프레인 글로벌 데이터 분석

연세대학교 빅데이터 대회

YMCS2

김태오 컴퓨터과학과
김현우 수학과
송민수 응용통계학과
신보현 응용통계학과

목차 / List

1장 팀 소개

2장 프레인 글로벌 회사소개

3장 원본 데이터 , 전처리 및 크롤링 데이터 설명

4장 EDA와 insight

목차 / List

5장 통계적 텍스트 마이닝

6장 수주 예측 모델링

7장 결론 및 한계점

1

팀 소개

빅데이터 대회에 함께 출전한 팀원들을 소개합니다.

팀 소개

YMCS2

김태오

컴퓨터과학 13학번

데이터 크롤링, 모델링

김현우

수학과 13학번

수주 예측 모델링

YMCS2

수학과 1명(Math) ,
컴퓨터학과(Computer) 1명 ,
응용통계학과(Statistics) 2명
으로 이루져있기에
YMCS2로 팀명을 지었습니다.

송민수

응용통계학과 14학번

분석, EDA

신보현

응용통계학과

통계적 텍스트 마이닝 모델링

2

프레인 글로벌 회사소개

저희가 다룬 데이터의 주인공을 소개합니다.

프레인 글로벌 회사 소개

What?

광고 대행업 / PR + Brain = Prain

How much?

2000년도 설립 / 180명 사원 / 중기업 / 매출액 295억원 / 직원 평균연령 30세
(2014년기준)

Amazing!

- 여준영 대표 : 홍보대행사 프레인으로 시작하여 계열사 14개를 거느린 PCG 그룹의 대표
- '2015 올해의 글로벌 PR' 세계59위, 국내 1위

3

원본 데이터, 전처리 및 크롤링 데이터 설명

저희가 다룬 데이터에 대해 설명하겠습니다.

원본 데이터 설명

변수

고객사

- 프레인의 고객사

문의일자 & 접수자

- 프레인이 해당건을 접수한 일자와 접수자가 나와있음

문의개요

- 접수건의 제목이라고 할 수 있는 내용
- 홍보내용에 대한 개략적인 내용 파악 가능

채널

- 접수가 들어온 경로
- 입찰/client/Grapevine/staff/tie-in/article/internet/media/ketchum/outbound

접수경로

- 구체적인 접수경로
- 공고 등등

접수내용

- 접수건의 구체적인 내용 (예산, 일자, 행사 등등)
- 고객사의 메일, 전화 내용 등의 형식

원본 데이터 설명

변수

미팅 및 REP & PT 리포트

- 미팅과 관련한 내용

Bid

- 경쟁PT/경쟁회사소개/수의계약/단독제안/기타

계약기간

- 고객사가 원하는 계약기간

파일명#1#2

- 프레인 데이터가 들어있는 한글파일 이름

Budget & 계약액

- Budget : 고객사가 제공하는 마케팅 금액
- 계약액 : 실제로 성사된 금액

진행

- 진행사항인데 수정이 반영이 안되서 drop(프레인 측 전달사항)

원본 데이터 설명

변수

경쟁사 & 최종선정사

- 경쟁PT인 경우에 해당하며 경쟁사와 최종선정사가 나와있음
- NA값도 많음

결과보고

- 해당 결과에 대한 간단한 설명

댓글

- 고객사와 프레인 내부 진행사항 관련 댓글

결과

- 수주에 성공했는지의 여부

결과구분

- 주로 Failed 일 때의 내부 피드백 내용

원본 데이터 전처리 과정

한글파일을 엑셀화 한 것이라 전처리할 것이 많음
옛날 데이터부터 있다보니 NA 값, 서로 다른 단위 존재

- 접수경로 : 직책 컬럼 생성, 접수 경로 그룹화
- Budget, 계약액 : 단위를 억으로 맞춤
- 경쟁사 : 형식 통일

- 필요 없는 컬럼 삭제

- 크롤링 데이터 추가

크롤링 데이터 설명

1. 고객사 관련 데이터 크롤링
목적 : 고객사를 파악하면 수주의 성공일 것이라 생각
From : nice, 사람인

< 추가 컬럼 >
기업명 / 소속그룹 / 업종 / 사업내용 / 기업형태 / 기업주소 /
자본금 / 매출액 / 영업이익 / 기업개요 및 비전

2. 프레인 관련 뉴스 4000여개 크롤링
목적 : 프레인의 기업 이미지
From : 네이버 뉴스

< 추가 컬럼>
제목/ 시간 / 언론사 / 내용

4

EDA와 insight

탐색적 자료 분석을 통하여 새롭게 알게 된 insight를
소개합니다.

EDA 와 insight

1) 고객사 워드클라우드

insight

- 중소기업, 대기업, 공기업 등 다양한 고객사 1970개
- 국가기관의 홍보대행 사항이 많으므로 이에 대한 준비 필요 (전담팀)
- 다양한 종류의 고객사에 맞추기 위해 프레인 내부에서의 팀 세분화 작업 필요

EDA 와 insight

1) 고객사 워드클라우드(success)

insight

- 상대적으로 횟수가 많은 고객사가 시각화 되었지만 실제로는 약 500개 정도의 고객사가 존재

EDA 와 insight

2) 문의개요 워드클라우드

insight

- 예상했던대로 홍보, 대행, 마케팅 등의 단어가 눈에 띔

EDA 와 insight

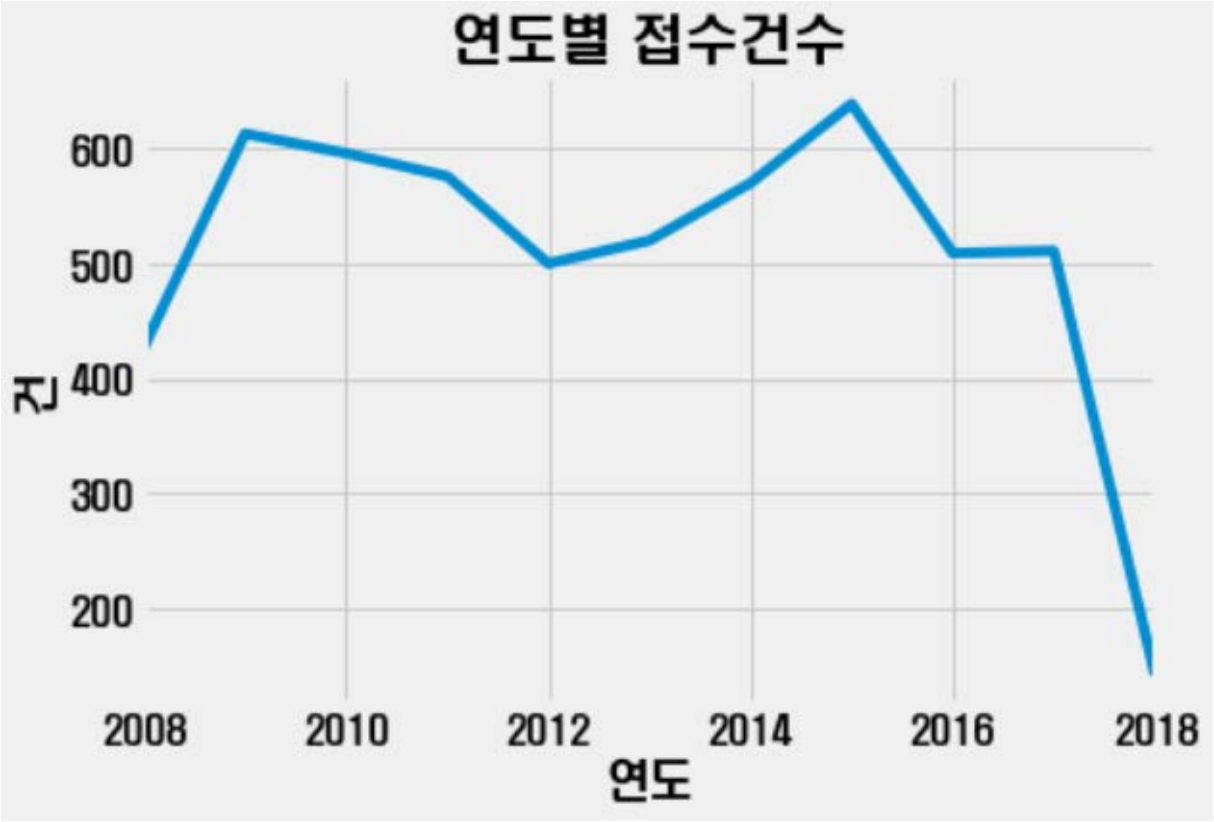
3) 접수내용 워드클라우드

insight

- 예상했던대로 홍보, 대행, 마케팅 등의 단어가 눈에 띔
- 언론, 매체, 미디어, 이벤트 등 고객사에서 원하는 구체적인 홍보 방향에 대한 정보를 갖고 있는 변수

