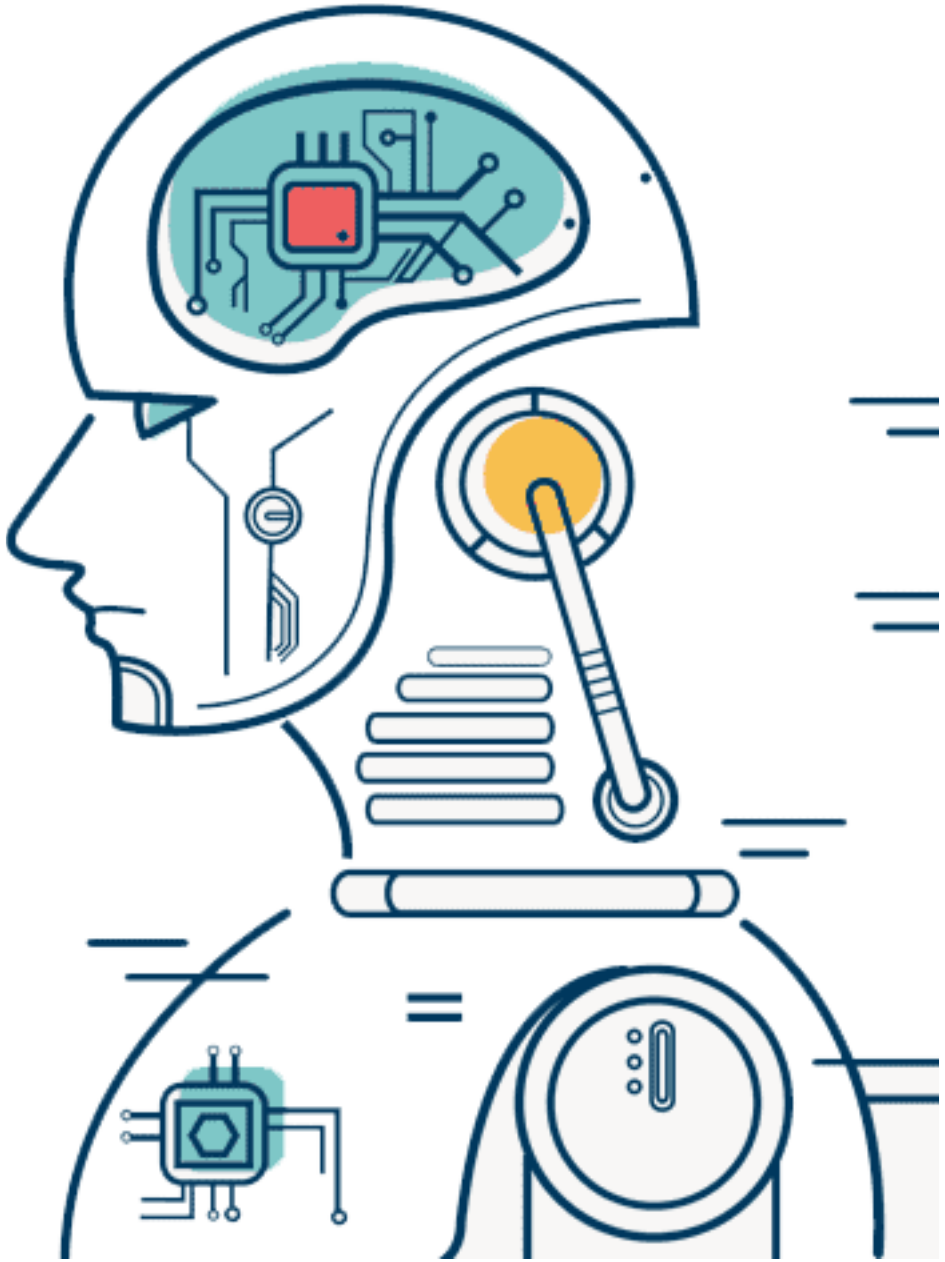
스마트인재개발원
Smart Human Resources Development

Machine Learning

Chapter_8 텍스트마이닝
(Text Mining process, BOW, TF-IDF)



김은영

학습목표

- 텍스트 마이닝의 개념과 프로세스를 이해한다.
- 텍스트 데이터를 이용해 머신러닝 학습을 수행할 수 있다.


