기업연계 멘토링 진행

차수명	1차	일자/시간	2021.10.19/10am
팀장명	김예찬	기업명	신한은행
주제	프로젝트 주제 및 방향성 검토		

세부 내용

[참여인원]

5명 전원 참석(김윤서, 김예찬, 방지환, 이규정, 정희진)

[주요 진행 내용]

- 프로젝트 주제 검토
- 사용하는 툴과 라이브러리
- 데이터 시각화 방법 등

[질문/답변 내용]

1. **명세를 해석한 결과 은행용 데이터 시각화 페이지 제작이 맞는지?**

→ 은행용 데이터 한정이 아닌, 빅데이터(larget data set)이 기준

→ 처음부터 분산 처리를 염두에 두고 주제를 잡았음

→ 이를 Spark 로 많이 하게 되는데 이번에 배우면 좋을텐데 여건이 안 될 경우 굳이 spark 분산처리하지 않아도 상관은 없습니다.

→ 하지만 프로젝트 메리트가 크게 있으려면 pyspark 무리가 없다면 100 기가정도 되는 large data 의 feature 를 도출하는 것은 어떨까?

2. Input 데이터가 주어지는지?
    - 주어진다면?
        - 어떤 데이터 형식, 내용인가?
    - 안 주어진다면?
        - csv 파일 같은 새로운 데이터를 넣으면 바로 시각화 해서 보여주는 것인지?
        - 데이터 분류가 필요한지? (부서 별로 터치하면 안되는 자료들에 따른 분류)
        - 이미 DB 에 저장되어 있는 데이터를 시각화 하는 건지?

→ 처음 데이터 제공하려고 했다가 절차 복잡함

kaggle 에서 공개하는 open data 를 dummy data 로 사용하면 좋을듯 (200-300 메가 용량의 데이터 혹은 값을 바꿔가면서 복제해서 input)

3. 분산 처리 관련 질문
    - *Feature 추출과 분산 처리 중 하나만 선택해야 된다면?*
    - 후순위로 둬도 되는지?
    - spark 를 꼭 사용해야 하는지?

→ Backend 에서 분산처리 되든 안되든 front-end 는 바뀔 것이 없으니 시각화를 먼저 학습해보자

→ 일반 data set 에서 특징 추출하는 것을 목적으로!

spark 추천 이유: backend 분산 처리에 가장 쉽고 효율이 높다. 다른 분산 처리하려면 오히려 어렵다.

4. 데이터셋 feature 추출
    - feature 는 어떤 것들을 추출해야 하는지

→ outlier 데이터 솎아내는 것이 중요 !!

(수치형 - 단순하게 최대 최소 평균뿐 아니라 Q1, Q2, Q3 값도 보여주고 상, 하위 얼마만큼을 outlier 로 만들지 정하기. 그 값들의 분호는 어떻다를 mini chart 로 보여주면 좋을 것 같음)

5. **데이터를 어떻게 시각화 하고 어느 정도까지 보여주는 게 좋을까요?**
    - (현재 행내에서 데이터분석가들이 사용하는 방법?)
    - 텍스트 데이터와 숫자형 데이터와 같이 다양한 타입에 따른 방법

→ 행내에서 사용하는 툴 (jupyter 로 전처리하고 확인하는 작업의 반복)

→ 시각화에서는 dataset 선택 후 거기 포함된 열을 쭉 보여주고 열 선택하면 그 선택된 열에 대한 type 및 분포 정보를 chart 로 만들어주는 것!

(ex 수치형 - 주식처럼 candle chart 도 좋고 여러 다양한 차트를 조합할 수 있다.)

6. **D3 chart 혹은 apexcharts 와 같은 추천해주실만한 시각화 라이브러리가 있나요?**
    - 은행 내 사용하는 특정 라이브러리가 존재하는지?

→ 정해져있진 않고.. 평소에는 e-chart ??

https://echarts.apache.org/en/index.html

7. **백엔드 스프링 대신 장고 써도 되는지?**

→ 상관없다..... 제약이 없음.....

<추가 질문>

Q 여러 데이터 중 특정 형식 지정이 되는지 or 어떤 데이터도 가능하게 해야 하는지

A 베스트는 parse 가 있어서 json, xml 등 데이터 형태를 지정할 수 있으면 좋은데 parse 만드는것도 어려울거야

일단 csv 형태로 생각하고 딜리미터를 선택하는 걸로 하자

Q csv 파일 염두에 두라고 하셨는데 그 안에 column 의 길이가 달라도 시각화가 가능한지?

A 같은 포맷의 데이터를 다루어야 하는 것

데이터 한 번 분석 후 저장 → 다시 볼 수 있어야 함

1 기가 미만의 데이터는 그럴 필요도 없지만 지금 바로 들어갈지 분산처리를 할지 분류해야 한다.

Q 금융 데이터를 선별해야 하는지

A 수치형이나 범주형 데이터가 많은 자료를 사용하자

Q 명세에 front-end 에 java-script 밖에 못 쓰는지?

A 뭘 써도 상관없다.

Q wire frame 점검

A 놓친 부분은 없되, 데이터 분석화면에서 min, max 등 팩트 기반 수치들과 outlier 분리한 자료 토대로 나온 수치를 분류하면 좋을 것 같다.

[멘토 전달사항]

- 프로젝트의 목표는 data set의 각 column들을 분석해서 outlier 선별하기.
- 유효 데이터 필터 + 필터 하기 전 후로 각각 min, max 같은 통계치 뽑아내기가 중요하다.

[건의사항]

- 신한은행 측과 연락을 이어갈 수 있는 소통 채널을 만들어주세요

> 이미 MM 채널이 있습니다.