<https://steemit.com/busy/@anpigon/5s1aam>

<https://steemit.com/busy/@anpigon/2>

<https://anpigon.github.io/blog/busy/@anpigon/3/>

형태소 분석 해서 남성 여성을 체크하는 머신러닝 기능

나의 경우 gcp 클라우드를 이용, Ubuntu16.04, nvidia-docker로 tesla k80 그래픽카드를 사용할 수 있는 환경을 구축할 수 있었다. Docker 이미지를 처음에 어떻게 설치하는 지 몰라서 엄청 많은 이미지를 다운로드 받았었다.

현재 주피터 노트북을 이용해서 tensorflow를 실행할 수 있는 환경이다.

KoNLPy 모듈 설치 하는 방법

<http://konlpy.org/ko/latest/install/>

우분투 설치 방법을 따랐다.

1. Install dependencies
2. *# Install Java 1.8 or up*
3. $ sudo apt-get install g++ openjdk-8-jdk python3-dev python3-pip curl
4. Install KoNLPy
5. $ python3 -m pip install --upgrade pip
6. $ python3 -m pip install konlpy *# Python 3.x*
7. MeCab 설치하기 (*선택사항*)
8. $ sudo apt-get install curl git

$ bash <(curl -s https://raw.githubusercontent.com/konlpy/konlpy/master/scripts/mecab

은전한닢……..

bash <(curl -s https://raw.githubusercontent.com/konlpy/konlpy/master/scripts/mecab.sh)

우분투 자바 홈 경로 설정

<https://lsjsj92.tistory.com/394>

도커 사용방법

<https://subicura.com/2017/01/19/docker-guide-for-beginners-2.html>

nvidia-docker run -d --shm-size 2G -it --name nagyeong-docker \

PREFIX=/root/anaconda3

Nvidia-docker run nvidia-ng

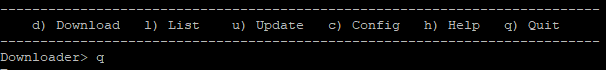
Nvidia-docker로 개발환경 세팅

<https://jybaek.tistory.com/791>

docker run nvidia-ng

$AVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64/bin/javac

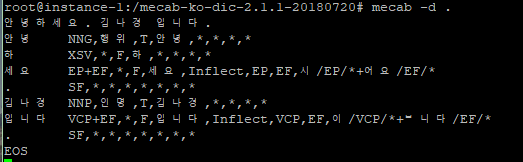
Nltk 다운로더 이용 전체 패키지 다운로드 완료



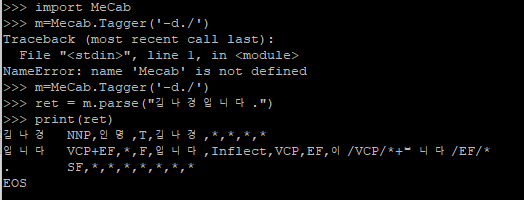
초보를 위한 도커

<https://subicura.com/2017/01/19/docker-guide-for-beginners-2.html>

bash <(curl -s https://raw.githubusercontent.com/konlpy/konlpy/master/scripts/mecab.sh)



은전한닢(mecab) 적용 완료



Python import 가능

나이브 베이즈 분류

스팸메일 필터링하는데 쓰이는 것

: 이것으로 성별을 알아내자

K-means 알고리즘

|  |
| --- |
| Review를 기준으로 k-means 알고리즘을 시행 |
|  |
|  |
|  |
|  |
| Label의 편향이 심하고, 우리가 원하는 맛에 대한 결과가 나오지 않는다.  맛과 관련 없는 배송 관련 리뷰 때문이라고 생각.  데이터를 맛에 대한 리뷰만 필터링하거나 다른 알고리즘을 찾아야 할 것. |

BERT를 사용한 모델링 과정

**: 관련 대량 코퍼스** -> **BERT** -> **분류를 원하는 데이터** -> **LSTM, CNN 등의 머신러닝 모델** -> **분류**

대량의 코퍼스를 Encoder가 임베딩하고(언어 모델링), 이를 전이하여 파인튜닝하고 Task를 수행합니다.(NLP Task)

**대량 코퍼스로 BERT 언어모델을 적용하고, BERT언어모델 출력에 추가적인 모델(RNN, CNN 등의 머신러닝 모델)을 쌓아 원하는 Task를 수행**

BERT란? 구글에서 개발한 자연어처리(NLP) 사전 훈련기술, 모든 자연어 처리 분야에서 좋은 성능을 내는 범용 언어 모델이다. Word2vec보다 단어 임베딩을 잘 할 수 있다고 한다.

