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 hotels in 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 one country didn't get logged. It just shows up as an empty field (""). Can you guess which country was that? How?
2. For each city, find the most likely city to be also searched for within the same session. You should develop a function that takes a city (or a list of cities) and returns the most likely next city.
3. There are 2 features describing each user (joining date and country). Are these features useful to predict the most likely city to be searched? How do they compare to the search history (i.e. previous city searched)?
4. How do you evaluate the performance of the prediction algorithm?

The project can be done with any language, but the code needs to be structured, easily readable and commented. All the code or functions used to answer the questions should be provided. In addition, write a short document 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