DATA
MINING
PATTERNS
SET
LEARNING
CUSTOMERS
ANALYSIS
MACHINE
BUSINESS
CUSTOMER
DECISION
PEOPLE
METHODS
INTELLIGENCE
PROCESS
EXAMPLE
LEARNED
OFFERS
INFORMATION
DATABASE
ALGORITHMS
OFFER
IDENTIFY
VISUALIZATION
E-MAIL
SOPHISTICATED
ARTIFICIAL
STRUCTURES
RULE
CITATION
NEW
SEND
WITHOUT
MIGHT
USE
STEP
GROUPS
TASK
DESIRED
APPLICATIONS
PREDICT
LIKELY
DETERMINE
MODEL
TARGET
PRODUCT
PRESENT
USING
TRAINED
DISCOVERY
SYSTEM
MANY
COMMON
NUMBER
HELP
PRE-PROCESSING
INVOLVES
ASSOCIATION
NEEDED
MODELING
KNOWLEDGE
WITHIN
LARGE
TEST
MARKET
E-MAILS
BASKET
RECORDS
MANAGEMENT
ALGORITHM
MODELS
TERM

Text Mining

텍스트 마이닝이란?

- 텍스트 데이터로부터 유용한 인사이트를 발굴하는 Data Mining 의 한 종류
- 정형 및 비정형 데이터를 자연어 처리방식(Natural Language Processing)과 문서처리 방법을 적용하여 유용한 정보를 추출/가공하는 것을 목적으로 하는 기술

* Data Mining : 빅데이터 안에서 규칙이나 패턴을 분석하여 가치 있는 정보를 추출하는 과정
* 자연어(natural language) 처리 : 자연어의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 과정

자연어(Natural language)란?

- 인간이 일상생활에서 사용하는 언어
- 인간이 정보를 전달하는 수단
- 특정 집단에서 사용되는 모국어의 집합(한국어, 영어, 일본어, 중국어 등)
- 인공언어와 대비되는 개념

* 인공언어?
  - 특정한 법칙들에 따라 적절하게 구성된 문자열들의 집합
  - 특정 목적을 위해 인위적, 의도적으로 만든 언어
  - 프로그래밍 언어(python, java 등), 형식언어(수학식)

자연어 처리 응용 분야

- 인간의 언어가 사용되는 실세계의 모든 영역
- 정보검색, 질의응답 시스템
  - Google, Naver, iphone siri, 갤럭시 bixby, IBM Waston
- 기계번역, 자동통역
  - Google 번역기, Naver Papago, ETRI 지니톡
- 문서작성, 문서요약, 문서 분류, 철자 오류 검색 및 수정, 문법 오류 검사 및 수정

Text Mining – 기업의 활용 사례

스마트인재개발원
Smart Human Resources Development

1. 지식경영
- 많은 양의 데이터 중 의미있는 데이터만 뽑아내어 효율적으로 관리

2. 사이버 범죄 예방
- 텍스트 마이닝을 이용한 범죄 예측, 예방 어플리케이션 등

3. 고객 관리 서비스
- 고객에게 빠르고 자동화된 응답을 제공하기 위해 활용 (챗봇)

4. 고객 클레임 분석을 통한 부정행위 탐지
- 보험회사는 텍스트 마이닝을 통해 사기를 방지하고 빠르게 클레임을 처리

5. 콘텐츠 강화
- 다양한 목적에 따라 그에 적합한 내용으로 정리하고 요약

6. 소셜 미디어 데이터 분석
- 해당 브랜드나 제품에 대한 다양한 의견과 감성 반응을 살펴봄

Text Mining – 기업의 활용 사례 : 지식경영

스마트인재개발원
Smart Human Resources Development

기관자금 강자 김홍곤 “텍스트마이닝 투자에 접목”

2019.01.28 10:31

AI기반 텍스트마이닝 단계 도입
투자자산 선정, 수익률 예측
로보어드바이저 · AI 활용 펀드 출시

김 상무는 “(AI가)리포트나 공시를 읽어내 특정 종목의 주가 상황과 상승여력을 알려주고 특정 종목을 얼마나 매수 · 매도하면 좋은지 판단해주는 ‘과학적(scientific) 투자’ 기반을 마련할 것”이라면서 “인공신경망과 유전자 알고리즘을 활용해 수익률이 개선되는 것을 확인했다”고 설명했다.

용한 정보를 찾아내는 기술이다.

김 상무는 “(AI가)리포트나 공시를 읽어내 특정 종목의 주가 상황과 상승여력을 알려주고 특정 종목을 얼마나 매수 · 매도하면 좋은지 판단해주는 ‘과학적(scientific) 투자’ 기반을 마련할 것”이라면서 “인공신경망과 유전자 알고리즘을 활용해 수익률이 개선되는 것을 확인했다”고 설명했다.

