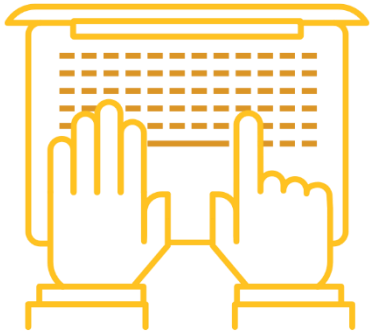


전문연구요원

포트폴리오

중앙대학교 대학원

전자상거래 및 인터넷 응용 연구실

김 준 호

Contents

자연어처리 관련 프로젝트 1

자연어처리 관련 프로젝트

국가 R&D 데이터 자연어 처리

컨셉 기반 빅데이터 큐레이션 기술을 이용한 R&D정보의 기술,산업화 분석 및 활용모델 연구

R&D 데이터(논문, 특허, 과제 이하 3P)로 부터 나오는 과학기술용어 간의 관계를 분석하여 컨셉 기반 용어 의미 확장 구축
R&D데이터에 대한 시계열/분야별 분석을 통해 용어들의 다양한 관점에서 용어에 대한 정보들을 제공할 수 있는 웹 서비스를 구현 및 동적식별체계 구축

사용 언어 및 자연어처리 알고리즘, 라이브러리

Java / Word2vec / Hannanum / LDA / TF-IDF

본인 담당 영역

데이터 수집 및 Word2vec을 활용한 용어 의미 확장 및 전반적인 동적식 별체계 알고리즘 구현

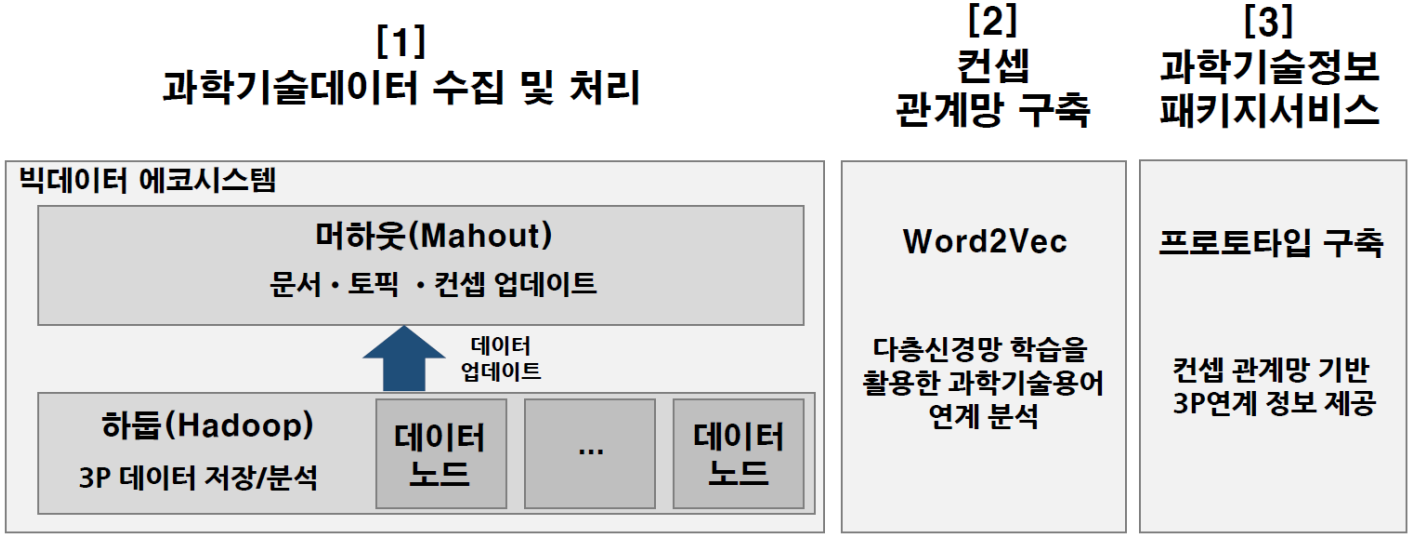
개발 기간

2015.04 ~ 2015. 12 (8개월)

참여 기관

KISTI (한국과학기술정보연구원), ETRI(한국전자통신연구원)

연구 과제의 흐름도



KeyWord	score
환자	0.48999488573057465
의료진	0.44072240323990747
진료	0.4277669235753152
병명	0.397889350108183
간호사	0.38388920086720324
검사명	0.35036120652255826
보호자	0.3480365885368003
처방	0.33660129957282753

KeyWord	score
다요소	0.48361841774967607
결정자	0.46561102372811686
결정권자	0.44711392171835784
정확	0.4201706477301294
다기준	0.40502095561305923
경영자	0.39252093250657133
대화자	0.3886219193246614
의사소통	0.38783841016385584
분석가	0.36764049303421464
합리	0.36385129112629755

[컨셉 확장 예시 그림]

'의사'에 대한 키워드를 2가지 컨셉으로 분류한 그림
(Doctor의 의사와 Decision의 의사로 컨셉 분류)

다음장에서 계속

자연어처리 관련 프로젝트

국가 R&D 데이터 자연어 처리 – 시계열 분석

Keyword (컨셉1)	Keyword (컨셉2)	Keyword (컨셉3)
애플	테블릿	폰
블랙베리	태블릿	어플
안드로이드	모바일	앱
앱스토어	넷북	마이크로소프트
구글	스마트	스마트패드
	안드로이드폰	스타일라이제이션
	셋톱박스	모바일폰
	휴대폰	인벤터
	노키아	어플리케이션
	블루투스	터치스크린

['아이폰'과 관련된 과학기술용어에 대한 질의 확장 결과]

컨셉1 = 제조사, 컨셉 2 = 아이폰 관련 카테고리, 컨셉 3 = 아이폰 관련 기술

2009-2010	2011-2012	2013-2014
	안드로이드	안드로이드
컴퓨터	손목	햅틱
제스처	인터페이스	휴대폰
컨트롤러	터치	디바이스
펜	구글	엔터테인먼트
비전	플랫폼	편리
수족관	유저	출시
인형	캡처	대중
애니메이션	관람차	전등
체험	플리케이션	보급

2009-2010	2011-2012	2013-2014
패널	패널	패널
용량식	용량식	
저항막	저항막	
정전식	정전식	
터치	윈드밀	두드림
터치키	광학식	정합키
전극배선	솔라셀	투명
게임기	스크린	커버글라스
터치부	표시체	시인
표시창	폴딩	감지셀

