 감성분석

감성분석

- 기업에 대한 한줄평을 이용해 문장의 감성(sentiment)을 예측하여 보도록 한다.

1. 한줄평 TDM만들기
 - 비교1: 토크나이져: CountVectoizer, TfidTransformer
 - 비교2: 토큰 추출 함수 ~ 명사만, 동사만, 명사와동사
 - 비교3: max_features 500 vs 1000 vs 1500 vs 2000

2. 로지스틱 회귀분석으로 감성분석
 x = 기업에 대한 한줄평
 y = 전처리를 통해 만든 sentiment_label

3. 모델링 결과
 train – AUC: 0.778, ACC: 0.706
 test – AUC: 0.746, ACC: 0.682

→

비교결과 사용한
파라미터

TfidTransfomer
extract_noun(명사만),
max_features 2000,

6 감성분석

- 기업에 대한 한줄평 감성분석 결과로 나온 단어별 가중치 표

부정단어

	토큰	가중치
1640	최악	-3.561258
875	소모품	-3.168608
1728	탈출	-2.961169
79	겉	-2.628799
1681	침몰	-2.627753
1409	전형	-2.579958
1098	연명	-2.458509
1490	주먹구구	-2.413682
1768	특근	-2.364092
714	부재	-2.344157
1437	정작	-2.334419
1396	전무	-2.303952
1876	하위	-2.215590
1447	제로	-2.203456
252	기계	-2.182372

긍정단어

	토큰	가중치
1417	젊음	1.664913
1480	종합	1.679944
705	부바부	1.684415
1745	토론	1.689897
842	선진	1.726874
1264	이커머스	1.795747
649	밸런스	1.833091
692	복지	1.908733
1520	즐거움	1.939390
909	수평	1.969914
192	국내	2.010613
1339	자율	2.099020
1474	존중	2.398305
1631	최고	2.533056
518	만족	2.780011

- 기업 한줄평에 대한 감성분석을 통해서 나온 토큰들의 데이터에서 긍정에 해당하는 상위 100개의 토큰이 기업별 한줄평에 몇 개가 나왔는지 세서 전체 평균보다 높은 기업들 선별.

  변수 part6_SentimentAnalysis_tokens_companies에 저장

- 감성분석 결과 선별된 기업의 수: 134

7 주제분석

주제분석

- 기업에 대한 한줄평을 이용해 빅데이터란 단어를 목표로 주제분석을 해본다.

1. 주제분석 분석방법 비교

- LSA vs NMF vs LDA
- 위의 세가지 방법을 모두 해보고 결과인 단어목록을 비교한 결과 생각한 바와는 다르게 LSA에서 회전도 하지 않고 병렬분석만 적용한 결과가 제일 나았다. 결과 비교로는 단어목록이 '빅데이터 ' 란 단어와의 연관성, 맥락, 단어의 분위기를 살펴보았다.

→

비교결과 사용한 파라미터

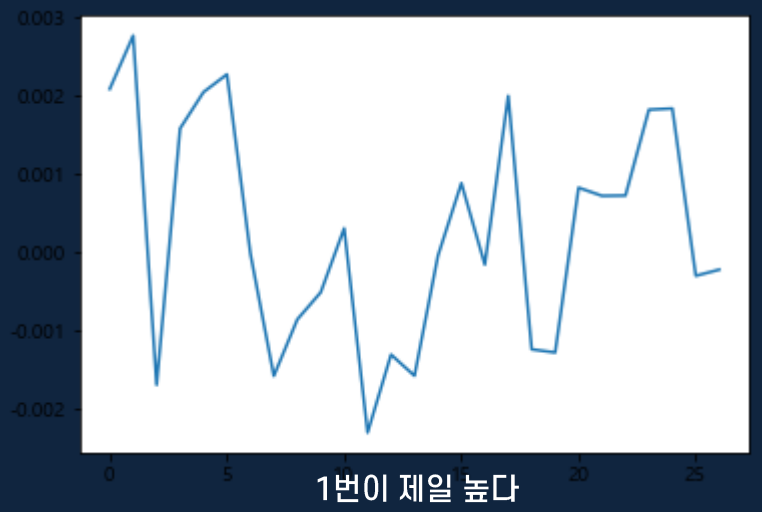
LSA(잠재의미분석)
TfidVectoizer,
max_features 2000,
n_components 27

2. 한줄평 TDM 만들기
- 비교1: 토크나이져: CountVectoizer, TfidVectoizer
- 비교2: max_features 1500 vs 2000 vs 2500 vs 3000 vs 3500 vs 4000
- 비교3: n_components = 27 (병렬분석결과)

7 주제분석

word = '빅데이터'

