47717: Data Mining
Homework #1: Sampling, Crawling
Due: **22 April, 11:59pm, on Moodle**

THE HEBREW UNIVERSITY OF JERUSALEM

Read the instructions **carefully** (that's a good idea in general).

- Each person submits their own theoretical part. The theoretical part should be a single file in pdf format only (no docx or jpg) named **ex1t_ID.pdf** (ID is your ID).
- If you are submitting handwritten answers, make sure they are crystal clear.
- Only **one** person from each group should submit the practical part code and answers. In the practical part submission link, the person who submits should enter the partner's name (using the add partner button). In addition, *both* people should write the name and id of their partners in the theoretical part pdf.
- For the practical part, the person who submits should submit a single ZIP file named **ex1p_ID.zip** (where ID is the ID of the person submitting). The zip should contain a folder named **code** and a folder called **output**.
- The meta question should be answered through a questionnaire. The link will be posted on the moodle.
- Points may be reduced for submissions that fail to comply.
- Make sure you follow the News forum for any updates.

‘

Problem 1 (Sampling).

Read each scenario and determine if each sample is a simple random sample or not. Explain why. If it is not SRS and you know the type of sample it is (e.g., convenience), write it down. Explain possible biases.

(a) A teacher wants to select five students from the class. Suppose that the classroom has six rows of chairs with five chairs in each row. The teacher assigns the rows the digits 1 through 6. She throws a die and selects all the students in the row corresponding to the number on the die in the sample.

(b) The teacher goes to the class WhatsApp group and picks the last five students that wrote messages.

(c) The teacher assigns each student a number from 1 to 30. The girls get the numbers 1 to 15 and the boys the numbers from 16 to 30. The teacher uses a random number table to select six two-digit numbers between 1 and 30, and select the corresponding students in the sample.

(d) The teacher asks everybody in class to reply, but makes clear it is optional. Half the students reply.

(e) There are fifteen boys and fifteen girls in a class. Each student's name is placed in a hat and the names are thoroughly mixed. Seven names are drawn and all names correspond to the boys in the class.

(f) There are fifteen boys and fifteen girls in the class. The teacher selects a sample of six students by using a random number table to choose 1 of the 15 boys, then 1 of the 15 girls, then a boy, then a girl, and so on until she has chosen 6 students.

(g) A toy designer is looking to develop a new toy for young children. A brainstorming session results in 15 possible new ideas. A full-scale prototype development and testing for each idea would not be cost effective, so the company decides to perform a preliminary study. One of the employees suggests drawing sketches of the ideas and taking them to his sonâs daycare to see which pictures get the most attention from the kids there.

Problem 2 (Creepy Crawlers).

In this question, you will use crawling to create a (tiny) dataset.
Crawling websites is impolite in general, so MAKE SURE you are following these rules:

- Start by downloading a single page to your computer. Only when you have it right (that is, you can open the local file, parse the html and save the output in the format you want), you can add a loop in and go for multiple pages.
- Make sure there is a significant delay between each request to the website (several seconds should do the trick).
- Be **really** careful. :)

Instructions: Pick a crawling task from the list below. For the sake of this homework, we only need **300 pages**, but your code should be able to crawl the entire site, if we remove that restriction. (In other words, do not hand-code URLs you want to crawl). You should submit three things:

(a) The code, inside a directory called **code**.

(b) Output should be **a JSON file** (indented for readability, please). This is the general format:

```
{
  "records": {
    "record": [
      {
        "id": "1",
        "url": "http://something",
        "field1": "value 1",
        "field2": "value 2"
      },
      {
        "id": "2",
        "url": "http://something",
        "field1": "value 1"
      },
      {
        "id": "3",
        "url": "http://something",
        "field1": "value 1",
        "field2": "value 2"
      }
    ]
  }
}
```

See task-specific fields.

(c) In addition, attach **one example**: a **screenshot** of one of the pages you crawled, plus the corresponding JSON. Put it in the output folder.

You can use existing code for a web crawler/scraper in your favorite language (or write your own, if you so choose). A part of the assignment is experimenting with tools. Consider using tool for extracting content from html, if html is too convoluted to get by with regexes (e.g., Beautiful Soup, lxml). If the website has dynamic content, consider using Selenium (a tool for automating browser activities).
Have fun. :)

Options:

- IMDB, TV and movies, pick one genre/theme (e.g., Post-Apocalypse, Dystopia, Zombie :)). `https://www.imdb.com/feature/genre/?ref_=nv_ch_gr`
  Fields: Name, Year, MovieRating (PG-13, ...), Length (minutes), Genres (string), Director, Stars, IMDBRating, NumberOfVotes, Gross (in USD), FullText (entire HTML, no cleaning, of the movie/TV show page). URL field should be the movie/TV show page (where you got FullText from).
- **Harder, 5 points extra credit**. AllRecipes (Egg-free category): `https://www.allrecipes.com/recipes/740/healthy-recipes/egg-free/`)
  Fields: Title, Creator, Rating, NumReviews, Ingredients (array of strings), Directions (array of strings), PrepTime (minutes), CookTime (minutes), TotalTime, Servings.
- **Easier, 5 points less**. Pinball database `https://www.ipdb.org`
  Fields: Name, AverageFunRating, Manufacturer, DateOfManufacture, Production (number of units), Theme, Specialty, NotableFeatures (string), MarketingSlogans (array of strings)
  Not all games have all fields. Make sure you have at least 50 with one of the last three fields.

Some fields might not be available for all the items – just don't include them in the output for these items.

Problem 3 (Meta). How long (in hours) did this assignment take? (Answer on the moodle, not in the writeup)