빅데이터의 이해

성공회대학교 열림교양대학 김종율 교수

빅데이터는 IOT, Cloud Computing, AI, 5G 와 같은 다양한 4차 산업의 미래 기술의 핵심입니다. 때문에 기술적으로 보자면 대부분의 IT 기술이 빅데이터 기술에 포함됩니다.

하지만 그 많은 기술들을 공부하기에는 시간이 부족하므로 빅데이터와 크게 연관된 기술을 위주로 강의를 진행하겠습니다.

먼저 빅데이터 기술, 데이터를 수집하여 우리에게 예쁘게 보여주는 과정에서의 기술에 대해 학습을 진행합니다.

첫 번째는 우리가 지난 시간에서 학습했던 내용인 데이터, 그 데이터를 어떻게 수집할 것인지에 대한 기술이 데이터 수집 기술입니다.

웹이나 일상 생활, 또는 업무, 공장 등에서의 다양한 환경에서의 다양한 유형의 데이터를 어떻게 수집할 것인가에 대한 내용이죠.

두 번째는 당연히 이 수집된 데이터를 어떻게 저장할 것인가에 대한 문제에 대한 해결책입니다.

어느 정도 많은 데이터, 한 가지 유형의 데이터 와 같은 경우는 저장하는데 아무 문제가 없습니다. 그런데 우리가 빅데이터의 3v, 4v, 5v 에 대해 배웠죠?

수십 테라 바이트 이상의 큰 데이터, 정형/비정형의 다양한 데이터, 실시간으로 쏟아지는 데이터를 어떻게 저장해야 하는지가 문제가 되었습니다.

세 번째는 이렇게 저장된 데이터를 가공하는 단계입니다.

이러한 가공은 3번째가 아니라 1,2 단계 사이에 위치하는 경우도 있습니다.

쌓여져 있는 데이터에서 필요한 항목을 선정하고, 분석에 활용하기 위해 가공하는 단계입니다.

네 번째는 이렇게 저장된 데이터를 어떻게 분석할 것인가에 대한 문제입니다.

스몰데이터 시대에는 아무 문제가 없었지만, 이렇게 크게 증가된 데이터를 어떻게 처리해야 할까요? 거기에서 파생된 기술이 빅데이터 분석 기술들입니다.

하둡, 스파크 와 같은 큰 데이터를 분석하기 위해서 분산처리하는 기술들이 발전하게 되었죠.

마지막 다섯 번째는 분석된 데이터를 사용자에게 시각화 하여 보여주는 기술입니다.

예를 들어 결과값을 이렇게 보여준다면, 어떻게 될까요?

```
_________________________________________________________________
embedding_2 (Embedding)      (None, None, 30)          175770
_________________________________________________________________
lstm_2 (LSTM)                (None, 128)               81408
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 129
=================================================================
Total params: 257,307
Trainable params: 257,307
Non-trainable params: 0
_________________________________________________________________
```

```
model_LSTM = load_model('best_model_lstm.h5')

model_LSTM.evaluate(test_X, test_Y)
```

