

Mid-Project

Applying traditional ML algorithms to text-based data (Text data analysis)

Course Overview

2021년 6월						
일	월	화	수	목	금	토
30	31	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	30	1	2	3
4	5	6				

2021년 7월						
일	월	화	수	목	금	토
20	21	22	23	24	25	26
27	28	29	30	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	30	31

- * 아래 커리큘럼의 세부 사항은 변동될 수 있습니다.
- * 진도 상황에 따라 1~2일 정도 차이가 발생할 수 있습니다.

파이썬 프로그래밍 기초 (프로그래밍의 3가지 축 / 파이썬 자료구조 등)
파이썬 정형 데이터 분석 (데이터 탐색 / 데이터 전처리 / 데이터 시각화)

파이썬을 활용한 데이터 수집 & 웹 스크레이핑 (+ 자동화 프로그램 개발)
파이썬 기반 텍스트 데이터 분석

Python 기초 수학 & 통계분석 (빈도분석 / 기술통계 / 교차검정 / 평균차이검정)

1차 세미 프로젝트 (데이터 수집 / 탐색 및 전처리 / 통계 분석 / 시각화 / 팀별 발표)

SQL 기초 프로그래밍 (Data Modeling / SQL CRUD / Adv. Techniques)

머신러닝 핵심 이론 & 주요 알고리즘 이론
파이썬 기반 머신러닝 알고리즘 실습 (Scikit-learn)
+ 데이터 분석 관련 직무 & 학습 리소스 소개

2차 세미 프로젝트 (Feature engineering & applying ML algorithms)

Mid-Project (텍스트 데이터 수집 / 데이터 전처리 / 각종 텍스트 기반 분석기법 적용)

딥러닝 핵심 이론 & 인공신경망 최적화 이론
파이썬 기반 딥러닝 알고리즘 실습 (Tensorflow & Keras)
+ 분야별 머신러닝 & 딥러닝 활용 사례 소개 + 각종 자동화 도구 실습

3차 세미 프로젝트 (데이터 수집 / 탐색 / 전처리 / 시각화 + ML&DL model tuning)

Available datasets in Korean

한국어 자연어 처리 데이터셋 목록

AI

한국어 자연어 처리 데이터셋 목록

Ym-LittleFox | 2020. 5. 24. 23:28

*** NLP / NLU 모델 학습을 위한 한국어 데이터셋 모음 ***
(8/1 UPDATE) AIHUB에 언어 인식기술 관련 멋진 데이터들이 대량 공개된 것 발견!! 특히 대화모델 학습용 데이터가 많이 포함되어 있습니다.

수어 데이터셋	텔리노스 대화 스크립트 데이터셋	한국어 대화 데이터셋
표준화 기반 일상 대화 데이터셋	대화형 한글 에이전트 데이터셋	어린이 음성 데이터셋
VRM 화상 데이터	한국어 감정 정보가 포함된 연속적 대화 데이터셋	한국어 감정 정보가 포함된 단발성 대화 데이터셋
인공지능 윤리 연구를 위한 비영향 테스트 데이터셋	실리싱달을 위한 텔레모달 데이터셋	

분류 분석 (감성분석/ 의도분류)