2009-2010	2011-2012	2013-2014
폰	폰	폰
터치	터치	터치
스마트폰	스마트폰	
	휴대폰	휴대폰
펜	게임기	스파클링
안드로이드	네비게이션	윈도우
리더기	태블릿	키패드
멀티터치	키보드	핸드폰
스타일러스	생동감	노트북
작업복	전자책	디스플레이

['아이폰 터치스크린'의 시계열 분석 결과 (순서대로 논문,특허,보고서)]

특허에서 '09~12년' 사이 '저항막'과 '정전기식'에 대한 공통 이슈들이 나타남. 이후 '두드림', '감지셀' 등의 확장 기능에 대한 용어들이 이슈로 나타남
= 즉, 터치스크린에 대한 기술이 바뀐것

다음장에서 계속

자연어처리 관련 프로젝트

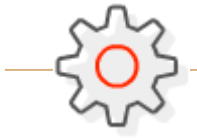
국가 R&D 데이터 자연어 처리 – 분야별 분석

논문	특허	보고서
소변	노화	기생충
피부병	질병	원인체
주사	국소	장염
모세	경구	바이러스
위염	항생체	결막염
확장증	외상	폐렴
신생아	지혜	감영증
증례	전염병	순환기
곰팡이	박테리아	면역학
외과	세균	요로

[질의어 '호흡기 감염'에 대한 질의 확장 결과]

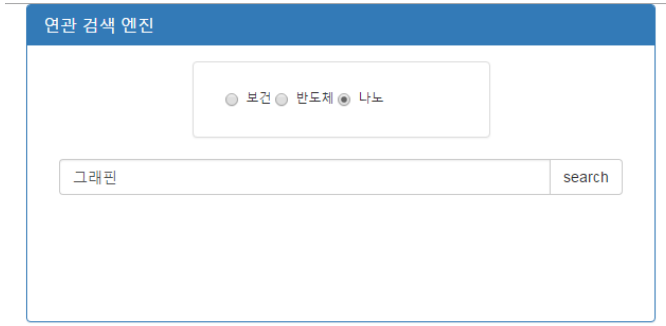
'호흡기 감염'과 관련하여 보고서/논문에서는 '호흡기 감염'과 관련된 증상을, 특허에서는 원인을 위주로 과학기술문서들이 작성됨

다음장에서 계속



자연어처리 관련 프로젝트

국가 R&D 데이터 자연어 처리 – 서비스 화면



단어 클러스터링1

KeyWord	score
탄소나노	0.6156603357510488
나노	0.6025743803432484
카본나노	0.4578274324787706
탄소나노튜브	0.4187742939615957
다중벽	0.38051342986363174

단어 클러스터링2

KeyWord	score
산화그래핀	0.5911343999930222
그래핀층	0.5800619474490502
나노선	0.5492343653713939
전자소자	0.4959700646398677
박막	0.48398457348443424
도핑	0.47588675779225187
나노점	0.4715361087501066
전도	0.46974148697606727
복합체	0.4636698346859004
레이어수	0.46278805031750736

단어 클러스터링3

KeyWord	score
박리법	0.5505110609076412
직성장	0.5020087491582538
그래핀막	0.45897745948569907
폴리메칠메타크	0.45338916347207187
나노와이어	0.4425242639814355
핀막	0.4321904750213929
산화아연	0.4286489649714113
동소체	0.4230881358542956
이종원소	0.417143675322848
폴러린	0.41254539837243487

시계열 분석 | 단계별 분석

이곳에 시계열 분석

논문 | 특허 | 연구보고서

논문 검색 결과

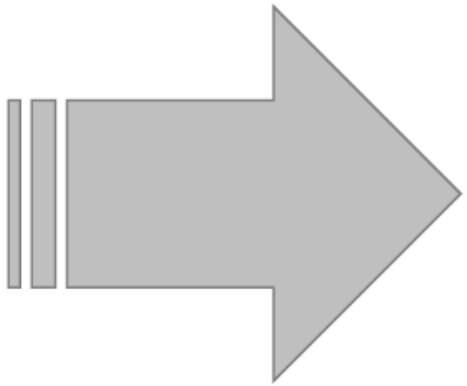
JAKO201523964821938	GPU 성능 저하 해결을 위한 내부 자원 활용/비활용 상태 분석 [조룡]	한국콘텐츠학회	한국콘텐츠학회논문지 = The Journal of the Korea Contents Association
JAKO201524453730188	리눅스 SSD caching mechanism 의 성능 비교 및 분석 [조룡]	한국스마트미디어학회	스마트미디어저널 = Smart media journal

다음장에서 계속

국가 R&D 데이터 자연어 처리 – 결과

연구 결과물

- 딥러닝 기반 과학기술용어 관계망 구축
- 시계열/단계별 과학기술용어 관계 분석
- 질의 확장 기반 과학기술문서 검색



기대효과

- 과학기술용어의 동적 식별 체계 구축
- 과학기술정보 흐름 추적 및 미래 기술 예측
- 사회문제 해결형 과학기술 패키지 정보 제공

자연어처리 관련 프로젝트 2

사용자의 법률 상담 서비스 편의 증진을 위한 법률 용어에 관한 연구

전문지식이 없는 일반인을 위해, 법률 용어와 일반 용어 간의 관계망을 형성하고 변환하는 연구
즉, 변호사에게 상담을 요청하기전에, 상담글을 작성하여 사용자와 유사한 사건이 있었는지 전문판례에서 자동적으로 찾아 사용자에게 보여줌 이를 통해 사용자는 변호사와 상담을 해야할지 고민을 덜어주고, 돈과 시간낭비를 최소화
유형화되어 있는 일상적인 법률문제의 경우에는 변호사와의 초기 상담에 의존하지 않고서도 국민 누구나 자신의 법률문제에 대한 1차적인 해법을 스스로 찾아 낼 수 있도록 함

사용 언어 및 자연어처리 알고리즘, 라이브러리
Java / Word2vec / Komoran / TF-IDF

본인 담당 영역
법률 용어와 일반 용어간의 관계망 형성 및 변환 알고리즘 구현

개발 기간
2016.04 ~ 2016.08 (4개월)

