Portfolio

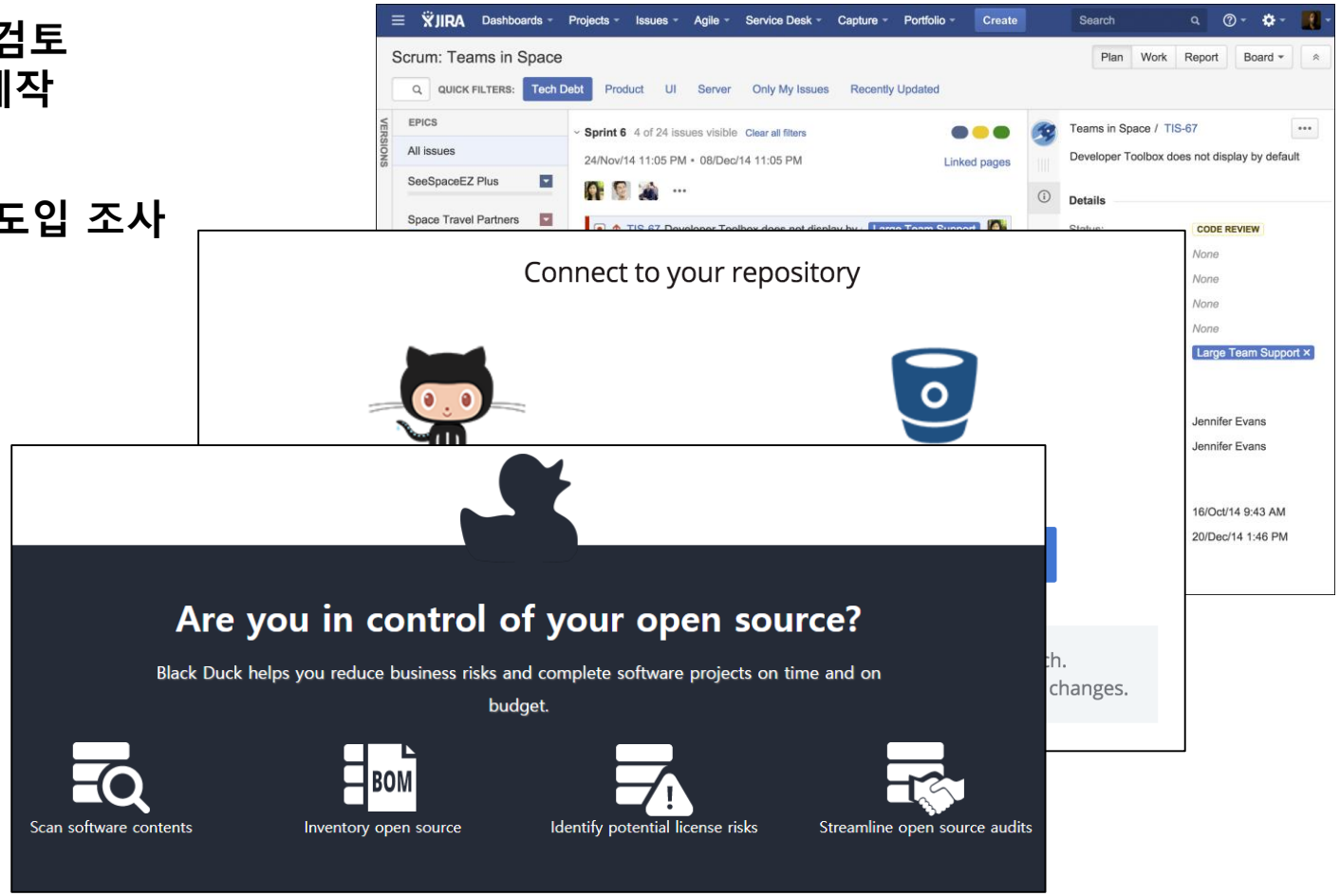
고려대학교 정보대학 컴퓨터학 석사

김우성

Portfolio

AhnLab 인턴 - 학부

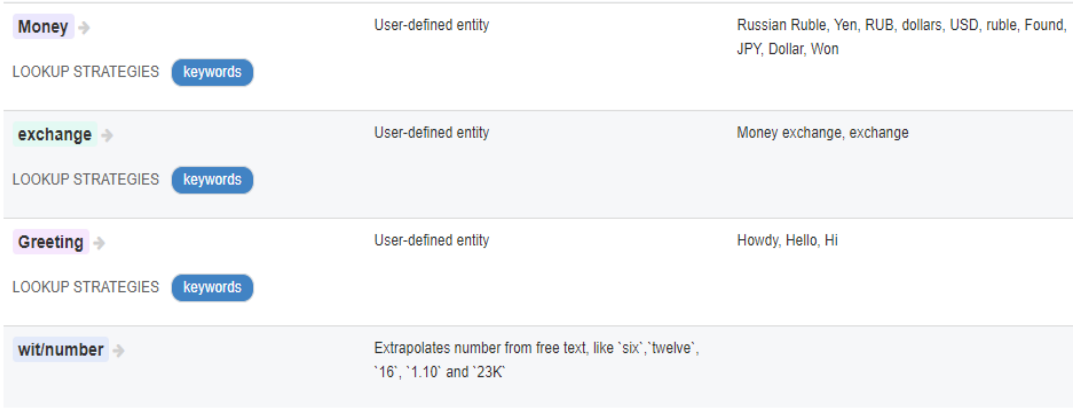
- 사내 소프트웨어 유지보수 및 프로그램 도입(2016)
  - 유지보수
    - 사내 사용 소프트웨어 버전 및 도입 검토
    - 사내 망에 소프트웨어 사용 설명서 제작
  - 프로그램 도입
    - OpenSource 저작권 검증 프로그램 도입 조사
    - JAVA 기반 코드 분석 프로그램 구현