EDA 와 insight

4) 연도별 변화

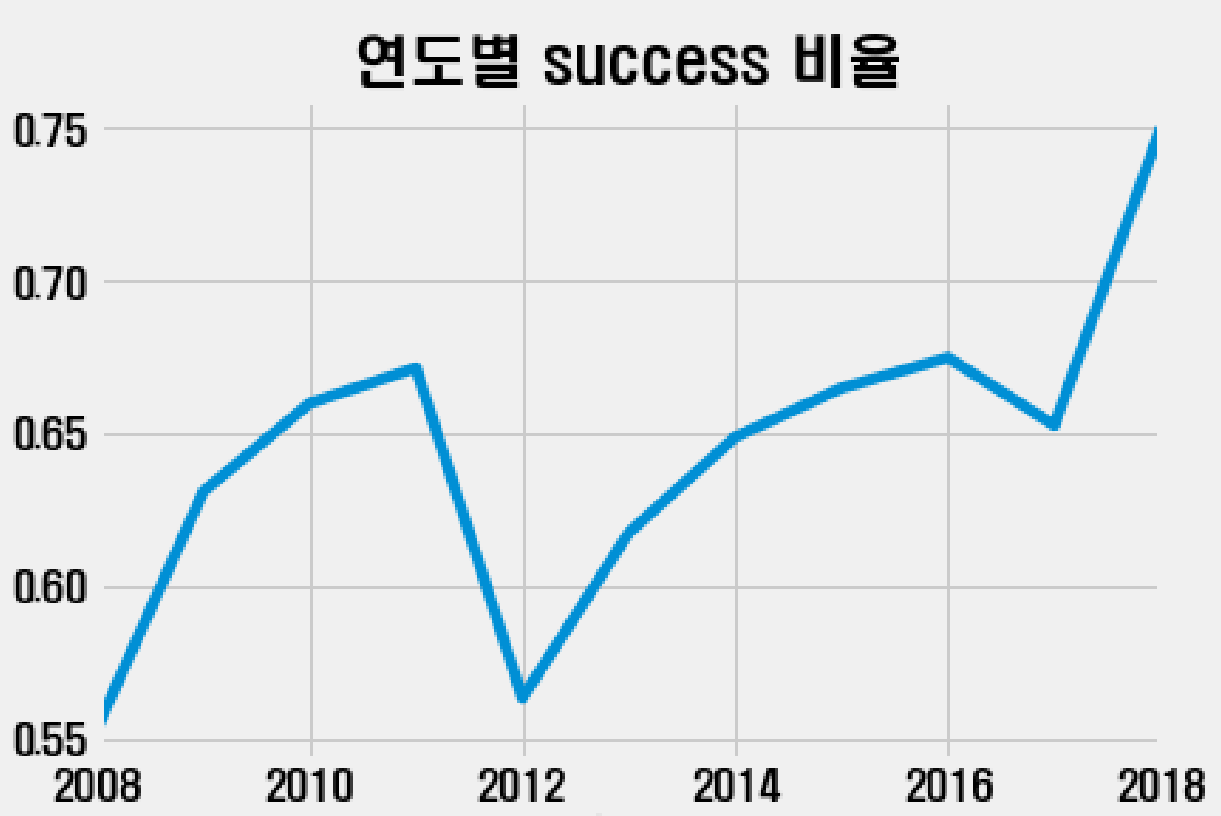


insight

- 연도별 프레인에 접수된 프로젝트 개수
- 연도별 상이한 차이는 없음

EDA 와 insight

4) 연도별 변화

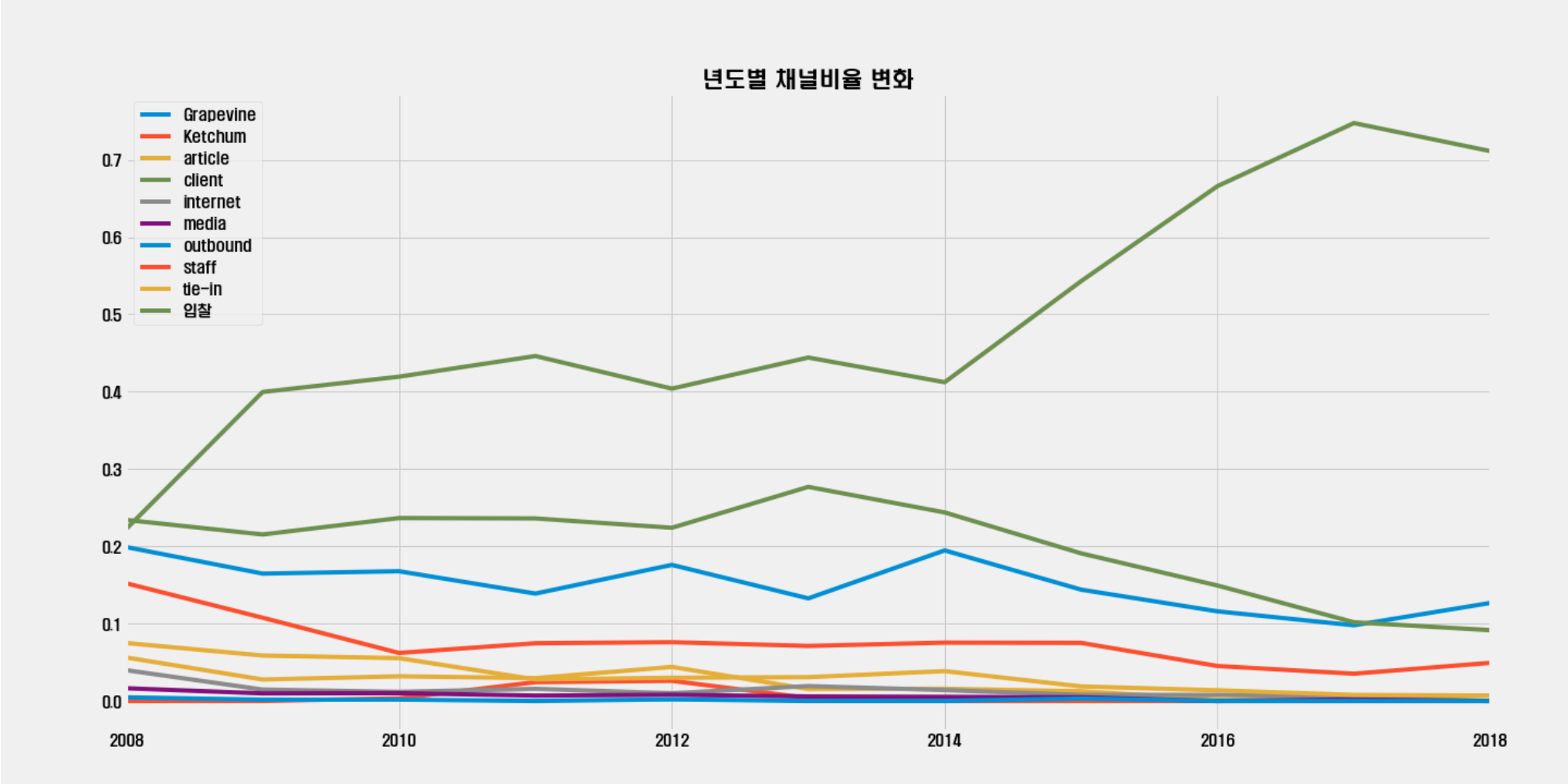


insight

- 연도별 프레인이 성공한 프로젝트 비율(success와 failed 만으로 비교)
- 성공의 비율은 상승함, 이에 따라 failed의 비율은 하락

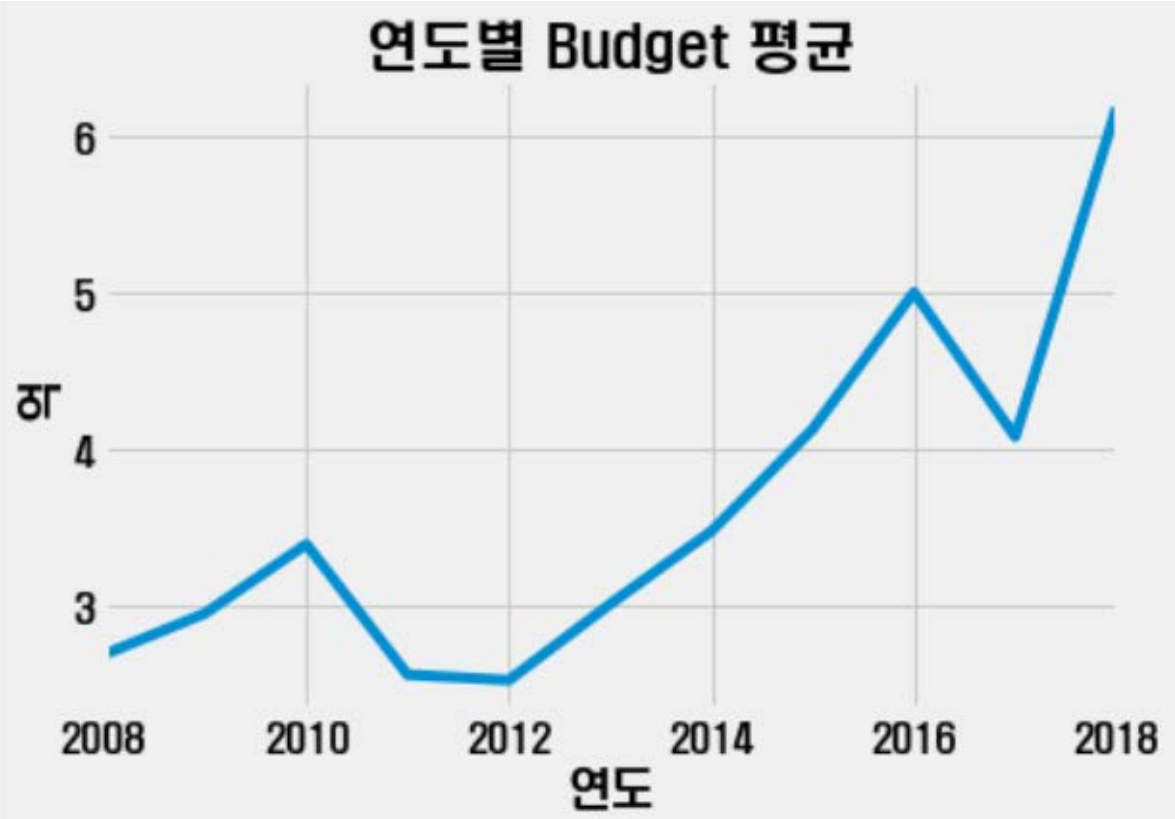
EDA 와 insight

4) 연도별 변화



EDA 와 insight

4) 연도별 변화



insight

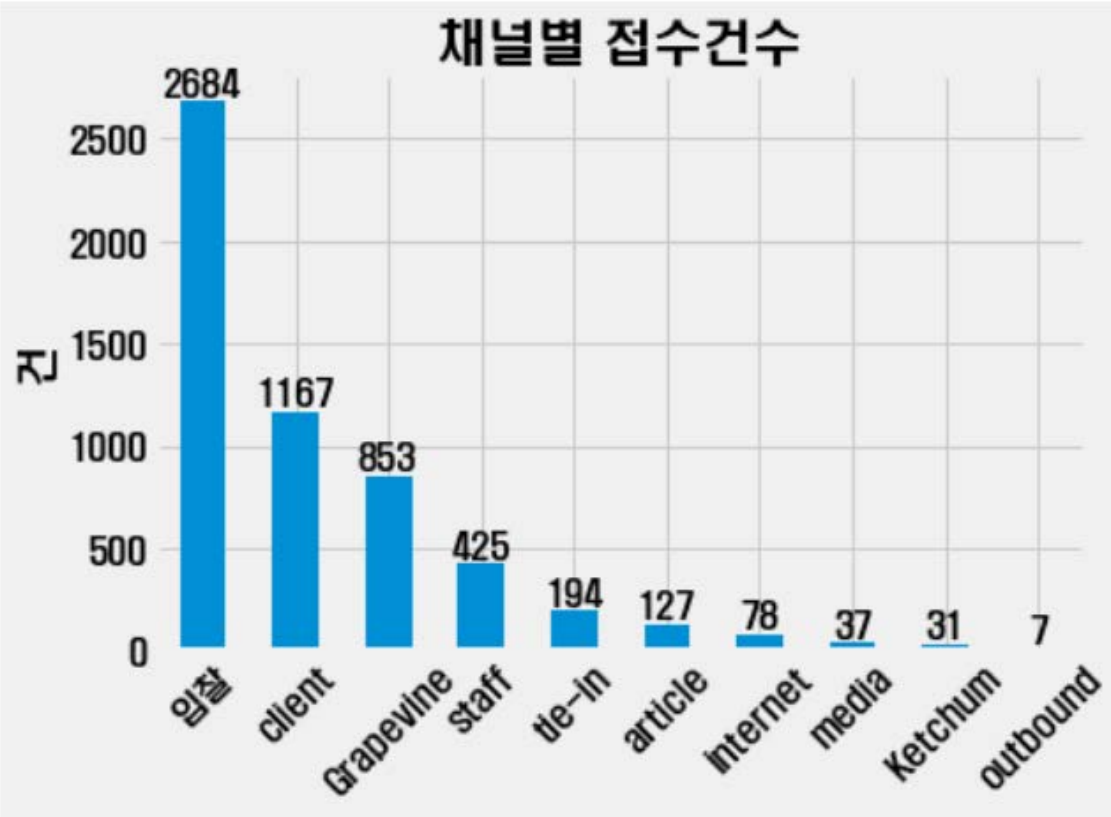
- 연도별 고객사에서 제시한 Budget의 평균
- NA의 값은 연도별로 골고루 분포되어 있기에 drop하여 연도별 변화추이강조

