Homework 2

Research seminar

Amal Yakubov

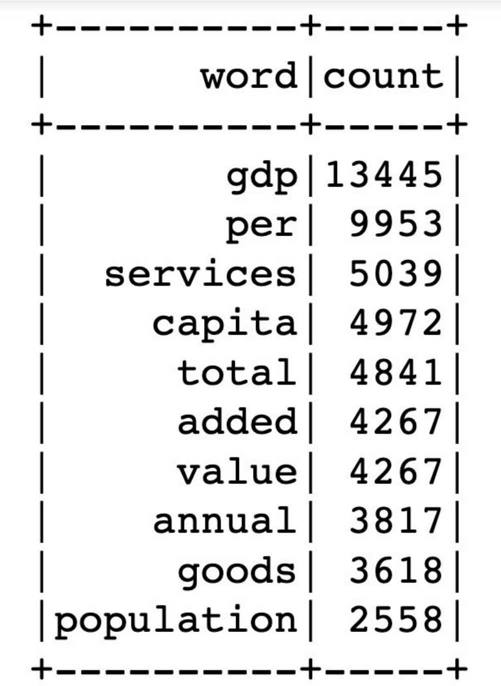
Data

10 different datasets were downloaded from several resources on different topics.
All of them were tokenized, the stop-words and everything except letters were deleted.

Analysis results

On the right are presented the most popular words.

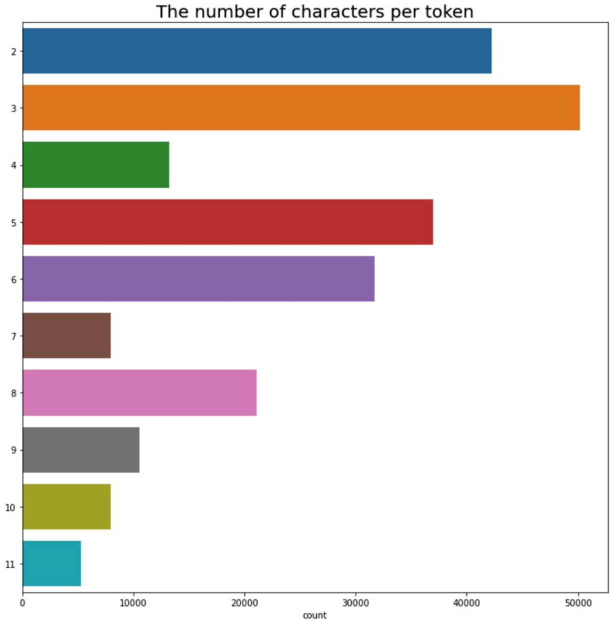
It is remarkable that words mostly used to analysis a country are “GDP” and “per”ю



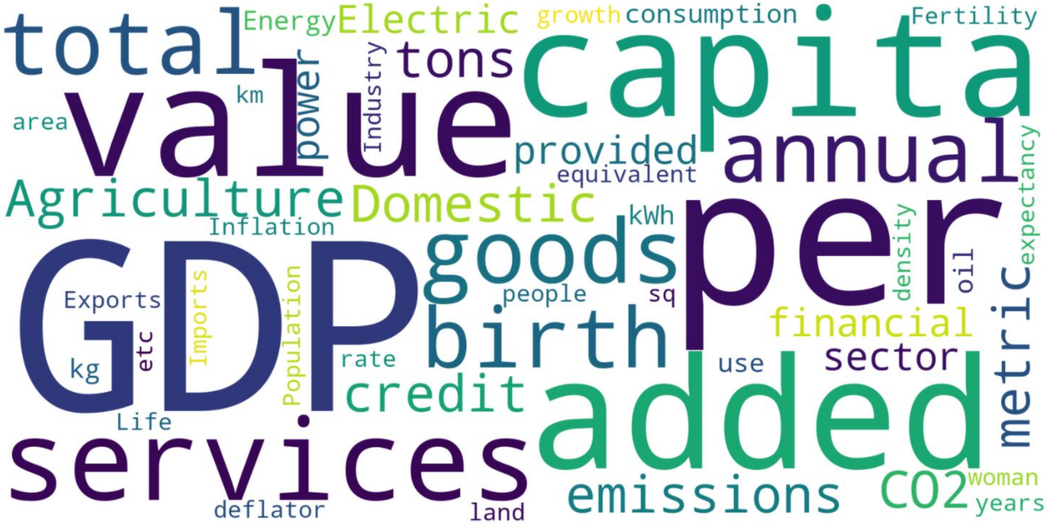
```
+----------+-----+
|      word|count|
+----------+-----+
|       gdp|13445|
|       per| 9953|
|  services| 5039|
|    capita| 4972|
|     total| 4841|
|     added| 4267|
|     value| 4267|
|    annual| 3817|
|     goods| 3618|
|population| 2558|
+----------+-----+
```

Analysis results

The number of characters per token is, actually, the length of the words in the dataset.



Word cloud


total
Energy
Electric
growth
consumption
Fertility
capita
value
tons
km
power
Industry
area
provided
annual
equivalent
Agriculture
Domestic
Inflation
kWh
per
goods
density
expectancy
oil
GDP
Exports
people
sq
Imports
Population
birth
financial
metric
etc
kg
rate
sector
use
credit
Life
services
added
emissions
CO2
woman
years
deflator
land

The end

Thank you!