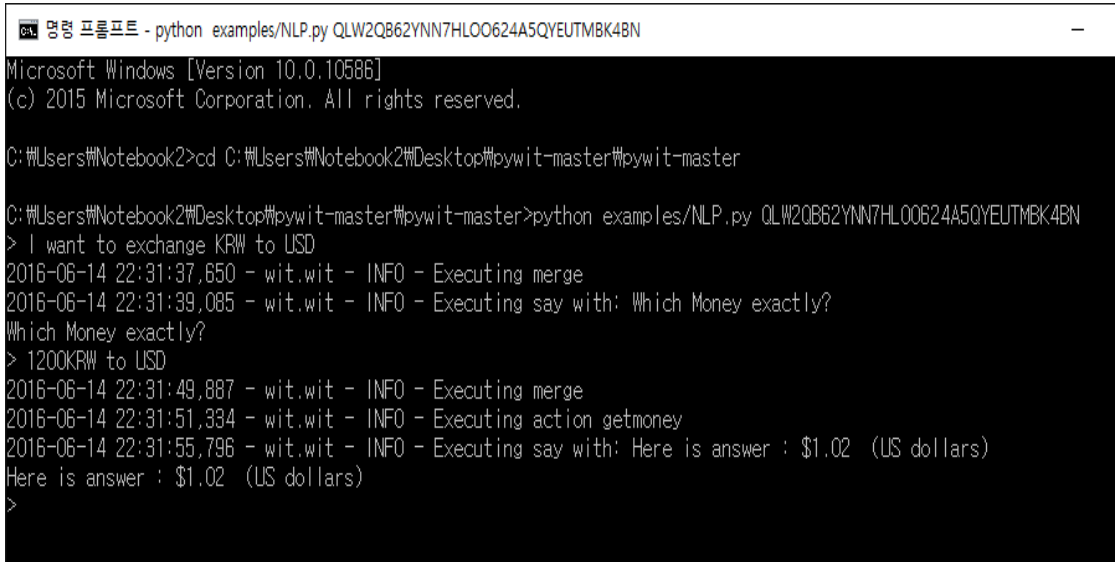


Portfolio

NLP Using Wit.ai - 학부

- Wit.ai 와 Wolfram Alpha를 사용한 환율 Dialog System(2016)
  - Python 언어 사용
  - Wit.ai
    - 규칙 기반과 기계 학습의 장점들을 결합한 시스템
    - Wit.ai 상에서 App을 생성
    - Entity 규칙을 학습 & 기계학습을 통한 상황 구분
  - 환율 Dialog System
    - Python을 기반으로 Wit.ai App 제작
    - 질문을 통하여 목적 파악
    - 환율의 경우 Wolfram Alpha를 사용하여 정보 획득
    - 지정된 방식으로 답변





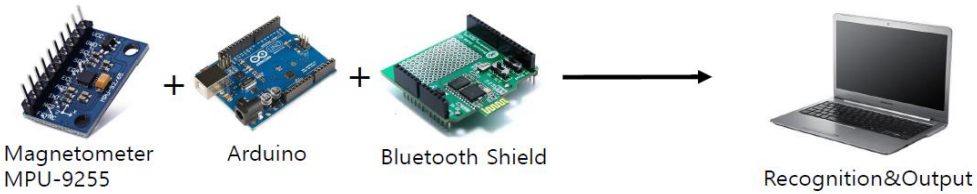
Link : https://github.com/adventure2165/NLP_currency_dialog

Portfolio

기계학습을 이용한 필기인식 장치 - 학부

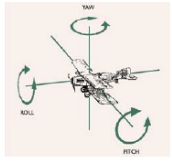
- Writing Recognition Device With Magnetometer 구현(2016)
  - Arduino – (전용)C언어, 프로그램 – JAVA 사용
  - 장비 구현
    - 아두이노, 지자기센서를 이용하여 3방향 움직임 파악
    - 블루투스 통신으로 일정 간격동안 장치의 좌표 전달
    - 좌표 데이터 CSV화
  - 데이터 수집 및 기계학습
    - 받아온 좌표를 PCA 기법을 통하여 2차원 데이터화
    - 좌표간 계산을 통해 이동 방향을 8개 숫자로 표현
    - 데이터 저장 및 Preprocessing 처리
    - WEKA 프로그램을 통한 MLP 학습

2. 장비 설명

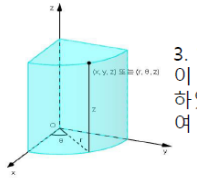


3. 이론 설명

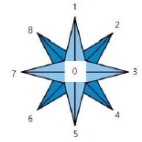
1. MPU-9255는 Hall Effect를 사용하여 지구상의 지자기의 세기를 측정하여 이를 이용하여 X, Y, Z축 의 값으로 전송을 하는 지자기 센서이다. 이를 사용하여 X, Y, Z축으로 얼마나 이동하였는지를 파악하여 이 값을 전송한다.