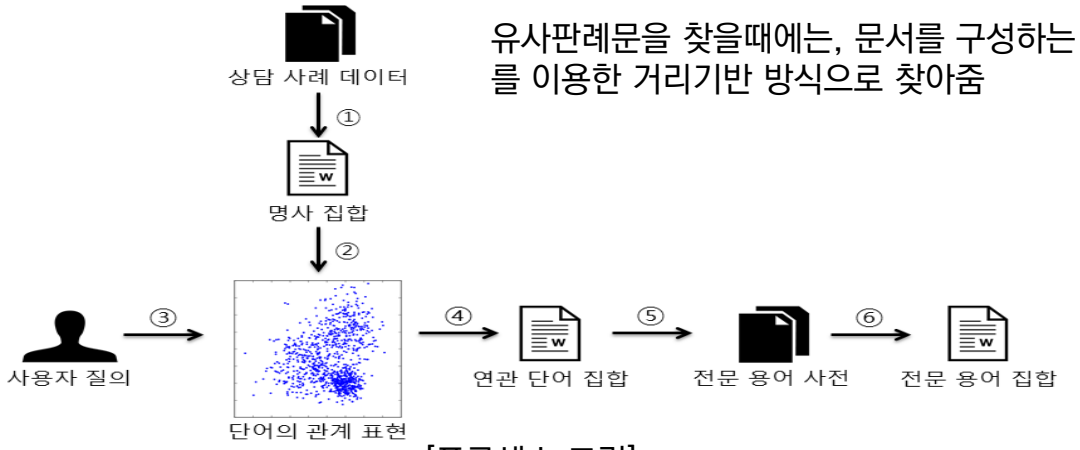
참여 기관
펄슨정보기술 ㈜ & 법률지능정보연구소

[용어 간의 관계망 형성 그림]

(위) 일반 용어, (아래) 법률 용어

설명 : 예를들어 사용자가 학교에서 왕따당하는 내용의 상담문이 있다면, 전문판례에서 이와 비슷한 내용의 판례를 찾아주고, 어떠한 판결이 나왔는지를 알려주어야함.
그러나, 전문판례에서는 일반인이 쓰는 용어를 쓰기 보다는, “학교폭력, 폭행, 언어폭력” 등의 법률 용어로 작성됨
따라서 이러한 일반인이 쓰는 용어를 자동으로 법률 용어로 바꿔준뒤, 유사한 판례문을 찾아줌

유사판례문을 찾을때에는, 문서를 구성하는 단어들의 가중치 평균 벡터를 이용한 거리기반 방식으로 찾아줌



[프로세스 그림]

다음장에서 계속

설명 : 단어망(일반용어와 법률용어가 혼재되어있는 단어망)에서 질의어(옥살이)와 가장 유사하다고 판단된 단어들을 나열한뒤, 이 리스트에서 판례에서 사용되지 않은 단어는 모두 제거
=> 이를 법률용어라고 판단

– '옥살이'와 가장 유사한 단어 벡터

키워드	유사도 값
감옥	0.507408
겁내	0.502993
감방	0.489679
술기운	0.487012
도둑놈	0.484532
시치미	0.481094
한국서부발전	0.480151
옥황상제	0.476802

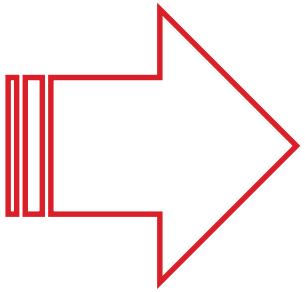
– 판례에서 사용되는 용어 선택 '옥살이'

키워드	유사도 값
감옥	0.507408
~~겁내~~	~~0.502993~~
~~감방~~	~~0.489679~~
~~술기운~~	~~0.487012~~
~~도둑놈~~	~~0.484532~~
~~시치미~~	~~0.481094~~
~~한국서부발전~~	~~0.480151~~
~~옥황상제~~	~~0.476802~~

다음장에서 계속

– ‘카톡’과 가장 유사한 단어 벡터

키워드	유사도 값
카카오톡	0.782731
캡쳐	0.709055
페북	0.697519
카카오스토리	0.667344
캡처	0.663798
문자	0.651157
메세지	0.649075
셀카	0.630966



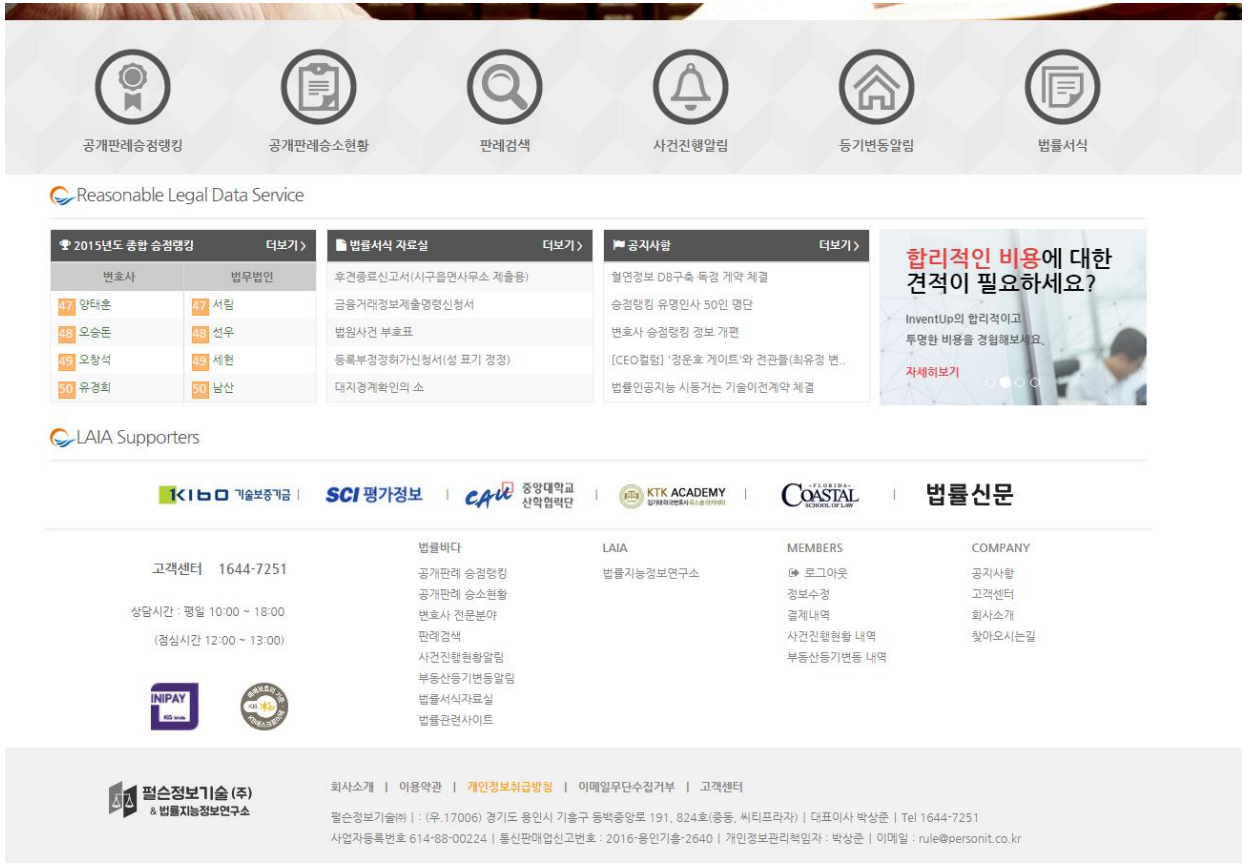
– 판례에서 사용되는 용어 선택 ‘카톡’

