Data Analysis Challenge

Goal

Analyzing user behavior within the same session is often crucial. Clustering users based on their browsing behavior is probably the most important step if you want to personalize your site.

The goal of this challenge is to build the foundation of personalization by identifying searches likely to happen together and cluster users based on their session searches.

Challenge Description

Company XYZ is an Online Travel Agent, such as Expedia, Booking.com, etc. They store their data in JSON files. Each row in the json shows all different cities which have been searched for by a user within the same session (as well as some other info about the user, see fields below).

That is, if a user goes to company XYZ site and looks for NY and SF within the same session, the corresponding JSON row will show his user id, some basic info about him and the two cities.

You are given the following tasks:

1. There was a bug in the code and for one specific country, the records don't have the country field logged. It just shows up as an empty field (""). The search sessions with a missing country either come from a country that is completely missing from the data, or from one of the countries that are logged in the data. Can you determine which country it is the most likely to be? Explain your hypothesis and the data analysis tasks you did to find the missing country.

2. Given a sequence of searched cities, find the most likely city or cities to be also searched next, within the same session. Your code should include a function that takes a list of 0 to *n* cities and returns the most likely next city or cities. Keep in mind that the goal is to call this function each time a user performs a search. It should therefore be fast to execute.

3. There are few features describing each user: user id, joining date and country. Are these features useful to predict the most likely city to be searched? How do they compare to the other features tried in Question 2 (i.e. previous cities searched)? Can the algorithm implemented in Question 2 be improved by making use of these features?

4. How did you measure the performance of the prediction algorithms from questions 2 and 3? What is your confidence that the measured score is accurate?

Keep in mind that you may “think outside of the box”: any improvement idea that is not specifically asked for in the questions can be considered.

Deliverables

1. All the code or functions used to answer the questions should be provided. The project can be done with any language, but the code needs to be structured, easily readable and commented.
2. Write a short document or notebook briefly explaining your answers to the questions.

Data

The file "city_search" contains a list of searches happening within the same session.

Fields

- **session_id** : session id. Unique by row
- **unix_timestamp** : unix timestamp of when the session started
- **cities** : the unique cities which were searched for within the same session by a user
- **user** : it has the following nested fields:
    - **user_id**: the id of the user
    - **joining_date**: when the user created the account
    - **country**: where the user is based

Example

Field	Value	Description
session_id	X061RFWB06K9V	unique identifier of the search session
unix_timestamp	1442503708	unix timestamp of when the session started. That means: Thu, 17 Sep 2015 15:28:28 GMT
cities	New York NY, Newark NJ	the user searched for hotels in two cities: NY and Newark
user_id	2024	id of the user
joining_date	2015-03-22	she joined the site on March, 22
country	UK	she is based in UK