2. 지자기센서로부터 전송받은 X, Y, Z값을 이용하여 아두이노를 휘두를 때의 움직임을 알 기 위하여 XY평면에서의 기준축으로부터의 돌아간 각도를 나타내는 Yaw를 구하였고 Z는 기존의 값을 사용하였다.



3. 얻어낸 Yaw값을 통하여 센서가 반지름이 100인 원기둥에 좌표들의 값들을 표현하였다. 이후, 이를 직교좌표계로 변환하여 좌표들의 X, Y, Z값을 얻어내었다.



4. 좌표계들의 점들의 움직임을 방향벡터로 표현하였다. 움직이지 않았을 때는 0으로, 시계방향으로 정북쪽을 1번으로 시작하여 8번까지 숫자를 매기어서 방향벡터를 설정하였다. 이를 통하여 각 점들의 움직임을 벡터로 표현하였다.

5. 방향벡터의 데이터들과 이 데이터가 어떤 글씨인지 알려주는 Feature를 입력한 후, Machine Learning 을 통하여 필기 인식에 대한 모델을 생성한다.

6. 이후 Device로 글씨를 그렸을 때 필기인식 모델에 따라서 글씨를 판독하여 결과를 산출한다. 우선적으로는 숫자 인식부터 진행하였다.

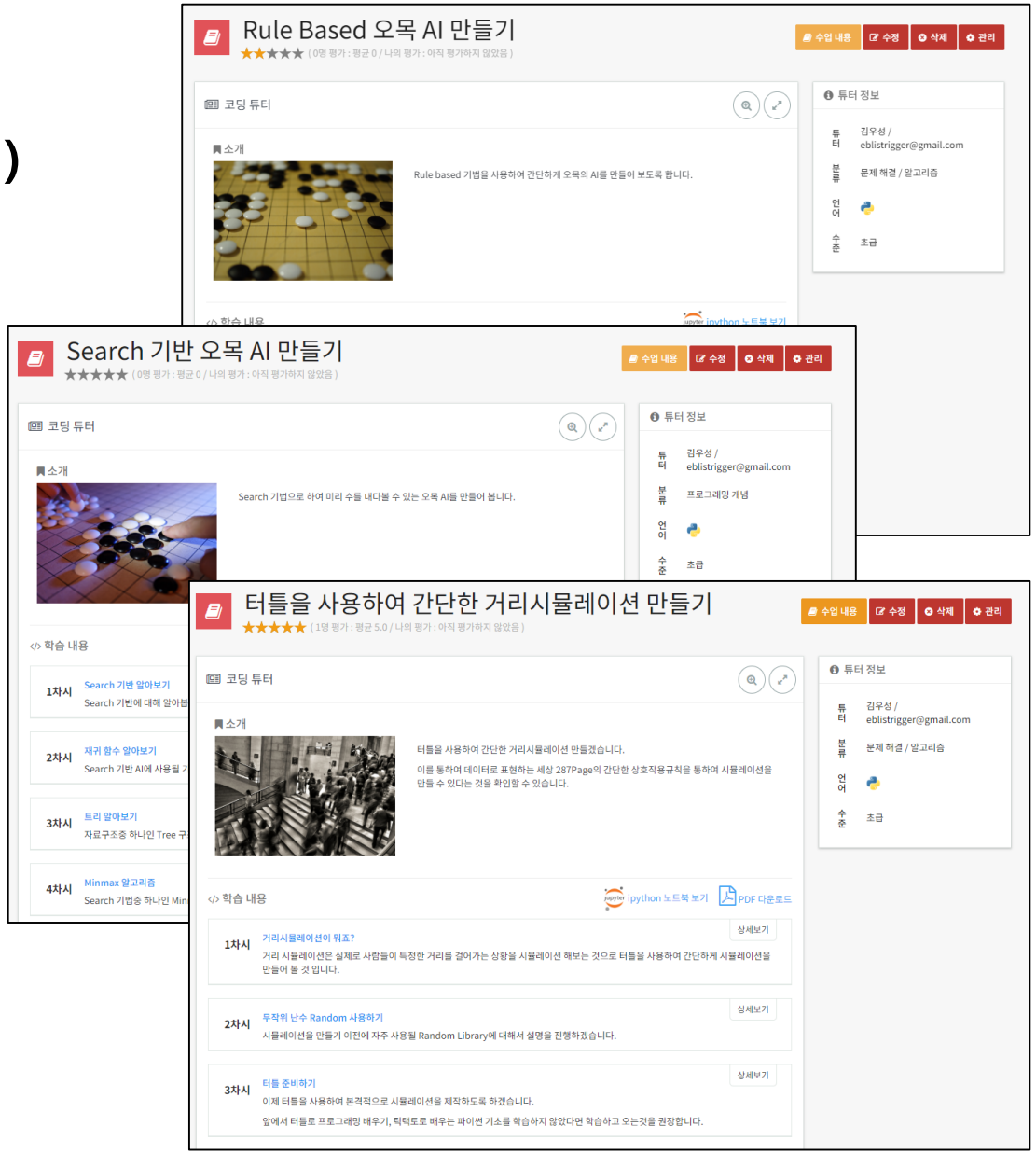
Link : https://github.com/adventure2165/graduate_project

