Assignment 4 -- Data Science Teams

There are two parts to this week's assignment.

1) Based on the feedback you've received (summarized at the end of this assignment), please update the proposed corpus and sampling methodologies for your tasks.

2) Generate the list of 20 questions for your domains, based on your final corpus and sampling procedures. We recommend you use R in order to get familiar with this platform, as we will be using R extensively during the data analysis phase of the project. A complete submission includes several pieces:

   a) **README.txt**. This file should fully document your corpus and sampling procedures from part (1) of the assignment. It should include a description of the domain, the corpus, the sampling procedure for the tasks, and the sampling procedure for the answer choices.

   b) **assets/ folder**. Create a folder containing any data (videos, audio, images) needed for the task.

   c) **tasks_name.tsv**. Each row in this tab-separated file should correspond to one of the 20 tasks. The columns should be:
      1. correct answer
      2. answer type: "multiple choice", "point estimate", or "map"
      3. asset type: "video"," map", " image", "audio", or "none"
      4. asset file: for images and videos, this should be the name of the image in the assets folder (e.g., car_1.png).
      5. possible answers: enumerate the possible answers for multiple-choice questions. If it is not a multiple-choice question, enter "no answers"
      6. randomize answers: either "true" or "false" indicating whether the multiple-choice answers should be randomized. If the answers are not randomized, they will appear in the order specified above. Answers should be randomized whenever there is not a helpful natural order to the choices.

         Please ensure that the format of your file is exactly as described above, as it must be read and interpreted by the web server.

Create a folder named after your domain. For example, if your domain is involves guessing country capitals, you may call the folder “country_capitals”. Please only use lowercase letters and hyphens (no spaces), and keep the name short but descriptive. Add the folder to the git repository, under the “data/tasks/” folder. Place the two files (README.txt and task_name.tsv) and the assets directory.

An example submission can be found in the repository at: data/tasks/country_shape.

Asset specifications

Images

- make sure your images have consistent dimensions (do not want one very large image, and the next very small)
- file type: jpg or .png

Videos & Audio

- length: 30 seconds or less

Feedback

Task 2 - Find a city in a map

- You can simplify your sampling procedure further, no need to calculate latitude /longitude, sampling randomly would be fine.

Task 3 - Find where the ball is in the magic trick

- We've looked through the videos you posted, great job in finding them. however, these were inconsistent and some of them were magic tricks. please make your own short (<=30sec) videos of the find the ball yourselves to ensure consistent, short videos. Some of the videos were magic tricks There is large variation in the

Task 4 - n/a

Task: 5: Guessing the age of a celebrity

- Ok, no changes needed. Great job!

Task 6: Predicting the height of buildings

- ok, good job!

Task 7- n/a

Task 8 - Box office collection

- Change input: include movie poster along with the movie title

Task 9 - IMDb ranking

- Input: the trailers are too long to show, use the poster of the move (image) and year of release
- the sampling methodology is not clearly defined. you will have to create a corpus of IMDb movies and then draw random samples from these. think about how will you scrape the IMDb site ? how will you construct the corpus (according to what metric, year of release, popularity, etc.)
- What do you mean by "user interest"?
- Please send us a revised version with the changes
- Answer type as integers is ok

Task 10 - Identify the number of Twitter follower of a particular (famous) person

- description: explain clearly to the participants what they will be doing. Clarify: what are “some discrete numbers” (tweets).
- sample 20 users, not 10
- MC answer sampling: could choose other candidates within some standard deviation away from the correct answer
-

Task 11 – Estimate the manufacture year of the car

- corpus - focus on Ford cars only from this corpus: https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=ford%20car%20models
- question type: integer,first year of manufacture, not range of 5 years
- answer choice - it is incorrect to always place the correct answer in the middle, as participants will notice this pattern. The ordering will be randomized by the web design team for all experiments

Task 12 - Predict the direction of the penalty shot

- not clear where you will find the penalty shot videos. please send us an email with an updated report including a link to the 20 videos you intend to use
- they will have to be short (<= 30sec)
- great job otherwise!

Task 13 - n/a

Task 14 - Estimating the age of Celestial bodies

- corpus: You do not need an interactive map. Include only static images of the constellations to the map of celestial bodies included.
- How is the list of craters relevant to the question?

Task 15 - n/a

Task 16 Guess the year a book was written

- We were thinking that it would be good to on top of showing the book’s title, we should also show the book’s author.
- Use as a corpus: http://www.modernlibrary.com/top-100/100-best-novels/ and get a random sample from this list.