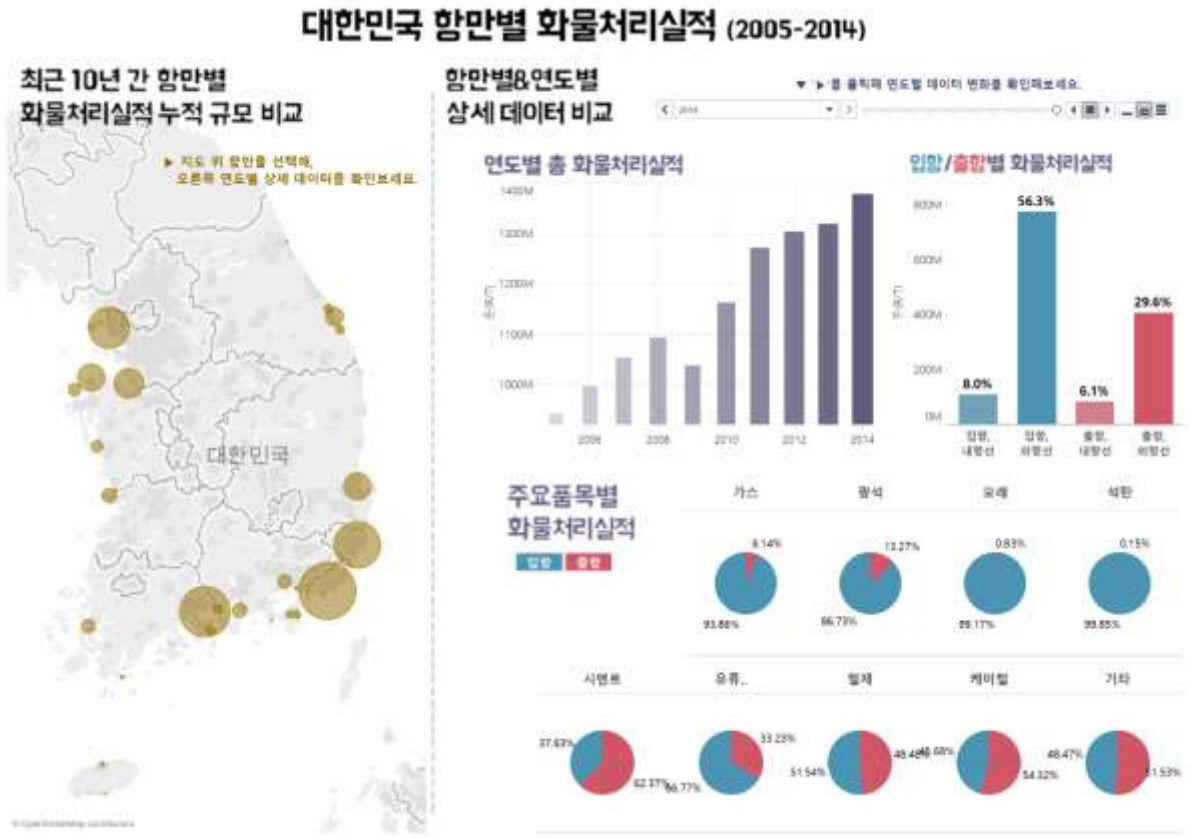
```
WARNING:tensorflow:Error in loading the saved optimizer state. As a result, your model is starting with a freshly in
8/8 [==============================] - 0s 10ms/step - loss: 0.4493 - acc: 0.8565
[0.4493107199668884, 0.856521725654602]
```

혼나겠죠?

그래서 사용자들이 이해하기 쉽게 시각화 해서 보내 주는 기술이 필요하게 됩니다.

예를 들어서 이렇게요.

훨씬 이해하기 쉽겠죠?



우선 첫 번째로 빅데이터 수집, 저장, 가공, 분석, 시각화 기술에 대해 알아보겠습니다.

1. 데이터 수집

데이터 수집 기술도 무척 많고 상용화된 프로그램도 많습니다.

우리는 우선 이러한 도구를 사용하지 않고 데이터를 수집하는 방법을 알아보겠습니다.

도구도 좋긴 하지만, 솔직히 일반적인 환경에서는 크게 의미가 없습니다.

오늘날 대부분의 서비스들이 웹에서 제공됩니다.

기업에서 직원을 채용하기 위한 구인 공고, 정부 등의 각종 공고나 문서, 예를 들어 각 구청(시/도)별 코로나 현황도 웹에서 제공되죠.

웹에는 많은 활용할 수 있는 많은 데이터가 있습니다.

예를 들어, 태양열 발전을 봅시다. IOT 센서에서 광량, 온도, 습도, 발전량 등의 데이터를 수집 했

다고 해도 미래를 예측하기 위해선 더 많은 정보가 필요하겠죠.

이러한 것들을 정부의 공공데이터 또는 웹에서 수집하여 기존의 데이터와 분석한다면 미래를 예측하는데 훨씬 효율적일 것입니다.