Portfolio

수업조교 활동 - 석사

- 데이터로 표현하는 세상 (2017-1학기, 2017-2학기)
  - 비전공자를 위한 오목 AI 과정
    - Python 기반 실습
    - Rule-based 기법
    - Min -Max 알고리즘
    - Alpha-beta Pruning 알고리즘
  - 생명체 모방 프로그래밍 과정
    - 단순한 규칙으로 생명체 모방 프로그래밍 소개
    - 번화가에서 충돌하지 않는 시뮬레이션 실습
- 모두를 위한 파이썬(2017-1학기)
  - 비전공자를 기초 Python 수업
  - Matplotlib등 라이브러리 사용법 안내

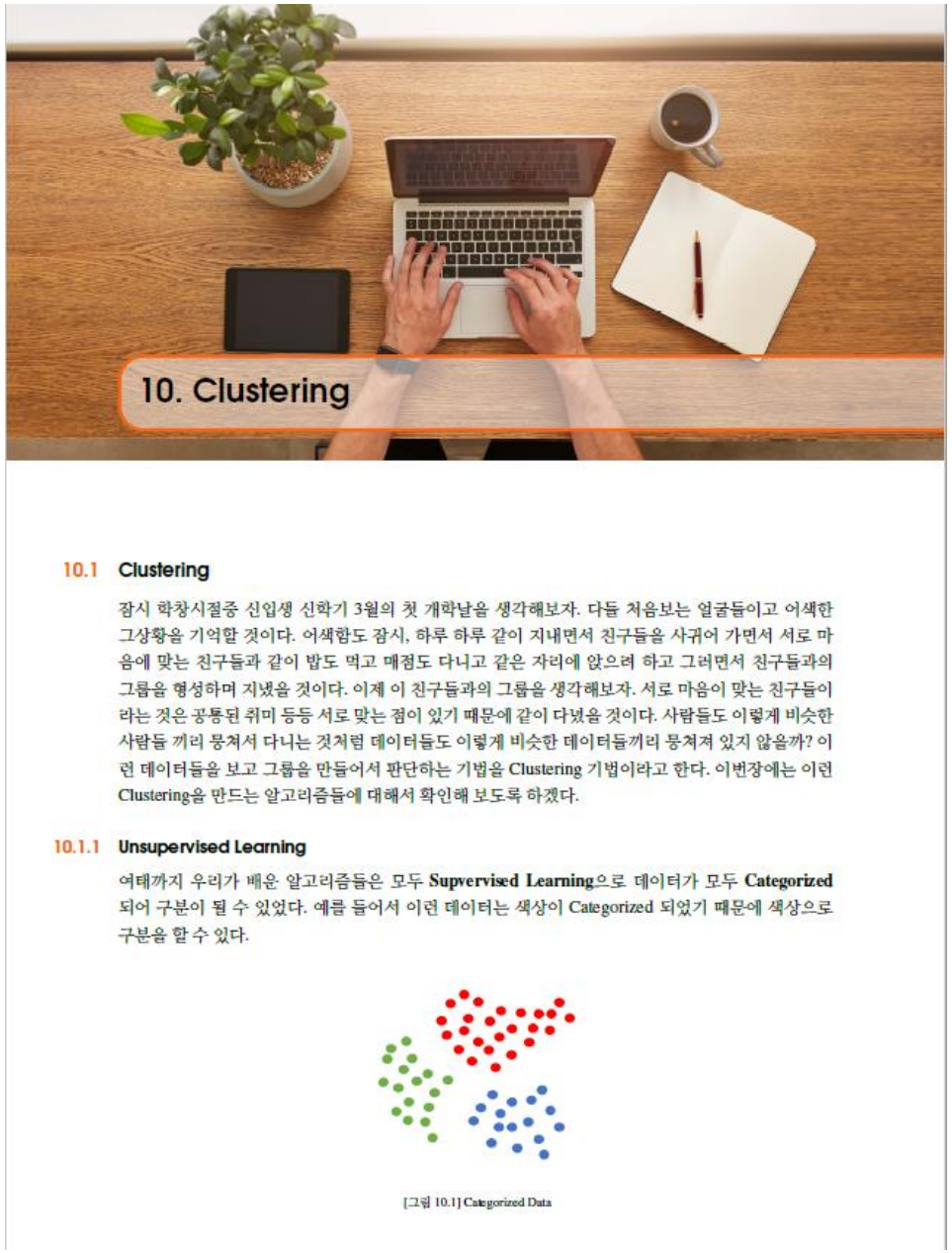
Link : https://everycoding.korea.ac.kr/tutor/grid/