이름	설명	링크
네이버 영화 리뷰	네이버 영화 리뷰 데이터에 대한 금/부경 라벨 데이터 - 학습 15만건 / 테스트 5만건	github

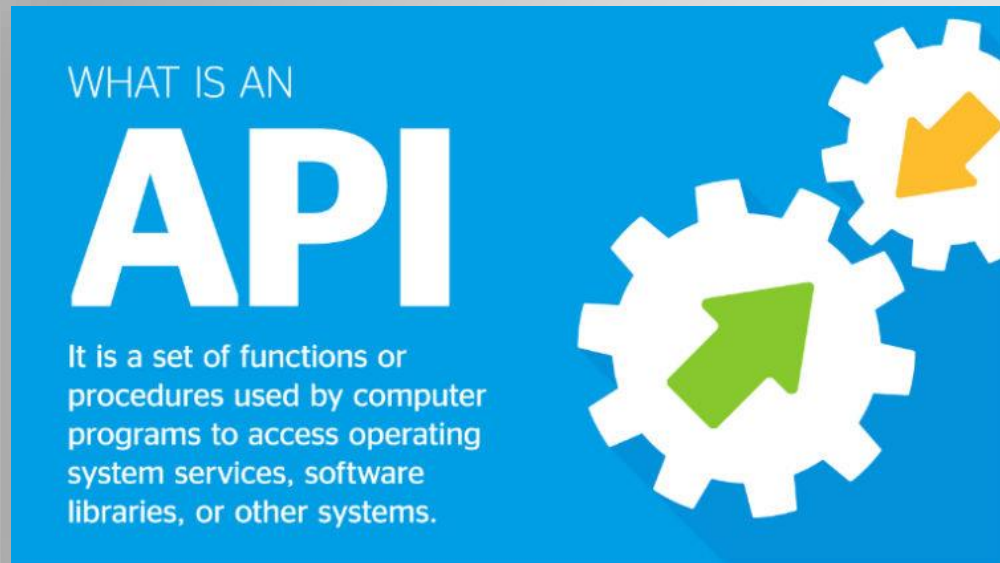
한국어 dataset 모음

코퍼스 명	용도	설명	링크
Naver sentiment movie corpus v1.0	분류	네이버 영화 리뷰 (긍정, 부정) 분류 라벨링 됨	https://github.com/e9t/nsmc
Chatbot_data	분류	채팅 대화 (일상,공정, 부정) 분류 라벨링 됨	https://github.com/songys/Chatbot_data
청와대 국민청원 사이트의 만료된 청원 데이터 모음	RAW	일자,카테고리,제목,내용 등 만료된 청원 Raw 데이터	https://github.com/akngs/petitions
Korean NER Corpus	NER	한국어 NER용 데이터 (NER, 형태소)	https://github.com/machinereading/KoreanNERCorpus
Korean		번역용 한국	

AwesomeKorean_Data

번호	데이터 종류	데이터 설명
1.	한국 정치인 뉴스 데이터 세트	-
2	청와대 국민청원 사이트의 만료된 청원 데이터 모음	
3	공공데이터포털 뉴스빅데이터	뉴스 데이터 'Kinds' 기반 분석 자료, 기사 메타 제공
4	챗봇용 대화 응답 세트	챗봇용 응답 쌍과 긍부정 태깅
5	영화추천시스템을 위한 데이터 세트	Synthetic dataset for recommender system created with Naver Movie rating system
6	육설데이터 세트	문장의 육설 여부를 분류한 데이터 세트
7	학습용 뉴스 댓글 데이터	BERT 모델과 학습에 이용한 11.62G 데이터를 모두 공개
8	AMR	문서요약에 대한 지침과 데이터 세트
9	네이버쇼핑, Steam 플랫폼 리뷰 데이터	감성분석(Sentiment Analysis)을 위한 제품 별, 게임 별 별점과 후기를 수집한 데이터셋