웹에서의 데이터 수집 방법은 대표적으로 크롤링과 스크래핑이 있습니다.

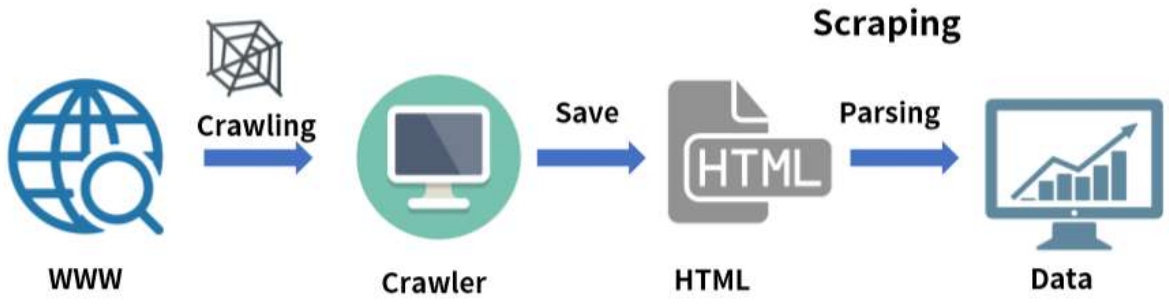
크롤링(Crawling)은 다양한 정보를 활용하기 쉽게 수집하는 행위이며, 크롤러는 크롤링을 하는 프로그램을 말합니다.

우리가 웹 크롤링(Web crawling)이라고 하면 웹에서 링크를 타고 다니면서 웹페이지를 수집하는 행위이죠.

예를 들어 구글봇이 전 세계의 웹사이트를 수집하여 검색 서비스를 제공하는 것이 크롤링이라고 볼 수 있습니다.

그러면 웹스크래핑(Web Scraping)은 무엇일까요? 웹사이트에서 정보를 추출하는 행위입니다.

예를 들어, 쇼핑몰 가격 비교를 위해 각 쇼핑몰 상품 페이지에서 상품 이름, 가격 등을 추출하는 행위가 스크래핑이죠.



더 쉽게 말하자면,

봇과 같이 계속 움직이면서 데이터를 수집하여 업데이트 하는 방식 → 크롤링

사용자가 수집 프로그램을 실행시키거나 배치 작업을 통해(예를 들어 하루 2번) 수집을 하는 방법은 스크래핑이라고 생각하면 됩니다.

그런데 현업에서도 대부분 스크래핑인데 크롤링이랍 부르는 경우가 많거든요. 그냥 비슷한 느낌이라고 생각해도 됩니다.

웹서비스 구조 설명

예제(학식) 설명

자율주행 자동차도 무척 많은 센서들을 통해 동작됩니다.

자동차가 스스로 자율 주행을 하려면 라이다, 카메라, 레이더와 같은 다양한 센서들이 앞의 차와의 간격, 옆 차, 뒷 차, 차선 등의 주행 환경을 스스로 인식하면서 주행하게 됩니다.

또한 도로의 환경, 공사 중, 사고 현황 등 끊임없이 정보를 주고받아야 합니다.

강의 자료에 있는 링크는 한 번씩 읽어 보시기 바래요

눈 셋 달린 자율주행차, 눈 하나는 사라진다?

https://biz.chosun.com/site/data/html_dir/2020/01/13/2020011303383.html

60년 만에 현실된 자율주행, 5G 기반 C-ITS로 완성될까?

https://www.e4ds.com/sub_view.asp?ch=11&t=0&idx=11017

라이다 수집 영상

https://www.youtube.com/watch?v=yKOdrA4_-1g

5G 추가 내용

4차 산업시대의 다양한 기술은 5G기술과 큰 연관이 있습니다.

빅데이터에서는 초고속 보다는 수 많은 Device와 연결 가능한 초연결과 저 지연의(1ms)의 5G 기술이 중요합니다.