- 시간이 지남에 따라 Budget의 평균값은 상승

- 이를 통해 홍보의 중요성이 커지고 있으며
  홍보대행회사의 경쟁력이 상승하고 있음을 예측가능. 프레인의 밝은 미래 ^^

EDA 와 insight

5) 채널



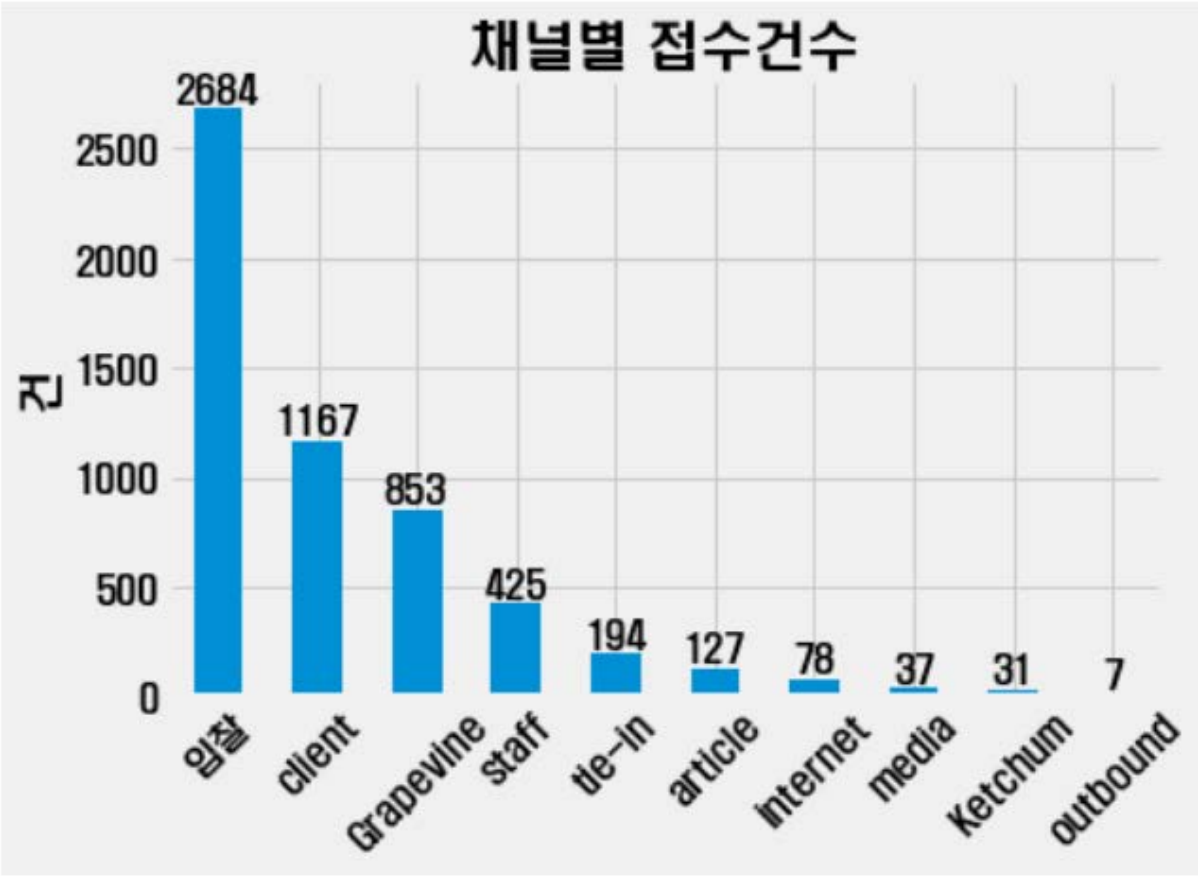
insight

- 채널별 접수건수
- NA값 없음

- article: 프레인의 신문/방송 기사를 보고 연락이 온 경우
- tie-in: 광고/행사, 전시 등 대행사를 통해 연락이 온 경우(예: 제일기획) 혹은 KPRCA 회원사 리스트를 보고 연락이 온 경우
- Grapevine: 프레인이 인지도가 높아 원래 알고 있어 연락이 온 경우 혹은 (기자가 아닌) 주변사람에게 추천 받은 경우
- staff: 프레이너(x-프레이너)의 개인 인맥에 의해서 연락이 온 경우
  * 최초 한번만 staff로 표기 하며 계약 후 추가 세일즈는 client로 표기
- internet: 대행사에 연락하기 위해 인터넷 검색을 하여 연락이 온 경우
- 입찰: 공식적으로 입찰이 난 경우
- media: 기자 소개/추천으로 연락이 온 경우
- outbound: 프레인에 먼저 연락이 온것이 아니라 프레인이 먼저 (전화,브로셔, 기타인맥을 통해) 연락을 취한 경우
- client: 현/과거 고객사로부터 연락이 온 경우 혹은 고객사 담당자가 회사를 옮긴 뒤 연락한 경우

EDA 와 insight

5) 채널

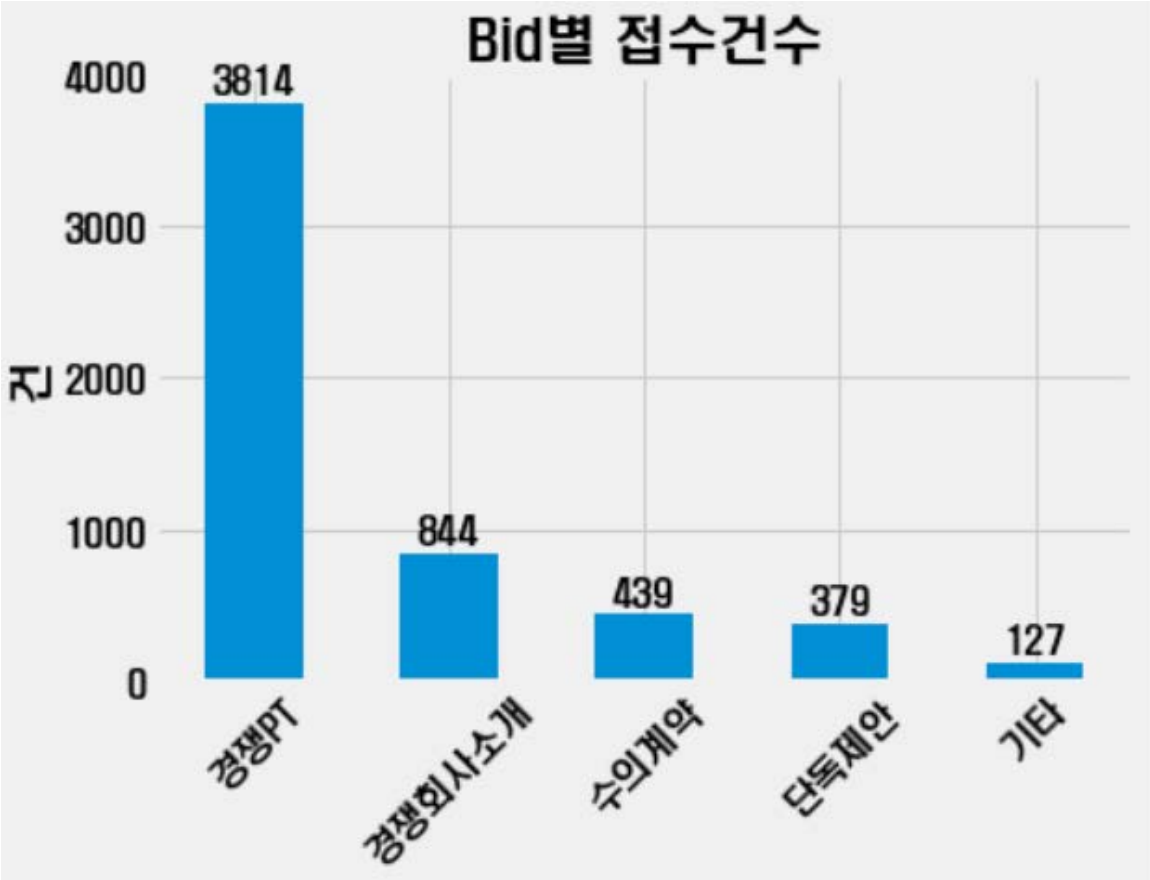


insight

- Client의 수가 많은 것으로 보아 평소 prain에 대한 고객사의 만족도가 높음을 알 수 있음
- Grapevine의 수가 많은 것으로 보아 프레인의 인지도가 높음을 알 수 있음
- Client, Grapevine, staff의 높은 수로 보아 PR시장에서 인맥의 중요성을 알 수 있으며 프레인이 뛰어난 네트워크 상황 을 알 수 있음
- 매체를 통한 홍보로 Article, internet, media의 수를 늘린다면 더 많은 접수 건수 예상

EDA 와 insight

7) Bid



insight

- Bid별 접수건수
- NA값 없음, 대다수가 경쟁PT

EDA 와 insight

8) 경쟁사 워드클라우드

insight

- 경쟁사 워드클라우드
- 경쟁사별로 분석을 통해 경쟁PT에서 우위를 점할 수 있다면 더 많은 수주를 딸 것으로 예상

EDA 와 insight

9) 기업개요 및 비전 (프레인 success)

insight

- 프레인이 계약에 성공한 기업들의 개요 및 비전을 크롤링하여 워드클라우드

EDA 와 insight

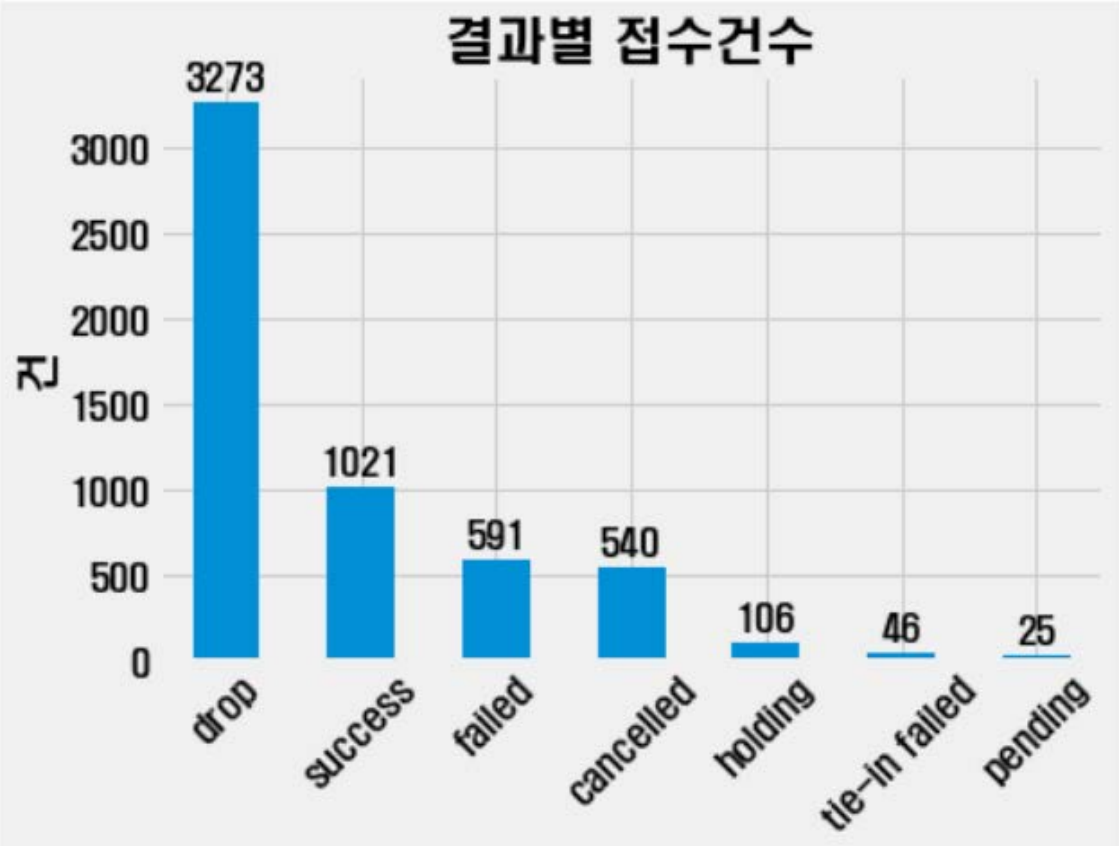
10) 결과 보고 워드클라우드

insight

- 결과 보고 컬럼의 내용은 주로 입찰경쟁에 실패한 경우 이에 대한 피드백을 적어놓은 경우가 대다수
- 성공, 실패의 요인을 상세히 적고 분석하는 프로세스를 만들고 prain 팀 전체가 공유할 수 있는 플랫폼을 만들면 업무에 있어서 효율적인 도움이 될 듯 함

EDA 와 insight

11) 결과



insight

- 결과별 접수건수
- NA값 없음, drop이 대다수(대부분 참여팀이 없기 때문)

- cancelled: 단순 문의건이나 요청 후 취소한 경우, 다시 연락이 오지 않을 것 같은 경우에 체크
- holding: 프로젝트 자체가 미뤄졌을 경우 (고객사측)
- pending: 대행사에서 미루는 경우

- Success와 failed에 초점을 맞추어서 분석
- 전체 성공률 평균(success / success + failed) = 0.63

EDA 와 insight

11) 결과 : 고객사의 업종별 성공률 (크롤링 데이터)

insight

고객사의 업종 분류에서 상위 15개(다수 분류)의 성공률을 살펴보자.

의약품 도매업	0.64
기타 인쇄업	0.91
통신 · 방송장비 및 부품 도매업	0.71
문화 및 관광 행정	0.64
전문 디자인업	0.83
광고 대행업	0.80
보건 및 복지 행정	0.52
협회·단체	0.70
부동산 임대업	0.58
전자상거래 소매 중개업	0.67
컴퓨터 , 소프트웨어 도매업	0.87
식품가공·개발	0.42
그 외 기타 협회 및 단체	0.47
정부·공공기관·공기업	0.65
자동차 세차업	0.56

전체 성공율 : 0.63 보다 낮은 업종

보건 및 복지 행정	0.52
부동산 임대업	0.58
식품가공·개발	0.42
그 외 기타 협회 및 단체	0.47
자동차 세차업	0.56

이 업종의 고객사를 PR할 때는 조금 다른 접근 방법이나 방향성을 고려할 필요가 있다.

이처럼 프레인은 성공율에 따라 업종을 분류하여 본인들의 전략을 수정한다면 더욱 성장할 것으로 기대된다.

