[K-Digital Training]
인공지능 통합과정

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

Daeyeon Jo
repositivator@gmail.com

Course Overview

* 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

Vm~ LittleFox | 2020. 5. 24. 23:28

*** NLP / NLU 모델 학습을 위한 한국어 데이터셋 모음 ***

(8/1 UPDATE) AI HUB에 언어 인식기술 관련 멋진 데이터들이 대량 공개된 것 발견!! 특히 대화모델 학습용 데이터가 많이 포함되어 있습니다.

분류 분석 (감성분석/ 의도분류)

이름	설명	링크
네이버 영화 리뷰	네이버 영화 리뷰 데이터에 대한 긍/부정 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,긍정, 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 NER 용 데이터 (NER, 형태소)	https://github.com/machinereading/KoreanNERCorpus
Korean		번역용 한국	

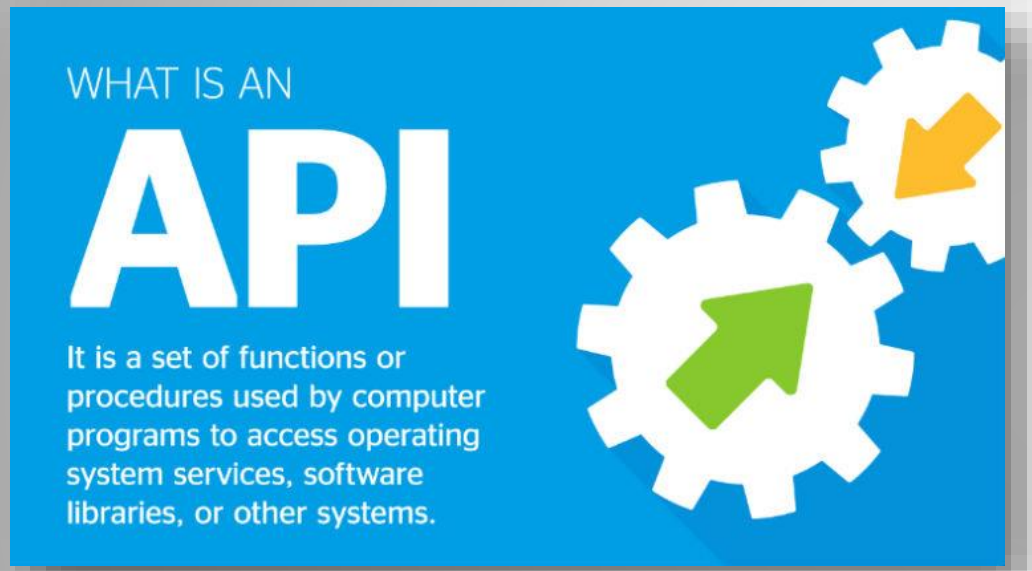
AwesomeKorean_Data

기타 주요 데이터

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	욕설데이터 세트	문장의 욕설 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

* **KLUE**: Korean Language Understanding Evaluation (**한국어 NLP 데이터셋** & pretrained-LM 베이스라인) @ https://j.mp/3woY9Iy & https://j.mp/3f7p0mU

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

The process of web scraping (detailed)

URL 분석 (패턴 존재 여부, query 종류)
ex. https://alldic.daum.net/search.do?q=happiness

↓

URL 구성 (str, 추후 자동화 고려)

HTTP Response 얻기 : urlopen(URL) **or** request.get(URL).content

HTML source 얻기 : BeautifulSoup(HTTP Response, 'html.parser')