키워드	유사도 값
카카오톡	0.782731
~~캡쳐~~	~~0.709055~~
~~페북~~	~~0.697519~~
~~카카오스토리~~	~~0.667344~~
~~캡처~~	~~0.663798~~
문자	0.651157
메세지	0.649075
~~셀카~~	~~0.630966~~

다음장에서 계속

자연어처리 관련 프로젝트

법률 용어 자연어 처리



[서비스 화면]

자연어처리 관련 프로젝트 3

자연어처리 관련 프로젝트

Naver Keyword analysis

대학교별 커뮤니티의 주된 키워드 및 관련 주제 검색

대학교 커뮤니티인 페이스북의 “OOO 대나무숲”데이터를 크롤링하여, 각 대학교별 주된 관심사를 분석하고, 이를 웹으로 구현하는 프로젝트
Naver Campus Hack Day 해커톤에서 진행

사용 언어 및 자연어처리 알고리즘, 라이브러리
Java / Python / Word2vec / Komoran / TF-IDF / Nods.js

본인 담당 영역
데이터 크롤링 및 키워드 분석과 관련주제 검색 시스템 개발

개발 기간
2017.05.26 ~ 2017.05.27

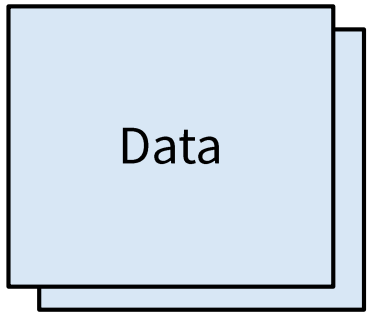
참여 기관
Naver

Kyunghee University Keyword & Words related to "연애"

다음장에서 계속

자연어처리 관련 프로젝트

Naver Keyword analysis



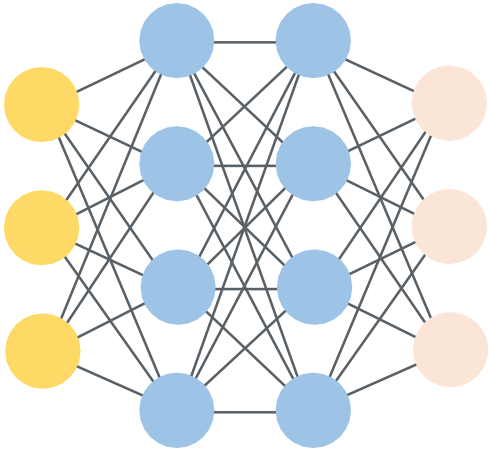
1. 각 대학교(서울대, 중앙대, 경희대)별 facebook 대나무숲 페이지의 글을 1만개씩 크롤링

2. Komoran 형태소 분석기를 통해 명사만 추출

3. TF-IDF를 통해 중요명사만 추출

이때 기준이 되는 TF-IDF는 전체적인 데이터 분포를 고려하여 결정함

기준 TF-IDF = $\log_2 average(tfidf)$



1. Keyword는 전처리를 모두 완료한 각 대학교별 데이터에서 DF를 통해 추출

2. 연관 주제는 Word2vec과 Cosine Similarity를 이용

Web

1. Node.js를 통해 Web으로 시각화

2. 적절한 변환을 통한 Word Cloud 구현

개인 프로젝트 1

- 본 프로젝트는 Python을 이용한 프로젝트 입니다. -
- 원하신다면 소스공개가 가능한 프로젝트임을 알려드립니다. -

개인 프로젝트

알고리즘 선택

동일 이미지 검색 프로그램 with Python

80,000장의 이미지에서 크롭된 이미지나, 다른 사이즈, 회전된 이미지를 빠른시간안에 찾아내는 프로젝트 (검색시간을 1초로 제한함)

개발 기간
2016.12 ~ 2017.01 (1개월)

https://github.com/taki0112/Image_similarity_with_elastic_search

개발 전, 고려한 알고리즘

SIFT
- feature 기반의 물체인식
- 정확률이 높음, 그러나 느리기 때문에 빅데이터 검색에 쓰이긴 어려움
- 특허(U.S. Patent 6,711,293)가 걸려있어, 상업적으로 이용하기 위해선 돈을 내야함
- 수정된 이미지(회전, 색상변경, 사이즈 변경 등)의 원본을 찾는데에는 좋음

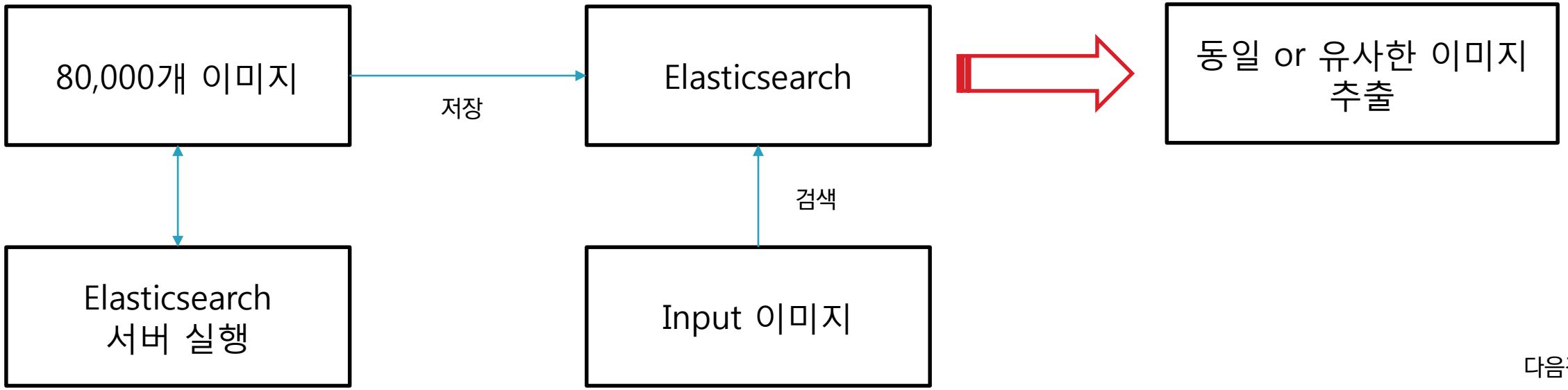
pHash
- Perceptual hash기반 방법으로, SHIFT보다 빠름
- rescaled 이미지의 원본이미지를 빠르게 찾을 순 있으나, 회전이나 위치가 심각하게 변경된것은 찾아낼 수 없음
- GPLv3 라이센스