자연어 데이터를 전처리할 때 씀.

* 구글이 미리 훈련해놓은 언어 모델. 3억개의 단어 (BookCorpus+Wikipedia Data)의 거대한 코퍼스를 정제하고, 임베딩하여 학습시킨 모델이다. 스스로 라벨을 만들고 준지도 학습을 수행한 모델임.

BERT 모델링

**관련 대량 코퍼스** -> **BERT** -> **분류를 원하는 데이터** -> **LSTM, CNN 등의 머신러닝 모델** -> **분류**

BERT 동작과정

Input =>

크롤러로 알아 낼 수 없는 것.

쿠팡 크롤러가 가장 양호한테 2가지가 문제다. => 사용자의 성별과 연령을 알기 힘들다는 것..

하지만, 이름을 닉네임으로 해서 끝 글자를 알아낼 수는 있었다. 그럼 끝 글자로 성별을 알아낼 수 있는 사람이라면..?

한글 이름은 한글로 알아내고, 영어이름은 영어로 예측할 수 있지 않을까? 대략적인 성별이라도..

이름으로 성별 알아내기 🡺 학습 데이터 만들자.

한국인의 이름 통계 주소 65065개의 이름

<https://koreanname.me/>

<https://koreanname.me/api/rank/2008/2020/1>

1~270 까지 요청 가능, => 원래 350까지 호출가능 but count : 0 빼니 270

요청 주소 1번에 100개씩 여, 남 순으로 들어있다. Json 방식

그럼, 총 35000개 정도의 남, 여 이름을 얻을 수 있을 것.

스크래핑을 사용할 것인가, 파싱을 할 것인가 알아서 선택하자.

<http://freesearch.pe.kr/archives/2459>

이름으로 성별구분 머신러닝 TEST

<http://www.ndsl.kr/ndsl/commons/util/ndslOriginalView.do?dbt=JAKO&cn=JAKO201305262617803&oCn=JAKO201305262617803&pageCode=PG11&journal=NJOU00423616>

나이브 베이지안 이용, 성별 구분 방법에 관한 논문

오레일리 파이썬으로 자연어처리 프로세싱 (논문과 위 블로그글의 코드는 여기서 파생된 것.)

<https://www.nltk.org/book/>

나이브 베이지안 관련

베이즈 추론이란?

베이즈 정리에서 나온 것.

조건부 확률 사건 B가 발생했을 때 사건 A가 발행할 확률

어떤 사건이 일어났을 때 다른 사건이 일어날 확률을 계산하는 정리.

**베이즈 정리는 새로운 정보에 대해 어떻게 대응하여 결과를 도출 해낼지를 알려주는 강력한 도구**

**나이브 베이즈 자연어 처리\_ 파이썬으로 자연어 처리 프로세싱**

<https://www.nltk.org/book/ch06.html>

A,e,I 로 끝나는 이름은 여자, k,o,r,s,t로 끝나는 이름은 남자이다.

분류모델을 만들자

labeled\_names = ([(name, 'male') for name in names.words('male.txt')] + [(name, 'female') for name in names.words('female.txt')])



Name값과, male 값을 한 쌍으로 하는 배열을 만드는 것. 그래서 결과가

(Male.txt의 값 , male) , (Male.txt의 값 , male) … 이렇게 나오는 것

Api가 json 형식으로 반환되기 때문에 가져오기 쉬웠음

* 27000\*2개의 남자, 여자 이름 txt파일을 얻을 수 있었다.

