채용공고 분석 및 테크트리 추천

201704227
이상수

목차

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1 주제 소개

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 1

주제 소개

주제 선택 배경 : 4학년 졸업을 앞두고 마주한 취업을 준비하기 위해 채용공고를 무작정 찾아보던 때가 있었다. 하지만 채용공고의 수는 많았고 과연 나는 이렇게 많은 기업 중 어느 기업이든 원하는 구직자의 모습을 갖추고 있는지 의문이 들었다. 그러자 채용공고를 하나하나 찾아보는 시간마저 아깝게 느껴졌고 모든 채용공고를 취합하여 기업이 원하는 인재상의 모습을 요약해주는 것이 있으면 좋겠다고 생각하였고 이 주제로 졸업 작품을 시작하기로 했다.

작품 시사점 : 수많은 구직자들이 자신의 분야를 살리기도 하고, 새로운 분야를 시도하기도 한다. 그때마다 새로운 채용공고를 모두 볼 수 없기에 이번 분석은 구직활동의 첫 시작의 길을 보여주는 역할을 할 것이다. 구직자에게 시간은 금이기 때문에 시간을 효율적으로 활용할 수 있도록 이 작품을 발전시키면 많은 사람에게 도움이 될 것이다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 1

프로젝트 진행 과정

9월 1주
주제 선정

9월 3주
데이터 수집

'원티드'사이트에서 HTML문서로 기록된 채용공고 데이터를 selenium을 활용하여 자동 수집

10월 1주
데이터 전처리

수집된 데이터를 데이터프레임으로 정리 및 정제

10월 3주
수집 & 전처리 보강

전처리된 데이터 세트에서 좀 더 많은 정보를 수집하도록 코드 확장

10월 5주
중간 보고서 제출

11월 1주
토큰화 및 불용어 제거

한국어 데이터 처리인 konlpy 패키지를 사용하여 토큰화 작업과 불용어를 제거

11월 2주
토픽모델링

토sklearn의 토픽모델링(LDA) 패키지를 불러와 모델링 작업

11월 3주
모델 혼잡도 및 시각화

토픽의 개수를 정하고 분류된 모델의 정보를 시각화

11월 4주
최종 정리

신입과 경력의 데이터를 비교하여 작업된 코드를 정리

ⓒSaebyeol Yu. Saebyeol's PowerPoint

2 데이터 수집 & 전처리

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 수집

패키지 설치 및 세팅

In [1]:
```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import Select
import time
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import requests
from konlpy.tag import Okt
from konlpy.tag import Kkma
import re
```

In [2]:
```
# 화면 확장

from IPython.core.display import display, HTML
display(HTML("<style>.container { width:90% !important; }</style>"))
```

패키지 설치

웹 크롤링 :

Webdriver – 파이썬으로 자동화를 위한 chrome

Selenium – 사이트의 HTML정보를 불러오는 패키지

Time – 대기 시간 조작

BeautifulSoup – url사이트 정보 수집

Konlpy – 한국어 데이터 전처리를 위한 패키지

Re – 불용어 제거를 위한 패키지

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 수집

원티드 웹스크래핑

로그인

In [4]:
```
driver.get('https://www.wanted.co.kr/') # 원티드 사이트 열기
time.sleep(1) # 로딩시간 1초

# 웹 창의 크기가 전체화면이어야 가능
login='//*[@id="__next"]/div[1]/div/nav/aside/ul/li[2]/button'
driver.find_element_by_xpath(login).click()
time.sleep(1)


login_kakao='//*[@id="__next"]/div/div/div/div[2]/form/div[2]/button[1]'
driver.find_element_by_xpath(login_kakao).click()
time.sleep(1)

tag_id = driver.find_element_by_xpath('//*[@id="input-loginKey"]')
tag_pw = driver.find_element_by_xpath('//*[@id="input-password"]')
# tag_id.clear()
# tag_pw.clear()
# time.sleep(1)
```

In [5]:
```
tag_id.send_keys('        ID        ')
time.sleep(0.5)

tag_pw.send_keys('  PW  ')
time.sleep(0.5)

login1='//*[@id="mainContent"]/div/div/form/div[4]/button[1]'
driver.find_element_by_xpath(login1).click()
time.sleep(2)
```

자동 로그인

Chrome webdriver을 사용하여 자동화 시작. 원티드 초기 화면 사이트를 들어가서 Xpath를 사용하여 클릭을 조정하여 아이디와 비밀번호를 자동입력 및 로그인.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 수집

채용사이트 들어가기

In [6]:
```
driver.get('https://www.wanted.co.kr/wdlist/518/655?country=kr&job_sort=company.response_rate_order&years=-1&locations=all')
```

서울지역 선택

In [7]:
```
driver.find_element_by_xpath('//*[@id="__next"]/div[3]/div/div/div[1]/div[1]/div/div[1]/button').click()
driver.find_element_by_xpath('//*[@id="MODAL_BODY"]/div[2]/div[1]/ul/li[2]/button').click()
driver.find_element_by_xpath('//*[@id="__next"]/div[3]/div/div/div[1]/div[1]/div[2]/div[1]/div[3]/button').click()
```

스크롤 다운

In [8]:
```
from selenium.webdriver.common.keys import Keys
import random

for c in range(0,30):
    driver.find_element_by_tag_name('body').send_keys(Keys.PAGE_DOWN)
    time.sleep(random.uniform(0.5,1))
```