EDA 와 insight

11) 결과 : 고객사의 기업형태별 성공율 (크롤링 데이터)

insight

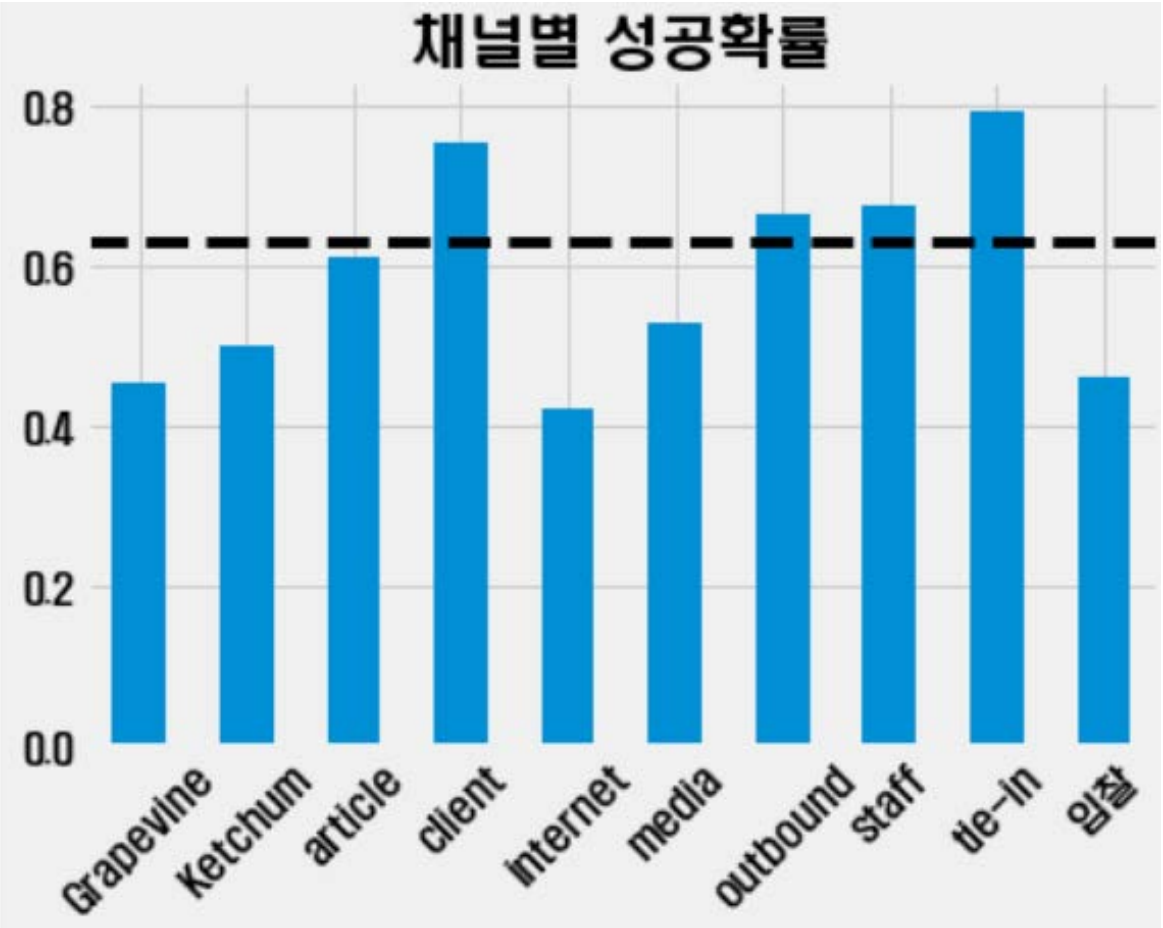
성공율

- 중견기업 0.46 < 대기업 0.61 < 중소기업 0.71
- 공사 공기업 0.60
- 외국인 투자기업 0.58
- 수출입기업 0.60

기업의 형태에 따른 특징을 데이터화하여 관리한다면
전략수립의 과정에 있어서 많은 도움을 받을 수 있을 거라고 생각한다.

EDA 와 insight

11) 결과 : 채널별 성공확률



insight

- client, staff, tie-in : 평균 성공율보다 높고 개수가 유의미한 채널
- Grapevine, 입찰 : 평균 성공율보다 낮고 개수가 유의미한 채널

- 개수에 대한 고민!!

5

통계적 텍스트 마이닝
Statistical text mining

텍스트 마이닝에서 통계적 관점 갖고 알고리즘을 설명합니다.

통계적 텍스트 마이닝(Statistical Text Mining)

정형화된 데이터 구조

	Var.1	Var.2	…	Var.n
Obs.1	Data.11	Data.12	…	Data.1n
Obs.2	Data.21	Data.22	…	Data.2n
…	…	…	…	…
Obs.m	Data.m1	Data.m2	…	Data.mn

Row는 관측치, Column은 변수를 의미하며 표의 형식을 갖춤.

각 행과 열이 만나는 지점은 해당 관측치의 해당 변수에 대한 값을 나타냄.

통계적 텍스트 마이닝(Statistical Text Mining)

프레인 데이터에서 제공한 데이터 중 텍스트 데이터를 어떻게 활용할까

- **비정형 데이터인 텍스트 데이터에 바로 통계적 기법을 적용하기는 힘듬**
- **전처리 과정을 통해서 정형화된 데이터 구조로 만든 후,**
- **통계적 기법 적용 및 시각화**

KONLP	tm	stringr	qgraph
전처리	**전처리**	**전처리**	**시각화**

통계적 텍스트 마이닝(Statistical Text Mining) [전처리]

	접수 내용
Obs.1	텍스트 데이터
Obs.2	텍스트 데이터
Obs.3	텍스트 데이터

텍스트 마이닝

	Term.1	Term.2	Term.3	Term.4
Obs.1				
Obs.2				
Obs.3				

각 관측치에서의 해당 용어가 나온 횟수(빈도)

문의한 고객이 무엇을 원했을까?
문의한 고객이 많이 사용한 단어는 무엇일까?
사용한 단어 사이의 관계는 어떻게 될까?

통계적 텍스트 마이닝(Statistical Text Mining) [전처리]

접수내용 칼럼

접수내용

1. 입찰공고명 : 쌀의 긍정적 이슈 확산을 위한 매체홍보
2. 사업금액(부가세포함) : 500,000,000원
3. 사업기간 : 계약체결일로부터 2018년 11월 30일까지
4. 입찰참가자격
1)「국가종합전자조달시스템 입찰참가자
록한 자
2) ｢국가를당사자로하는계약에
3) 공동계약(공동이행방식)이 가능
5. 입찰마감일시
1) 제안서 제출마감 : 2018. 4. 9 (월) 16:0
2) 전자 입찰마감 : 2018. 4. 9 (월) 16:00
6. 문의사항

1) 제안에 관한 사항 : 소비문화실 여동욱 과장 ☎ 044-861-8862
2) 입찰에 관한 사항 : 재무관리실 신원상 대리 ☎ 044-861-8782

주요 사업내용

1. 2018년 사업목표 및 전략제시(종합실행계획)

- 효과적인 과업 수행과 홍보 효과 극대화를 위한 KPI 제시 및 달성을 위한 과업별 구체적 전략 제시

Dear. 프레인글로벌,
주식회사 설빙의 이제훈입니다.

다음은 저희가 기대하는 업무 스콥입니다.

1. 보도자료 작성/배포 (단신, 기획) - 평균 월 12건
2. 미디어 케어 (기자 미팅 외) - 평균 월 8회 미팅
3. 위기관리 (사전, 사후, 상시)
4. 주요 미디어 피칭 (공중파TV, 종편방송TV, 경제지, 온라인지, 주간지 외_ 우선순위)
5. 기타 포토행사, 이슈메이킹 행사 등 제안 (별도)

상기 업무를 중심으로 설빙의 얼굴이 되어주실 것에 대해
귀사의 검토의사가 있으시다면 우선 짧게 아래의 내용에 대해서 메일 피드백을 부탁드립니다.
바로 연락 드리겠습니다. (~3/16,금)

주요 사업내용 + 형식 없는 접수내용
= (새로운) 접수내용 칼럼

통계적 텍스트 마이닝(Statistical Text Mining) [전처리]

결과 칼럼 값

```
> unique(sales$결과)
[1] drop          ing           success       cancelled     failed        holding
[7] tie-in failed pending
Levels:  cancelled drop failed holding ing pending success tie-in failed
```

결과 칼럼 값이 success인 행

나눠서 전처리

결과 칼럼 값이 cancelled, failed, holding, tie-in failed, pending인 행

각각의 경우에 대해 나타나는 단어 및 서로의 관계를 살펴봄

통계적 텍스트 마이닝(Statistical Text Mining) [전처리]

	접수 내용
Obs.1	
Obs.2	
Obs.3	

KONLP, tm, stringr 패키지를 통한 명사추출

	접수내용의 명사
Obs.1	
Obs.2	
Obs.3	

통계적 텍스트 마이닝(Statistical Text Mining) [전처리]

	주요 사업내용 명사
Obs.1	
Obs.2	
Obs.3	

KONLP 패키지를 통해 Corpus 생성 후, TermDocumentMatrix 생성

```
> inspect(noun.dtm.suc)
<<DocumentTermMatrix (documents: 1022, terms: 12349)>>
Non-/sparse entries: 41619/12579059
Sparsity           : 100%
Maximal term length: 47
Weighting          : term frequency (tf)
Sample             :
     Terms
Docs  관련 대행사 미팅 선정 예산 예정 요청 제안 진행 홍보
  181    2      0    0    2    0    3    7   21    3    0
  251    0      0    0    0    0    0    0    0    2    4
  296    0      0    0    0    0    1    0    1   11    1
  337    1      0    1    1    1    0    0    1    0    0
  540    0      0    0    1    1    0    0    0    0    0
  616    9      0    0    1    1    0    0    2    0   31
  720    0      0    1    1    0    0    0    1    4    0
  79     2      0    1    0    0    0    2    5    0   19
  998    4      0    0    1    0    1    0    2    0   10
  999    0      0    1    2    1    0    0    8    1    1
```

```
> dim(noun.dtm.suc)
[1]  1022 12349
```

1022개의 관측치(success)와 12349개의 칼럼(명사)

통계적 텍스트 마이닝(Statistical Text Mining) [전처리]

고차원 데이터

```
> dim(noun.dtm.suc)
[1]  1022 12349
```

PCA??

전체 관측치에서 나오는 횟수가
100번 이상인 단어만 선택

```
> dim(sales.suc.100)
[1] 1022   70
```

통계적 텍스트 마이닝(Statistical Text Mining) [사회망 분석 SNA]

```
> sales.suc.100[1:5,1:10]
     Terms
Docs 가능 국내 기존 매체 브랜드 예산 정도 제안 제품 진행
   1    1    2    2    1      2    3    1    4    2    1
   2    0    0    0    0      0    1    0    3    0    0
   3    0    0    0    0      0    0    0    1    0    1
   4    0    0    1    0      0    0    0    0    0    0
   5    0    0    0    0      0    0    0    2    0    1
```