Web scraping for text data



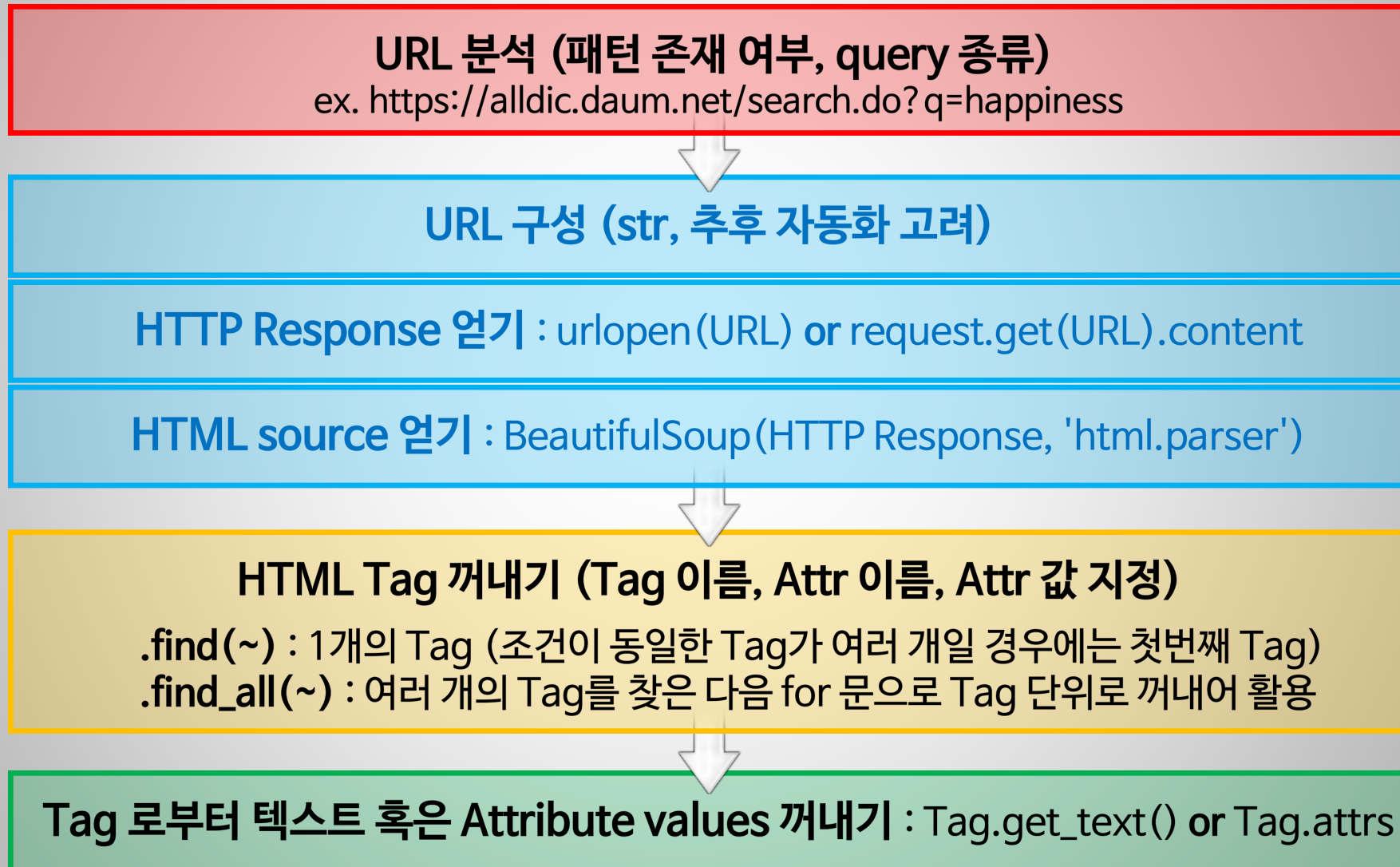
Use APIs & Web scraper

- APIs (Twitter, Facebook, Instagram, etc)
- Bots (Web crawler, Web scraper)

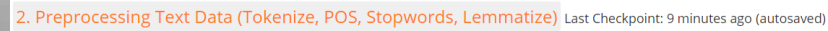
* 여기어때, 야놀자 DB 무단수집 위법 판결 @ <https://j.mp/3fgxi9Q> + 크롤링과 저작권 침해 고소 진행 일대기 @ <https://j.mp/3k5vbbl>
* Listly (크롬 확장프로그램 for 웹크롤링) @ <https://j.mp/2LSb8kh> / 네이버 크롤링 라이브러리 Kocrawl (날씨/미세먼지/지도/맛집/맞춤법) @ <https://j.mp/2CbdRA8>
* Web Scraping Tool & Web Data Extractor : ScrapeStorm @ <http://j.mp/2Y4porj> / Octoparse @ <https://j.mp/3o5i23q> / Automatio @ <https://automatio.co>

본 교안은 K-Digital 교육을 위해 제작되었으며, 교육 외 배포/게시/공개를 금합니다.