Libpuzzle
- pHash와 방식이 비슷하나, 속도가 더 빠름
- rescaled, 회전된 이미지의 원본이미지를 찾을 수 있음
- BSD 라이센스(free하게 이용 가능)
- 관련 논문 : AN IMAGE SIGNATURE FOR ANY KIND OF IMAGE

다음장에서 계속

Elasticsearch
- 5.1.1 서버를 이용

Image-match
- Libpuzzle기반 파이썬 라이브러리
- Elasticsearch와 쉽게 연동하여 이용이 가능
- numpy+mkl, scipy, image-match, pillow, pycairo, cairocffi scikit-image 라이브러리 사전설치 필요



다음장에서 계속

개인 프로젝트

image-match안에서 수정이 필요한 소스코드

Goldberg.py

- 배열에 넣는 과정에서, int형으로 바꿔주어야함.. 소스코드 수정 필요

```
394         for i, x in enumerate(x_coords):        # not the fastest implementation
395             lower_x_lim = int(max([x - P/2, 0]))
396             upper_x_lim = int(min([lower_x_lim + P, image.shape[0]]))
397             for j, y in enumerate(y_coords):
398                 lower_y_lim = int(max([y - P/2, 0]))
399                 upper_y_lim = int(min([lower_y_lim + P, image.shape[1]]))
400
401                 avg_grey[i, j] = np.mean(image[lower_x_lim:upper_x_lim, lower_y_lim:upper_y_lim])  # no smoothing here as in the paper
402
403
404         return avg_grey
```

Signature_database_base.py

- distance_cutoff 값에 따라, 정확한 유사이미지가 나옴. 그러나 값이 너무 작아도 안좋음 0.45(default)가 제일 적당함

```
113     def __init__(self, k=16, N=63, n_grid=9,
114                  crop_percentile=(5, 95), distance_cutoff=0.45,
115                  *signature_args, **signature_kwargs):
116         """Set up storage scheme for images
117
```

다음장에서 계속

Elasticsearch

- elasticsearch.bat으로 서버를 실행하여, 오른쪽 그림의 cmd창처럼 나와야함

Pycharm

- 80,000개의 이미지를 아래의 주소에 넣은뒤, Elasticsearch 서버에 저장함

```
sample_image = "C:/goods_classify_11st_sample"
```

- 찾을 이미지는 crop_image변수에 저장
- 결과로 나오는 사진들은, 보기 쉽게 save_folder에 저장하게 둠
  - 이 폴더는 없을경우, 자동적으로 생성하도록 함

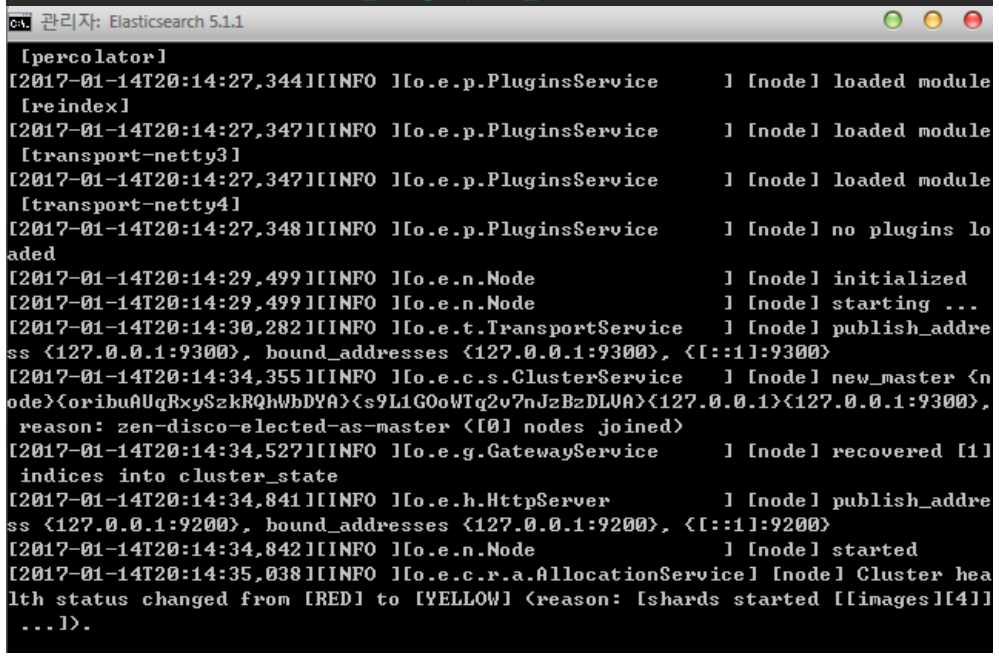
```
crop_image = "C:/test/2966_75.jpg" #원본이미지의 25% 사이즈

save_folder = "C:/test/original_folder" #결과이미지들을 저장할 폴더
folder_make(save_folder)
```

- 본 실험에서는 원본이미지, 25%이미지, 50%이미지, 75%이미지를 실험에 사용

- 0.3초 정도 검색시간이 걸림

- 이후, Flask란것을 공부해 웹 연동까지 코딩

```
for(path, dir, files) in os.walk(sample_image) :
    for filename in files :
        path_file = os.path.join(path,filename)
        ext = os.path.splitext(filename)[1] # 파일 확장자
        if ext in extensions :
            if(os.path.getsize(path_fil e) == 0) :
                pass
            else :
                ses.add_image(path_file)
```