수집 전 세팅

'원티드'사이트는 스크롤을 내리는 만큼 채용공고가 늘어나기 때문에 가장 밑까지 스크롤을 내려야 한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 2

데이터 수집

데이터세트 만들기

In [17]:
```
JOB = pd.DataFrame(columns = ['이름','회사정보','주요업무', '자격요건', '우대사항','혜택 및 복지','URL', '태그정보', '스택'])
```

In [18]:
```
# 첫 칸에 빈 칸이 들어와서 2번째 리스트부터 수집 but 누락되는 것이 있음(리스트가 1개인 경우)
# REAL

for i in range(0,len( url_list)):
    driver.get(url_list[i])
    response = requests.get(url_list[i])
    html = response.text
    bs = BeautifulSoup(html, 'html.parser')

    # 주요업무 스크레핑
    a = str(bs)[str(bs).find('₩n₩n주요업무') + 8 : str(bs).find('₩n₩n자격요건') -2].replace('₩n','')
    if len(a) < 2000:
        JOB.loc[i, '주요업무'] = a
    else:
        JOB.loc[i, '주요업무'] = ''

    # 자격요건
    b = str(bs)[str(bs).find('₩n₩n자격요건') + 8 : str(bs).find('₩n₩n우대사항') -2].replace('₩n','')
    if len(b) < 2000:
        JOB.loc[i, '자격요건'] = ''.join(b)

    # 우대조건 스크레핑
    c = str(bs)[str(bs).find('₩n₩n우대사항') + 8 : str(bs).find('₩n₩n혜택') -2].replace('₩n','')
    if len(c) < 2000:
        JOB.loc[i, '우대사항'] = ''.join(c)

    # 혜택 및 복지
    d = str(bs)[str(bs).find('₩n₩n혜택 및 복지') + 11 : str(bs).find('company_name') -3].replace('₩n','')
    if len(d) < 2000:
        JOB.loc[i, '혜택 및 복지'] = ''.join(d)

    # URL
    JOB.loc[i, 'URL'] = url_list[i]

    # 태그정보
    tag = ''
```

데이터 수집

JOB이라는 데이터 프레임을 만들고 수집할 컬럼명을 만든다. 가끔 HTML구조가 달라서 수집에 오류가 생기기 때문에 수집된 데이터의 길이가 2000자 이상이면 오류라고 판단하여 수집을 하지 않는다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 2

데이터 수집

In [42]:
```
JOB
```

Out [42]:

	이름	회사정보	주요업무	자격요건	우대사항	혜택 및 복지	URL	태그정보	스택
0	[엔터플] DBA 채용	엔터플은 임직원의 Work \u0026 Life balance를 소중하게 생각합니다...	• DBMS(Mysql/MariaDB) 운영 및 관리• DB 모니터링, 장애 대응•...	• 실무 3년차 이상 또는 그에 준하는 실력을 갖추신 분• 클라우드 DBMS(AWS...	• 대규모 Mysql의 안정적인 운영 경험이 있으신 분• Mysql 외 RDBMS(...	엔터플은 임직원의 Work \u0026 Life balance를 소중하게 생각합니다...	http://www.wanted.co.kr/wd/53460	연봉업계평균이상 인원급성장 50명이하 설립4~9년 육아휴직 수평적조직 스타트업 식...	Git AngularJS Azure BackboneJS GraphQL iOS Lin...
1	[디플래닉스 (Dplanex)] Data Engineer 채용	▷ 혜택 및 복지• 최신 사양 기기 지원 : 원하시는 컴퓨터 사양에 맞추어 준비해 ...	• 교보 그룹의 데이터와 외부 데이터를 활용하여 다양한 관점에서 분석이 가능하도록...	• 3년 이상의 데이터 엔지니어링 실무 경력이 있으신 분• 클라우드 서비스(AWS,...	• 클라우드 플랫폼을 활용해 데이터 설계/구축/운영 업무를 수행하신 분• 대용량 데...	▷ 혜택 및 복지• 최신 사양 기기 지원 : 원하시는 컴퓨터 사양에 맞추어 준비해 ...	http://www.wanted.co.kr/wd/119829	출산휴가 스타트업 IT, 컨텐츠	Java Python Scala SQL AWS GCP
2	[배럴아이] FPGA Engineer 채용	• 자율출근제 시행(08:00~10:30 출근)•건강검진 지원•매월 마지막 주 금요...	• 서초음파 영상시스템 신호처리, FPGA 설계• ADC, Serdes, LVDS,...	• 학력 : 대졸 이상• 경력 : 책임 급• 성별 : 무관• 디지털 신호처리	• 초음파영상 시스템 유경험자 우대• 의료기기 개발 경험자 우대\t• Matla...	• 자율출근제 시행(08:00~10:30 출근)•건강검진 지원•매월 마지막 주 금요...	http://www.wanted.co.kr/wd/86872	식비 건강검진 전문, 과학기술	
3	[아도바] Data engineer 채용	무료 종합건강검진, 교육/도서비 지원, 경조금/경조휴가,점심식대(야근시, 저녁식대...	• 데이터 마트 개발 및 운영• 분석 모형 서빙을 위한 파이프라인 개발 및 운영...	• 경력 1년 이상 • DBMS 운영 관리 경험• ETL 설계 및 구축 경험•...	• 단순 반복 작업을 자동화하여 효율적으로 운영한 경험을 가지고 계신 분• In...	무료 종합건강검진, 교육/도서비 지원, 경조금/경조휴가,점심식대(야근시, 저녁식대...	http://www.wanted.co.kr/wd/126565	퇴사율5%이하 50명이하 설립4~9년 간식 수면실 휴게실 생일선물 건강검진 IT,...	Azure Python R SQL Data Analysis Tableau AWS S...
4	[스콥정보통신] AI(인공지능), 빅데이터 개발자 채용	[Culture] • 워라벨을 중요시 합니다. 야근 문화 없음 (9to6 지향) ...	• 빅데이터, AI, 데이터 엔지니어링 분석 및 처리• 데이터 파이프라인 개발과 데...	• Tensorflow, PyTorch, Keras 등 머신러닝/딥러닝 프레임워크 ...	• 로그시스템 및 데이터마트 Structure 설계 경험 • Hadoop MR, H...	[Culture] • 워라벨을 중요시 합니다. 야근 문화 없음 (9to6 지향) ...	http://www.wanted.co.kr/wd/129440	인원급성장 퇴사율 6~10% 51~300명 설립10년이상 성과급 커피 사내카페 안...	Git Github Android iOS Linux MySQL Oracle Pyto...
...	...	...	...	...	...	...	...	...	...
368	[하렉스인포텍] 머신러닝 엔지니어 (추천엔진 개발)	• 하렉스인포텍의 연봉제도는 기존의 연공서열 중심의 승진 및 호봉제	• 자연어처리 딥러닝 기반 추천엔진 솔루션 개발• 다국어 구매 기록을 활용한 추개인	• Python 프로그래밍 실무 경력 (1년 이상)• 딥러닝 지식 및 개발 경험•	• 추천 엔진 개발 유경험자• Transformer 기반 딥러닝 개발 유경험	• 하렉스인포텍의 연봉제도는 기존의 연공서열 중심의 승진 및 호봉제	http://www.wanted.co.kr/wd/111502	연봉상위1% 퇴사율5%이하 50명이하 설립10년이상 자기계발 인공지능	Git Linux MySQL Python SQL AWS Docker