The process of web scraping (detailed)



The process of data analysis for text data



File Edit View Insert Cell Kernel Help

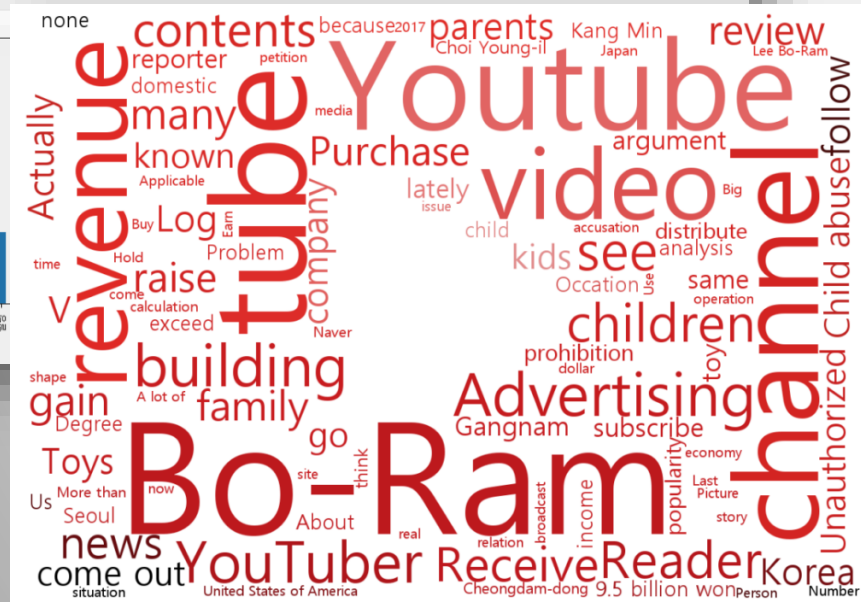
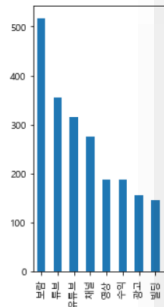
Text data Preprocessing

- nltk library(Natural Language Toolkit)를 이용하여 Text Processing을 위한 전처리를 실습한다.

1. 영어 문장 토큰화하기

```
In [ ]: 1 # Test preprocessing을 위해 nltk package를 import!
        2 import nltk

In [ ]: 1 #
        2 #
        3 result.plot(kind='bar', legend=False, figsize=(15,5))
        4 # 그림 사이즈를 변경하고 싶을 경우 figsize=(가로, 세로)를 변경합니다.
        5 # 기타 그래프 관련 옵션은 https://goo.gl/YNeJGt에서 확인하고 적용할 수 있습니다.
        6
        7 plt.show()
```



텍스트 데이터를 str 자료형으로 준비

Tokenize (형태소 분석)

POS Tagging (Part-of-speech, 품사 표시)

Stopwords 제거 (불용어 제거)