Portfolio

수업조교 활동 - 석사

- 인공지능, 기계학습(2017-1학기, 2017-2학기)
  - 인공지능&기계학습 강의자료 제작
    - Numpy, Matplotlib, pandas, Seaborn 사용
    - Sci-kit learn 사용
      - Regression
      - Decision Tree
      - Perceptron / MLP
      - SVM
      - Naïve bayes
      - Clustering
    - Dimension Reduction
    - Anaconda, Jupyter notebook
      - 설치 및 사용법



Portfolio

수업조교 활동 - 석사

- Project - KDD Cup 1999 Data 분석(2017-2학기)
  - 데이터 분석
    - 네트워크 침입 데이터
    - 데이터 종류, 데이터 타입, Null 값 확인
  - 데이터 전처리
    - 데이터 Normalizing
    - Feature Selection
  - 기계학습을 통한 침입 Predict(Ensemble Model 포함)

기계학습 기법 별 정확도

KNN : 0.73180144612518294

Linear Regression : 0.32706767093764999

Logistic Regression : 0.75455795590648977

Decision Tree : 0.73317659583906314

SVM : 0.67373464046488929

Perceptron : 0.63607328217184933

MLP : 0.7571308166614914

Naive bayes : 0.63811382690857477

random forest : 0.73535021958035751

adaboost : 0.57813955551612473

Gradient Boosting : 0.73073681408863067

Stacking : 0.76489375859468567

KDD Cup 1999 Data

Abstract

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Information files:

- task_description. This is the original task decription given to competition participants.

Data files:

- kddcup.names A list of features.
- kddcup.data.gz The full data set (18M; 743M Uncompressed)
- kddcup.data_10_percent.gz A 10% subset. (2.1M; 75M Uncompressed)
- kddcup.newtestdata_10_percent_unlabeled.gz (1.4M; 45M Uncompressed)
- kddcup.testdata.unlabeled.gz (11.2M; 430M Uncompressed)
- kddcup.testdata.unlabeled_10_percent.gz (1.4M;45M Uncompressed)
- corrected.gz Test data with corrected labels.
- training_attack_types A list of intrusion types.
- typo-correction.txt A brief note on a typo in the data set that has been corrected (6/26/07)

Portfolio

외부 강의 조교 활동 - 석사

- Python을 사용하는 Information Retriever
- 한국어 문서(뉴스기사) 간단한 검색 엔진 구현
  - Crawler
    - Requests
    - Beautiful Soup
    - URLParse
  - Retriver
    - Data Preprocessing
    - NLTK
    - KoNLPy
    - TF-IDF 적용

Crawler에 의해 모인 기사 파일

0000006296.txt	2017-07-31 오후...	텍스트 문서	12KB
0000020426.txt	2017-08-18 오후...	텍스트 문서	7KB
0000161087.txt	2017-08-18 오후...	텍스트 문서	3KB
0000554045.txt	2017-07-31 오후...	텍스트 문서	4KB
0000985133.txt	2017-08-30 오후...	텍스트 문서	2KB
0002157720.txt	2017-07-31 오후...	텍스트 문서	16KB
0002322559.txt	2017-07-31 오후...	텍스트 문서	5KB
0002700418.txt	2017-08-18 오후...	텍스트 문서	4KB
0002746247.txt	2017-08-18 오후...	텍스트 문서	4KB
0002746265.txt	2017-08-18 오후...	텍스트 문서	2KB
0002849523.txt	2017-08-30 오후...	텍스트 문서	3KB
0002871757.txt	2017-08-02 오후...	텍스트 문서	8KB
0002897176.txt	2017-08-18 오후...	텍스트 문서	4KB
0003302049.txt	2017-08-02 오후...	텍스트 문서	6KB
0003589511.txt	2017-08-02 오후...	텍스트 문서	3KB
0003927194.txt	2017-08-30 오후...	텍스트 문서	4KB
0004047210.txt	2017-08-02 오후...	텍스트 문서	4KB
0004047391.txt	2017-08-02 오후...	텍스트 문서	5KB
0004057400.txt	2017-08-18 오후...	텍스트 문서	5KB
0008097508.txt	2017-07-31 오후...	텍스트 문서	2KB
0008127872.txt	2017-08-18 오후...	텍스트 문서	4KB
0009444392.txt	2017-07-31 오후...	텍스트 문서	2KB
0009444913.txt	2017-07-31 오후...	텍스트 문서	4KB
0009445099.txt	2017-07-31 오후...	텍스트 문서	6KB
0009482544.txt	2017-08-18 오후...	텍스트 문서	5KB

```
print('{}[{:.5f}] - {}...'.format(docname, similarity, newscontent[:50]))
```