개인 프로젝트 2

- 본 프로젝트는 Tensorflow를 이용한 프로젝트 입니다. -
- 원하신다면 소스공개가 가능한 프로젝트임을 알려드립니다. -

개인 프로젝트

시스템 수행 구조

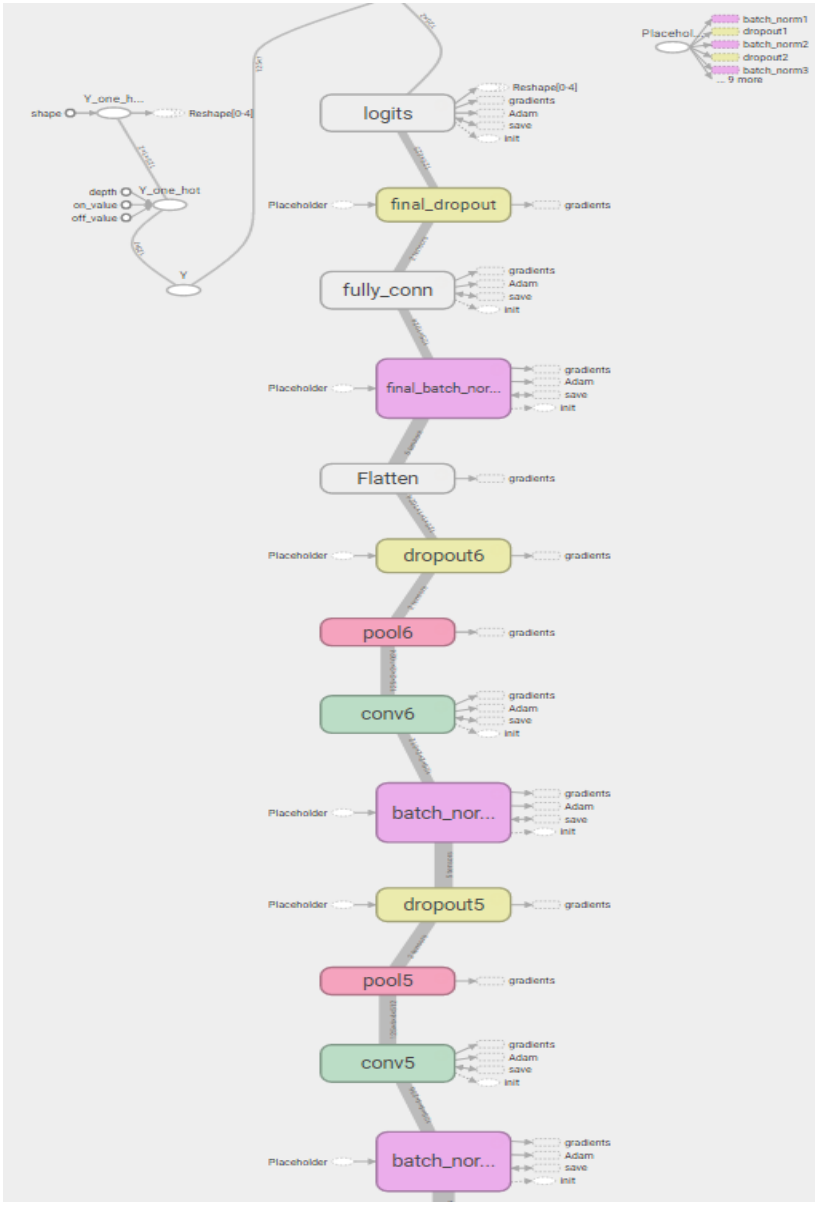
Image classification with kaggle data

Convolutional Neural Network를 통해
고양이와 개의 이미지를 분류하는 Tensorflow

개발 기간
2017.04 ~ Present

사용 기술
- Convolutional Neural Network
- Dropout
- Batch Normalization

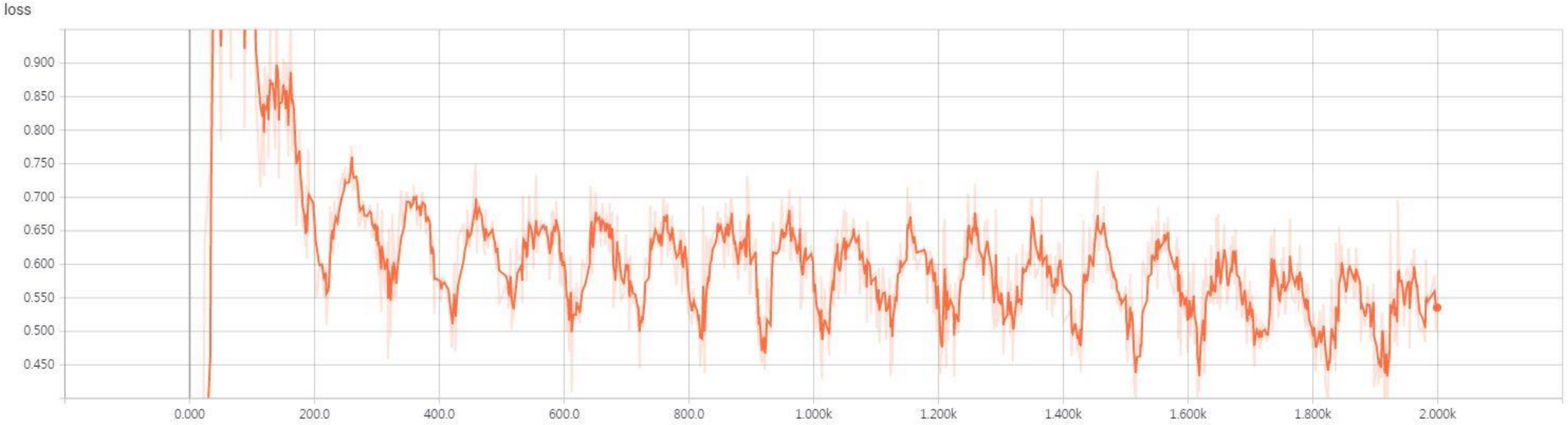
https://github.com/taki0112/Image_classification_CNN-Tensorflow



다음장에서 계속

결과

- 그림과 같이 loss는 잘 줄어들음
- 정확한 accuracy는 test data를 라벨링 하지 못하여, 아직 측정은 못함



개인 프로젝트 3

- 본 프로젝트는 Tensorflow를 이용한 프로젝트 입니다. -
- 스타트업을 하는 지인을 돕고자 알고리즘 연구차 진행한것으로, 소스코드 공개는 일부 가능합니다. -

