Text Mining

Chapter 7 텍스트 마이닝 (Text mining)

START

스마트인재개발원
Smart Human Resources Development

학습 목표

- 텍스트 마이닝에 대해 알 수 있다.
- 텍스트 마이닝이 사용된 사례를 알 수 있다.
- 텍스트 마이닝의 응용 기술을 알 수 있다.
- 텍스트 구조에 대해 알 수 있다.
- 텍스트 마이닝 프로세스에 대해 알 수 있다.

텍스트 마이닝 이란?

- 텍스트 마이닝은 정형 및 비정형 데이터를 자연어 처리 방식(Natural Language Processing)과 문서처리 방법을 적용하여 유용한 정보를 추출하여 가공하는 것을 목적으로 하는 기술

- 텍스트 마이닝은 데이터로부터 유용한 인사이트를 발굴하는 데이터 마이닝(Data Mining), 언어를 정보로 변환하기 위한 자연어처리, 정보검색 등 다양한 분야가 접목되어 발전한 학문, 기술

텍스트 마이닝 사례- 챗봇

완드 9번 카드가 나왔어

오늘 명호의 매력지수는... 2점이네

오늘 바쁘다 바빠 명호

해야 할 일이 엄청 많아서 그거만 해도 시간이 훌쩍 흘러갈 거야

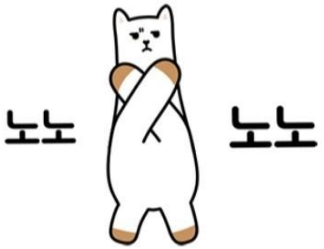
피곤하고, 정신적 여유가 없으니 매력을 보여줄 틈도 없다 😭

오후 10:51

그렇구나

게다가 너무 힘드니까 네 태도도 다소 방어적이라

호감을 가지고 다가왔던 사람도 삐빅! 명호 마음 입장 불가 당할 거야



현생이 우리 명호 매력 발산 방해하네

얼른 바쁜 거 지나가고 여유 좀 찾으면 금세 또 돌아올 테니 걱정하지 마

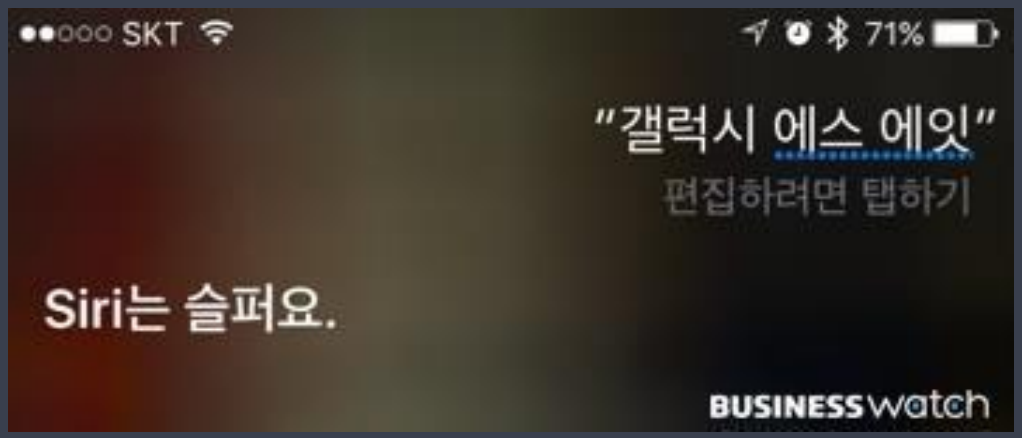
해석은 여기까지야

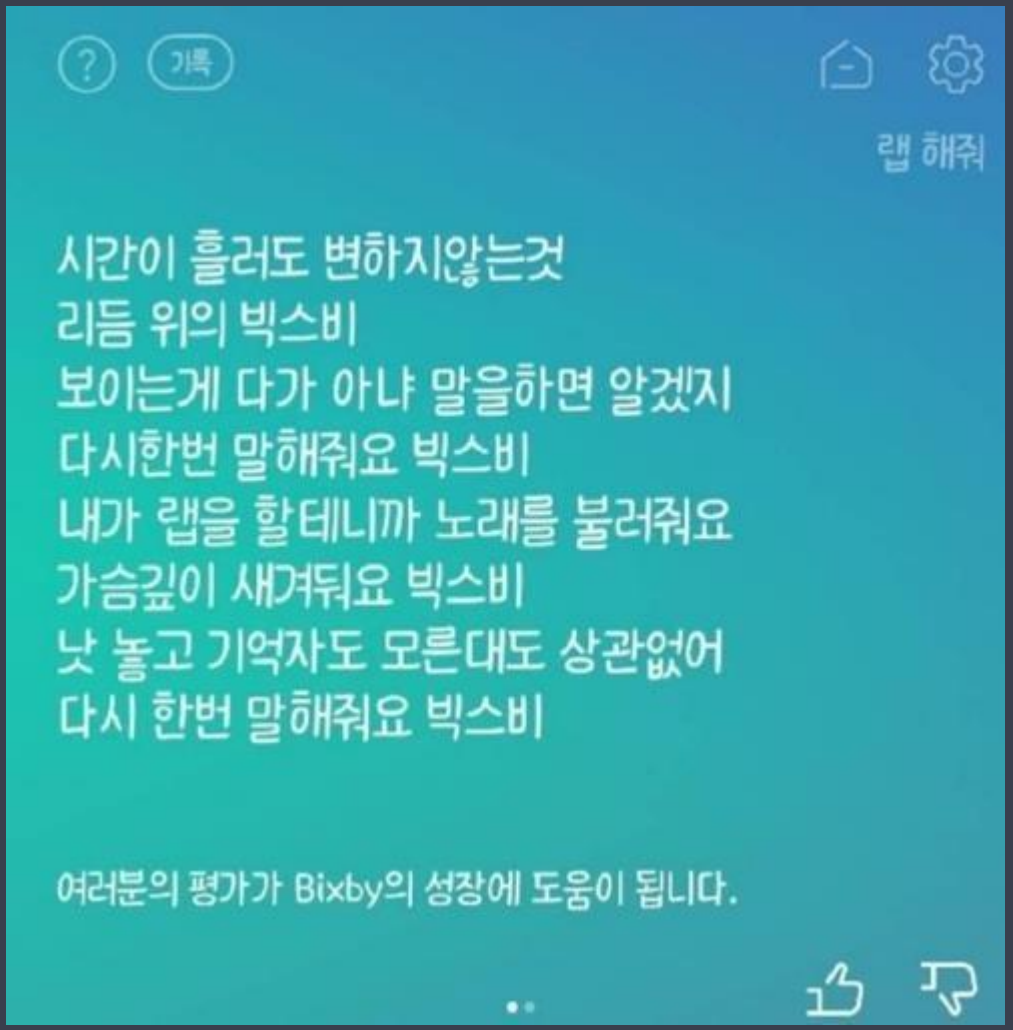
오늘의 매력지수 보고 나니 어때?

오후 10:51

💩

텍스트 마이닝 사례- 인공지능





텍스트 마이닝 사례 – 기업에서의 분석 사례

1. 지식 경영(Knowledge management)
   - 많은 양의 데이터 중 의미 있는 데이터만 뽑아내고 효율적으로 관리 할 수 있다.

