Mining the Insights of Stack Overflow Developer Survey

Team members:

Mohammad Imrul Jubair, CSCI 5502

Mohsena Ashraf, CSCI 5502

Cornelius Onimisi Adejoro, CSCI 5502

Motivation – Problem Domain

- High demand for programming or software development jobs.

- Need to understand the market!
  - For both applicants and employer

- But market is huge. **How to analyze?**
- Looking at a sample >> Online Community >> ***Stack Overflow***
  - **Stack Overflow Developer Survey (SODS)**

Motivation – what is SODS

- Survey on users to find how they learn and level up, which tools they are using, and what they want.

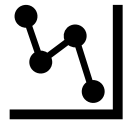
  Sample questions:

  - `What is the primary operating system in which you work?`
  - `What are the primary version control systems you use?`

- In May 2022 over 70,000 developers participated in SODS
- Publicly available.

Motivation – Research Questions

- ***Trends:*** Temporal behavior of salary in tech-industries for different countries?
- ***Correlation:*** Online course platform vs. job status?
- ***Frequent Pattern Analysis:*** Frequently preferred programming frameworks? Tools?
- ***Cluster:*** Based on age groups and salaries?
- ***Classifications:*** Given the preferred PL and education information can we classify user to an estimated salary range?

Literature Survey

- **Used *Stack Exchange*, *web-scraping*, *own survey*:**
  - Peruma et al. 2021, Brooke 2021, Ragkhitwetsagul et al. 2021, Moutidis et al. 2021, Fu et al. 2020, Wu et al 2019, Ahmed et al. 2017
- **Used *SODS*:**
  - SODS 2020: Dada et al. 2022 >> *Frequencies and visualization*
  - SODS 2011 – 2018: Nivala et al. 2020 >> *Current situation in the SO*
  - SODS 2017: Ford et al. 2017 >> *Women parity*
- No correlation, FPA, clustering, classification [to the best of our knowledge]

Proposed Work

1. Data Preprocessing:
    - Measure data quality by using statistical description
    - Data cleaning and handling missing data
    - Data integration (for multiple years)
    - Data reduction
    - Data Transformation (if needed)
2. Data Warehousing:
    - For making a specific and concise view of 'only useful' data for a particular purpose
    - Well Organized

Proposed Work – contd.

3. Data Visualization:
    - Visualize the data distribution
    - Visualizing relationships between different attribute

4. Trend Analysis:
    - Analysis of different attributes
        - e.g., programming languages, or operating systems usage over the years

Proposed Work – contd.

5. Correlation Analysis:
    - Analyzing the correlation between different attributes
    - Correlation coefficient, Chi-square test, Lift Measure
        - e.g., demographics vs programming languages

6. Frequent Pattern Analysis:
    - Mine frequent patterns
    - Using Apriori Algorithm, and FP-growth tree
        - e.g., which PL are frequently used by developers, or different ethnic?

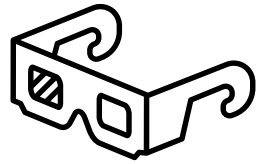
Proposed Work – contd.

7. Clustering:
    - Extracting meaningful clusters from the SODS dataset
        - e.g., education level vs. salaries

8. Classification:
    - Classification based on different attributes
        - e.g., classifying salary group based on given experience, work-mode preference, and education information of an unknown user

Evaluation

- We are not sure yet
- Tentative:
    1. Accuracy
    2. Precision
    3. Recall
    4. MSE
    5. F1 score
    6. Speed
    7. Robustness
    8. Scalability
    9. Interpretability
    10. Goodness of rules



Milestones

SL	Project Work	Estimated time laps to finish*
1	Data Preprocessing	Week 8
2	Data Warehousing	Week 9
3	Data Visualization	Week 9
4	Trend Analysis	Week 10
5	Correlation Analysis	Week 12
6	Frequent Pattern Analysis	Week 14
7	Clustering	Week 15
8	Classification	Week 16

* by the end of the week

The END!!!

Comments and Suggestions?