|  |
| --- |
|  |
|  |
|  |
| 나이브 베이지안 분류로 인한 정확도는 최대 73퍼센트 정도.. 어떻게 정확도를 높일 수 있을 까? |
| 하지만, 쿠팡에서 제공하는 이름은 처음, 끝 자리 글자 뿐이기 때문에 높은 정확도를 기대하기 어려울지도 모른다.. |
|  |
| 이름은 차츰 알아보기로 하고.. BERT와 데이터 전처리를 이용, 맛 리뷰를 걸러내고, 그 맛 리뷰만을 가지고, 필터링을 해야 한다.  그러려면 일단, 맛 리뷰인지 확인하는 분석 모델 1, 맛 리뷰 만을 가지고 필터링을 하는 모델 1개 => 총 2개의 모델이 필요 할 것.  우선 데이터를 많이 모아놓는 것이 중요하다. => 본격 작업 스케줄링 :    플라스크이용, api 서버 구현 test, json 확인 : 2일 (**월화**)  모든 사이트 크롤링 1번씩 돌려놓아 데이터를 준비한 후, : 최소 3일 (**수목금**)  <https://niceman.tistory.com/193?category=1009824a>  <https://niceman.tistory.com/101>  주말 : 덜 구현한 거 있으면 이 때 추가적으로 기간 맞춰보기   * 내가 생각한 항목이랑 웹 & 앱 담당한 팀원 항목이랑 일치하는 지 확인 + 회의 |
| BERT 모델 => 일부러 시간을 2주 정도 다른 것 보다 길게 잡음.  이유 : 논문도 좀 찾아보고 데이터 전처리와 파인튜닝 하는데 시간이 좀 걸릴 것 같다.  영어로 된 글이 많아 읽는데 시간이 걸리고, 학습 시키는데 시간이 많이 걸린다고 하니 2배로 시간을 배분 했다. BERT 파인 튜닝에 관한 글은 케글이나 블로그 글 등에서 많이 찾아 볼 수 있기 때문에 집중해서 한다면 가능 할 것. |
| 5.4~5.10 : 실제로 aws에서 파이썬 코드를 돌려보자. 환경 때문에 안되는 것이 있다면 맞춰보고 디버깅 해가면 될 것. |
| 5.11 이후 : 예상치 못한 오류나 기한을 못 맞출 때를 대비해서 완충 시간을 배정. |

<https://mc.ai/%eb%94%a5%eb%9f%ac%eb%8b%9d%ec%9c%bc%eb%a1%9c-%eb%8f%99%eb%84%a4%ec%83%9d%ed%99%9c-%ea%b2%8c%ec%8b%9c%ea%b8%80-%ed%95%84%ed%84%b0%eb%a7%81%ed%95%98%ea%b8%b0/>

<https://mc.ai/%EC%9E%90%EC%97%B0%EC%96%B4%EC%B2%98%EB%A6%ACnlp-17%EC%9D%BC%EC%B0%A8-%EB%A1%9C%EC%9D%B4%ED%84%B0-%EB%89%B4%EC%8A%A4-%EB%B6%84%EB%A5%98%ED%95%98%EA%B8%B0/>

로이터 뉴스 분류

BERT 순서

1. 텍스트 전처리 => 데이터 셋 준비 : pre-training(선행학습, 사전훈련, 전처리과정)

가중치와 편향을 초기화 시키는 방법이다. 레이블 된 데이터로 지도학습을 수행해야 함.(fine tuning)

<https://github.com/codertimo/BERT-pytorch>

파이토치를 사용, bert 모델을 사용할 수도 있고,

텐서플로우를 사용하는 경우, 구글 BERT 공식 코드를 사용할 수 있다.

* 토큰화 해서 word corpus로 만들어 주어야 한다.

1. Fine tuning

기존 학습 모델 기반으로 아키텍쳐를 나의 데이터에 맞게 변형하고 이미 학습된 모델 가중치로부터 학습을 업데이트 하는 것.

* 이미 존재하는 모델에 추가 데이터를 투입해서 파라미터를 업데이트하는 것이므로,
* 정교한 파라미터 이어야만 한다. 오버피팅이 일어나거나 전체 파라미터가 망가질 수 있기 때문

<https://lv99.tistory.com/34>

케라스 BERT 공식 깃허브

<https://github.com/kpe/bert-for-tf2>

|  |
| --- |
| 케라스 오픈채팅방에서 건진 정보 |
| BERT를 사용하여 한글로된 위키같은 데이터를 학습시키기 위해 어떤 전처리가 필요할까요? BERT를 학습시키기 위한 한글 파일의 포맷이 어떻게 되나요? 단순한 csv파일 포맷은 아닐것 같은데요..불용어나 ?.-:\*&^/~#!@같은 전처리는 제외하구요..  목적은 cdqa를 이용하여 사내 매뉴얼 등을 학습시켜서 간단한 QA시스템을 구현해보고자 하는것입니다. |
| 이 문서에서 1편에서 3편까지 보시면 힌트를 찾으실 수 있을 것 같습니다.  1편  <https://blog.pingpong.us/dialog-bert-1/>  2편  <https://blog.pingpong.us/tokenizer/>  3편  <https://blog.pingpong.us/dialog-bert-pretrain/> |

4/13

플라스크 프레임 워크 api 서버 만들기

<https://velog.io/@decody/%ED%8C%8C%EC%9D%B4%EC%8D%AC-Flask%EB%A1%9C-%EA%B0%84%EB%8B%A8-%EC%9B%B9%EC%84%9C%EB%B2%84-%EA%B5%AC%EB%8F%99%ED%95%98%EA%B8%B0>

SVM 알고리즘 << ??

파이썬 페이스북 크롤링 방법?