2021-2학기 빅데이터 프로그래밍 Term Project:

빅데이터 분석 시스템 설계 및 개발

1. 개요

Hadoop 기반의 빅데이터 처리 기술을 활용하여 대용량의 데이터를 수집, 분석, 활용 할 수 있는 시스템을 개발하라. 필요에 따라 Hadoop, HDFS, MapReduce, Spark, Pig, Hive, Kafka, Flume, Sqoop, RDBMS, NoSQL 등 다양한 도구나 서비스를 활용가능하다.

2. 제출 내용

- 프로젝트 제안서
  - 문제 정의: 프로젝트의 명확한 주제와 목표
  - 실행계획
    - 데이터 수집 방법
    - 데이터 분석 방법
    - 시스템 아키텍처
- 최종 보고서
  - 문제 정의: 프로젝트의 명확한 주제와 목표
  - 시스템 아키텍처: 전체적인 SW 스택의 구성과 데이터의 흐름 (Pipeline), 저장 방법 포멧 등에 대한 구체적인 설계내용. 해당 기술 스택을 사용한 이유나 아키텍처의 설계 이유 등을 명확히 설명해야 함
  - 데이터 수집 방법: 데이터 수집을 위한 SW의 소스 코드 및 방법, 전략 등 기술
  - 데이터 분석 방법: 데이터 분석을 위한 절차, SW의 소스 코드 및 전략 등 기술, 수집한 데이터의 전처리 내용
  - 데이터 분석 결과: 데이터 분석으로 얻어진 결론, 흥미로운 분석 내용 등 기술
  - 추가적인 확장 가능성: 추가적인 확장 아이디어나 가치
- 발표자료: 발표 동영상 제출. 길이 10분(Youtube 링크로 공유)
- 소스코드 및 관련 자료는 github을 통해서 공유하는 것을 기본으로 합니다.

3. 프로젝트 주제 예

- 스포츠 데이터 수집 및 분석: 야구, 축구, 농구, LoL ...
- 자산 정보 데이터 수집 및 분석: 주가정보, 기업정보, 공시정보, 암호화폐, 부동산 등
- 공공데이터: http://www.data.go.kr , http://www.data.gov
- 상품정보: 가격, 품목

5. 주의사항

- 데이터 수집 단계가 반드시 필요하며, 이를 통하여 충분한(수백KB~GB) 데이터를 확보하여야 합니다.

- 모든 단계는 최대한 자동화되어 이후에도 동일한 데이터 분석이 쉽게 이루어질 수 있도록 해야 합니다.
- Hadoop 등의 빅데이터 관련 기술을 사용하여야 합니다.
- 인터넷, 서적 등에서 있는 내용을 그대로 따라하기만 하면 안되며, 차별점이 있어야 합니다.
- 참고한 글이나 자료가 있다면 반드시 출처를 밝혀야 합니다.

6. 채점 기준

- 데이터를 수집하는 방법의 자동화
- 데이터의 전처리 및 저장 방법
- 데이터 분석 방법
- 빅데이터로 확장 가능성
- 수집한 데이터의 양이나 난이도
- 배점: 프로젝트 완성도 30점, 발표 20점, 보고서 50점

7. 참고 자료:

- NLTK: Python NLP Toolkit, https://www.nltk.org/
- KoNLPy: https://konlpy-ko.readthedocs.io/ko/v0.4.3/start/
- Scrapping: https://www.slideshare.net/lucypark/the-beginners-guide-to-54279917
- Pandas: https://pandas.pydata.org/
- 파이썬으로 영어와 한국어 텍스트 다루기: https://www.lucypark.kr/courses/2015-dm/text-mining.html
- Natural Language Processing and Big Data: Using NLTK and Hadoop - Talk Overview: http://www.datacommunitydc.org/blog/2013/05/nlp-and-big-data-using-nltk-and-hadoop-talk-overview
- Python's Natural Language Took Kit (NLTK) and Hadoop - Part 3: http://www.datacommunitydc.org/blog/2013/05/nltk-hadoop
- News Authors Personality Detection - End-to-end data science & engineering platform with Nifi, Kafka, SAM, Druid, Superset, Hive, Zeppelin, Spark, https://community.hortonworks.com/articles/214051/news-authors-personality-detection-end-to-end-data.html
- PYTHON을 이용한 데이터 수집(Crawling): https://ericnjennifer.github.io/python_crawling
- 나만의 웹 크롤러 만들기: https://beomi.github.io/gb-crawling/
- taspinar/twitterscraper: https://github.com/taspinar/twitterscraper
- https://github.com/YBIGTA/Deep_learning/blob/master/Scratch%20ML/%EC%99%95%EC%B4%88%EC%A7%9C%EC%9D%98%2B%EB%8D%B0%EC%9D%B4%ED%84%B0%2B%EC%A0%95%EB%B3%B5%EA%B8%B0%2B%231%2B%EB%8B%A4%EC%A7%9C%EA%B3%A0%EC%A7%9C%2B%ED%81%AC%EB%A1%A4%EB%A7%81.md
- (Python) 부동산 데이터 수집, 크롤러 제작하기: http://jangwon.me/python/2018/02/15/(Python)-%EB%B6%80%EB%8F%99%EC%82%B0-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%88%98%EC%A7%91-%ED%81%AC%EB%A1%A4%EB%9F%AC-%EB%A7%8C%EB%93%A4%EA%B8%B0/
- 오픈 API를 통한 공공데이터 수집: https://medium.com/@whj2013123218/%EC%98%A4%ED%94%88-api%EB%A5%BC-%ED%86%B5%ED%95%9C-%EA%B3%B5%EA%B3%B5%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%88%98%EC%A7%91-e1dd0ad203b6
- Pandas를 이용한 Naver금융에서 주식데이터 가져오기: 출처: http://excelsior-cjh.tistory.com/109?category=975542 [EXCELSIOR]

- Using Python, what's the best way to get stock data?: https://www.quora.com/Using-Python-whats-the-best-way-to-get-stock-data
- Stock Analysis in Python: https://towardsdatascience.com/stock-analysis-in-python-a0054e2c1a4c