개인 프로젝트

시스템 수행 구조

이상형 매칭 방법론

여러 직업, 연봉, 취미, 이상형, 자기소개서 등을 기반으로 사귈것같은 (결혼할것같은) 사람들을 매칭시키는 프로젝트

개발 기간
2017.04 ~ Present

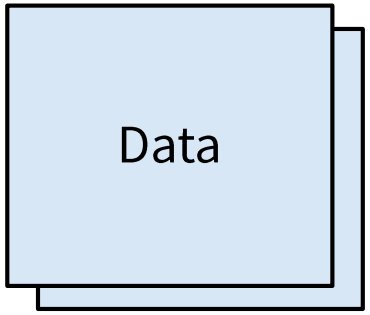
사용 기술
- Tensorflow
- Deep Neural Network
- Relu activation
- Adam Optimizer

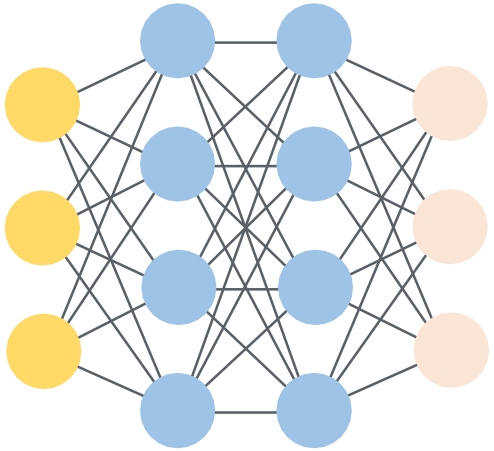
회원 데이터
회원번호
기본 인적정보(이름, 성별, 나이)
직업
재산
가족관계
연봉
학력
부모님 직업
건강상태
특기, 취미
성격검사 결과
자기소개서
이상형
연애경험

다음장에서 계속

개인 프로젝트

시스템 수행 구조





Hypothesis

남자 : 회원 데이터

여자 : 회원 데이터

결혼여부 데이터

weight

loss

vanishing problem

overfitting

training

Xavier initialization

Sigmoid cross entropy

Relu activation function

Drop out

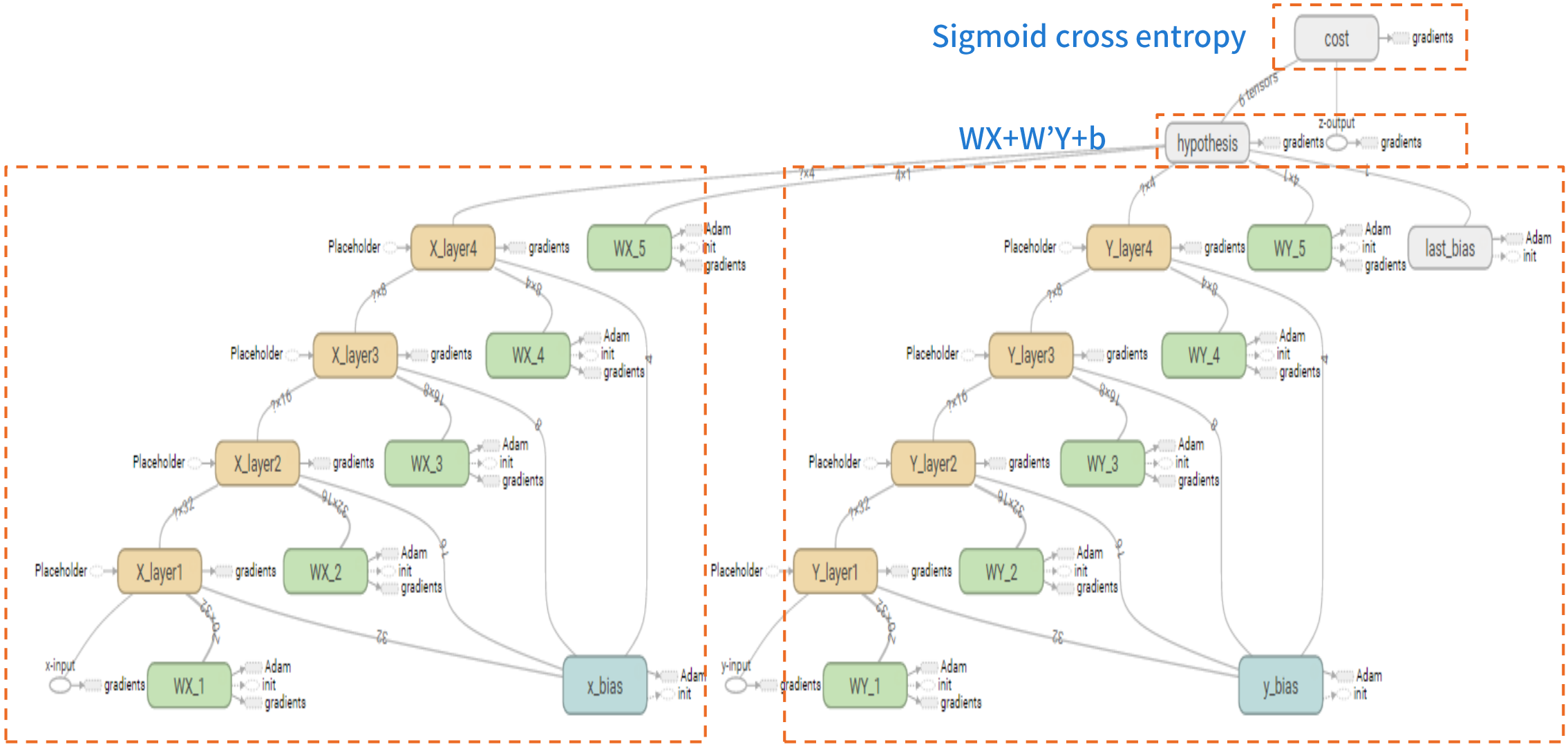
AdamOptimizer

$H = WX + W'Y + b$

Sigmoid(H)

0 ~ 1

다음장에서 계속


Sigmoid cross entropy
WX+W'Y+b
cost
gradients
6 tensors
hypothesis
z-output
gradients
gradients
Placeholder
X_layer4
gradients
WX_5
Adam
init
gradients
Placeholder
X_layer3
gradients
WX_4
Adam
init
gradients
Placeholder
X_layer2
gradients
WX_3
Adam
init
gradients
Placeholder
X_layer1
gradients
WX_2
Adam
init
gradients
x-input
gradients
WX_1
Adam
init
gradients
x_bias
Adam
init
Placeholder
Y_layer4
gradients
WY_5
Adam
init
gradients
last_bias
Adam
init
Placeholder
Y_layer3
gradients
WY_4
Adam
init
gradients
Placeholder
Y_layer2
gradients
WY_3
Adam
init
gradients
Placeholder
Y_layer1
gradients
WY_2
Adam
init
gradients
y-input
gradients
WY_1
Adam
init
gradients
y_bias
Adam
init