↓

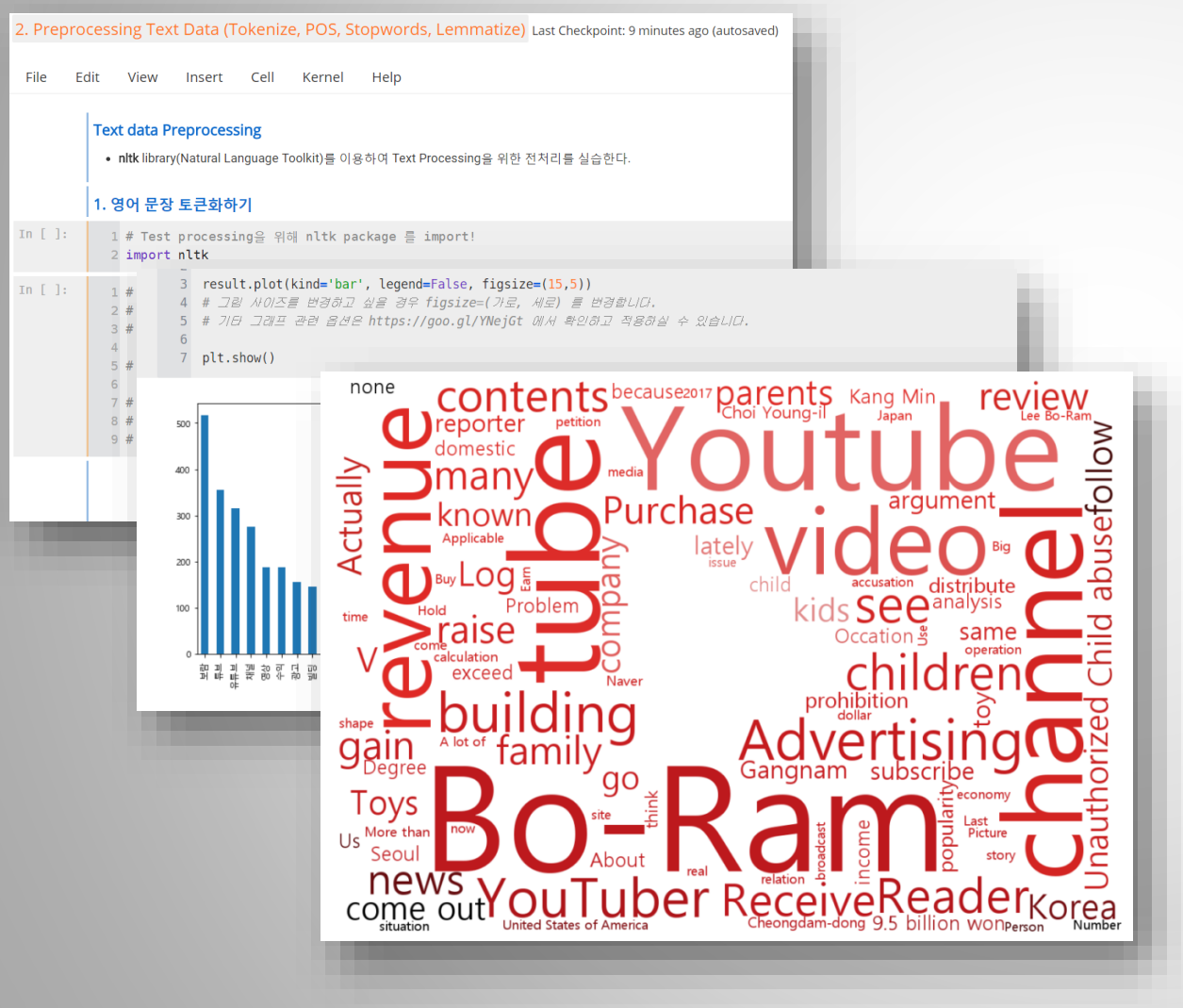
HTML Tag 꺼내기 (Tag 이름, Attr 이름, Attr 값 지정)

.find(~) : 1개의 Tag (조건이 동일한 Tag가 여러 개일 경우에는 첫번째 Tag)
.find_all(~) : 여러 개의 Tag를 찾은 다음 for 문으로 Tag 단위로 꺼내어 활용