수집된 데이터(경력직)를 보면 특수기호가 많은 것을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 2

데이터 전처리

데이터 토큰화

데이터 전처리

In [45]:
```
new_necessary = []
for i in necessary:
    if i == '':
        continue
    new_necessary.append(re.sub(r'[^A-Za-z0-9가-힣ㅣs+]', '', i))
```

In [46]:
```
new_necessary
```

Out[46]:
```
[' 실무 3년차 이상 또는 그에 준하는 실력을 갖추신 분 클라우드 DBMSAWSAzureNaver Cloud등 운영 경험이 있으신 분 MySQLMariaDB 설치 백업복구 HA DR 업무 경험이 있으신 분 사
과 관심이 있는',
 ' 3년 이상의 데이터 엔지니어링 실무 경력이 있으신 분 클라우드 서비스AWS GCP Azure 등 활용 경험이 있는 분    AWS 의 경우 Glue EC2 S3 Athena Lambda Redshift EMR 등 Pyth
한 가지 이상의 언어로 개발이 가능한 분 데이터 분석가와 원활한 소통을 위한 쿼리 작성 능력을 갖춘 분 새로운 기술에 대한 관심과 호기심이 많으신',
 ' 학력  대졸 이상 경력  책임 급 성별  무관 디지털 신호처리 ',
 '  경력 1년 이상   DBMS 운영 관리 경험  ETL 설계 및 구축 경험  데이터 분석 프로젝트를 수행하면서 성과를 만들어보신 분  클라우드AWSMS Azure 등 환경에서 데이터 분석 모
가능자  데이터 분석 및 처리 여러 데이터베이스 사용 경험 및 운영 경험이 있으신',
 ' Tensorflow PyTorch Keras 등 머신러닝딥러닝 프레임워크 사용 경험자  Python  Java  Scala  Kotlin  Go 등 최소 하나 이상의 개발 언어에 능숙한 분 클라우드 환경에서의 개
그에 준하는 지식 보유 협업 대상과의 원활한 커뮤니케이션 ',
 ' Computer Science 수학 통계 혹은 관련 분야의 학위 혹은 동등한 조건을 갖추신 분 SQL 문법의 이해 및 사용 경험이 있으신',
 ' 실시간 및 배치성 대용량 데이터테라단위 처리 개발 능력 대용량 데이터 아키텍쳐 설계 및 운영 능력 분산 시스템 설계 및 운영 능력 데이터 분석 및 모니터링을 포함한 ETL 파
경험 관련 경력 8년 수준최소 3년 이',
 ' 학력  전문대졸 이상 경력  관련 분야 2년  7년 혹은 이에 상응하는 실력을 갖추신 분 PythonR 등 한개 이상의 컴퓨터언어에 익숙한 분 머신러닝 프레임워크TensorflowPytorch
은 분 머신러닝 Pipeline 설계 및 구축 경험이 있는',
 ' 다양한 가설을 쉽고 빠르게 검증하실 수 있는 분을 찾고 있어요 복잡한 분석 결과도 쉽게 전달해주실 수 있는 분을 찾아요 아무리 긴 논리적인 대화도 어려움 없이 나누실 수 있
핵심 지표를 설계하고 모니터링을 위한 대시보드를 구축해보신 분을 찾아요 SQL 작성 능력이 빠르고 정확하신 분을 찾아요 Python R 등 코드 레벨에서 제품 분석이 편하신 분을 찾
 ' 데이터 엔지니어 관련 업무 경력이 5년 이상이거나 그에 준하는 역량을 갖고 계신 분을 찾고 있어요 Kafka SparkHadoop 등 분산 처리 프레임워크 활용에 능숙한 분을 찾고 있어
부터 운영까지 A to Z를 주도적으로 개발해본 경험이 있는 분을 찾고 있어요 대용량 분산 처리를 경험해 보신 분을 찾고 있어',
```