남자 X

여자 Y

석사 논문

석사 논문

짧은 글 문서 의도 파악

문서 의도 파악

- 문서의 의도를 나타내는 핵심 구절(Core passage)을 찾고, 이를 통해 문서간의 유사도를 계산
- 이를 위해 다양한 주제를 가지고 있는 네이버 지식인 데이터를 학습 데이터로 선정
- 이후, 유사한 문서가 얼마나 의도를 반영하는지 설문조사를 통해 평가

짧은 글에서의 핵심 구절 추출을 통한 문서의 의도 파악

사용 언어 및 알고리즘

Java / Word2Vec / TF-IDF / Twitter형태소 분석기 / Komoran

설명

최근 트위터, 페이스북, 커뮤니티 등과 같은 마이크로 블로깅 서비스(Microblogging services)의 인기로 인해 짧은 길이의 문서가 10억개씩 쌓이고 있음

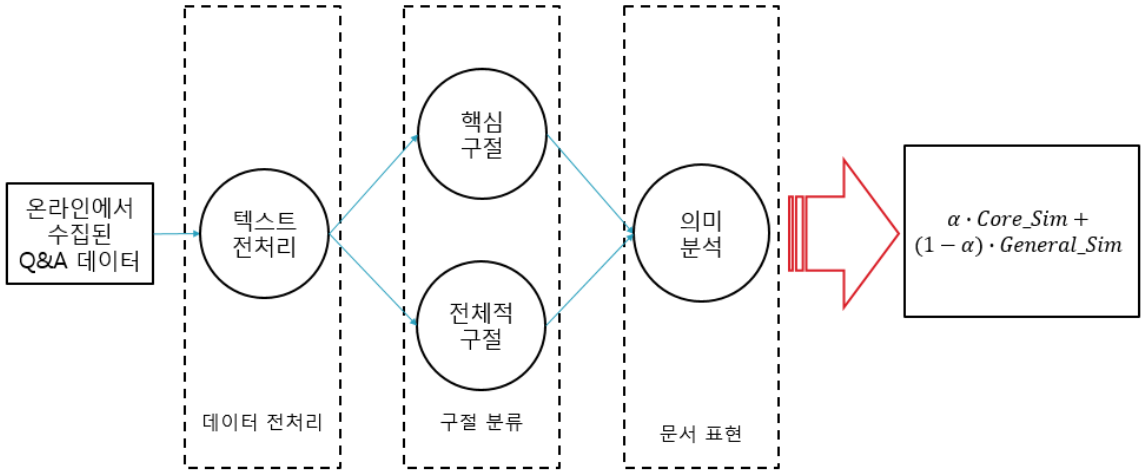
-> 정보가 많다보니, 다양한 주제에 관련된 문서에 대한 검색과 조직화를 어렵게 함

-> 따라서 질의문과 유사한 단어들이 사용된 질의를 찾는게 아닌, 의도에 맞는 유사 문서 결과를 제공해야할 필요성이 있음

-> 기존방법인 Paragraph2vec의 경우 단문에서 좋은 성능을 보이지 못함

-> 따라서 본 논문에서는 핵심구절(Core Passage)을 찾아, 의도를 파악해냄

-> 기존에 존재하는 여러 방법에 비해 정확도가 80% 향상이 됨



[제안방법의 구조도]

석사 논문 外 외부 발표 논문

석사 논문 外 외부 발표 논문

의사 결정 지원 시스템

The Online Community Based Decision Making Support System for Mitigating Biased Decision Making

온라인 집단(일베, 오유)의 의견 성향 분류(보수,진보)

사용 언어 및 알고리즘
Java / Sentiment Analysis / TF-IDF / Naïve bayes

본인 담당 영역
Sentiment Analysis

게재지
AIP Conference Proceedings
http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.4965400

설명
인터넷 공간의 의견의 분포는 편향되어 있을 수 있다. 이를 인지하지 못하면, 인터넷을 통한 의견 수렴 과정에서는 편향적으로 의견 분포를 받아들일 수 있다. 따라서 인터넷의 의견의 성향을 분류하고, 긍부정을 구분지어줌으로써 사용자의 의사 결정을 지원해 줌

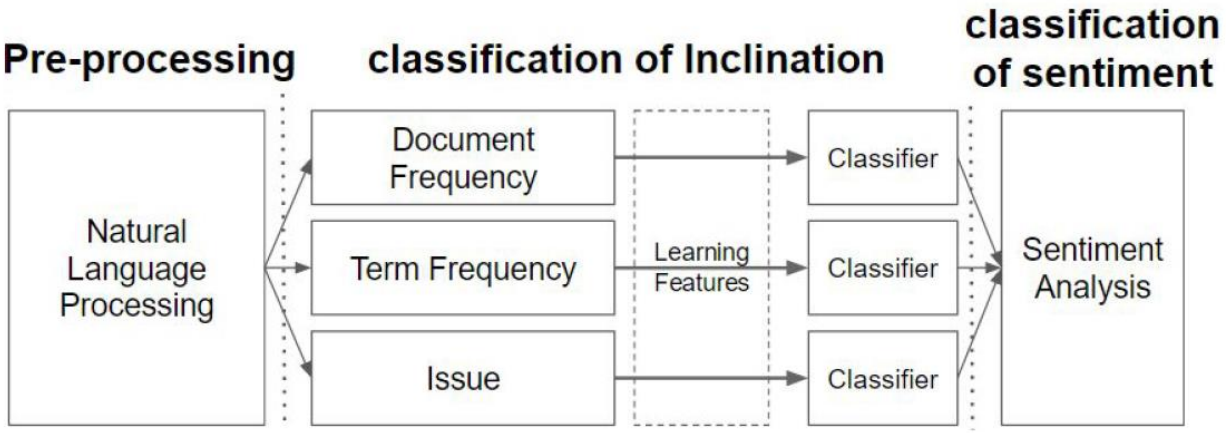


FIGURE 1. The structure of proposed method.

['대학원'에 대한 보수,진보 의견 결과]

Thank you