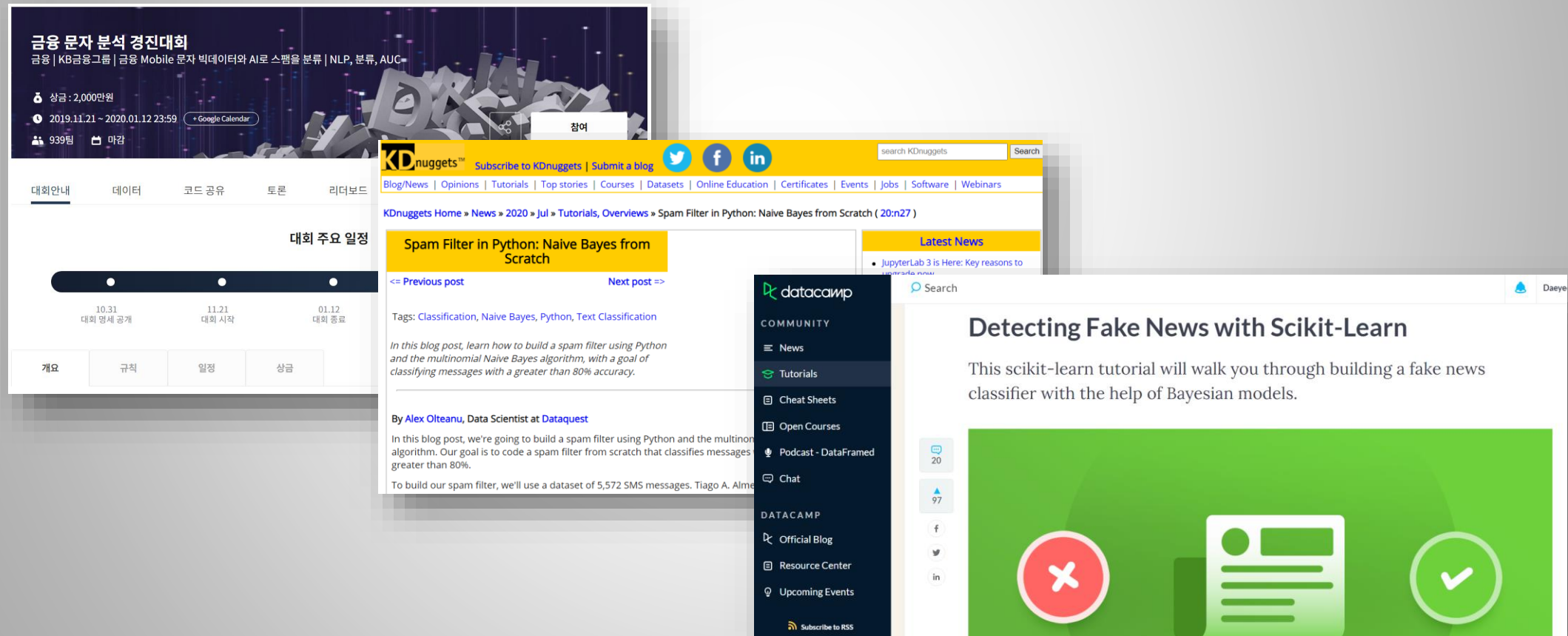
단어 갯수 카운팅 & 단어 사전 생성

단어 사전 기반 데이터 시각화

(+ 머신러닝/딥러닝 모델 적용)

1) Binary classification for text data

- (한) **스미싱 문자메시지** 분류 (TF-IDF & LightGBM) @ <http://j.mp/2SxBKqb>
- (영) Naive Bayes 기반 **Spam SMS** 분류기 구현 @ <https://j.mp/2B03D5x>
- (영) Naive Bayes 기반 **Fake news** 분류기 구현 @ <https://j.mp/2YiG5zu>



2) Similarity analysis for text data

- (한) **스타트업 복지 혜택** 기반 유사도 분석 (TF-IDF & Cosine similarity) @ <https://j.mp/34K5Cak>
- (영) **영화 줄거리** 기반 유사도 분석 (TF-IDF & Cosine similarity) @ <https://j.mp/35sXqLp>
- (한) **카카오지도 리뷰** 기반 맛집 추천 (Count vectorizer & Cosine similarity) @ <https://j.mp/3egb1ZF>



제멋대로인 텍스트, 손 쉽게 정돈 / 나열하기 - 코사인 유사도

🔑 "김 사원, 그 내역서에서 A항 내용 비슷한 것들만 쪽 뽑아
"네? 데이터가 주관식 응답에 2000개인데요...?"

스타트업의 사내 복지 데이터를 크롤링에 성공한 김 사원.
하지만 '자기계발 복지' 항목이 A회사랑 '비슷한 회사들만 걸러!