주어진 단어와 주제별 관련도



1번이 제일 높다

1번 주제와 강한 관계를 갖는 단어들

	word	loading
1473	있는	0.453799
315	기업	0.314584
1640	좋은	0.272571
1503	있음	0.237042
479	다양한	0.231689
152	경험을	0.133921
1456	일할	0.128065
86	개발자가	0.125675
1480	있다	0.103320
638	많이	0.099514
631	많은	0.089811
760	배울	0.081169
1647	좋음	0.078707
1908	하지만	0.073758
1983	회사입니다	0.072847

- 기업 한줄평에 대한 주제분석을 통해서 `빅데이터`와 관련된 주제를 찾고, 해당 주제와 강한 관계를 가진 단어목록을 추출함.
  그리고 `빅데이터`와 단어목록이 기업별 한줄평에 몇개가 나왔는지 세서 전체 평균보다 높은 기업들 선별.

- 변수 part7_topicAnalysis_topic_companies에 저장

- 주제분석 결과 선별된 기업의 수: 81

8 최종 기업선별 및 결론

기업 교집합

- 각 분석을 통해 나온 결과물을 이용해 긍정의 분위기가 반영된 기업들을 각 분석별로 선별하였다.

1. part4_Visualization_positive_companies : 기업별 긍정의 비율이 50%가 넘는 회사들의 집합
2. part5_WordCloud_positive_companies : WordCloud에 나타난 장점단어들의 갯수가 전체 평균보다 높은 회사들의 집합
3. part6_sentimentAnalysis_tokens_companies : 감성분석을 통해 추출한 긍정 토큰의 갯수가 전체 평균보다 높은 회사들의 집합
4. part7_topicAnalysis_topic_companies : 주제분석을 통해 빅데이터와 가장 관련있는 주제를 구하고 그 주제와 강한 관계를 갖는 단어목록의 갯수가 전체 평균보다 높은 회사들의 집합

- part4~part6을 통해서 긍정의 분위기가 반영된 기업들을 선별하였고, part7에서 주제분석을 통해 빅데이터와 관련된 기업들을 선별하였다.

- 이렇게 선별된 기업들을 교집합을 하여 목표인 빅데이터와 관련 있으면서 현장 업무 분위기 또한 좋은 기업들을 구해내는 것이다.

- 교집합한 결과 총 35개의 기업들이 선정되었다.