데이터 전처리

데이터 중 a~Z, 숫자 0~9, 한글 가~힣, + 까지 데이터는 살리되 특수기호([, (, * 등)는 제거한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 2

데이터 전처리

```
In [48]: okt = Okt()
         nn = []
         for i in new_necessary:
             a = okt.morphs(i)
             nn.append(listToString(a))
```

```
In [49]: nn
```

```
Out[49]: ['실무 3년 차 이상 또는 그 에 준 하는 실력 을 갖추신 분 클라우드 DBMSAWSAzureNaver Cloud 등 운영 경험 이 있으신 분 MySQLMariaDB 설치 백업 복구 HA DR 업무 경험 이 있으
         대한 열정 과 관심 이 있는',
          '3년 이상 의 데이터 엔지니어링 실무 경력 이 있으신 분 클라우드 서비스 AWS GCP Azure 등 활용 경험 이 있는 분 AWS 의 경우 Glue EC 2 S 3 Athena Lambda Redshift EMR 등 Py
         중 한 가지 이상 의 언어 로 개발 이 가능한 분 데이터 분석 가 와 원활한 소통 을 위 한 쿼리 작성 능력 을 갖춘 분 새로운 기술 에 대한 관심 과 호기심 이 많으신',
          '학력 대졸 이상 경력 책임 급 성별 무관 디지털 신호처리',
          '경력 1년 이상 DBMS 운영 관리 경험 ETL 설계 및 구축 경험 데이터 분석 프로젝트 를 수행 하면서 성과 를 만들어 보신 분 클라우드 AWSMS Azure 등 환경 에서 데이터 분석 모델
         가능 자 데이터 분석 및 처리 여러 데이터베이스 사용 경험 및 운영 경험 이 있으신',
          'Tensorflow PyTorch Keras 등 머신 러닝 딥 러닝 프레임워크 사용 경험 자 Python Java Scala Kotlin Go 등 최소 하나 이상 의 개발 언어 에 능숙한 분 클라우드 환경 에서의 가
         나 그 에 준 하는 지식 보유 협업 대상 과의 원활한 커뮤니케이션',
          'Computer Science 수학 통계 혹은 관련 분야 의 학위 혹은 동등 한 조건 을 갖추신 분 SQL 문법 의 이해 및 사용 경험 이 있으신',
          '실시간 및 배치 성 대 용량 데이터 테라 단위 처리 개발 능력 대 용량 데이터 아키텍쳐 설계 및 운영 능력 분산 시스템 설계 및 운영 능력 데이터 분석 및 모니터링 을 포함 한
         및 운영 경험 관련 경력 9년 수준 최소 3년 이',
          '학력 전문 대졸 이상 경력 관련 분야 2년 7년 혹은 이 에 상응 하는 실력 을 갖추신 분 PythonR 등 한개 이상 의 컴퓨터언어 에 익숙한 분 머신 러닝 프레임워크 TensorflowPyto
         가 높은 분 머신 러닝 Pipeline 설계 및 구축 경험 이 있는',
          '다양한 가설 을 쉽고 빠르게 검증 하실 수 있는 분 을 찾고 있어요 복잡한 분석 결과 도 쉽게 전달 해주실 수 있는 분 을 찾아요 아무리 긴 논리 적 인 대화 도 어려움 없이 나
         아요 직접 핵심 지표 를 설계 하고 모니터링 을 위 한 대시보드 를 구축 해보신 분 을 찾아요 SQL 작성 능력 이 빠르고 정확하신 분 을 찾아요 Python R 등 코드 레벨 에서 제품
         찾고 있어',
          '데이터 엔지니어 관련 업무 경력 이 5년 이상 이 거나 그 에 준 하는 역량 을 갖고 계신 분 을 찾고 있어요 Kafka SparkHadoop 등 분산 처리 프레임워크 활용 에 능숙한 분 을
```

```
In [64]: stop_words = 'and in or 있도록 하신 다루고 하는데 있어서 필요한 없으신 분이면 좋겠어요 졸업 졸업예정자 졸업예정 아 휴 아이구 아이쿠 아이고 어 나 우리 저희 따라 의해 을

         stop_words = set(stop_words.split(' '))
         nnn = []
         for i in nn:
             word_tokens = okt.morphs(i)
             result = [word for word in word_tokens if not word in stop_words]
             nnn.append(listToString(result))
```

데이터 토큰화

Konlpy 패키지 중 Okt를 사용하여 형태소를 남겼다. 또한 불용어를 직접 입력하여 자주 등장하지만 의미없는 단어들을 제거한다.