Task 17 – Guess the age of a tree

- the Wikipedia corpus does not seem to have images associated with the list of trees
- the other corpus looks more relevant
- be very carful when choosing the pictures, from briefly skimming through them, the images seemed to vary significantly in clarity, lighting dimensions
- please try to find a set of consistent pictures
- perhaps try to find an alternative corpus you cannot extract sufficient from this one
- methodology: start with a simpler procedure that just samples uniformly
- if that does not work well, consider the stratified sampling approach you proposed

Task 18: - Capacity of the container

- It should not be multiple choice. People should give a point estimate.
- The picture has to have a reference frame. If you can get the dimensions from Amazon and put beside the picture a good reference frame like a coin then it would be fine. Otherwise you could collect around 30 containers, pick 20 at random and then take pictures of them with a reference beside them.

Task 19 – Total Calories Consumed

- You should specify the corpus of how are you choosing the food.
- We think it would actually be interesting to show images of food better than just asking for an edible.

Task 20 - Number of retweets of a specific person

- ok no changes, good job!

Task 21 - Predict stock prices, gold oil

- Nice submission quality.
- You only will be able to ask one question so in this case it should be the person's prediction of the price in 15 days. This way you can recover both the direction and the percentage.

Task 22 - Predict the distance between two cities

- do not show the radius of the earth as input

- otherwise, great job!

Task 23 - n/a

Task 24 - Word Meaning

- which part of the corpus will you be using ? the anc corpus for instance, has a lot of sub-categories
- please submit a clearly-defined sampling methodology

Task 25 - n/a

Task 26 - n/a

Task 27 - n/a

Task 28 - Determining the language script of a text

- Good job.
- Just we think it would be better if instead of multiple choice there was a scroll down menu with all available options to choose from. So in the next submission you should put this as the output.

Task 29 - Spam mail detection

- Try this as corpus: http://plg.uwaterloo.ca/~gvcormac/spam/. From this corpus randomly sample.

Task 30 - Identify the band

- Use a corpus like: http://www.billboard.com/articles/list/2155531/the-hot-100-all-time-top-songs to pick the songs and sample from it. It should be more straightforward. You can use any corpus of top 100 songs from all times.

Task 32 - Solve Chess Puzzle

- Use 5 MC answers (out of which 1 is correct), otherwise good job!

Task 33 - Identify the currency of a country

- We really appreciate your enthusiasm but would like for the randomization to be simpler.
- You should choose countries randomly given only one dimension for the case of weights.

Task 34 - n/a

Task: 35: Predict the result of a cricket match

- Methodology: not clearly defined. How will you sample 20 games ?
- remove “No response” from the option set

Task 36 - Price of painting

- You should only ask for how much the user thinks the painting was priced at an auction.
- You should not exclude famous paintings since it would still be hard to guess the price.
- Given a corpus of auction prices of paintings just choose a random sample. We should not care if the painting is well known or if the artist is unknown.

Task 37 - Weight of an object

- It would be better if the type of answer were not multiple choice in order to have more granularity. Instead people could just submit the weight.
- We loved the webpage you posted and would be a good way to select random objects, but we think that the images would make more sense if you took them yourselves from objects you could measure their weight. So if the webpage gives you a tie you could take the picture of a tie and weight that tie in order to get the correct ground truth.

Task 40 - n/a

Task 39 - Predict the price of the given product on Amazon

- Ok, nice job!

Task 40- Success of a Business Idea:

- We think it would be better for you to randomly choose products from kickstarter and show a picture and some summary from it. The selection would be random among the kickstarter products. You could limit yourselves to gadgets.

Task 41 - Identify a country by its flag

- Ok, nice job!

Task 42 - Guess the order of historical events

- Do not consider any future events

- The corpus you have found may be difficult to work with, it is very extensive and includes many events that are not likely to be known by the general population.
- Try to find a more concise list of “100 most important historical events”
- please revise the corpus and send us an updated version of your report
- The methodology you have devised is well-thought-out, however if you pick a smaller corpus, you will be able to perhaps simplify it even more and not have to condition on time period at all

Task 43 - Guess the popularity of a brand

- You should have created the corpus for this task. In that case write all sources you used for the creation of this corpus.
- The point of this task is to show users two products from different brands from other countries and have them guess which is the most popular one.

Task 44 - n/a

Task 45- n/a

Task 46 - Predicting number of goals in a football match

- ok, no changes needed. Great job!
- Really liked the quality of your submission.

Task 47 - Predicting the mood of a person

- ok, no changes needed. Great job!

Task 48 - n/a

Task 49 - Guess the length of a river

- Excellent submission, we really liked the quality of it.
- The problem with simple sampling is that we feel a lot of the rivers are totally unknown. So we were thinking of two solutions. Either coming up with the image of each river shown in the same scale. Or use the top 20 most popular rivers if that corpus can be found somewhere.
- Please tell us if you think of other solution to overcome this problem.

Task 50 - Estimate countries by landmass

- Methodology: perhaps experiment with population density as a weight as well, and see which approach works better (ie. Ensure there is sufficient variation in the questions)

Task 51- n/a

Task 52- n/a

Task 53- n/a

Task 54- n/a

Task 55- n/a