2. 사이버 범죄 예방(Cybercrime prevention)
   - 텍스트 마이닝을 이용한 범죄 예방 어플리케이션 등

3. 고객 관리 서비스(Customer care service)
   - 고객에게 빠르고 자동화된 응답을 제공하기 위해 활용

텍스트 마이닝 사례 – 기업에서의 분석 사례

4. 고객 클레임 분석을 통한 부정행위 탐지(Fraud detection through claims investigation)
   - 보험회사는 텍스트 마이닝을 통해 사기를 방지하고 빠르게 클레임을 처리

5. 콘텐츠 강화(Content enrichment)
   - 다양한 목적에 따라 그에 적합한 내용으로 정리하고 요약

6. 소셜 미디어 데이터 분석(Social media data analysis)
   - 해당 브랜드나 제품에 대한 다양한 의견과 감성반응을 살펴볼 수 있다.

텍스트 마이닝 기술 영역

텍스트 분류
(Text Classification)

감성 분석
(Sentiment Analysis)

텍스트 요약
(Summarization)

텍스트 군집화 및 유사도 분석
(Clustering)

텍스트 마이닝(Text mining) - 데이터 수집

스텐포드 대학의 앤드류 마스가
수집한 영화 리뷰 텍스트 분석

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

텍스트 데이터의 구조

말뭉치 >> 문서 >> 단락 >> 문장 >> 단어 >> 형태소

말뭉치 >> 문서 >> 단어

- 말뭉치 : 분석을 위해 수집된 문서들의 집합

- 말뭉치는 여러 개의 문서가 존재
- 문서는 여러 개의 단락으로 구성
- 단락은 여러 개의 문장으로 구성
- 문장은 여러 개의 단어로 구성
- 단어 중 여러 개의 형태소로 구성

텍스트 마이닝 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

텍스트 마이닝 분석 프로세스 - 텍스트 데이터 수집

- 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 정보 제공 사이트

텍스트 마이닝 분석 프로세스 - 텍스트 전처리

- 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

- 정제(cleaning) : 가지고 있는 코퍼스로부터 노이즈 데이터를 제거

- 정규화(normalization) : 표현 방법이 다른 단어들을 통합시켜서 같은 단어로 만듦

텍스트 마이닝 분석 프로세스 - 토큰화

- 토큰화 (tokenization) : 주어진 코퍼스(corpus, 말뭉치)에서 토큰(token)이라 불리는 단위로 나누는 작업(공백기준, 형태소기준, 명사기준)
- 기준은 분석 방법에 따라 다르다.
- 감성 분석한다면, 감성을 나타내는 품사가 동사, 형용사 쪽에 가깝기 때문에 형태소 분석기를 사용해서 동사, 형용사 위주로 추출한다.

텍스트 마이닝 분석 프로세스 - 특징 값 추출

- '중요한 단어'를 선별하는 과정
- '중요한 단어'로서의 특징은 적은 수의 문서에 분포되어 있어야 하고, 문서 내에서도 빈번하게 출현해야 한다.
- 특정 텍스트를 통해 문서를 구분 짓는 것이기 때문에 어떤 단어가 모든 문서에 분포되어 있다면 이는 차별성 없는 단어

텍스트 마이닝 분석 프로세스 - 데이터 분석

- 머신러닝
  - Linear Regression
  - Logistic Regression
  - Random Forest
  - XGBoost
- 딥러닝
  - CNN
  - RNN
  - LSTM
  - GRU

텍스트 마이닝(Text mining) : ex10

영화리뷰 텍스트 데이터 전처리

텍스트 마이닝(Text mining)

토큰화(tokenize)의 종류

- 단어(word) 단위
- 글자(character) 단위
- n-gram 단위

텍스트 마이닝(Text mining)

n-gram 단위

- n개의 연속된 단어를 하나로 취급하는 방법

- 예를 들어 “러시아 월드컵”이라는 표현을 “러시아”와 “월드컵” 두 개의 독립된 단어로만 취급하지 않고 두 단어로 구성된 하나의 토큰으로 취급한다.

  - n=2 경우를 bi-gram이라고도 부른다.
  - 단어의 개수가 늘어난 효과를 얻는다.