- 데이터는 정돈된 것이 아닌, **날것의 텍스트 데이터** 라면?
- 데이터를 눈으로 읽으며 비슷한지 판단하기엔 **왕이 너무**



00. 컨텐츠 기반 추천시스템 - TF-IDF를 이용한 추천시스템#2

Python notebook using data from T Academy Recommendation2 · 121 views · 20d ago

```
(11367, 0.15712074193481165),  
(1916, 0.15288512626542436),  
(3039, 0.1433450408051554),  
(483, 0.13765225108436677),  
(11573, 0.1337032693869044)]
```

```
In [14]:  
sim_scores = [(movie2id[i], score) for i, score in sim_scores[0:10]]  
sim_scores
```

```
Out[14]:  
[('Toy Story 3', 0.5262275451171008),  
(('Toy Story 2', 0.463276799830381),  
(('The 40 Year Old Virgin', 0.2797390476075632),  
(('The Champ', 0.20078538664316947),  
(('Rebel Without a Cause', 0.18287334034120212),  
(('For Your Consideration', 0.15712074193481165),  
(('Condorman', 0.15288512626542436),  
(('Man on the Moon', 0.1433450408051554),  
(('Malice', 0.13765225108436677),  
(('Factory Girl', 0.1337032693869044)]
```



우리 동네 맛집 추천엔진 직접, 쉽게 만들기 (크롤링과 코사인 유사도)

👉 이상한 외국 영화 추천 알고리즘 실습은 그만!
기본적인 라이브러리로 우리 동네 식당 · 카페 추천 프로그램을 만들어 보자

머신러닝을 배우면 자연스럽게 추천 알고리즘에 대해서 공부하게 됩니다.
하지만 추상적이고 다른 나라 언어로 된 예제로 공부하면 영 지루하겠죠?
당장 우리 일상의 데이터를 활용해 직접 추천 엔진을 구현해봅시다. 🤖

3) Sentiment analysis & Topic modeling (both of them are classification problem)

- (한) 네이버 영화 리뷰 데이터 Sentiment 분석 (TF-IDF & SGD Classifier) @ <https://j.mp/2Vs8jl6>
- (영) Game of Thrones 대본 기반 Sentiment 분석 @ <http://j.mp/2S2AbSa>
- (한) LDA 기반 트위터 메시지 토픽 모델링 @ <https://j.mp/35rGYLn>

네이버 영화 리뷰 감성 분류

IMDb 영화 리뷰 데이터셋과 비슷한 네이버 영화 리뷰 데이터셋(<https://github.com/e9t/nsmc>)을 네이버 영화 사이트에 있는 리뷰 20만 개를 모은 것입니다. 네이버 영화 리뷰 데이터셋 깃허브에서 직접 폴더에 데이터셋을 넣어 놓았습니다.

20만개 데이터 중 15만개는 훈련 데이터셋으로 ratings_train.txt 파일에 저장되어 있고 5만개는 리뷰의 길이는 140을 넘지 않습니다. 부정 리뷰는 1~4까지 점수를 매긴 리뷰이고 긍정 리뷰는 6~10까지 긍정 리뷰는 약 50%씩 구성되어 있습니다.

한글은 영어와 달리 조사와 어미가 발달해 있기 때문에 BoW나 어간 추출보다 표제어 추출 방식이 더 태소 분석을 하기 위한 대표적인 패키지는 konlpy와 soynlp입니다. 두 패키지를 모두 사용해 네이버 언어 이 예제를 실행하려면 konlpy와 soynlp가 필요합니다. 다음 명령을 실행해 두 패키지를 설치합니다.

```
In [ ]: !pip install konlpy soynlp
```

다음 konlpy, pandas, numpy를 임포트합니다.

```
In [1]: import konlpy
import pandas as pd
import numpy as np
```

코랩을 사용하는 경우 다음 코드 셀의 주석을 제거하고 실행하세요.

```
In [ ]: #!wget https://github.com/rickiepark/python-machine-learning-book-3rd-edition/raw/master/
```

감성 분류를 시작하기 전에 훈련 데이터셋과 테스트 데이터셋을 각각 판다스 데이터프레임으로 읽어 ratings_train.txt 파일은 하나의 리뷰가 한 행을 구성하며 각 필드는 탭으로 구분되어 있기 read_csv())는 기본적으로 콤마를 기준으로 필드를 구분하므로 delimiter='\\t'으로 지정하여 탭으로 된 문자열을 그대로 유지하기 위해 keep_default_na 매개변수를 False로 지정합니다.

Game of Thrones: Exploratory and Sentiment Analysis

How did Tyrion Lannister 'Dominate' the whole series?