Text Mining – 기업의 활용 사례 : 고객 관리 서비스

스마트인재개발원
Smart Human Resources Development

Text Mining – 기업의 활용 사례 : 콘텐츠 요약

스마트인재개발원
Smart Human Resources Development

텍스트 마이닝 기술을 활용한 농산물 주문 정리 시스템

→ **비정형 주문 메시지를 간단하게 편집, 저장**하여 농가의 주문 처리 시간을 단축

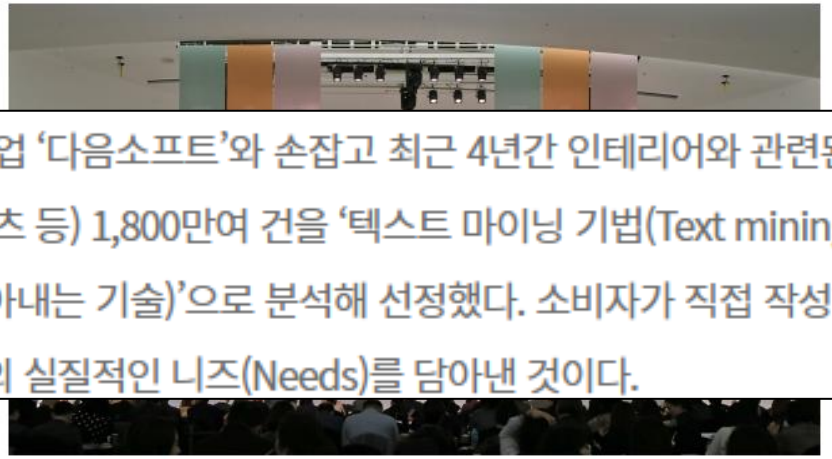
ex) 실제 사용자가 주문 내용을 복사해서 붙여넣기만 하면 일정한 형식의 주문서가 자동으로 생성되는 방식

Text Mining – 기업의 활용 사례 : 소셜 미디어 데이터 분석

스마트인재개발원
Smart Human Resources Development

현대L&C, '텍스트 마이닝 기법' 통해 2020년 인테리어 트렌드 공개

2019.11.05 10:58 by 유선이



특히, 올해는 빅데이터 전문기업 '다음소프트'와 손잡고 최근 4년간 인테리어와 관련된 소비 트렌드 자료(인테리어 전문 블로그 및 커뮤니티 콘텐츠 등) 1,800만여 건을 '텍스트 마이닝 기법(Text mining, AI를 이용해 문자로 구성된 데이터에서 유용한 정보를 찾아내는 기술)'으로 분석해 선정했다. 소비자가 직접 작성한 인테리어 관련 콘텐츠들의 연관 관계를 분석해 소비자들의 실질적인 니즈(Needs)를 담아낸 것이다.

사진=현대L&C

현대L&C는 지난 4일 서울 동대문 디자인 플라자에서 건설·인테리어업계 관계자들을 초청해 인테리어 트렌드 세미나 '인트렌드(Intrend) 2020·2021'를 열고, 내년 인테리어 트렌드를 공개했다고 5일 밝혔다.

올해로 10회째를 맞는 '인트렌드'는 현대L&C가 다음해 인테리어 트렌드를 제안하는 세미나로, 건자재 업계에선 유일하게 지난 2017년부터 빅데이터를 분석해 인테리어 트렌드를 제시하고 있다.

Text Mining 영역

텍스트 분류
(Text Classification)

감성 분석
(Sentiment Analysis)

텍스트 요약
(Summarization)

텍스트 군집화 및
유사도 분석
(Clustering)

스탠포드 대학의 앤드류 마스가 수집한 영화 리뷰 텍스트 분석

http://ai.stanford.edu/~amaas/data/sentiment/

Text Mining – Process : 텍스트 데이터의 구조

스마트인재개발원
Smart Human Resources Development

말뭉치 >> 문서 >> 문단 >> 문장 >> 단어 >> 형태소

- **말뭉치(corpus)** : 분석을 위해 수집된 **문서**들의 집합
- 말뭉치는 **여러 개의 문서**로 구성
- **문서**는 **여러 개의 문단**으로 구성
- **문단**은 **여러 개의 문장**으로 구성
- **문장**은 **여러 개의 단어**로 구성
- **단어**는 여러 개의 **형태소**로 구성