Correlation Matrix

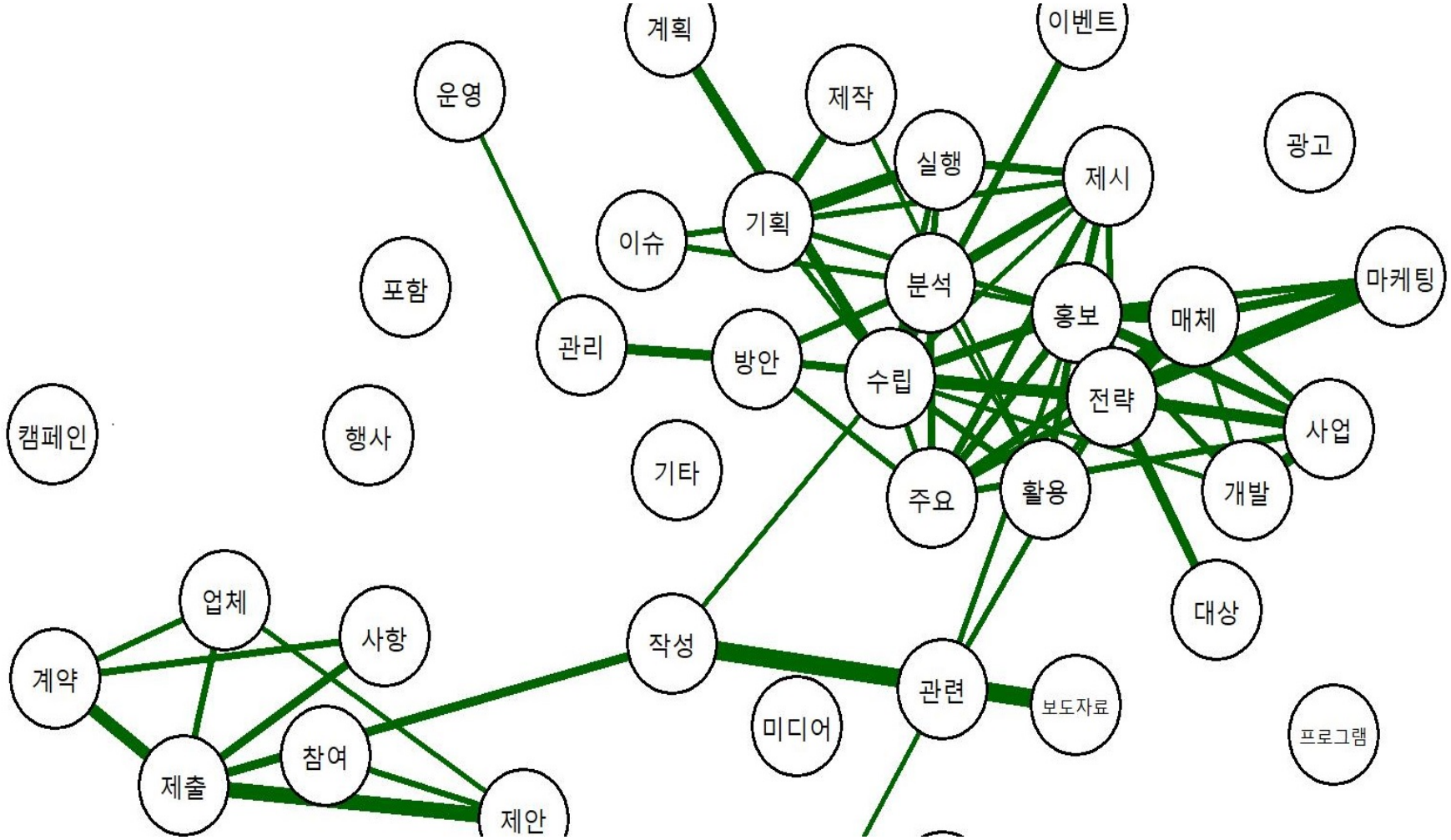
```
> cormat.suc[1:5,1:5]
              가능         국내         기존         매체       브랜드
가능   1.00000000 0.062682157 0.015854334 0.02870697 0.02415953
국내   0.06268216 1.000000000 0.005464289 0.13602769 0.08463179
기존   0.01585433 0.005464289 1.000000000 0.01580801 0.06025626
매체   0.02870697 0.136027693 0.015808014 1.00000000 0.13873847
브랜드 0.02415953 0.084631794 0.060256259 0.13873847 1.00000000
```

통계적 텍스트 마이닝(Statistical Text Mining) [사회망 분석 SNA]

상관계수가 0.3 이상인 변수만 골랐음

qgraph 패키지

Success

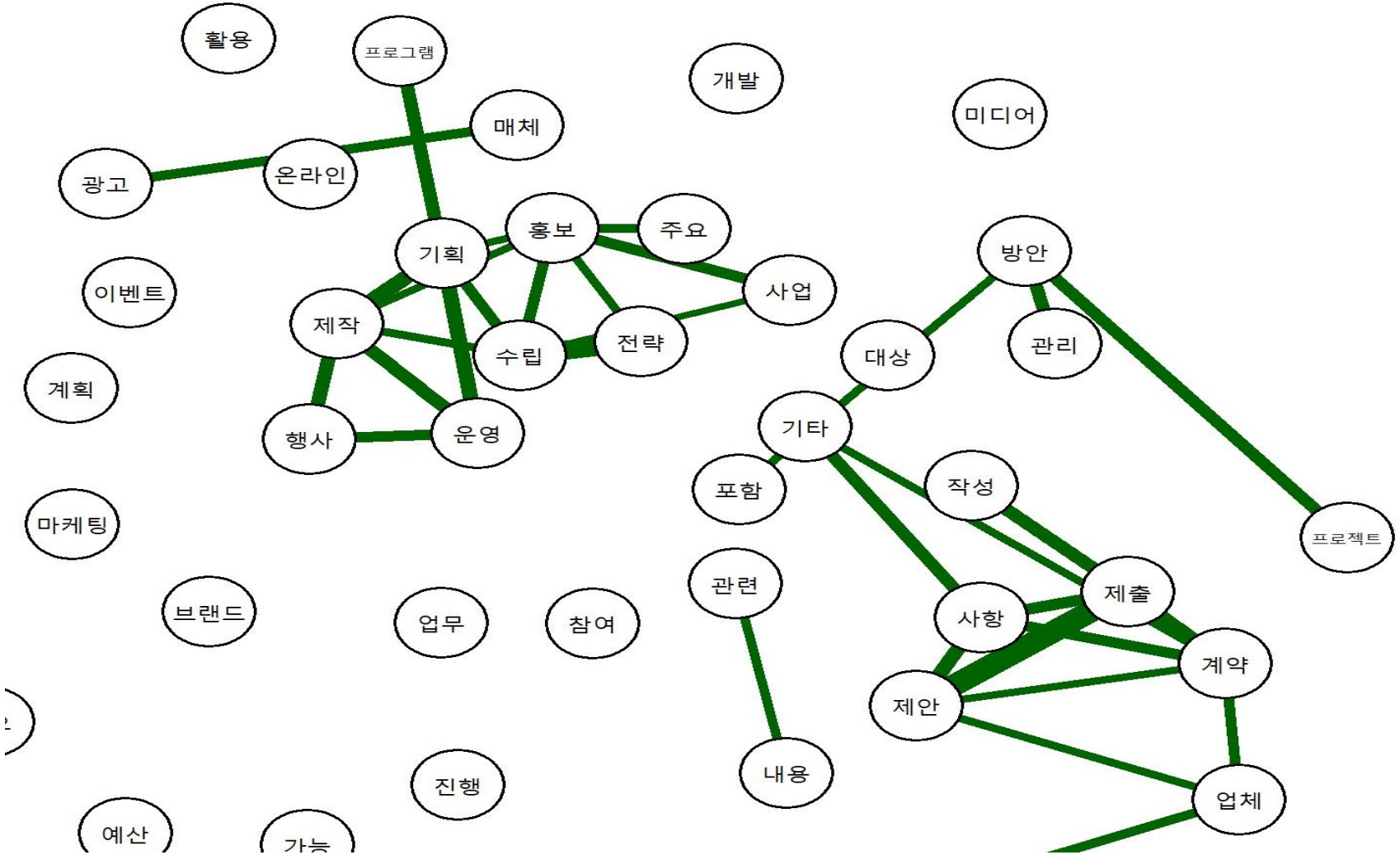


통계적 텍스트 마이닝(Statistical Text Mining) [사회망 분석 SNA]

상관계수가 0.3 이상인 변수만 골랐음

qgraph 패키지

Fail 및 기타 등



통계적 텍스트 마이닝(Statistical Text Mining) [사회망 분석 SNA]

```
sort(cent.suc$OutDegree, decreasing = TRUE)
수립      10.694732
홍보      10.693297
전략      10.544943
주요      10.245823
제안       9.247244
매체       8.853639
작성       8.716944
기획       8.665699
분석       8.567153
활용       8.512040
관련       8.074791
실행       7.941906
기타       7.884240
방안       7.718467
대상       7.692298
제시       7.615048
내용       7.605453
사업       7.450259
개발       7.057381
사항       6.972346
보도자료   6.880400
관리       6.720327
제작       6.717830
마케팅     6.484792
선정       6.313082
미디어     6.209478
이슈       6.139843
업체       6.099512
진행       6.096927
제출       6.053327
광고       6.000601
일정       5.992439
이벤트     5.856791
온라인     5.831964
포함       5.760779
운영       5.748018
```

Success

```
sort(cent.fail$OutDegree, decreasing = TRUE)
홍보      7.611303
수립      7.595637
포함      7.413731
제안      7.284859
제작      7.121479
기타      7.012973
주요      6.999771
기획      6.788362
전략      6.767379
관련      6.742840
사항      6.526631
운영      6.224634
작성      6.045685
제출      5.998892
사업      5.971261
내용      5.944567
매체      5.872581
진행      5.726015
대상      5.515107
참여      5.461049
온라인    5.460929
아래      5.365413
행사      5.360542
방안      5.288236
선정      5.208491
관리      5.107598
프로그램  4.989117
업무      4.984991
계약      4.920450
의뢰내용  4.880452
예정      4.840221
개발      4.758314
이벤트    4.683294
일정      4.651525
대행사    4.631096
마케팅    4.629054
광고      4.595177
요청      4.567643
계획      4.465630
미팅      4.443838
```

fail 및
기타 등

통계적 텍스트 마이닝(Statistical Text Mining) [회귀분석]

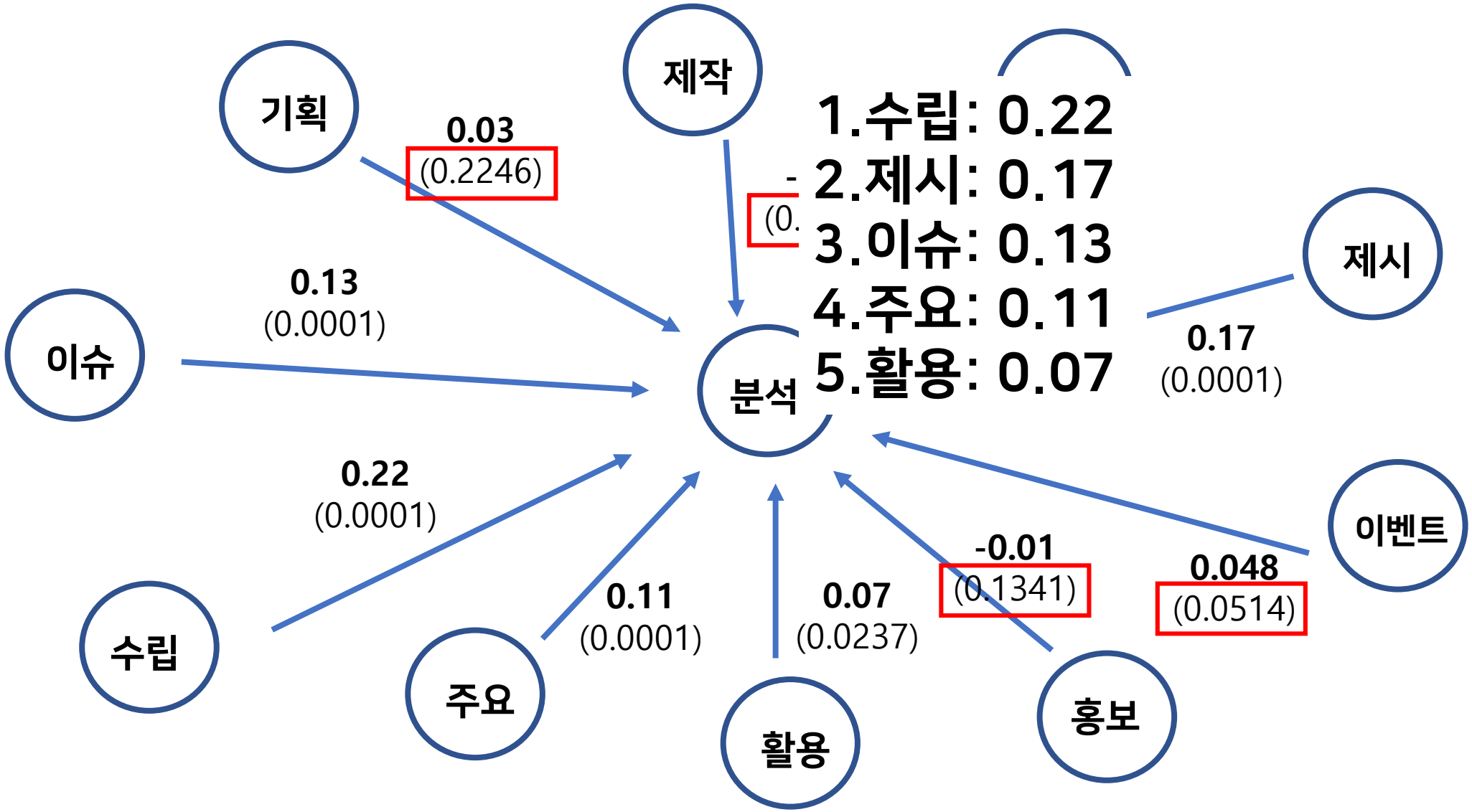
success에만 높은 차수로 나타나는 용어에 어떤 용어들이 영향을 미치는지 분석