Alben Tumanggor [Follow](#)
Nov 22, 2019 · 13 min read ★



1) Binary classification for text data

- (한) **스미싱 문자메시지** 분류 (TF-IDF & LightGBM) @ <http://j.mp/2SxBKqb>
- (영) Naive Bayes 기반 **Spam SMS** 분류기 구현 @ <https://j.mp/2B03D5x>
- (영) Naive Bayes 기반 **Fake news** 분류기 구현 @ <https://j.mp/2YiG5zu>

2) Similarity analysis for text data

- (한) **스타트업 복지 혜택** 기반 유사도 분석 (TF-IDF & Cosine similarity) @ <https://j.mp/34K5Cak>
- (영) **영화 줄거리** 기반 유사도 분석 (TF-IDF & Cosine similarity) @ <https://j.mp/35sXqLp>
- (한) **카카오지도 리뷰** 기반 맛집 추천 (Count vectorizer & Cosine similarity) @ <https://j.mp/3egb1ZF>

3) Sentiment analysis & Topic modeling (both of them are classification problem)

- (한) **네이버 영화 리뷰** 데이터 Sentiment 분석 (TF-IDF & SGD Classifier) @ <https://j.mp/2Vs8jl6>
- (영) **Game of Thrones 대본** 기반 Sentiment 분석 @ <http://j.mp/2S2AbSa>
- (한) LDA 기반 **트위터 메시지** 토픽 모델링 @ <https://j.mp/35rGYLn>

* Windows OS mecab 설치 @ <http://j.mp/2OXlkGX>

* Colab mecab 설치 @ <https://j.mp/2lehSqA> & <https://j.mp/3eBGO7S>

* KoNLPy 형태소 분석 클래스 비교 @ <https://j.mp/3sdZeBZ> & <https://j.mp/3brctJt>

* 스팸 이메일 분류기 만들기 (Naive Bayes 설명) @ <https://j.mp/3g0QQ1Q>

* 토픽 모델링을 위한 LDA (Latent Dirichlet Allocation) 설명 @ <https://j.mp/2FdEnL3> & <https://j.mp/2LjhSHU>

* Kaggle 노트북 도커 이미지 (with 한글폰트/한글자연어처리패키지/형태소분석기(mecab) 등) @ <https://j.mp/3z8eXGe>

본 교안은 K-Digital 교육을 위해 제작되었으며, 교육 외 배포/게시/공개를 금합니다.

수업 관련 공지사항

* 앞서 설명한 예시 이외에도 텍스트 데이터를 다룬 분석이면 무엇이든 가능합니다!

* 데이터 수집 방법 / 전처리 & 시각화 방법 / Model 선택 및 적용 여부 모두 자유입니다.

* Part 3/4/5 에서 배운 지식들을 최대한 적용하는데 초점을 맞춰주세요. (웹 크롤링 필수 X)

* 발표 시 포함할 사항 : 분석 주제 / 데이터 소개 / 전처리 방법 / 분석 프로세스 / 결과 해석
발표 시 제출할 사항 : 발표 자료 (ppt or pdf) / 전체 코드 (.ipynb, 주석 포함) / 원본 데이터

* **7/12 월요일 14:30** : 팀별 발표 및 질의응답 (15~20분 내외/팀, **최대 20분**)
: **7/12 (월) 14:20 전까지** 발표 자료 & Jupyter notebook(+원본 데이터) 제출 @ 슬랙 DM

* 1차/2차 세미프로젝트 발표자는 발표 X & 도움이 필요할 경우 슬랙 채널에서 호출

수업 관련 공지사항

1팀 : 강민정, 박건우, 이병준, 이혜민

2팀 : 민정현, 강원석, 이규호, 조윤정, 최용수

3팀 : 전가은, 박민수, 박정재, 임주란, 조성곤

4팀 : 정소연, 김주연, 이준동, 황준우

5팀 : 주리아, 권산하, 박용민, 안성훈

6팀 : 류범상, 이소연, 이용석, 한창환

[K-Digital Training]
인공지능 통합과정

End of Document