형태소 : 일정한 의미가 있는 가장 작은 말의 단위
ex) 첫사랑 → 첫, 사랑
ex) 애늙은이 → 애, 늙은이

Text Mining – Process : 분석 프로세스

- 텍스트 데이터 수집 : SNS/뉴스/블로그 등 텍스트 데이터 수집
- 텍스트 전처리 : 컴퓨터가 이해하기 쉽게 텍스트를 변환하는 과정
- 토큰화 : 단어단위로 나누는 과정
- 특징 값 추출 : 중요한 단어를 선별하는 과정
- 데이터 분석 : 머신러닝, 딥러닝 등 분석 모델 사용

Text Mining – Process : 분석 프로세스

스마트인재개발원
Smart Human Resources Development

- Crawling을 이용한 Web 데이터 수집(SNS/블로그/카페 등)
- 빅카인즈(BIG Kinds) 뉴스 데이터 제공 사이트
- NDSL(www.ndsl.kr) : 국내외 논문, 특허, 연구보고서 통합 정보제공 사이트

Text Mining – Process : 분석 프로세스

스마트인재개발원 Smart Human Resources Development

- 전처리는 용도에 맞게 텍스트를 사전에 처리하는 작업
- 궁극적으로 '중요한 특징 값'을 선택하는 것이 중요
- 오탈자 제거, 띄어쓰기 교정
- 불용어 제거 : 데이터에서 큰 의미가 없는 단어 제거 (ex 음, 뭐, 아 등)
- 정제(cleaning) : 가지고 있는 코퍼스(corpus, 말뭉치)로부터 노이즈 데이터를 제거

  corpus : 수집된 문서들의 집합, noise data : 등장 빈도가 적은 데이터, 의미 없는 특수문자 등
- 정규화(normalization) : 표현 방법이 다른 단어들을 통합하여 같은 단어로 만듦
- 어간 추출 (Stemming) : 단어의 핵심 부분(어간)만 추출 (ex 먹다, 먹고, 먹지, 먹어서 등)
- 표제어 추출 (Lemmatization) : 유사한 단어들에서 대표 단어를 추출 (am, are, is는 be동사)

Text Mining – Process : 분석 프로세스

- 토큰화 (tokenization) : 학습을 위해서 주어진 코퍼스(corpus, 말뭉치)에서 토큰(token)이라 불리는 단위로 나누는 작업(공백기준, 형태소기준, 명사기준, 단어기준 등)

  말뭉치 : 수집된 문서들의 집합
- 단위의 기준은 분석 방법에 따라 다름
- 감성 분석을 한다면, 한글에서 감성을 나타내는 품사가 동사, 형용사 쪽에 가깝기 때문에 한글 형태소 분석기를 사용해서 동사, 형용사 위주로 추출

Text Mining – Process : 분석 프로세스

스마트인재개발원
Smart Human Resources Development

- '중요한 단어'를 선별하는 과정
- '중요한 단어'로서의 특징은 적은 수의 문서에 분포되어 있어야 하고, 문서 내에서도 빈번하게 출현해야 함(TF-IDF)
- 특정 텍스트를 통해 문서를 구분 짓는 것이기 때문에 어떤 단어가 모든 문서에 분포되어 있다면 개수가 많더라도 차별성 없는 일반적인 단어(ex - a, the)

Text Mining - Process : 분석 프로세스

스마트인재개발원
Smart Human Resources Development

- 머신러닝
  - Linear Regression
  - Logistic Regression
  - Random Forest
  - XGBoost

- 딥러닝
  - CNN
  - RNN
  - LSTM
  - GRU

토큰화(tokenize)의 종류

- 단어(word) 단위 : **텍스트를 단어로 나누고 각 단어를 하나의 벡터로 변환**
- 문자(character) 단위 : **텍스트를 문자로 나누고 각 문자를 하나의 벡터로 변환**
- n-gram 단위: **텍스트에서 단어나 문자의 n-gram을 추출하여 n-gram을 하나의 벡터로 변환**

n-gram 단위

- n개의 연속된 단어를 하나로 취급하는 방법

- 예를 들어 “춘천 마라톤”라는 표현을 “춘천”와 “마라톤” 두 개의 독립된 단어로만 취급하지 않고 두 단어로 구성된 하나의 토큰으로 취급

  - n=2 경우를 bi-gram이라고도 부름
  - 단어의 개수가 늘어난 효과
  - 연속된 단어가 갖는 의미

토큰화(n-gram)