텍스트 마이닝(Text mining)

토큰화 (n-gram)

텍스트 : “어제 러시아에 갔다가 러시아 월드컵을 관람했다”

단어 토큰 : {“어제”, “러시아”, “갔다”, “월드컵”, “관람”}

2-gram 토큰 : {“어제 러시아”, “러시아 갔다“, “갔다 월드컵”, “월드컵 관람”}

텍스트 마이닝(Text mining)

토큰화 (n-gram)

- n-gram을 허용하면 토큰화 대상의 수가 크게 증가한다.
- 토큰화 한 결과를 수치로 만드는 방법
  - ➢ 원 핫(one-hot) 인코딩
  - ➢ BOW(단어모음)
  - ➢ 단어벡터(Word Vector)방법

텍스트 마이닝(Text mining)

원핫(one-hot) 인코딩

- 토큰에 고유 번호를 배정하고 모든 고유번호 위치의 한 컬럼만 1, 나머지 컬럼은 0인 벡터로 표시하는 방법

텍스트 : “어제 러시아에 갔다가 러시아 월드컵을 관람했다”
토큰사전 : {“어제”:0, “러시아”:1, “갔다”:2, “월드컵”:3, “관람”:4}

어제 = {1,0,0,0,0}
러시아 = {0,1,0,0,0}
갔다 = {0,0,1,0,0}
월드컵 = {0,0,0,1,0}
관람 = {0,0,0,0,1}

텍스트 마이닝(Text mining)

BOW(Bag of Word, 단어모음)

- "문장"을 하나의 벡터로 만드는 방법

단어사전 : {"어제":0, "오늘":1, "미국":2, "러시아":3, "갔다":4,
"축구":5, "월드컵":6, … , "중국":4999}

Text_1 : "어제 러시아에 갔다가 러시아 월드컵을 관람했다"를 BOW로 표현
(문장에 들어있는 단어의 컬럼만 1로나머지 컬럼은 0으로 표현.)

문장번호	0	1	2	3	4	5	6	…	4999
Text_1	1	0	0	2	1	0	1		0
Text_2	0	0	0	0	0	0	0		0

텍스트 마이닝(Text mining) : ex10 추가

BOW를 활용한 분석

텍스트 마이닝(Text mining) : ex10 추가

BOW를 활용한 분석
(min_df, max_df, n_gram, pipeline,
grid search)

텍스트 마이닝(Text mining)

tf-idf(term frequency-inverse document frequency)

- tf란 단어가 각 문서에서 발생한 빈도
  (단어가 등장한 '문서'의 빈도를 df라 한다.)
- 적은 문서에서 발견될수록 가치 있는 정보라고 할 수 있다.
- 많은 문서에 등장하는 단어일수록 일반적인 단어이며 이러한 공통적인 단어는 tf가 크다고 하여도 비중을 낮추어야 분석이 제대로 이루어질 수 있다.
- 따라서 단어가 특정 문서에만 나타나는 희소성을 반영하기 위해서 idf(df의 역수)를 tf에 곱한 값을 사용한다

텍스트 마이닝(Text mining)

tf-idf(term frequency-inverse document frequency)

- tf(d,t) : 특정 문서 d에서의 특정 단어 t의 등장 횟수.
- df(t) : 특정 단어 t가 등장한 문서의 수.
- idf(t) : df(t)에 반비례하는 수.

$$\mathrm{idf}(d,t) = \log \frac{n}{1+\mathrm{df}(t)}$$

텍스트 마이닝(Text mining)

tf-idf(term frequency-inverse document frequency)

- tf(d,t) : 특정 문서 d에서의 특정 단어 t의 등장 횟수.
- df(t) : 특정 단어 t가 등장한 문서의 수.
- idf(t) : df(t)에 반비례하는 수.

$$\text{tfidf}(w, d) = \text{tf} \times \left(\log\left(\frac{N+1}{N_w+1}\right) + 1\right)$$

텍스트 마이닝(Text mining) : ex10 추가

tf-idf를 활용한 분석 및 시각화