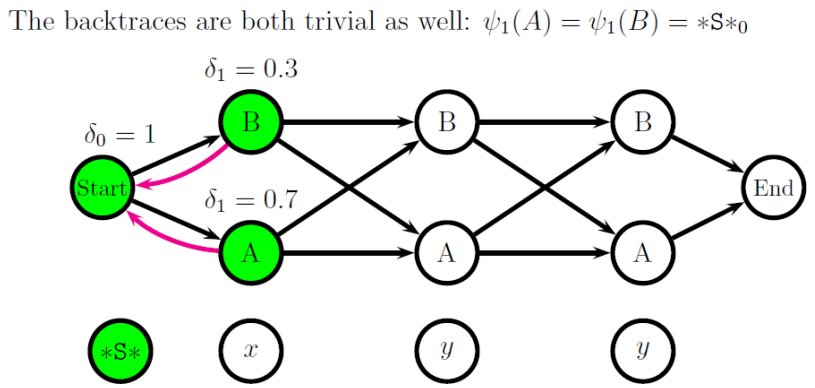
```
북한
Euclidian
Query - 북한
IndexTerm - dict_keys(['북한'])
0009483552.txt[0.59946] - 전자발찌 끊고 도주한 40대 오리무중…현상금 1천만원 본문 내용 플레이어 플레이어 오류를 ...
0002322559.txt[0.00000] - 北 2차 도발美 北 실전배치땐 '정권교체 작전' 가능성 본문 내용 플레이어 플레이어 오류를...

Cosine
Query - 북한
IndexTerm - dict_keys(['북한'])
0002322559.txt[0.43049] - 北 2차 도발美 北 실전배치땐 '정권교체 작전' 가능성 본문 내용 플레이어 플레이어 오류를...
0009483552.txt[0.10021] - 전자발찌 끊고 도주한 40대 오리무중…현상금 1천만원 본문 내용 플레이어 플레이어 오류를 ...
```

Query : "북한" 일 때 Retrieval 된 결과

Portfolio

NLP(Hidden Markov model -> Viterbi algorithm) - 석사

- Python 사용
- Hidden markov model 구현해보기
  - 특정 태그에 대해서 HMM 모델을 사용
  - 특정 단어 뒤에 태그가 나올지 안 나올지 예측
  - 10-fold cross validation을 통한 validation
- Viterbi Algorithm
  - 참고자료를 기반으로 한 구현
  - 예시 데이터 기반
  - 참고자료에는 오류가 있어서 수정 내용 추가




```
In [2]: states = ["Sunny","Rainy"]
        start_prob = {'Sunny': 0.7, 'Rainy': 0.3}
        transition_prob = {
            'Sunny' : {'Sunny': 0.2, 'Rainy': 0.7,"End" : 0.1},
            'Rainy' :   {'Sunny': 0.7, 'Rainy': 0.2,"End" : 0.1},
        }
        emission_prob = {
            'Sunny' : {'o': 0.4, 'x': 0.6},
            'Rainy' : {'o': 0.3, 'x': 0.7},
        }

In [3]: sentencee = "oxx" # 편의상 구름이 없으면 o, 구름이 있으면 x로 표현함.
        a = Viterbi(sentencee,states,start_prob,transition_prob,emission_prob)

In [4]: a
Out[4]: 'Sunny Rainy Sunny'
```

Link : https://github.com/adventure2165/NLP_HMMandViterbi

Portfolio

Personal Research Proposal(제안서) - 석사

- **Lip Reading Project**
  - Deep Learning을 이용한 Lip reading 연구
  - Visemes 분석
  - Speech Recognition 보조 를 통한 인식률 향상
  - 한국어 처리의 새로운 보조 수단 목표

Lip Reading 연구 계획

Lip Reading이란?

- Lip Reading(Visual speech recognition, 구화)은 화자의 입술 모양을 분석하여, 어떠한 발언을 하는지 파악하는 방법임.
- 사람들이 발화를 하는데 반드시 입이 움직이게 되어 있으며, 이는 필연적으로 발화를 하는데 입모양이 반드시 생기게 된다는 점을 의미함.
- 이와 같이 언어를 사용하는데 있어서 입모양은 필수 불가결 적인 요소 이고, 이를 중요하게 활용하고 있는 분야들이 존재함.
- 입모양이 중요하게 사용되는 분야는 아래와 같음.
  - 청각 장애를 지닌 사람들에게 수화와 더불어 다른 사람과의 의사소통의 방법으로 써 구화가 교육되고 있음.
  - 외화나 애니메이션에서 보다 자연스러운 연기 및 몰입도를 위하여 입모양에 신경 쓰고 있음.
  - 언어 교육 분야에서 발음을 교육하는데 있어 발음 기호와 더불어 입모양 사진을 사용하고 있음.
  - 아나운서와 같이 정확한 발음을 요구하는 직업에서는 표준 입모양을 통하여 정확한 발음을 추구하고 있음.
  - 언어 교육에 있어서 그림과 같이 발음에 따른 입모양 사진을 통하여 발음의 방법을 학습자가 보다 쉽게 이해할 수 있도록 하고 있음.

