2/19(수요일) TFIDF, R크롤링

2020년 2월 19일 수요일 오전 9:15

산점도 · TFIDF

TF * IDF

$\frac{n}{1+DF}$ ← 24 총 / ← 단어등장

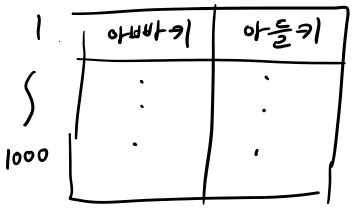
분산

$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \Rightarrow \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})$

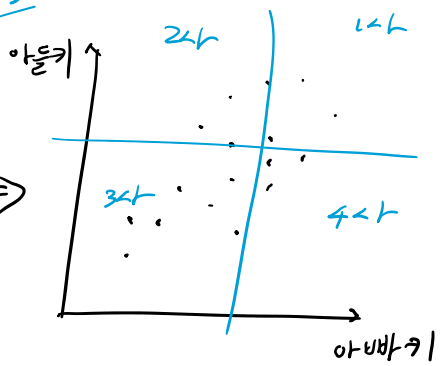
정사각형면적

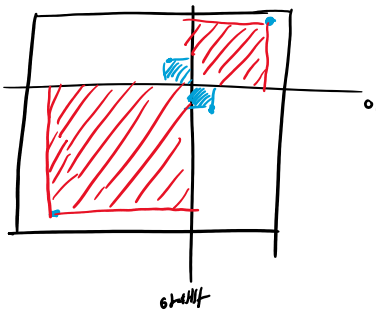
아빠, 아들 $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

공분산



⇒ 산점도 ⇒





단위 : cm or m ⇒ 표준화

두 변수 표준화 ⇒ 공분산 ⇒ 상관계수(-1~1)

아빠키, 아들키를 평균, 표준편차를 이용하여 표준화

각 $\frac{data - 평균}{표준편차}$

표준화된 아빠키 : $x' = \frac{(x_i-\bar{x})}{s_x}$

〃 아들키 : $y' = \frac{(y_i-\bar{y})}{s_y}$

$q_{x'y'} = \frac{1}{1000-1}\sum_{1000-1}^{1000}(x'-\bar{x})(y'-\bar{y})$

$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\cdot\frac{(y_i-\bar{y})}{s_y} = \frac{q_{xy}}{s_x s_y}$ $\quad -1 \le r_{xy} \le 1$

상관계수

- 모수와 비모수 (평균, 분산)
  - 모수적 : 알려진 확률분포를 기반으로 모수를 추정
    - 모델 / 정규분산

- 다수서열회귀모델 : x, y 사이의 관계가 직선으로 표현

모델 정규분산

• 단순선형회귀모델 : X, Y 사이의 관계가 직선으로 표현
정규분포를 가짐 $y = Wx + b$

• 비모수적 모델 : KNN 확률 분포를 따지지 않음

상관계수

① 연속형 & 연속형(모수적) – 피어슨
- 두 변수가 모두 정규분포를 따른다는 가정
- -1 ~ 1, 0에 가까우면 무상관
1 or -1 가까우면 상관

② 연속형 & 연속형(비모수적) – 켄달(순위), 스피어만(순위)
순서형도 가능
- 두 변수가 정규성을 따르지 않는 경우
- 스피어만 상관계수는 순위를 내어 상관계수를 추정

• 켄달 상관계수: 두 변수 순위 비교 ⇒ 연관성
concordant pair

5명 사람	사람 A	C	B	D	C	D	E
키	1	<	2	3	4	5	
몸무	3	<	4	1	2	5	

ex) 켄달 상관계수가 1인 경우

키	1	2	3	4	5
몸	1	2	3	4	5

켄달 -1인 경우

키	1	2	3	4	5
몸	5	4	3	2	1

ex) 6 4

$$\text{켄달 상관계수} = \frac{C - D}{C + D}$$

C: concordant pair
D: concordant pair가 아닌 수
C + D = 10

$\frac{2}{10} = 0.2$ 상관관계 낮다.

```r
install.packages("rvest")#r크롤링위해
library(rvest)
library(dplyr) #require(패키지명)

#문제
#네이버or신문사or포털 등
#기사 추출->tfidf 구성->상관계수
#기사 키워드 추출
#경제신문 ex)오늘, 어제, 그제, ...
#웹 스크래핑

#1. 포털, 신문사, 언론사,...
#TEXT -> DTM -> TFIDF 구성 -> 상관계수

#2.한글 논문 문서
#DTM => TFIDF 구성
#TF는 높고, IDF가 낮은 단어 추출

#doc1~doc10
#corpus->dtm->tfdif->cor

#read_html(),
#html_node()/html_nodes(),
#html_text()
```

