CS685: Data Mining

Assignment 2 (100 marks)

Due on: 1st November, 2021, 11:00pm

Explore the India Census 2011 website. In particular, we will need to use the data files on language available from: `https://censusindia.gov.in/2011Census/Language_MTs.html`. Out of these, look at categories C-17, C-18, and C-19.

Submit all the necessary components of your data as a single `zip` file named `rollno-assign2.zip` in the `hello.iitk.ac.in` portal within the deadline.

If you do not follow the naming conventions, marks for that question will be automatically 0 (zero). For outputs pertaining to states, use state-codes (present in data-files).

While the programs/scripts should be named `.sh` files, you can invoke any program from within the shell file. The programs should run in the Linux operating system.

All the output files should be sorted by the first field.

1. (10 marks) Find the percentage of population of India who speaks (a) only one language, (b) exactly two languages, and (c) three languages or more. Find this for the states and union territories of India (i.e., overall). Display each part on a separate line in output.

   Call this `percent-india.csv` and script/program to generate this `percent-india.sh`.

2. (10 marks) Repeat the above question for separately males and females speaking three languages or more. For which states and union territories, are the ratios significantly different between males and females? Do a statistical test and report the p-value. Output should contain columns `male-percentage, female-percentage, p-value`. Call this `gender-india.csv` and script/program to generate this `gender-india.sh`.

3. (10 marks) Repeat the above question for urban and rural population separately. Report p-value as above. Output should contain columns `urban-percentage, rural-percentage, p-value`. Call this `geography-india.csv` and script/program to generate this `geography-india.sh`.

4. (10 marks) Find the top-3 states where the ratio of population speaking three languages or more to exactly two languages is the best. Find the worst-3 states as well. The output should contain 6 rows displaying top-3 states (higher to lower ratio) first and then worst-3 states (lower to higher ratio). Call this `3-to-2-ratio.csv` and script/program to generate this `3-to-2-ratio.sh`.

   Repeat the question for the ratio of exactly two languages to only one language. Call this `2-to-1-ratio.csv` and script/program to generate this `2-to-1-ratio.sh`.

5. (10 marks) Find the age group in India that has the highest percentage of people speaking three languages or more. Do this for all the states and union territories as well. Output this as rows having the following columns: `state/ut`, `age-group`, `percentage`. Call this `age-india.csv` and script/program to generate this `age-india.sh`.

6. (10 marks) Find the literacy group in India that has the highest percentage of people speaking three languages or more. Output this as rows having the following columns: `state/ut`, `literacy-group`, `percentage`. Call this `literacy-india.csv` and script/program to generate this `literacy-india.sh`.

7. (10 marks) Divide India into six regions:

   - North: JK, Ladakh, PN, HP, HR, UK, Delhi, Chandigarh
   - West: RJ, GJ, MH, Goa, Dadra & Nagar Haveli, Daman & Diu
   - Central: MP, UP, CG
   - East: BH, WB, OR, JH
   - South: KT, TG, AP, TN, KL, Lakshadweep, Puducherry
   - North-East: AS, SK, MG, TR, AR, MN, NG, MZ, Andaman & Nicobar

   For each region of India, report the top three most spoken languages. Output should contain the following columns: `region`, `language-1`, `language-2`, `language-3`. Call this `region-india.csv` and script/program to generate this `region-india.sh`.

8. (10 marks) For overall India and for each state and union territory, find the combination of age group and literacy group that has the highest ratio of population that can speak 3 or more languages. Output columns as: `state/ut`, `age-group`, `literacy-group`, `ratio-of-3`.

   Repeat this for highest ratio of population that speaks exactly 2 languages, and only 1 language. Output columns as: `state/ut`, `age-group`, `literacy-group`, `ratio-of-2` and `state/ut`, `age-group`, `literacy-group`, `ratio-of-1`.

   Call this `age-literacy.csv` and the script/program `age-literacy.sh`.

9. (10 marks) For the outputs of age-literacy combination group in the last question that can speak 3 or more languages, find the ratio of male:female population (normalized by overall male:female population). Are these significant? Report p-values. Output columns as: `state/ut`, `age-group`, `literacy-group`, `male-ratio`, `female-ratio`, `male-female`, `p-value`.

10. (10 marks) Write a manual that describes how to use your code. Include all the programs, their plugins, and dependencies needed to run the program. Include a top-level script `assign2.sh` that runs the entire assignment. Call this manual `README.txt`.