분석, 활용 용어에 연결된 용어들을 SNA 시각화에서 파악

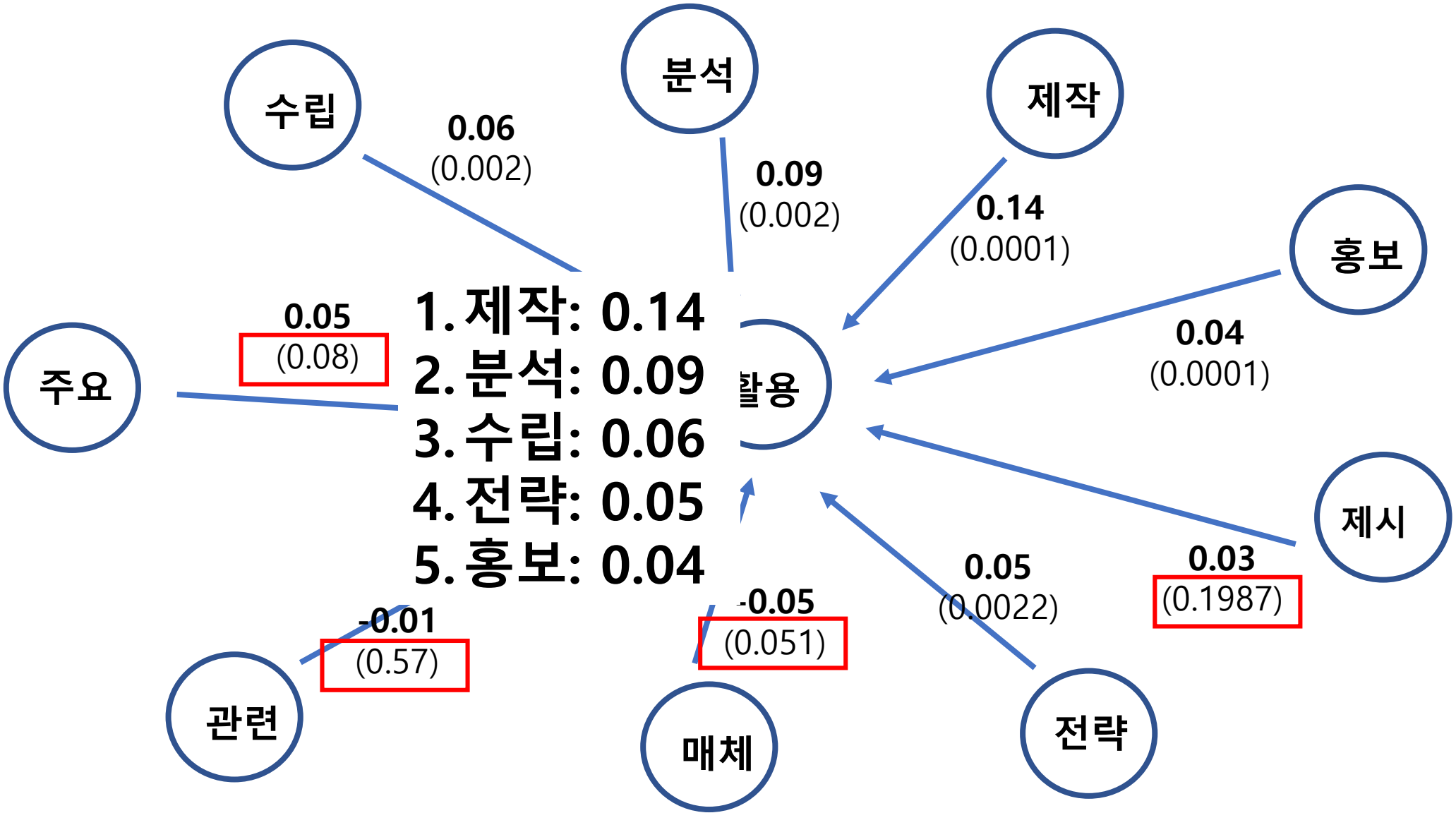
회귀분석 실시

Dependent	independent
분석	이슈, 기획, 제작, 실행, 이벤트, 제시, 홍보, 활용, 주요, 수립
활용	주요, 수립, 분석, 제작, 홍보, 제시, 전략, 매체, 관련

통계적 텍스트 마이닝(Statistical Text Mining) [회귀 분석 – 분석 용어]



통계적 텍스트 마이닝(Statistical Text Mining) [회귀 분석 – 활용 용어]



1. 제작: 0.14
2. 분석: 0.09
3. 수립: 0.06
4. 전략: 0.05
5. 홍보: 0.04

6

수주 예측 알고리즘

공고와 고객사의 정보를 바탕으로 수주를 예측합니다.

수주 예측 알고리즘

최대 82.6%의 예측율을 가진 수주 예측 알고리즘 생성

과정

1. 전처리
2. Word2vec 으로 단어별 벡터생성
3. Rep2vec 으로 문서별 벡터생성
4. Autoencoder 를 통한 차원축소
5. 다양한 머신러닝, 딥러닝 모델 학습을 통해 수주 확률 예측
6. 수주 예측 알고리즘 어플리케이션 제작

수주 예측 알고리즘

1) 전처리

1. 문의개요 / 접수내용 / 사업내용 -> regex(정규표현식)을 통해 한글만 남김

```
# 접수내용
# 한국말만 남겨주기

file2["접수내용_regex"] = file2["접수내용"].map(lambda x : re.sub(r'[^가-힣 ]', ' ', x) if type(x)==str else "")
file2["접수내용"] = file2["접수내용"].map(lambda x : x.replace("-", "").split("₩n") if type(x)==str else "")
```

수주 예측 알고리즘

1) 전처리

2. 접수경로 / 채널 / Bid / 소속그룹 / 업종 / 기업형태 -> dummy화

```
# 채널
# 카테고리로 전환 -> onehot

file2 = pd.concat([file2, pd.get_dummies(file2["채널"])], axis=1).drop("채널", axis=1)
file2.head(1)
```

수주 예측 알고리즘

1) 전처리

3. Budget -> regex를 통한 전처리 및 정규화

```
# 채널
# 카테고리로 전환 -> onehot

file2 = pd.concat([file2, pd.get_dummies(file2["채널"])], axis=1).drop("채널", axis=1)
file2.head(1)
```

4. 결과 -> success or not 으로 masking

수주 예측 알고리즘

1) 전처리

5. 자본금 / 매출액 -> 기업유형별 평균값으로 null값을 채워줌

```
# 자본금
# regex
# NaN 값을 기업유형별 평균으로 처리 해주기 및 정규화

temp1 = file2[file2["기타"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp2 = file2[file2["공사/공기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp3 = file2[file2["1000대기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp4 = file2[file2["재단법인"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp5 = file2[file2["비영리단체/협회"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp6 = file2[file2["대기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp7 = file2[file2["외부감사법인"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp8 = file2[file2["중견기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp9 = file2[file2["중소기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp10 = file2[file2["외국인 투자기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp11 = file2[file2["주식회사"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp12 = file2[file2["학교/교육기관"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp13 = file2[file2["수출입 기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp14 = file2[file2["병원/의료기관"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp15 = file2[file2["사단법인"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp16 = file2[file2["코스피"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp17 = file2[file2["연구소"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp18 = file2[file2["코스닥"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp19 = file2[file2["외국 법인기업"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp20 = file2[file2["협동조합"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp21 = file2[file2["금융기관"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp22 = file2[file2["병역특례 인증업체"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp23 = file2[file2["유한회사"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp24 = file2[file2["코넥스"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()
temp25 = file2[file2["사회복지기관"] == 1]["자본금"].dropna().map(lambda x : change_money(x)).mean()

file2["자본금"] = file2["자본금"].map(lambda x : change_money(x) if type(x) == str else "")
file2["자본금"] = file2.apply(lambda x : money(x) if x["자본금"]=="" else x["자본금"], axis=1)
file2["자본금"] = (file2["자본금"] - file2["자본금"].mean())/file2["자본금"].std()
file2.head(1)
```

수주 예측 알고리즘

1) 전처리

6. 맞춤법 검사

```
def chcking(x) :
    global cnt
    if tpye(result) == str :
        try :
            result = spell_checker.check(x).as_dict()["checked"]
        except :
            result = ""
            print(cnt)
            cnt+=1
    else :
        result = ""
        print(cnt)
        cnt+=1

    return result
```

```
test = "아버지가방에들어가신다"
result_test = spell_checker.check(test).as_dict()["checked"]
print(result_test)
```

아버지가 방에 들어가신다

7. 명사화

수주 예측 알고리즘

1) 전처리

- One-hot encoding

단어를 컴퓨터에서 표현하는 여러가지 방법 중 하나로, N개의 단어가 있으면 각 단어는 한 개의 요소에 대하여 1의 값을 가지는 N차원 벡터가 됨

단어를 쉽게 표현할 수 있다는 장점이 있지만, 단어 간의 유사도, 상관관계를 고려할 수 없다는 단점이 존재

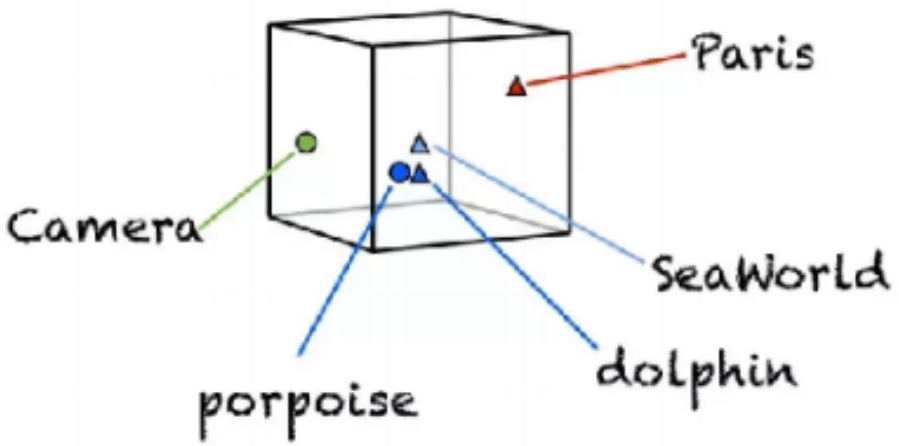
Categorical Feature		f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Louise	=>	1	0	0	0	0	0	0	0	0	0
Gabriel	=>	0	1	0	0	0	0	0	0	0	0
Emma	=>	0	0	1	0	0	0	0	0	0	0
Adam	=>	0	0	0	1	0	0	0	0	0	0
Alice	=>	0	0	0	0	1	0	0	0	0	0
Raphael	=>	0	0	0	0	0	1	0	0	0	0
Chloe	=>	0	0	0	0	0	0	1	0	0	0
Louis	=>	0	0	0	0	0	0	0	1	0	0
Jeanne	=>	0	0	0	0	0	0	0	0	1	0
Arthur	=>	0	0	0	0	0	0	0	0	0	1

수주 예측 알고리즘

2) Word2vec

단어를 컴퓨터에서 표현하는 여러가지 방법 중 하나로, 단어를 embedding을 시켜 벡터화

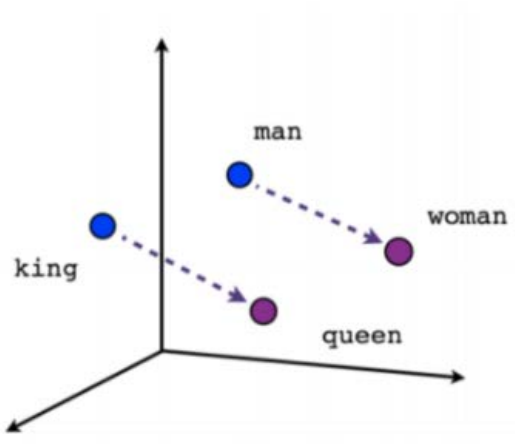
Embedding이란, 단어를 N차원 공간에 mapping하는 방법으로, 주변 맥락을 고려하여 mapping을 시키기 때문에 의미적으로 비슷한 단어는 비슷한 공간에 위치하게 됨. 임의의 벡터가 아니라 단어들 사이의 관계가 고려된 벡터이기 때문에 +,- 등의 연산이 정의될 수 있으며, 학습과정에 단어 자체의 의미도 포함시킬 수 있게 됨



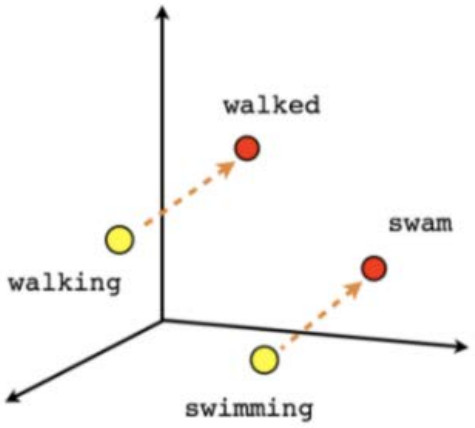
수주 예측 알고리즘

2) Word2vec

King – man + woman = queen
Seoul – korea + japan = tokyo



Male-Female



Verb tense

우리는 Pretrained된 word2vec을 이용하여 명사화된 단어들을 모두 벡터로 전환시킨다.

