BiWeekly meetings (2 November-15 November biweek)

Agenda

First Week

- Team will be informed about Latest version of the scraper.
- Ml/Dl needed input structure for the scraper will be discussed.
- Database table structure by the needs of the ML/DL will be discussed.
- Design of the scraping spiders will be determined. ### Second Week
- Mail from supervisor will be discussed
- Ml/Dl needed input structure for the scraper will be discussed.
- Database table structure by the needs of the ML/DL will be discussed.
- Design of the scraping spiders will be determined. ## Summary ### First Week
- Team will be informed about Latest version of the scraper.
  - 2 types of parsers are currently used in our spider
  - first for storing taking urls
  - second for taking full information
  - Some of the URLs are chosen to be not scraped because of the robots txt(As we know its not legally binding but we chose to not to over the users for now so if we dont decide to we dont need to look at legality)
  - Takes about 5 minutes to scrape 13xx products
- Ml/Dl needed input structure for the scraper will be discussed.
  - Trendyol varying factors
    - There are new items added and removed each day(how to handle?)
    - price and star count are time dependant features
    - There are a lot of static features
  - How it is wanted to be handled
    - We will scrape N urls and only scrape those N urls in every price scraping interval
    - In the case of an product's discontinuation of sales the item's price information will be given as NULL or NONE
    - Our spider needs to be decomposed into 3 spiders to handle these needs
      - First spider will collect only URLs(will be used once)
      - Second spider will collect static features(like name,colour,length,OS,camera of a smartphone)(will be used once)
      - Third spider will collect dynamic time-varied features(price,number of stars)
- Database table structure by the needs of the ML/DL will be discussed
  - There will be 2 tables that corresponds to 2nd and 3rd spiders
    - First table(Second Spider) will store static features(like name,colour,length,OS,camera of a smartphone)
      - Second table will not be updated after its creation

        - A lot of the features can be NULL since trendyol does not show all features
    - Second table(Third Spider) will collect dynamic time-varied features(date,price,number of stars) ### Second Week
- Hosting of web server
  - We were told that we could use our department's web server hosting services. We discussed whether we would need such a server since we are planning to deploy a web server. The reason we are considering not to use our department's web server is that member of our team who are responsible for our database and web server hasn't reached a conclusion about this topic yet. We may also want to consider using hosting services which we are more familiar and more mainstream.
- Mail from supervisor will be discussed
  - We have written our first web scraper for Trendyol but as our supervisored informed us STOCKMOUNT (the firm we are working with) has more data on sites n11 and Hepsiburada, we might need to write two scrapers for n11 and HepsiBurada
  - Stockmount also informed us that they have yet to decided on a product for our ML/DL models but they have informed us cosmetics are one of products groups that have the highest traffic.
  - Since our partner firm hasn't decided on product we will need to revise our web scraper in the future
  - Since different products may need different models we will also need to revise our ML/DL models in the future
  - We have also decided to split our scraper to 3 parts so that is decoupled, where there parts are collectors of URLs, static data and time-varied data
- We have prepared our retrospective document 1
  - Since the document is self-explanatory explaining it here would be redundant
- Scraper architecture
  - We decided to continue with the already decided 3 spider architecture on Trendyol spider and if there are any new spiders we would do the same.
  - We will modify our spider's output structure after STOCKMOUNT confirms the product they want.
  - For the next 2 items we have decided to talk to our instructor to clarify the changed needs according to STOCKMOUNT and only then continue with below items(There is a high chance only one of them will be implemented):
    - New n11 spider will be created
    - New HepsiBurada spider will be created