) 크롤링위한 기능

```r
#tv.naver.com/jtbc.youth
url_tvcast<-"http://tv.naver.com/jtbc.youth"
html_tvcast<-read_html(url_tvcast,encoding = "UTF-8")
#class가 title인 부분 안에 있는 a태그의 내용 추출
html_tvcast %>% html_nodes(".title a")
#html_node():매칭된 요소 하나 추출
#html_nodes():모든 요소 추출(class, tag)
html_tvcast %>%
  html_nodes(".title a") %>%
  html_text()
tvcast_res<-html_tvcast %>%  # 여기 코드 다시
  html_nodes(".title a") %>%
  html_text()
tvcast_res
tvcast_df<-html_tvcast %>%  # 여기 코드 다시확인
  html_nodes(".title a") %>%
  html_text() %>%
  data.frame()
str(tvcast_df)
```

```
> #html_nodes():모든 요소 추출(class, tag)
> html_tvcast %>%
+   html_nodes(".title a") %>%
+   html_text()
 [1] "[워맨스 스페셜]나도 이런 연ㅇ..아니 우정 하고 싶다..."
 [2] "[워맨스 스페셜]나도 이런 연ㅇ..아니 우정 하고 싶다..."
 [3] "[메이킹] '똥차' 고두영 시원하게 때려잡는 하메들!"
 [4] "[메이킹] 자그마치 윤선배가 사온 어묵이란 말입니다..!!"
 [5] "[청춘어록] 이 시대의 청춘들에게 바칩니다.. 괜찮다고"
 [6] "'여행.. 같이 가자고는 안 하네(삐죽)' 윤선배, 알고 보니 밀당고수!?"
 [7] "선배 왜 웃어요'-'? 난 너만 보면 웃음이 나~^_^♥ (흑심 폴폴)"
 [8] "혜수를 위한 '선의의 거짓말' 이런 하메들 어디 있나요?"
 [9] "죽으려고 했는데.. 아무리 생각해도 죽고 싶지가 않아요"
[10] "사라진 박혜수, 같이 살면서 어떤 기분인지 몰랐어.. 속상해"
```

```r
#https://en.wikipedia.org/wiki/Student%27s_t-distribution
url_t<-"https://en.wikipedia.org/wiki/Student%27s_t-distribution"
html_t<-read_html(url_t,encoding = "UTF-8")
#class가 title인 부분 안에 있는 a태그의 내용 추출
html_t %>% html_nodes(".wikitable") %>%
  html_table()

#html_text():텍스트 추출
#html_name():attribute 명을 추출
#html_childer():하위 요소 추출
#html_tag():태그명 추출
#html_attrs:속성을 추출
```

```r
my.text.location<-"C:/JMOh/refer_data/ymbaek_refer"
mypaper<-VCorpus(DirSource(my.text.location))
mypaper
mykorean<-mypaper[[19]]$content
#전처리
library(stringr)
mykorean
mytext<-str_replace_all(mykorean,"[[:lower:]]","")
mytext
mytext<-str_replace_all(mytext,"\\(","")
mytext<-str_replace_all(mytext,"\\)","")
mytext<-str_replace_all(mytext,"’","")
mytext<-str_replace_all(mytext,"‘","")
mytext<-str_replace_all(mytext," · ","")
removePunctuation(mytext)
mytext
```

```
> mytext
[1] "본 논문의 목적은 언론학 교육과정 개선논의의 등장배경과 역사를 서술하고 그 필요성과 방향성을 제시하는 것이다. 본 논문에서는 컴퓨터 프로그래밍 언어와 데이터 수집관리분석재현과 같은 데이터 과학관련 지식과 기술의 필요성을 강조하며 이를 언론학의 교과과정에 첨가융합해야 한다는 주장의 등장배경과 필요성을 전반적으로 개괄소개하였다. 또한 제도주의 이론에  언론학 교과과정을 제도로 파악하였으며, 제도를 둘러싼 이해관계자들로 교수진, 학생, 학부모를 비롯한 일반인, 다른 학문분과들을 논의하였다. 이를 통해 기존의 교과과정과 새로운 교과과정이 특정 집단내부 혹은 집단 간 갈등을 일으킬 수 있으며, 이러한 갈등을 최소화시키고 협력가능성을 증대시킬 수 있는 방안을 추상적 수준에서나마 제안해 보았다."
```

```r
#명사 추출
noun.mytext<-extractNoun(mytext)
noun.mytext
table(noun.mytext)
```

```
> noun.mytext
 [1] "논문"         "목적"         "언론학"       "교육과정"     "개선"
 [6] "논의"         "등장"         "배경"         "역사"         "서술"
[11] "필요"         "성"           "방향"         "성"           "시하"
[16] "것"           "논문"         "컴퓨터"       "프로그래밍"   "언어"
[21] "데이터"       "수집"         "관리"         "분석"         "재현"
[26] "데이터"       "과학"         "관련"         "지식"         "기술"
[31] "필요"         "성"           "강조"         "이"           "언론학"
[36] "교과과정"     "첨가"         "융합"         "주장"         "등장"
[41] "배경"         "필요"         "성"           "전반"         "적"
[46] "개괄"         "소개"         "제도"         "주의"         "이론"
[51] "언론학"       "교과과정"     "제도"         "파악"         "제도"
[56] "이해관계자"   "들"           "교수진"       "학생"         "학부모"
[61] "비롯"         "한"           "일반"         "학문분과"     "들"
[66] "논의"         "이"           "기존"         "교과과정"     "교과과정"
[71] "특정"         "집단"         "내부"         "집단"         "등"
[76] "수"           "등"           "최소"         "화시"         "키"
[81] "협력가능성"   "증대"         "수"           "방안"         "추상"
[86] "적"           "수준"         "제안"         "해"
> table(noun.mytext)
noun.mytext
      강조       개괄       개선         것       과학       관련       관리
         1          1          1          1          1          1          1
  교과과정     교수진   교육과정       기술       기존       내부       논문
         4          1          1          1          1          1          2
      논의     데이터         들         등       등장       목적       방안
         2          2          2          2          2          1          1
      방향       배경       분석       비롯       서술         성       소개
         1          2          1          1          1          4          1
        소        소중        소집        시하      언론학         언어         역사
```