Nouns가 아닌 morph를 쓴 이유는 명사가 아니어도 중요한 의미가 있을 수 있기 때문이다.(ex 능숙한)

ⓒSaebyeol Yu. Saebyeol's PowerPoint

3

데이터 분석

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 분석

토픽모델링

```
In [66]: from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
         from sklearn.decomposition import LatentDirichletAllocation

In [67]: count_vect = CountVectorizer()

In [68]: fect_vect = count_vect.fit_transform(nnn)
```

분석기법

Sklearn의 LDA를 사용하는데 벡터 기반의 분석기법이라 sklearn의 CountVectorizer 함수를 사용하여 자주 데이터의 벡터를 계산한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 분석

```
In [70]: count_vect.vocabulary_

Out[70]: {'실무': 1389,
          '3년': 16,
          '또는': 1061,
          '실력': 1388,
          '갖추신': 799,
          '클라우드': 1886,
          'dbmsawsazurenaver': 163,
          'cloud': 112,
          '운영': 1502,
          '경험': 830,
          '있으신': 1628,
          'mysqlmariadb': 442,
          '설치': 1303,
          '백업': 1183,
          '복구': 1208,
          'ha': 291,
          'dr': 196,
          '업무': 1439,
```

분석기법

Sklearn의 LDA를 사용하는데 벡터 기반의 분석기법이라 sklearn의 CountVectorizer 함수를 사용하여 자주 데이터의 벡터를 계산한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 분석

```
In [73]: for i in range(1, 7):
             lda = LatentDirichletAllocation(n_components=i, random_state=0)
             lda.fit(fect_vect)
             print(i, lda.perplexity(fect_vect))
             #CountVectorizer객체 내의 전체 word의 명칭을 get_feature_names()을 통해 추출
             feature_names = count_vect.get_feature_names()

             #토픽별 가장 연관도가 높은 word를 15개만 추출
             display_topics(lda, feature_names, 15)
```

분석기법

LDA기법을 사용하되 몇 개의 토픽을 정하면 좋을지 모르기 때문에 1~6개의 토픽을 순차적으로 적용 및 출력

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 분석

```
1 598.8261750521872
Topic # 0
경험 데이터 개발 있으신 경력 운영 대한 있는 python 분석 보유 능력 구축 관련 이해
2 597.4303651007262
Topic # 0
경험 데이터 개발 있으신 운영 경력 대한 보유 있는 python 구축 분석 관련 이해 처리
Topic # 1
능력 분석 있는 데이터 data 업무 python 경력 커뮤니케이션 해결 sql 활용 논리 가능한 관련
3 607.329783484323
Topic # 0
경험 데이터 개발 있으신 운영 경력 대한 보유 python 구축 있는 분석 관련 이해 처리
Topic # 1
능력 있는 데이터 분석 러닝 업무 경력 해결 경험 python 관련 이해 활용 커뮤니케이션 sql
Topic # 2
data of with to experience python skills years strong the work spark etl 대졸 하시는
4 634.2948186989411
Topic # 0
경험 데이터 개발 있으신 운영 대한 경력 보유 있는 python 분석 구축 처리 이해 관련
Topic # 1
능력 데이터 있는 분석 업무 경험 관련 경력 해결 이해 커뮤니케이션 활용 러닝 sql python
Topic # 2
data of with to experience years skills strong the 경력 python spark work 연구개발 이해도
Topic # 3
경험 개발 있으신 운영 경력 python 보유 구축 또는 java 2년 대한 언어 시스템 사용
5 651.8330752994709
Topic # 0
경험 데이터 개발 있으신 운영 대한 있는 경력 보유 구축 이해 python 처리 sql 서비스
Topic # 1
능력 데이터 있는 분석 경험 활용 있으신 업무 sql 해결 경력 python 커뮤니케이션 서비스 논리
Topic # 2
data of with to experience years skills strong the work spark python 코드 etl 가능한
Topic # 3
경험 개발 있으신 python 경력 운영 또는 구축 보유 2년 대한 java 업무 언어 기본
Topic # 4
경력 관련 데이터 분석 개발 경험 학력 대한 보유 3년 전공 러닝 공학 사용 역량
6 662.9228540311725
Topic # 0
경험 데이터 개발 대한 경력 보유 이해 운영 있으신 있는 python 사용 구축 sql 3년
Topic # 1
능력 데이터 분석 있는 경험 있으신 python sql 커뮤니케이션 업무 활용 해결 서비스 소통 다양한
Topic # 2
data of with to experience years skills strong the work spark python etl knowledge 없는
Topic # 3
경험 개발 있으신 운영 보유 python 경력 구축 업무 또는 java 능력 대한 지원 2년
Topic # 4
경력 관련 경험 분석 학력 개발 대한 보유 전공 운영 3년 공학 학사 데이터 또는
Topic # 5
경험 데이터 있으신 개발 운영 처리 활용 있는 용량 필요해요 구축 서비스 역량 python 환경
```

분석기법

혼잡도는 토픽의 수가 가장 적은 1개가 가장 낮지만 혼잡도의 수치가 절대적인 판단의 기준은 아님.

개인적으로 가장 적당하게 배치된 토픽의 수는 2개로 토픽모델링을 했을 때 가장 잘 나뉘어진 것 같다고 생각한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