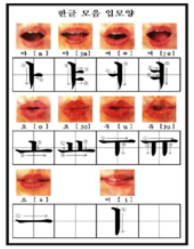
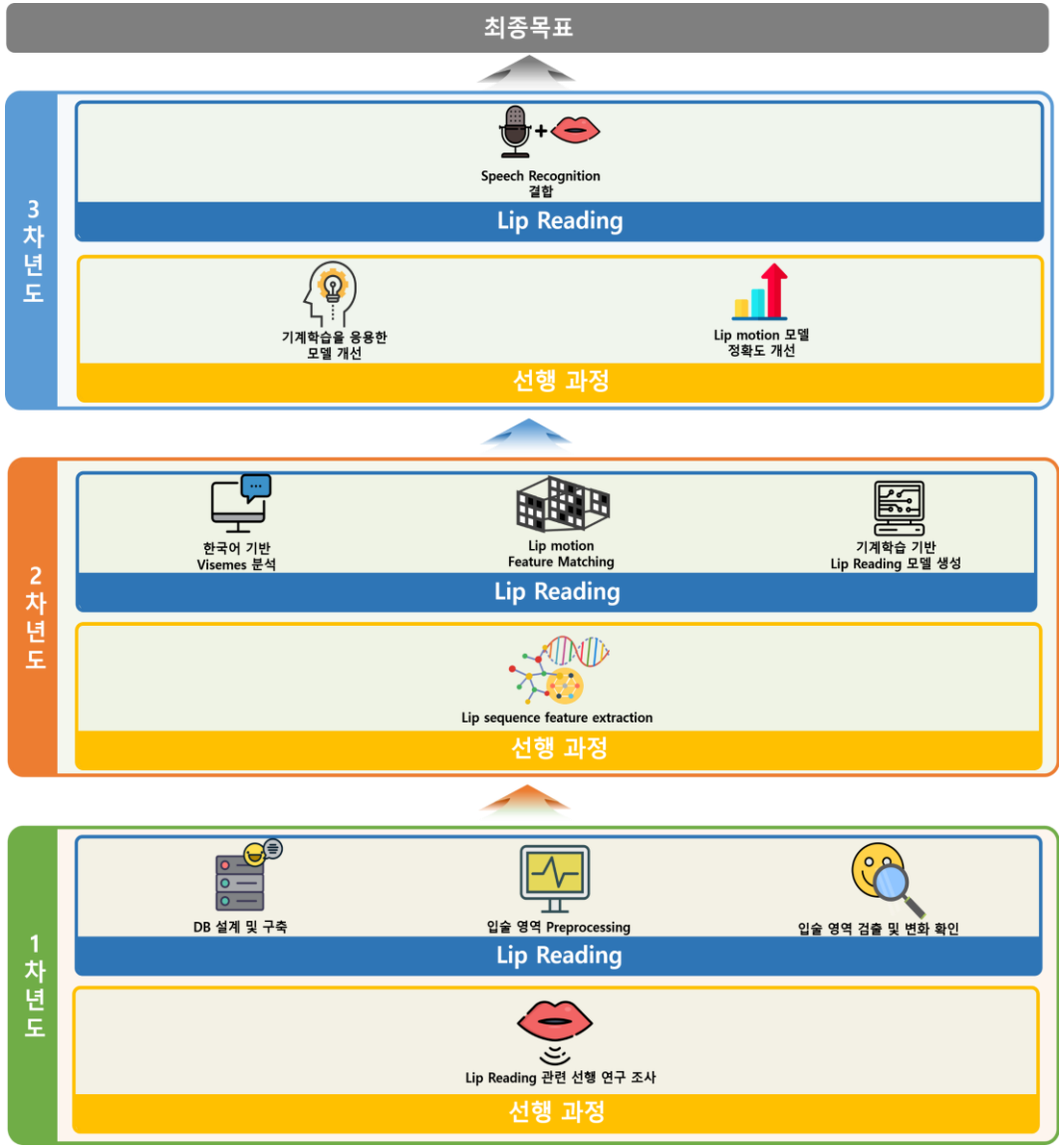


Figure 1. 한글 모음 입모양

Lip Reading 연구 계획 (3년 기준)

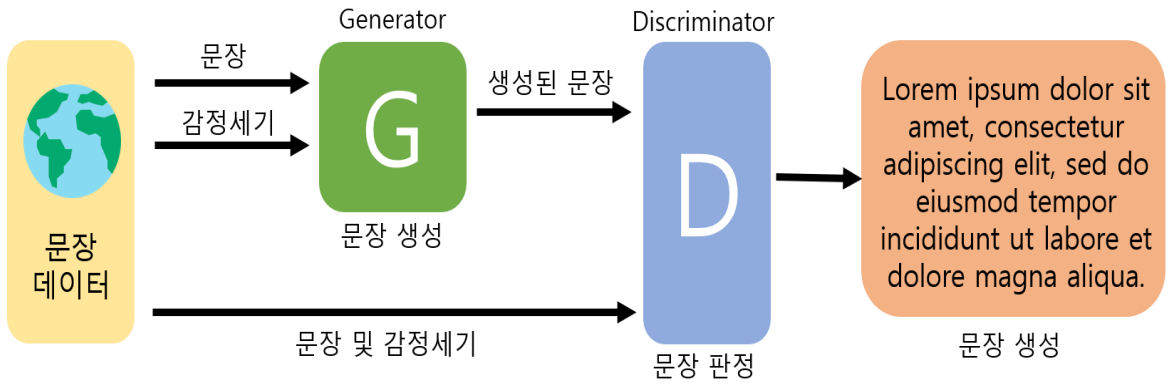
구분	Lip Reading
최종목표	● Lip Reading 연구를 통하여 음성 인식의 정확도를 높이고, 보조적으로만 사용되었던 기술에 대해 새롭게 조명한다.
1차년도 (2018)	● Lip Reading 선행 연구 조사 및 개선 연구 ● Lip Reading실험을 위한 DB 설계 및 구축 ● 이미지/동영상에서 입술 영역에 대한 전처리 연구 ● 이미지/동영상에서 입술 영역에 대한 검출/입술모양 변화에 대한 연구
2차년도 (2019)	● Lip sequence feature extraction 알고리즘 연구 ● 독립된 단어 단위의 한국어 Visemes 분석 ● 문장 속 Lip motion과 Visemes 간 연관 분석 ● DNN, CNN, RNN을 이용한 입술 모양 인식 모델 연구 및 생성
3차년도 (2020)	● 기계학습 응용 기법을 사용한 입술 모양 인식 모델 개선 ● 음성/화자 인식 모델과 연동하여 Lip reading 모델 확장 ● ensemble 기법을 활용한 통합 인식 모델 개발