수주 예측 알고리즘

3) Rep2vec
- TF-IDF

문서 내에서 각 단어가 가지는 중요도를 표현하는 지표로 사용
각 문의개요 / 접수내용 / 사업내용 등 텍스트가 사용된 컬럼에 대해서 모든 명사의 중요도를 계산

TF-IDF : 한 문서 내에서 얼마나 등장했는 지와 문서내의 문장에서 얼마나 자주 등장하느냐를 연산에 포함시킴.
문서에 자주 등장하는 단어는 중요한 단어이지만, 모든 문장에 등장하는 단어는 '나 ' 처럼 의미가 없는 단어이기 때문에 두 특성을 모두 고려

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

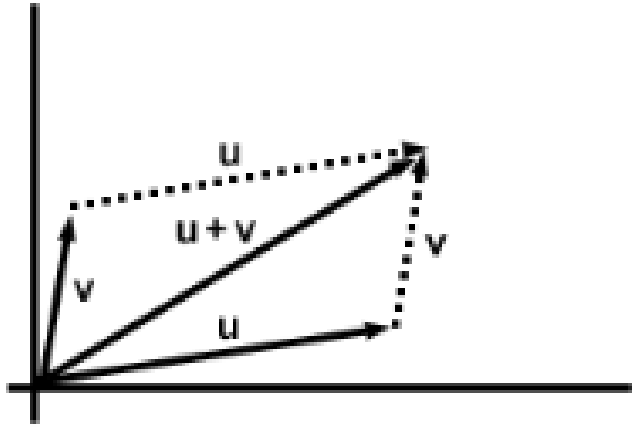
수주 예측 알고리즘

3) Rep2vec

현재 문서의 한 단어에는 임베딩 된 벡터와 가중치 값이 각각 하나씩 존재하는데, 이 둘을 곱한 벡터들을 다시 모두 더해 준 후 총 가중치 값만큼 나눠줌, 즉, 가중합

이렇게 한다면, 전체적인 문서의 벡터가 만들어지는데, 중요한 단어의 방향으로 벡터의 방향이 정해지게 된다.

Report vector = Σ (tf_idf x embedding vector) / Σ tf_idf

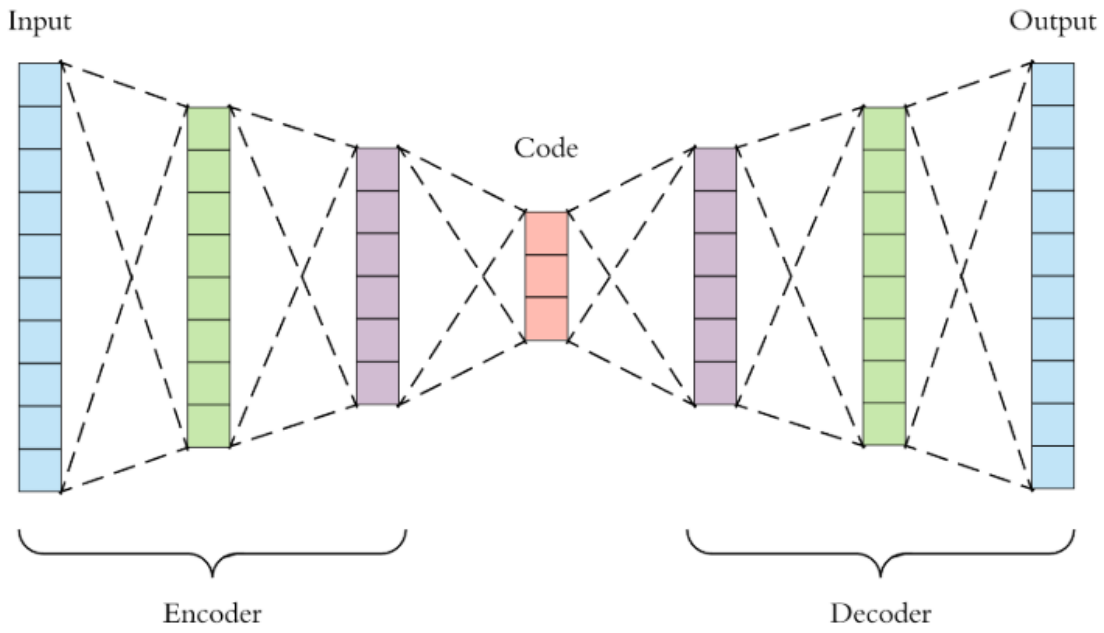


수주 예측 알고리즘

4) Dimension reduction by AutoEncoder (vs PCA)

Dummy화와 embedding을 통하여 만들어진 column의 수가 총 738개이므로 차원을 감소시켜줘야할 필요성이 존재

원래의 벡터를 압축하는 encoder와 이 압축된 벡터를 다시 복원하는 decoder로 구성
원래의 벡터를 효과적으로 저차원의 벡터로 압축 가능

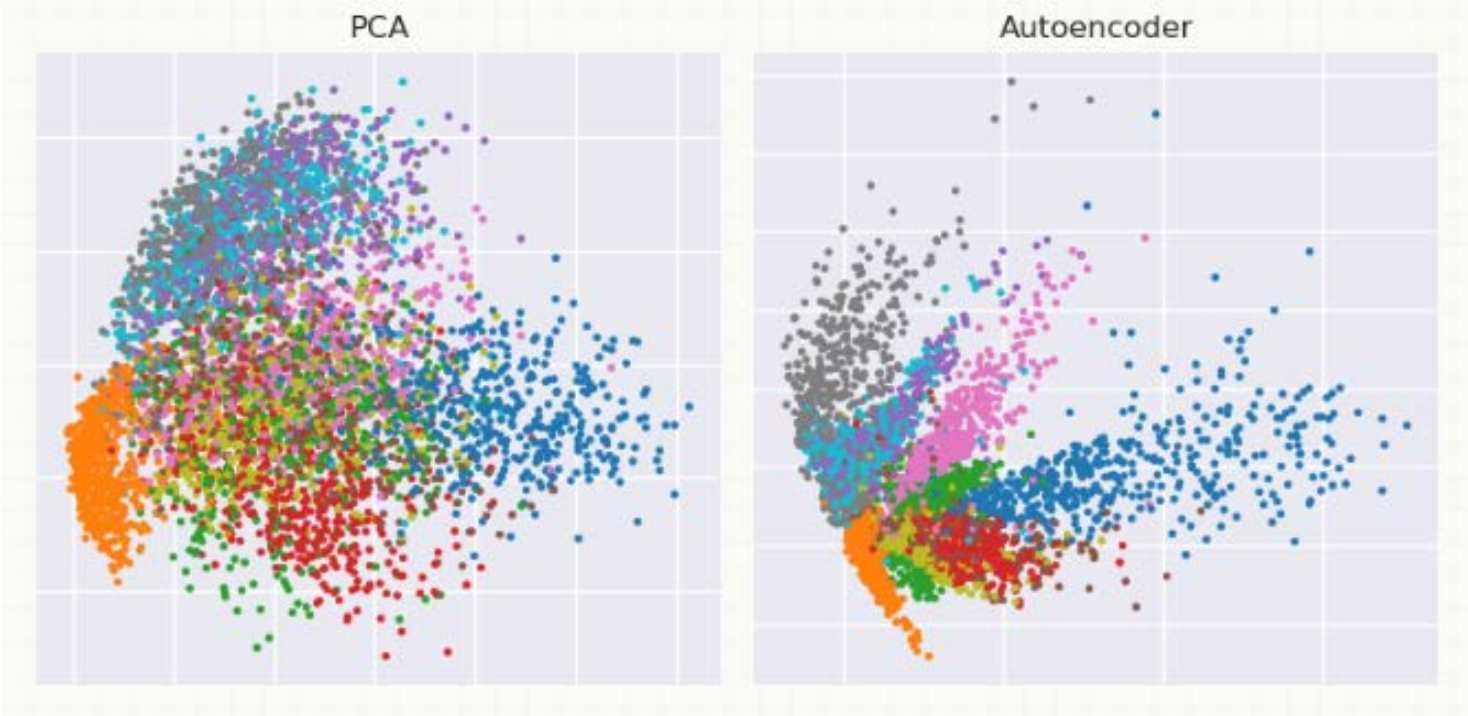


수주 예측 알고리즘

4) Dimension reduction by AutoEncoder (vs PCA)

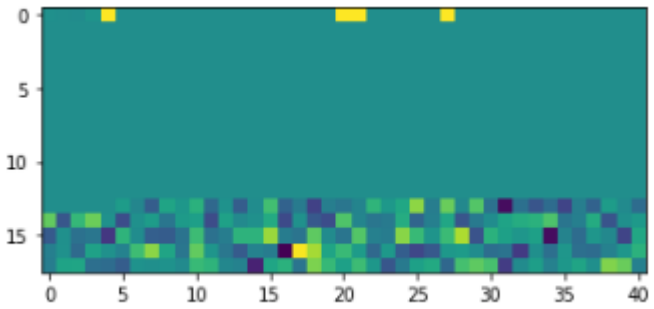
흔히 사용되는 차원감소법으로는 PCA가 있지만, 딥러닝 기법으로 사용되는 AutoEncoder가 정보를 압축하는 것에 더 효과적이므로 PCA가 아니라 AutoEncoder 사용

738 차원의 데이터를 50차원의 데이터와 100차원의 데이터로 각각 압축

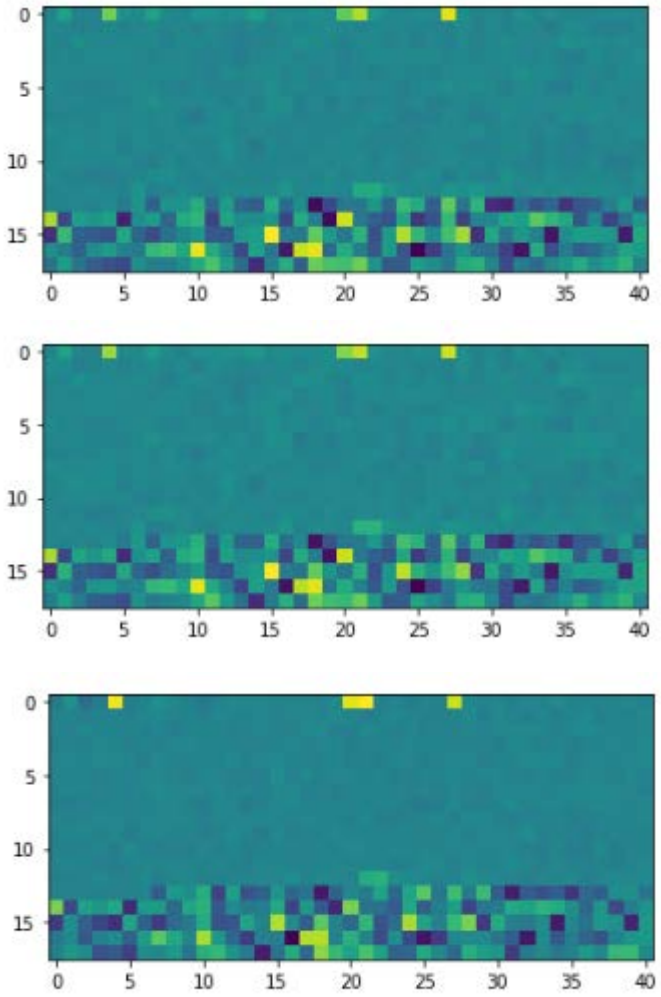


수주 예측 알고리즘

4) Dimension reduction by AutoEncoder (vs PCA)



원본 데이터



학습진행

N=100

N=1000

N=2000

수주 예측 알고리즘

4) Dimension reduction by AutoEncoder (vs PCA)

	company	success	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	3
0	농림수산식품교육문화정보원	0	0.014481	0	-0.013918	0.041313	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	
1	농림수산식품교육문화정보원	0	-0.167399	0	-0.013918	0.041313	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	
2	디디비코리아(DDB korea)	0	0.000000	0	-0.013918	0.041313	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	
3	산림청	0	-0.203775	0	-0.006795	0.049534	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	
4	건강보험심사평가원	0	-0.029170	1	-0.011275	-0.283299	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	
5	HS AD	1	0.000000	0	-0.013918	0.041313	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	
6	과학기술일자리진흥원	0	-0.160124	0	-0.012669	0.035681	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	
7	서울특별시	0	-0.131023	0	-0.018359	0.025978	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	
8	신세계조선호텔	1	-0.203775	1	-0.145318	-0.181467	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	
9	액트투 테크놀로지스	1	-0.339093	0	-0.013918	0.041313	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	