↓

Tag 로부터 텍스트 혹은 Attribute values 꺼내기 : Tag.get_text() **or** Tag.attrs

The process of data analysis for text data



텍스트 데이터를 str 자료형으로 준비

Tokenize (형태소 분석)

POS Tagging (Part-of-speech, 품사 표시)

Stopwords 제거 (불용어 제거)

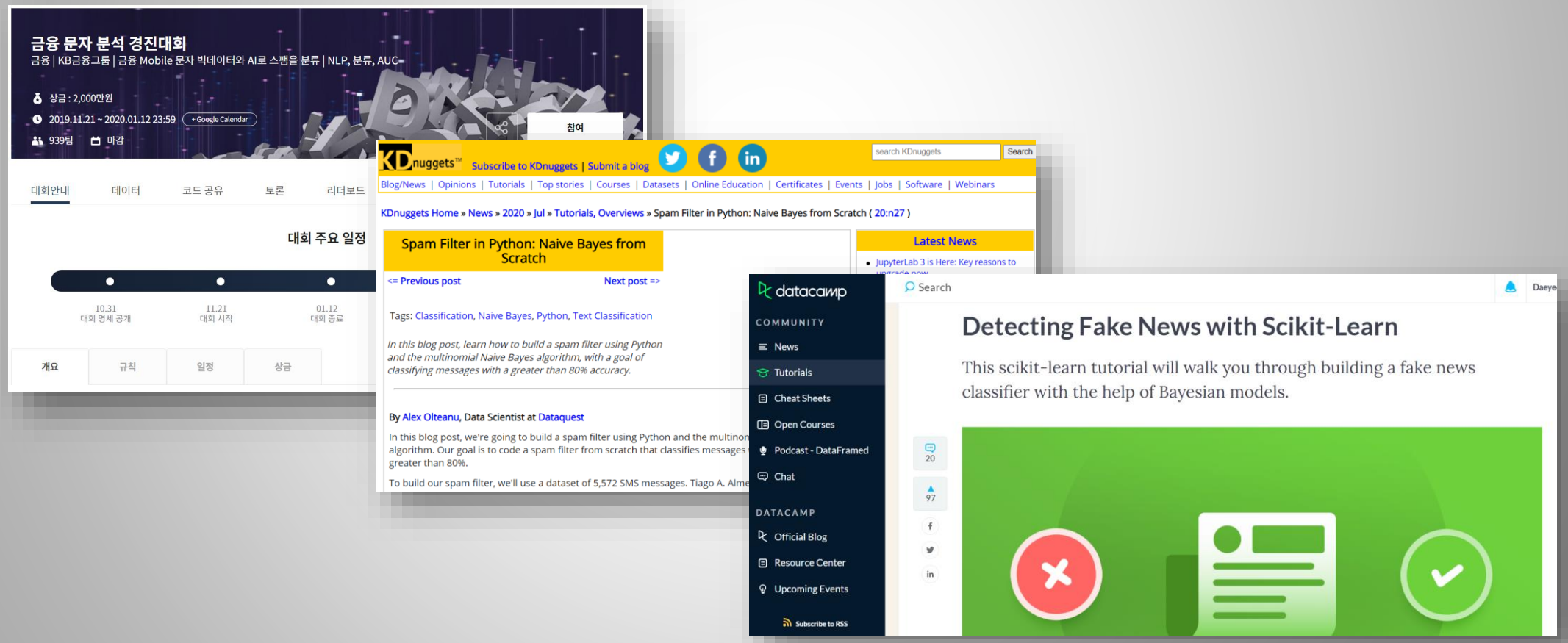
단어 갯수 카운팅 & 단어 사전 생성

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

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

* Mecab → Twitter (Okt)
* SGDClassifier → Other classification models

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/2Vs8jI6
- (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/2IehSqA & 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

* 매 정시마다 메인 회의실에서 출석 체크 진행합니다.

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 적용**하는데 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