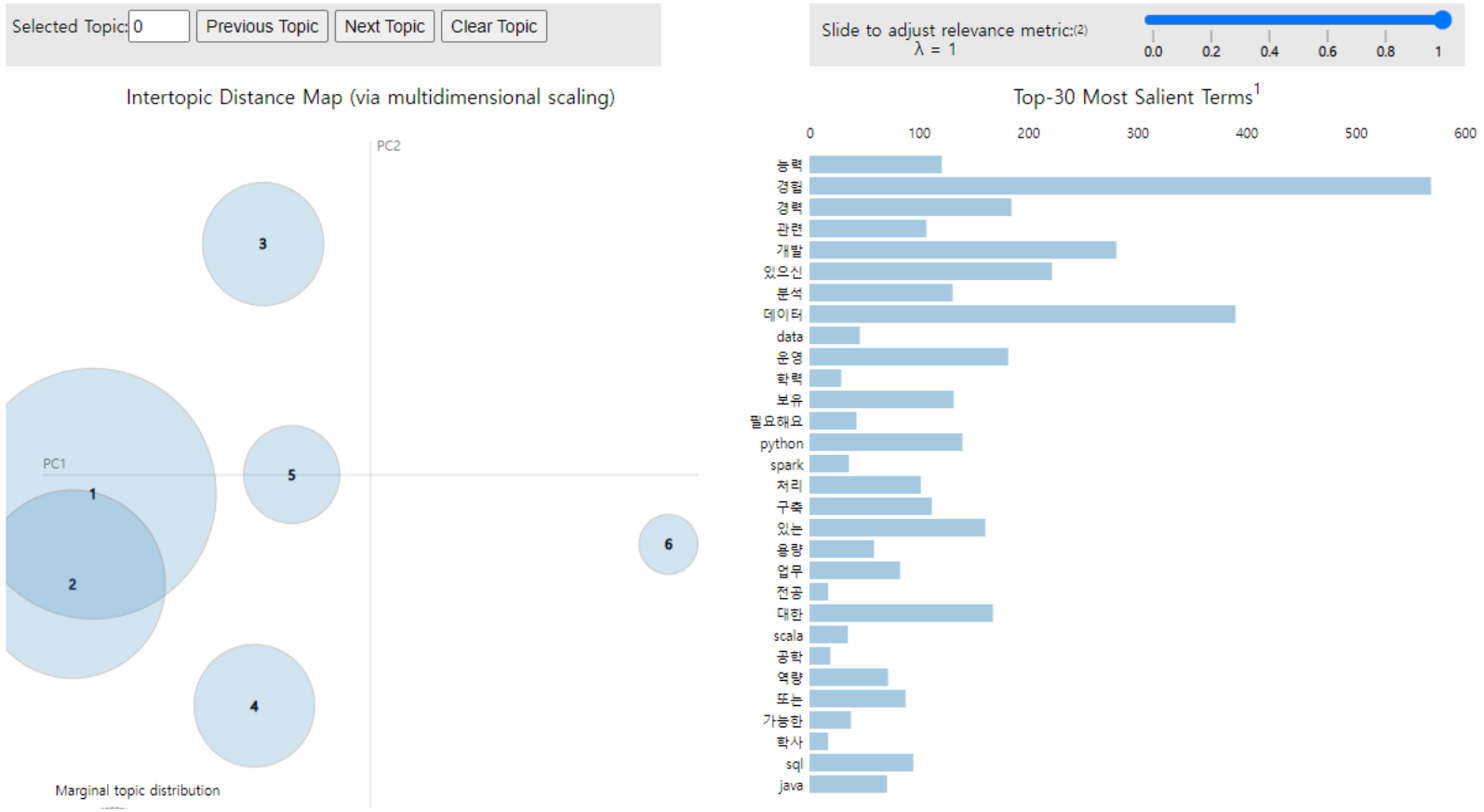
데이터 분석

In [74]:
```
import pyLDAvis.sklearn  # sklearn의 ldamodel에 최적화된 라이브러리

pyLDAvis.enable_notebook()
vis = pyLDAvis.sklearn.prepare(lda, fect_vect, count_vect)
pyLDAvis.display(vis)
```

```
C:₩Users₩user₩AppData₩Roaming₩Python₩Python37₩site-packages₩pyLDAvis₩_prepare.py:247: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except f
'labels' will be keyword-only
  by='saliency', ascending=False).head(R).drop('saliency', 1)
```

Out [74]:



대표적인 LDA 시각화 기법.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 분석

In [96]:
```
from sklearn.decomposition import NMF

n_topics = 2
nmf = NMF(n_components=n_topics,random_state=0)
topics = nmf.fit_transform(fect_vect)
top_n_words = 5
t_words, word_strengths = {}, {}
for t_id, t in enumerate(nmf.components_):
    t_words[t_id] = [count_vect.get_feature_names()[i]
                     for i in t.argsort()[:-top_n_words-1:-1]]
    word_strengths[t_id] = t[t.argsort()[:-top_n_words-1:-1]]

t_words
```

```
C:₩ProgramData₩Anaconda3₩lib₩site-packages₩scipy₩linalg₩decomp_qr.py:20: DeprecationWarning: `n
elf. Doing this will not modify any behavior and is safe. When replacing `np.int`, you may wish
rrent use, check the release note link for additional information.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20
  kwargs['lwork'] = ret[-2][0].real.astype(numpy.int)
C:₩ProgramData₩Anaconda3₩lib₩site-packages₩scipy₩linalg₩decomp_qr.py:20: DeprecationWarning: `n
elf. Doing this will not modify any behavior and is safe. When replacing `np.int`, you may wish
rrent use, check the release note link for additional information.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20
  kwargs['lwork'] = ret[-2][0].real.astype(numpy.int)
```

Out[96]:
```
{0: ['경험', '있으신', '개발', '운영', '구축'], 1: ['데이터', '분석', '있는', '능력', '보유']}
```

분석기법

LDA와 비슷하지만 다른 NMF이라는 비음수행렬을 이용하여 해봤을 때도 결과는 비슷하다. 2개의 토픽이 가장 잘 나뉘어졌다고 판단되었고 경력직의 경우 1.개발 운영의 경험 2.데이터 분석의 능력 두가지를 요구한다고 LDA와 NMF 두가지 방법을 통해 살펴보았다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 3

