EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT
Seminar für Sprachwissenschaft

Data Science for Linguists
Summer 2025
Johannes Dellert

Assignment 03:
Geographic Distribution of Languages in Pandas

handed out: 7 May, 14:00
to be submitted by: 14 May, 12:00

In this sheet, you will use Pandas in order to explore data from the very popular open-source language database Glottolog [Hammarström et al., 2024], finding out various potentially interesting facts about the geographical distribution of the world's languages.

Task 1: Loading and Exploring a Dataset in Pandas

Create a new Jupyter notebook and import `pandas`. Load the Glottolog language database from the file `languages_and_dialects_geo.csv` into a `DataFrame` object. Explore the dataset and gain a first overview:
a) What are the columns and their value ranges?
b) How many entries does this database of languages and dialects have?
c) What is the full inventory of macroareas into which this dataset partitions languages?
d) How many languages and dialects have an ISO 639-3 code associated with them?
e) For how many languages and dialects do we have latitude and longitude data?

Task 2: Visualising Simple Distributions

a) Drop all the dialects (as opposed to languages) from the dataframe, we are not going to be interested in dialects in any of the subsequent tasks.
b) Use Seaborn to visualise the distribution of languages across macroareas in a useful format.
c) We want to visualise which percentage of languages in each macroarea have ISO 639-3 codes associated with them. For this purpose, create a new dataframe containing the relevant counts (doable in three lines using vectorisation on two prepared views). Create a stacked bar plot not in Seaborn (that would be complicated), but using the `plot` method of a new Pandas dataframe (check the documentation!). Can you draw any conclusion from the differences between areas?

Task 3: Exploring Extreme Locations

a) Extract name and macroarea of the northernmost and the southernmost languages for which we have data.
b) Which percentage of the world's languages are spoken in the tropics (between 23.43619°N and 23.43619°S)?
c) The following formula yields spherical distances in kilometres, approximating the length of the shortest surface arc between geographic coordinates (φ_1, λ_1) and (φ_2, λ_2), where φ is latitude and λ is longitude:

$$\arccos(\sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos(\lambda_2 - \lambda_1)) \cdot 6371$$

Use vectorised functions to implement this formula, then rank all other languages by their distances to the northernmost language you established in a). Is the language furthest from the northernmost language the southernmost language, as we might suspect? Why (not)?

Task 4: Density of Languages

a) Use the `pd.Series.quantile()` function to find out the latitude range which covers the central half of the world's languages. Is it symmetric around the equator?
b) Estimate the area covered by each macroarea by computing the distances (see Task 3c) between 1000 random pairs of languages from the area, taking the two largest distances a and b, and assuming they indicate the lengths of the semi-major and semi-minor axes of an ellipsis-shaped area (formula: $\pi \cdot \frac{a}{2} \cdot \frac{b}{2}$).
c) Estimate the densities of languages in each macroarea by dividing the number of languages in the areas by the sizes computed in b), and visualise the differences in a bar plot using Seaborn.

Task 5: Gaps in Longitude and Neighbour Density

a) Between which two languages do we find the largest gap in longitude, i.e. the longest stretch in west-east direction which is not inhabited by any other language?
b) Add a new column to your dataframe in which you store how distant each language is from its closest neighbour, again using the distance formula from 3c). Plot the distributions for the macroareas as a boxplot in such a way that outliers are visible. Are these outliers actually the most isolated languages? Why (not)?

References

[Hammarström et al., 2024] Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2024). Glottolog 5.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://glottolog.org.