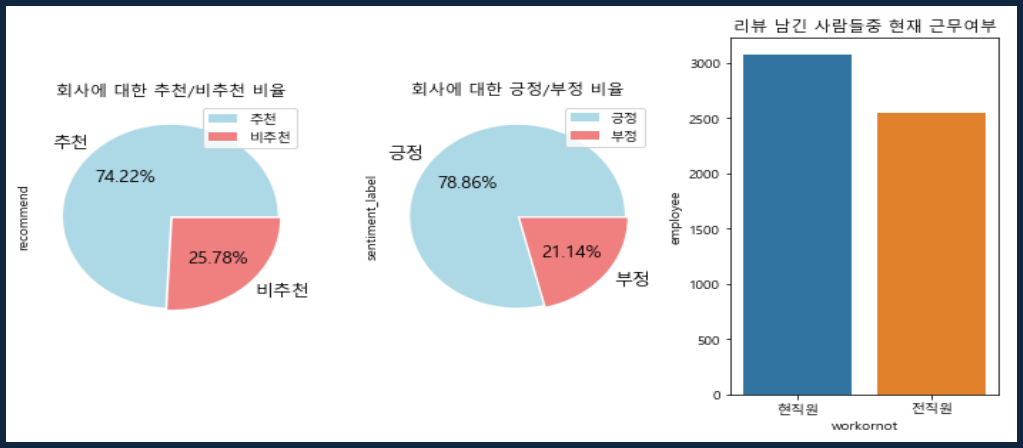
8 최종 기업선별 및 결론

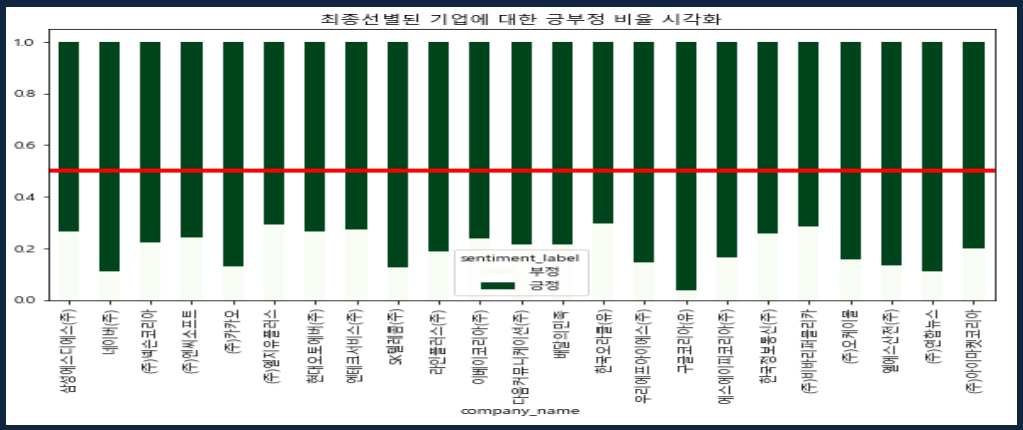
최종선별

- 선별된 35개의 회사들 중에서 긍정비율이 0.7보다 높은 12개의 회사 최종선별, 시각화

- 최종선발된 기업들, 긍정비율순

• 1. 구글코리아(유)
• 2. 네이버(주)
• 3. (주)연합뉴스
• 4. SK텔레콤(주)
• 5. (주)카카오
• 6. 라인플러스(주)
• 7. (주)넥슨코리아
• 8. 이베이코리아(주)
• 9. (주)엔씨소프트
• 10. 삼성에스디에스(주)
• 11. 엔테크서비스(주)
• 12. (주)비바리퍼블리카





8 최종 기업선별 및 결론

결론

- 텍스트 분석 결과 의도하였던 대로 `빅데이터`와 관련 있으면서
  `실제 현장의 업무 분위기 또한 좋은` IT회사들을 선별해낼 수 있었다.

- 시각화, 장단점 단어구름, 감성분석을 하면서 IT부서의 분위기가 어떠한지 어떠한 장점,
  단점들이 있는지 그러한 요소들을 내가 판단하였을 때 아울러 잘 받아들이고 일할 수
  있을지 생각하여 볼 수 있었다. 또 주제분석을 하여 빅데이터와 관련된 주제, 단어목록을
  추출하면서 단어목록에 나온 단어들이 전반적으로 장점 구름단어에 있는 단어들과
  유사하여 다행이었다.

- 선별된 35개의 기업들 중에서 긍정비율이 0.7보다 높은 회사들을 최종 선별하여
  총 12개의 기업들이 나왔다. 해당 기업들은 모두 한번씩 들어본 규모 있는 회사들이었다.
  빅데이터를 다루고 부서의 환경 및 분위기가 좋은 곳은 역시 규모가 있는 회사들이니
  선별된 결과가 납득이 갔다.

- 취업할 때 이번 프로젝트로 도출된 회사들을 살펴보고
  나에게 맞을 것 같은 회사들에 지원하여 보도록 해야 겠다.

9 자기평가

항목	점수	평가근거
서론	2/2	서론을 통해 다루고자 하는 주제, 현황, 문제점 등을 파악할 수 있다.
데이터 수집	2/3	lxml과 selenium 모두 활용. 필요에 의해 자동로그인 과정도 넣음. 동일한 태그명에 대해서 nth-of-type()을 이용. 수집과 전처리 동시에 진행함.
전처리	2/3	수집하며 전처리 거침. 수집후 EDA와 감성분석에 활용할 변수 생성. 불필요한 칼럼을 제거하고 NaN값을 처리함.
단어 빈도	3/3	파라미터별 비교분석 과정을 거치고 한 글자 단어를 제거하는 함수도 추가하여 목적에 잘 부합하도록 단어 빈도를 제시함. 단어구름을 주제에 맞게 사람 이미지를 사용함. 잡 플래닛에서 제공하는 기능을 함수로 추가 구현함. 단어구름을 만들고 단어목록을 이용해 주제와 부합하는 데이터를 추출하는 과정을 거침
감성 분석	3/3	파라미터별 비교분석 과정을 거치고 주제와 잘 부합하는 감성 분석을 하였음. 더 나아가 감성분석 결과 나온 토큰들을 활용하여 주제와 부합하는 데이터를 추출하는 과정을 거침
주제 분석	3/3	LSA, NMF, LDA를 모두 적용하여 결과를 비교하고 결과가 가장 주제와 부합한 LSA를 선택하여 진행. 파라미터별 비교분석 과정을 거치고 목적에 맞도록 빅데이터와 관련된 주제 분석을 진행. 더 나아가 주제 분석 결과 나온 빅데이터와 연관 있는 단어목록을 이용하여 주제와 부합하는 데이터를 추출하는 과정을 거침
결론	2/2	위 과정에서 추출한 데이터를 교집합하고 시각화하는 과정을 거침. 서론에서 의도한 바와 같이 주제와 목적에 부합하는 결론을 잘 제시하였음.
합계	17/19	

THANK YOU