# THIS IS WORK IN PROGRESS

Only d08\_wilcoxon.ipynb, d08\_wilcoxon.ru and d08\_student.ipynb are finalized by the web-interface is not ready yet.

## New in this version:

- Changed the test from one-sided to two sided version and added a continuity correction.

# Phrase frequencies in Google Books

It was a study project on the Data Scientist upskilling course at School 21 (https://edu.21-school.ru/). Project task was to find an API in the Internet and create a web-service using Flask, FastAPI or a Telegram bot.

I had worked as a translator for many years so I did a project in linguistics. One of the tools I often use in my work to analyze word usage is Google Ngrams (<https://books.google.com/ngrams>). It builds plots that can be used to compare frequencies of different phrases. The comparison is qualitative and it is not always easy to say which is more frequent as in the example on the home page. It is not also possible to say whether the apparent different is statistically significant.

The idea is to provide an interface where a user can enter the same data (language, two phrases, time interval, etc., except smoothing which is not used) and get both the plot and an answer which phrase is more common and whether the difference is statistically significant.

This version supports only some corpora: only the 2019 versions of the English, French, German, Italian, Russian and Spanish corpora.

The web-service gets frequencies for each year in the interval and compares the using the Wilcoxon signed-rank test with continuity correction.

### Files:

- d08\_0.ipynb – original Jupyter notebook

- d08\_0.R – same algorithm in R

- d08\_1.ipynb, d08\_1.py – notebook and python script with class definition for the MVP version

- p-value\_phrases-main.zip – web-interface

This version is hosted at