데이터 분석

경력

1 598.8261750521872
Topic # 0
경험 데이터 개발 있으신 경력 운영 대한 있는 python 분석 보유 능력 구축 관련 이해
2 597.4303651007262
Topic # 0
경험 데이터 개발 있으신 운영 경력 대한 보유 있는 python 구축 분석 관련 이해 처리
Topic # 1
능력 분석 있는 데이터 data 업무 python 경력 커뮤니케이션 해결 sql 활용 논리 가능한 관련
3 607.329783484323
Topic # 0
경험 데이터 개발 있으신 운영 경력 대한 보유 python 구축 있는 분석 관련 이해 처리
Topic # 1
능력 있는 데이터 분석 러닝 업무 경력 해결 경험 python 관련 이해 활용 커뮤니케이션 sql
Topic # 2
data of with to experience python skills years strong the work spark etl 대졸 하시는
4 634.2948186989411
Topic # 0
경험 데이터 개발 있으신 운영 대한 경력 보유 있는 python 분석 구축 처리 이해 관련
Topic # 1
능력 데이터 있는 분석 업무 경험 관련 경력 해결 이해 커뮤니케이션 활용 러닝 sql python
Topic # 2
data of with to experience years skills strong the 경력 python spark work 연구개발 이해
Topic # 3
경험 개발 있으신 운영 경력 python 보유 구축 또는 java 2년 대한 언어 시스템 사용
5 651.8330752994709
Topic # 0
경험 데이터 개발 있으신 운영 대한 있는 경력 보유 구축 이해 python 처리 sql 서비스
Topic # 1
능력 데이터 있는 분석 경험 활용 있으신 업무 sql 해결 경력 python 커뮤니케이션 서비스 논
Topic # 2
data of with to experience years skills strong the work spark python 코드 etl 가능한
Topic # 3
경험 개발 있으신 python 경력 운영 또는 구축 보유 2년 대한 java 업무 언어 기본
Topic # 4
경력 관련 데이터 분석 개발 경험 학력 대한 보유 3년 전공 러닝 공학 사용 역량
6 662.9228540311725
Topic # 0
경험 데이터 개발 대한 경력 보유 이해 운영 있으신 있는 python 사용 구축 sql 3년
Topic # 1
능력 데이터 분석 있는 경험 있으신 python sql 커뮤니케이션 업무 활용 해결 서비스 소통 다
Topic # 2
data of with to experience years skills strong the work spark python etl knowledge 없는
Topic # 3
경험 개발 있으신 운영 보유 python 경력 구축 업무 또는 java 능력 대한 지원 2년
Topic # 4
경력 관련 경험 분석 학력 개발 대한 보유 전공 운영 3년 공학 학사 데이터 또는
Topic # 5
경험 데이터 있으신 개발 운영 처리 활용 있는 용량 필요해요 구축 서비스 역량 python 환경

신입

1 480.8169420809848
Topic # 0
경험 데이터 개발 있으신 대한 이해 있는 능력 분석 또는 보유 활용 관련 python 사용
2 512.3638581741152
Topic # 0
있는 데이터 경험 업무 기술 분석 프로젝트 필수 이해 엔지니어링 대한 지식 학력 해결 경력
Topic # 1
경험 데이터 개발 있으신 대한 이해 능력 관련 보유 또는 분석 활용 python 사용 역량
3 536.9700846789873
Topic # 0
있는 경험 데이터 업무 개발 이해 대한 기술 지식 또는 엔지니어링 능력 해결 보유 러닝
Topic # 1
데이터 경험 있으신 이해 대한 능력 분석 개발 활용 python 보유 있는 관련 설계 사용
Topic # 2
경험 개발 있으신 또는 관련 학사 경력 역량 sql 필수 가능 학위 사용 예정자 전공
4 558.9323598282491
Topic # 0
있는 책임감 새로운 업무 러닝 기술 무관 학력 커뮤니케이션 도전 거부 뛰어나신 열정 주도 대한
Topic # 1
데이터 개발 이해 경험 소통 업무 대한 러닝 가능한 분석 기본 공유 위해 python 이를
Topic # 2
필수 지원 인터뷰 운영 경험 서류 가능 sql 신입 database 해보신 튜닝 cloud mysql 환경
Topic # 3
경험 데이터 있으신 개발 대한 이해 있는 능력 보유 분석 또는 관련 사용 활용 역량
5 588.1152280025622
Topic # 0
예정자 학사 거부 뛰어나신 가능 능력 코딩 찾습니다 개선 aiml 커브 요원 말은 좋겠어 전반
Topic # 1
데이터 개발 경험 업무 소통 이해 대한 러닝 기본 공유 위해 python 이를 분석 전공자
Topic # 2
운영 경험 sql 신입 가능 database 튜닝 해보신 cloud 무관 서비스 보유 aws 능숙하게 활용
Topic # 3
데이터 경험 있는 있으신 개발 대한 이해 능력 분석 활용 기술 관련 또는 보유 python
Topic # 4
경험 개발 있으신 설계 대한 역량 db 데이터 java 사용 이해 가능하신 학력 보유 필요해요
6 606.9720187310644
Topic # 0
예정자 교육 학사 db 전공 컴퓨터공학 무관 이미지 가능 수학 컴퓨터 코딩 가능하신 집요함 discussion
Topic # 1
데이터 개발 업무 소통 경험 대한 위해 이를 있게 분석 이해 능력 기본 전공자 갖추신
Topic # 2
운영 경험 예정자 sql database 학사 해보신 튜닝 가능 신입 cloud aws 서비스 활용 보유
Topic # 3
데이터 경험 능력 있으신 분석 있는 대한 관련 이해 활용 보유 기술 python 또는 sql
Topic # 4
경험 있으신 개발 대한 사용 언어 이해 보유 역량 가능하신 java 업무 db 설계 최소
Topic # 5
경험 개발 있는 데이터 경력 이해 러닝 대한 필요해요 학력 필수 엔지니어링 또는 지원 설계