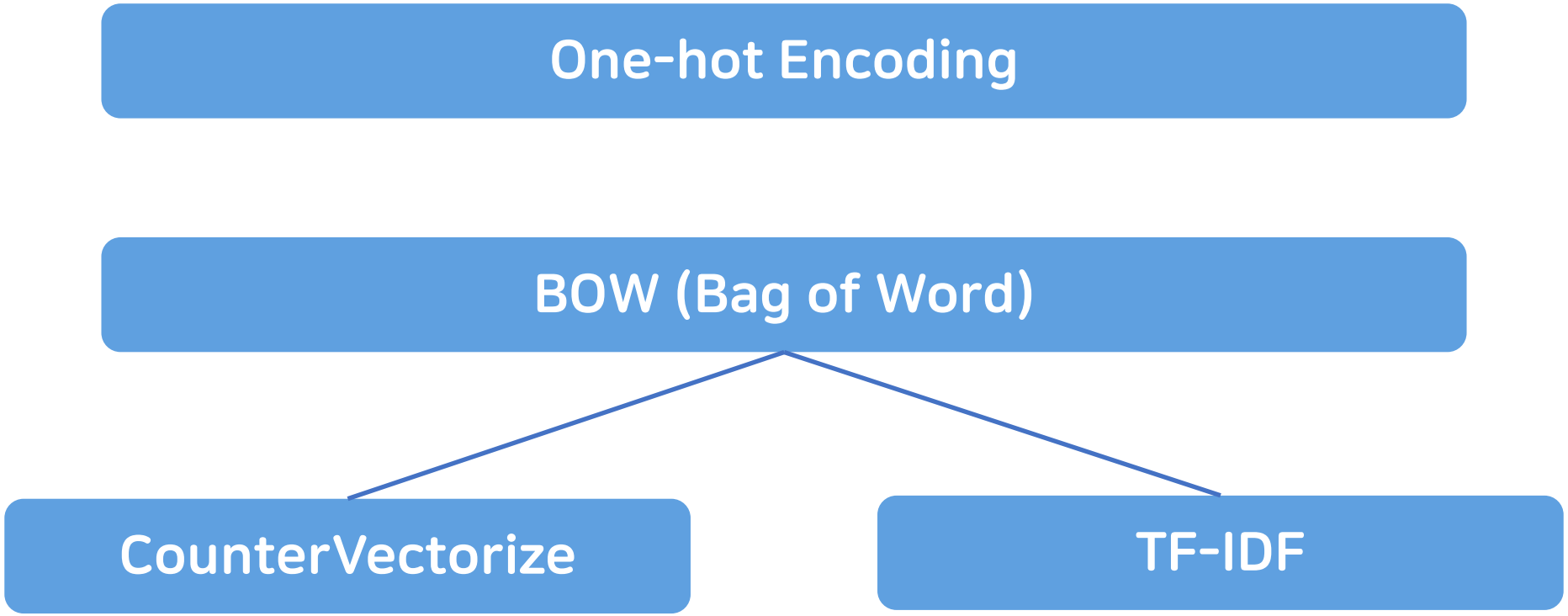
텍스트 : “어제 춘천에 갔다가 춘천 마라톤을 관람했다”

단어 토큰 : {“어제”, “춘천”, “갔다”, “마라톤”, “관람”}

2-gram 토큰 : {“어제 춘천”, 춘천 갔다“, “갔다 마라톤”, “마라톤 관람“}

토큰화(n-gram)

- n-gram을 허용하면 토큰화 대상의 수가 크게 증가
- 데이터 분석을 위한 토큰과 결과를 수치로 만드는 방법



Text Mining – Process : One-hot Encoding

- 토큰에 고유번호를 배정하고 모든 고유번호 위치의 한 컬럼만 1, 나머지 컬럼은0인 벡터로 표시하는 방법

어제 춘천에 갔다가 춘천 마라톤을 관람했다

↓

단어사전 {“어제”:0, “춘천”:1, “갔다”:2, “마라톤”:3, “관람”:4}

어제 {1,0,0,0,0}
춘천 {0,1,0,0,0}
갔다 {0,0,1,0,0}
마라톤 {0,0,0,1,0}
관람 {0,0,0,0,1}

CounterVectorize

- 단어들의 문맥이나 순서를 무시하고, 빈도수를 기반으로 벡터화 시키는 방식

단어사전

{"어제":0, "오늘":1, "강릉":2, "춘천":3, "갔다":4,
"달리기":5, "마라톤":6, … , "광주":4999}

입력문장

어제 춘천에 갔다가 춘천 마라톤을 관람했다

벡터화

0	1	2	3	4	5	6	…	4999
1	0	0	2	1	0	1	0	0

단어의 순서나 중요도를 고려하지 않기 때문에 문맥의 의미를 반영하기 힘듦

Text Mining – Process : BOW(Bag of Word, 단어모음)