수주 예측 알고리즘

4) Dimension reduction by AutoEncoder (vs PCA)

	company	success	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
0	농림수산식품교육문화정보원	0	0.000000	0.0	1.324221	0.0	1.840877	0.0	0.000000	1.096918	0.012100	1.280952	0.0	1.535885	1.113207	0.0	0.0	0.0	0.139537	1.
1	농림수산식품교육문화정보원	0	0.000000	0.0	0.891234	0.0	1.041974	0.0	0.000000	1.065878	0.000000	1.046751	0.0	1.316150	1.437465	0.0	0.0	0.0	0.379287	0.
2	디디비코리아(DDB korea)	0	0.000000	0.0	0.433915	0.0	0.258729	0.0	0.000000	0.242707	0.000000	0.733972	0.0	0.906406	0.497926	0.0	0.0	0.0	0.342969	0.
3	산림청	0	0.000000	0.0	1.863061	0.0	2.357804	0.0	0.116938	1.473594	0.000000	1.755706	0.0	2.308227	2.103565	0.0	0.0	0.0	0.849136	1.
4	건강보험심사평가원	0	0.000000	0.0	2.258312	0.0	2.111263	0.0	0.000000	1.995230	0.000000	1.606607	0.0	2.507540	1.827975	0.0	0.0	0.0	0.532287	1.
5	HS AD	1	0.000000	0.0	0.886202	0.0	0.527191	0.0	0.000000	0.289491	0.000000	0.906178	0.0	1.157479	0.666918	0.0	0.0	0.0	0.387662	1.
6	과학기술일자리진흥원	0	0.144345	0.0	2.195486	0.0	2.114045	0.0	0.038365	1.935194	0.000000	1.853759	0.0	2.845155	2.788848	0.0	0.0	0.0	0.786304	1.
7	서울특별시	0	0.000000	0.0	1.623837	0.0	1.591860	0.0	0.050852	1.351055	0.000000	1.816893	0.0	2.167738	2.377781	0.0	0.0	0.0	0.633916	1.
8	신세계조선호텔	1	0.000000	0.0	1.213029	0.0	0.942216	0.0	0.000000	1.433594	0.019104	0.632248	0.0	2.525195	2.098996	0.0	0.0	0.0	0.461764	1.
9	액트투 테크놀로지스	1	0.020583	0.0	1.268462	0.0	0.990949	0.0	0.000000	1.344337	0.015537	1.015698	0.0	2.066495	1.865912	0.0	0.0	0.0	0.088058	1.

수주 예측 알고리즘

5) Classification

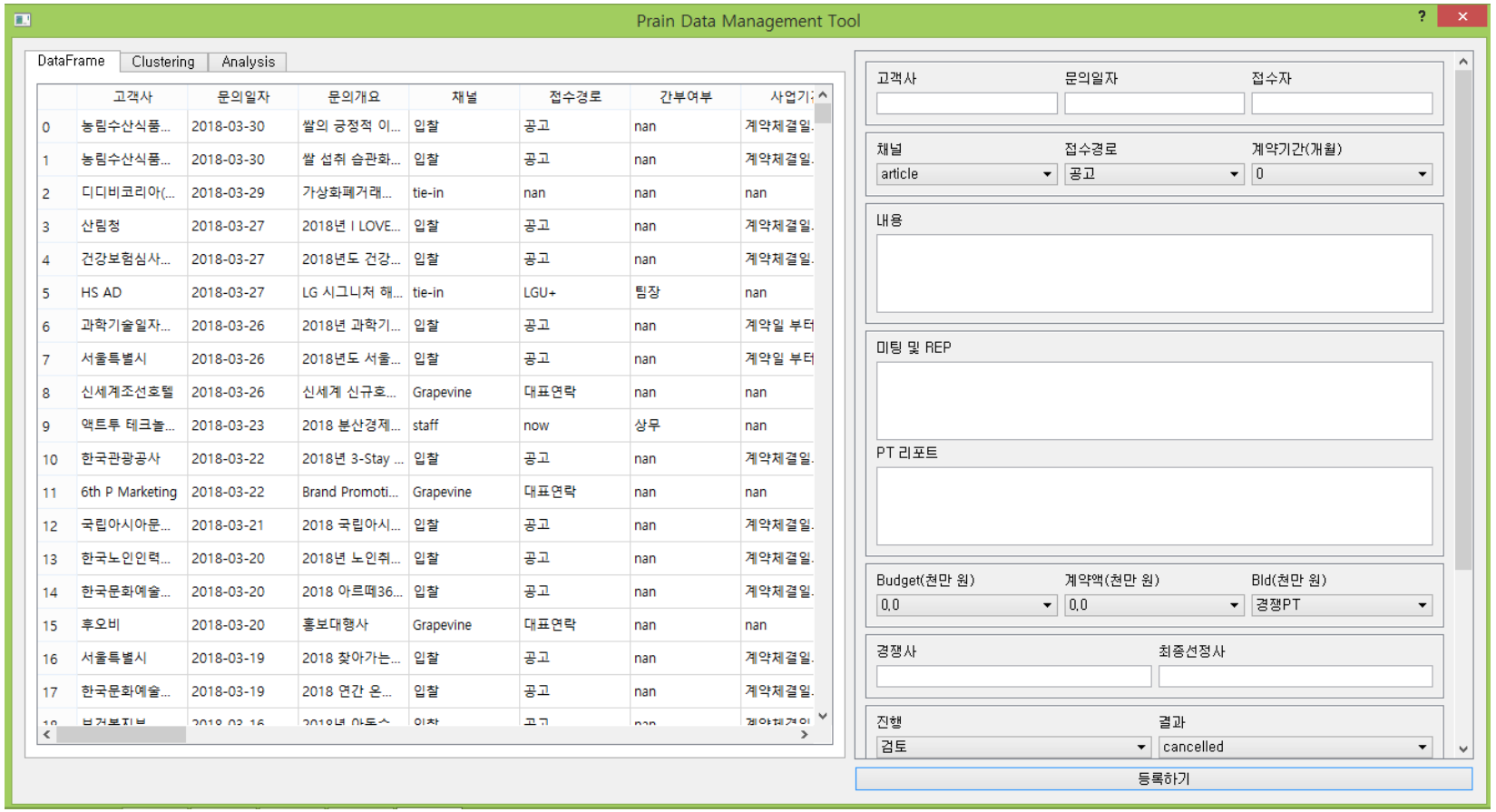
Encoding된 vector를 이용하여 ml 및 dl 모델 적용
(grid 5-fold cv 를 통한 parameter 조정)

Logistic : 0.764
Random forest : 0.753
Xgboost : 0.760
MLP : 0.826

수주 예측 알고리즘

6) 어플리케이션

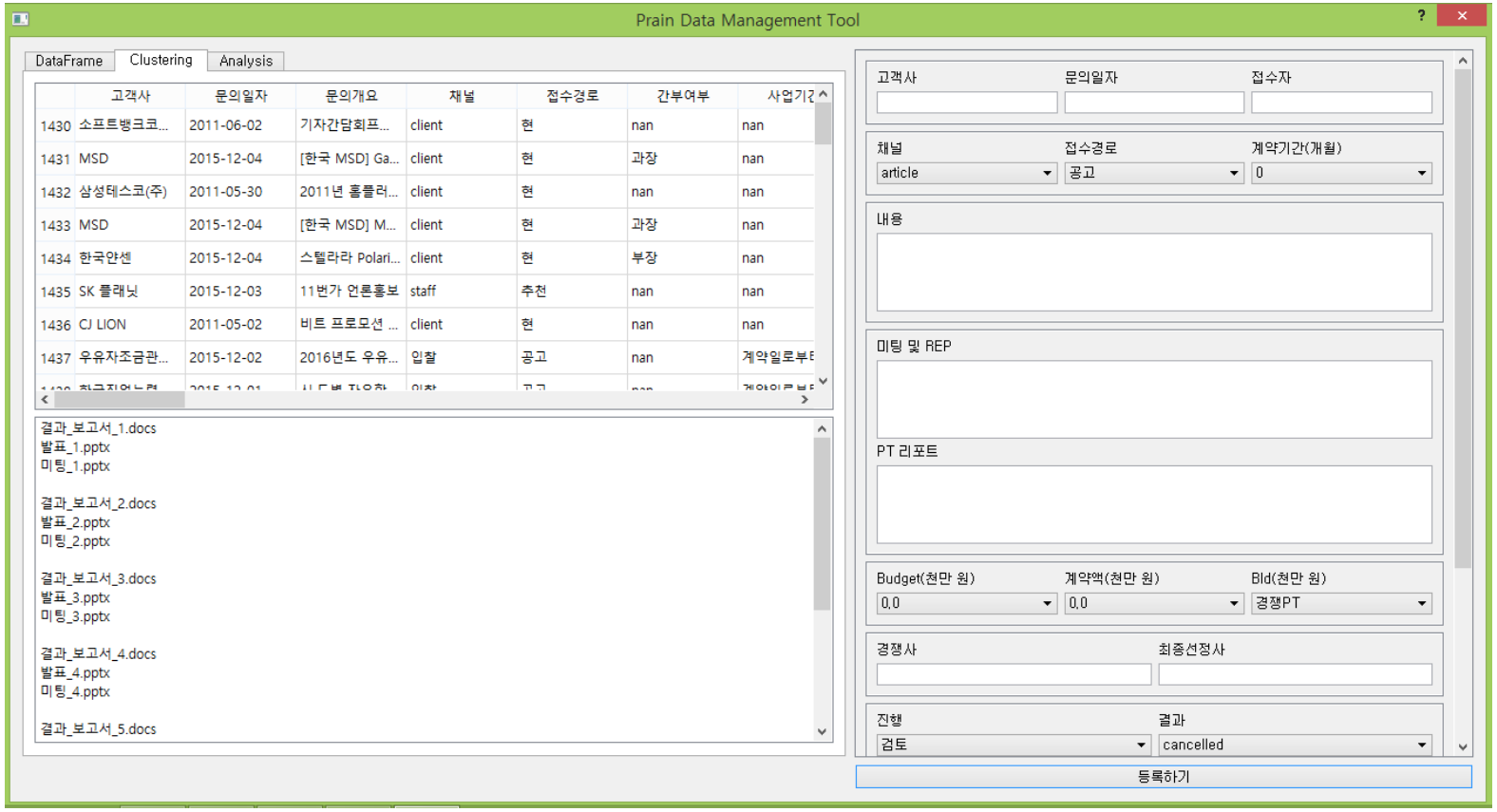
UI – dataframe : 새로운 데이터를 등록할 수 있는 화면



수주 예측 알고리즘

6) 어플리케이션

UI – clustering : 새로 등록된 데이터와 같은 클러스터에 있는 이전의 공고, 즉, 비슷한 공고를 추천해주는 화면. 클러스터에 해당하는 보고서, ppt 등을 한 번에 볼 수 있고 관리할 수 있음



수주 예측 알고리즘

6) 어플리케이션

UI – analysis : 새로 등록한 데이터의 수주확률과 분석 결과를 보여주는 화면

수주 예측 알고리즘

6) 어플리케이션

Program Architecture

‘정리된’ 데이터를 등록
-> 맞춤법 검사 -> 명사화 -> rep2vec 생성 -> AutoEncoder로 압축
-> Clustering 및 수주확률 예측
-> 분석결과 제시

7

결론 및 한계점

프로젝트를 진행하면서 알게 된 결론 및 한계점을 소개합니다.

결론 및 한계점

결론

- 경쟁 PT에서 성공의 비율이 증가, 고객사의 Budget상승을 통해 PR 시장의 가치 상승 예상 가능
- Grapevine의 성공률 증가를 위해 inbound로 걸려오는 문의를 전담하는 업무 프로세스가 필요
- 고객사가 제출한 접수내용에 대한 텍스트 분석을 통해 고객사가 사용한 단어들과 관련성 분석
- 텍스트 데이터를 통해 수주 예측 가능성 제고

한계점

- 데이터 수집 변수 추가 : 고객사 정보, 수주 성공실패 시의 피드백
- 실제 고객사에 제출한 자료 부재 : STM 모델링, 수주 예측 모델링에서 활용 가능
- 수주 예측 모델 보완점 :수주예측 모델에서 실제 FP, TN를 통해 한계점을 언급!!
- 고객사 관련 크롤링을 하려했으나 시간 부족으로 못함

For Prain !

고객사 그룹화

고객사 DB 확보

피드백 정보
공유 강화

감사합니다