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 분석

경력

In [96]:
```
from sklearn.decomposition import NMF

n_topics = 2
nmf = NMF(n_components=n_topics,random_state=0)
topics = nmf.fit_transform(fect_vect)
top_n_words = 5
t_words, word_strengths = {}, {}
for t_id, t in enumerate(nmf.components_):
    t_words[t_id] = [count_vect.get_feature_names()[i]
                     for i in t.argsort()[:-top_n_words-1:-1]]
    word_strengths[t_id] = t[t.argsort()[:-top_n_words-1:-1]]

t_words
```

```
C:₩ProgramData₩Anaconda3₩lib₩site-packages₩scipy₩linalg₩decomp_qr.py:20: DeprecationWarning: `n
elf. Doing this will not modify any behavior and is safe. When replacing `np.int`, you may wish
rrent use, check the release note link for additional information.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20
  kwargs['lwork'] = ret[-2][0].real.astype(numpy.int)
C:₩ProgramData₩Anaconda3₩lib₩site-packages₩scipy₩linalg₩decomp_qr.py:20: DeprecationWarning: `n
elf. Doing this will not modify any behavior and is safe. When replacing `np.int`, you may wish
rrent use, check the release note link for additional information.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20
  kwargs['lwork'] = ret[-2][0].real.astype(numpy.int)
```

Out[96]:
```
{0: ['경험', '있으신', '개발', '운영', '구축'], 1: ['데이터', '분석', '있는', '능력', '보유']}
```

신입

In [47]:
```
from sklearn.decomposition import NMF

n_topics = 2
nmf = NMF(n_components=n_topics,random_state=0)
topics = nmf.fit_transform(fect_vect)
top_n_words = 5
t_words, word_strengths = {}, {}
for t_id, t in enumerate(nmf.components_):
    t_words[t_id] = [count_vect.get_feature_names()[i]
                     for i in t.argsort()[:-top_n_words-1:-1]]
    word_strengths[t_id] = t[t.argsort()[:-top_n_words-1:-1]]

t_words
```

```
C:₩ProgramData₩Anaconda3₩lib₩site-packages₩scipy₩linalg₩decomp_qr.py:20: DeprecationWarning: `np.int`
elf. Doing this will not modify any behavior and is safe. When replacing `np.int`, you may wish to us
rrent use, check the release note link for additional information.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-not
  kwargs['lwork'] = ret[-2][0].real.astype(numpy.int)
C:₩ProgramData₩Anaconda3₩lib₩site-packages₩scipy₩linalg₩decomp_qr.py:20: DeprecationWarning: `np.int`
elf. Doing this will not modify any behavior and is safe. When replacing `np.int`, you may wish to us
rrent use, check the release note link for additional information.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-not
  kwargs['lwork'] = ret[-2][0].real.astype(numpy.int)
```

Out[47]:
```
{0: ['경험', '있으신', '개발', '대한', '이해'], 1: ['데이터', '분석', '활용', '보유', 'python']}
```

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4

분석 활용

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 4

분석 활용

신입

신입사원 요구사항 :
- 프로세스 이해
- 개발 경험을 통한 이해
- 언어에 대한 이해

경력

경력직 요구사항 :
- 데이터 개발 운영 경험
- 모델 러닝 해결 경험
- 커뮤니케이션을 이용한 업무 수행

ⓒSaebyeol Yu. Saebyeol's PowerPoint

분석 활용 (신입 지원의 경우)

첫째	이해가 가장 중요!(프로세스, 개발 언어 등)
둘째	이해된 내용을 기반으로 개발 경험
셋째	다양한 경험을 통한 개발 언어 친밀감

ⓒSaebyeol Yu. Saebyeol's PowerPoint

분석 활용 (경력 지원의 경우)

첫째	업무 경험이 중요!
둘째	운영 및 구축을 할 수 있는 능력
셋째	해결할 논리를 커뮤니케이션하는 능력

ⓒSaebyeol Yu. Saebyeol's PowerPoint

개발자 느낌의
데이터 엔지니어

잘들어라„
주석은 '다는' 것이고,
코드는 '짜는' 것이다„„

신입 데이터 엔지니어의 최소 자격 요건은
프로세스에 대한 이해

이해와 경험을 바탕으로

나의 것으로 만드는 것이 필요

ⓒSaebyeol Yu. Saebyeol's PowerPoint

작품을 진행하며…

이번 작품을 진행하며 좋아하는 분야를 좀 더 구체적으로 알게 되었다. 이번에 분석한 작품을 바탕으로 앞으로의 취업을 준비하여 기업이 원하는 구직자의 모습으로 원하는 직장에 들어갈 수 있도록 노력하자!

ⓒSaebyeol Yu. Saebyeol's PowerPoint