Portfolio

GAN Based Emotional Text Generation- 석사

- **Emotional Text Generation**
  - **Semeval 2018 Task1. Affect in Tweets 데이터 사용**
  - **Conditional SeqGAN을 사용**
  - **감정세기(0~3)에 맞게 원시적인 감정 문장 생성**
  - **BLEU 점수를 이용한 문장이 제대로 생성되었는지 파악**
  - **한국정보처리학회 논문 및 석사 졸업논문(예정)**



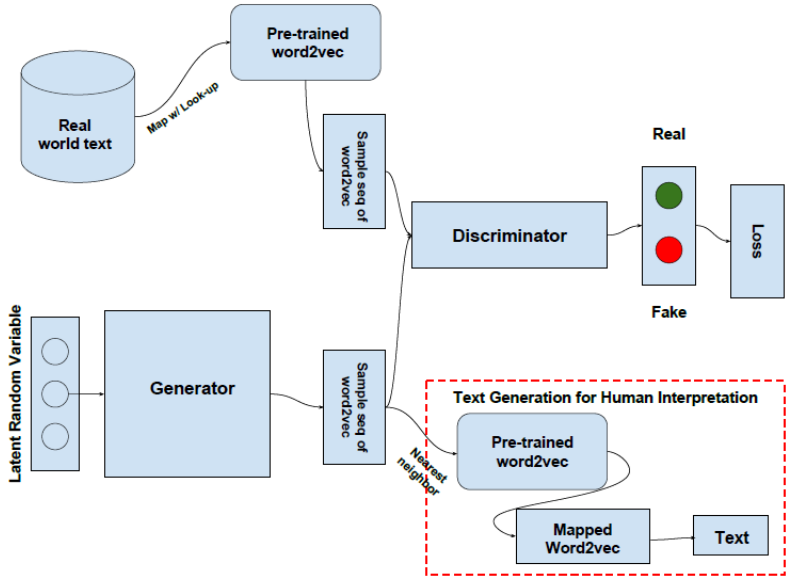
Link : https://github.com/adventure2165/ConditonalSeqGAN

Intensity Class	Generated Tweet
0	<user> <user> <user> what a idea #success id a backs i #kik me
0	me i the to fuming a free massage a backs a about hurt roac h was massage #worry
1	<user> <user> <user> was their want in was ass game to fee ls hell both heat <user> crashing
2	<user> i <user> told <user> glowing a most least in talk eat and just the mind and just this #revenge was #mad #upset pl ayers
3	3 <user> <user> so always trusted #anger in by blooded <us er> an absolute piece an affront <user> #furious the

Portfolio

GAN2VEC - 석사

- GAN2VEC 논문구현
  - Word2vec를 기반으로 한 GAN 문장생성 논문
  - 문장을 단어별로 분리
  - 분리된 단어들을 Word2vec 모델을 통해 벡터로 변환
  - 문장을 Word2Vec 벡터의 Stack화
  - DCGAN, InfoGAN을 기반으로 하여 학습 및 문장 생성
  - CSE-MU 데이터 사용
  - 논문:
    Generative Adversarial Networks for text using word2vec intermediaries



Link : https://github.com/adventure2165/GAN2vec

Portfolio

추후 개인 연구 주제

- GAN for NLP
  1. 자연어 데이터를 GAN에 적합하게 변환
     - 자연어는 discrete한 데이터라 미분이 불가능
     - 예시 – 펭귄과 타조는 표현 가능한데, 그 사이의 존재는 어떻게 표현하는가?
     - 이러한 특성으로 인해 GAN 적용이 상당히 어려움
     - 그렇다면, 자연어 문장을 Continuous하게 바꿀 수 있을까?
     - 예시 - Categorical 데이터를 Continuous하게 만드는 Gumbel Softmax
  2. GAN을 자연어에 맞게 개조
     - GAN을 자연어에 맞게 개조하는 시도는 다양함
     - 예시 – SeqGAN, TextKD-GAN…
     - 대신 강화학습을 적용하여 문제를 해결하거나 다양한 시도가 진행중
     - 이 외에 다른 기법에 대한 적용 및 연구 진행 목표

Thank You