스마트인재개발원 Smart Human Resources Development

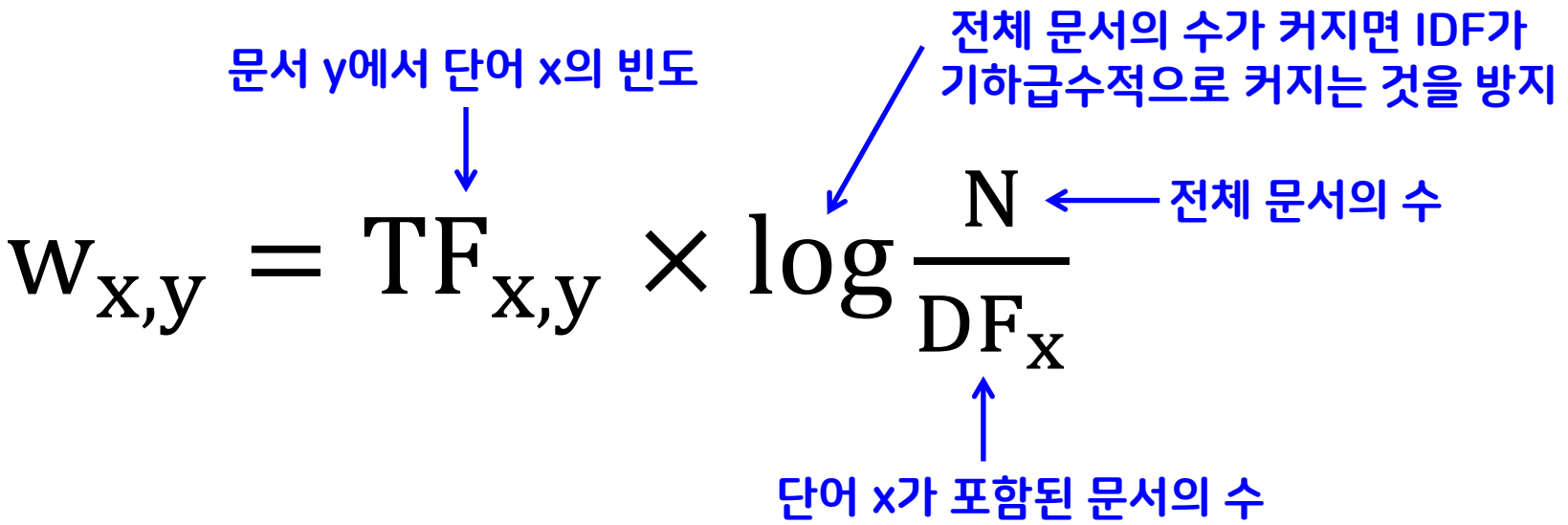
TF-IDF(term frequency-inverse document frequency)

- 개별 문서에서 자주 등장하는 단어에는 높은 가중치를 주되,
  모든 문서에 자주 등장하는 단어에는 페널티를 주는 방식(단어의 중요도를 반영)

- TF : 단어가 각 문서에서 발생한 빈도
- DF : 단어가 등장한 문서의 수

- 적은 문서에서 발견될수록 가치 있는 정보
- 많은 문서에 자주 등장하는 단어일수록 일반적인 단어 (ex 나, 그, 했다 등)
- 단어가 특정문서에만 나타나는 희소성을 반영하기 위해서 TF에 DF의 역수(IDF)
  를 곱한 값을 사용

TF-IDF (term frequency-inverse document frequency)



네이버 영화리뷰 데이터 감성 분석

네이버 영화 리뷰 데이터 분석

- 다운로드 링크 : https://github.com/e9t/nsmc/

- 총 200,000개 리뷰로 구성된 데이터로 영화 리뷰에 대한 텍스트와 해당 리뷰가 긍정인 경우 1을 부정인 경우 0으로 표시한 레이블로 구성

- 훈련 데이터에 해당하는 ratings_train.csv와 테스트 데이터에 해당하는 ratings_test.csv를 다운로드