

Exercise 1 – Python environment, pandas, data sources (submission date: April 22nd at 12:00pm)

This exercise checks an initial understanding of Python and pandas.

You can submit it in pairs. Please

Use pandas on a Hebrew file

1. Go through the installation process listed in the file `install1.pdf` on the course Moodle.
2. activate the `hebnlp` environment.
3. Download the file `"israel_places_latin_names.csv"` from the course Moodle.
4. `cd` into the course notebooks directory.
5. run the command:
`jupyter notebook`
6. Create a new `hebnlp` notebook from the browser.
7. Read the `places` file into the notebook using `pandas`.
8. Write Python functions in the notebook to answer the following questions:
 - a. What are the 5 most frequent district names in Israel?
 - b. What are the 5 most frequent council names in Israel?
 - c. Break the settlement names into whitespace-separated tokens by using `split()`. What are the 5 most frequent tokens in settlement names?
 - d. Create a dataframe containing just the settlements in the "באר שבע" district. What are the 5 most frequent tokens in these settlement names?
 - e. Write the answers to a through d as a CSV file in the following structure:
`question,frequency,name`
`a,40,פרטצ'יה`
`a,27,אוטופיה`
...

גרילנדה, 14, b

...

צוף, 5, d

שני, 2, d

The file's name should be settlement_summaries.csv

- f. Submit the notebook and the result csv file.