Ира Пантелеева, 152 группа, HW2

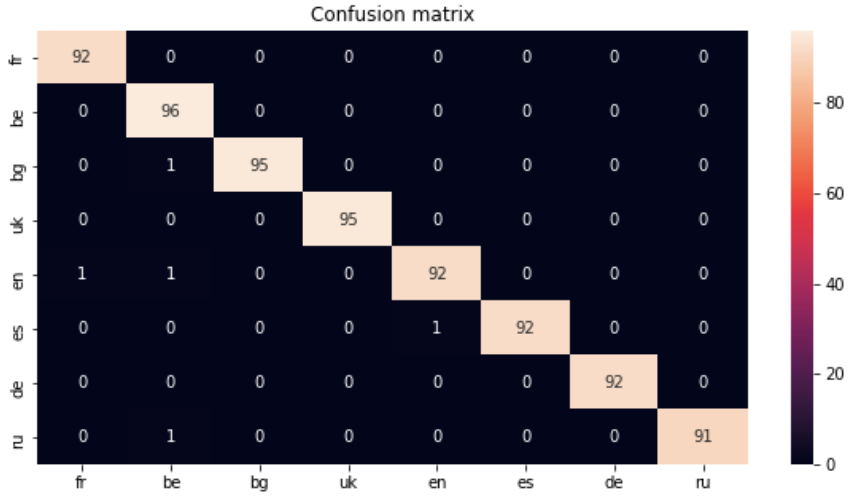
Отчёт по определению языка в тексте

Корпусный метод работает лучше. У него процент ошибок меньше и, соответственно, точность выше, чем у n-граммного. На confusion matrix видим, что путаются языки с одинаковым алфавитом (кириллическим / латиницей). Корпусный (словарный) метод лучше: мы отфильтровали повторяющиеся n-граммы, то есть в список попали менее частноные n-граммы (стоит учесть, что в языках больше совпадений n-грамм, чем слов). Возможно, это сказалось на точность второго метода.

```
Процент ошибки у корпусного метода: 0.006666666666666667
Процент ошибки у n-граммного метода: 0.05333333333333334
```

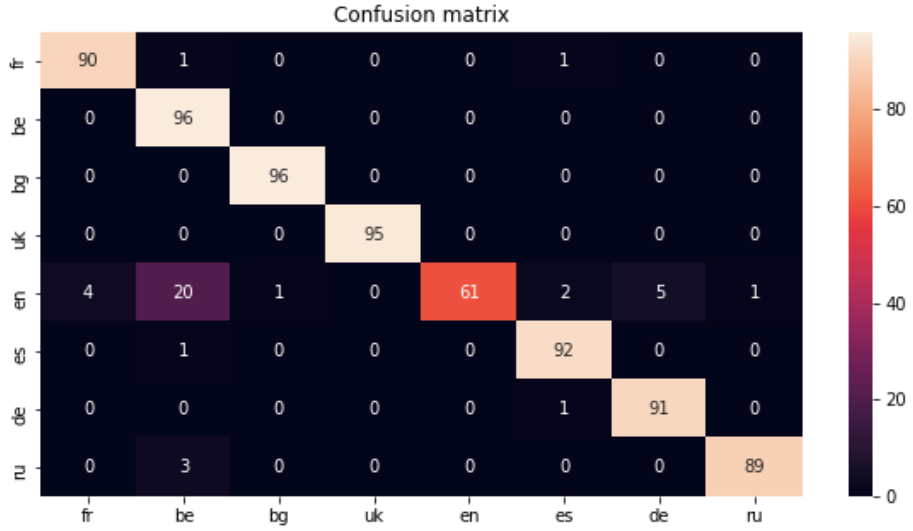
Корпусный (словарный) метод

	precision	recall	f1-score	support
be	0.97	1.00	0.98	96
bg	1.00	0.99	0.99	96
de	1.00	1.00	1.00	92
en	0.99	0.98	0.98	94
es	1.00	0.99	0.99	93
fr	0.99	1.00	0.99	92
ru	1.00	0.99	0.99	92
uk	1.00	1.00	1.00	95
avg / total	0.99	0.99	0.99	750



n-граммный метод

	precision	recall	f1-score	support
be	0.79	1.00	0.88	96
bg	0.99	1.00	0.99	96
de	0.95	0.99	0.97	92
en	1.00	0.65	0.79	94
es	0.96	0.99	0.97	93
fr	0.96	0.98	0.97	92
ru	0.99	0.97	0.98	92
uk	1.00	1.00	1.00	95
avg / total	0.95	0.95	0.94	750


Confusion matrix
fr
be
bg
uk
en
es
de
ru
90 1 0 0 0 1 0 0
0 96 0 0 0 0 0 0
0 0 96 0 0 0 0 0
0 0 0 95 0 0 0 0
4 20 1 0 61 2 5 1
0 1 0 0 0 92 0 0
0 0 0 0 0 1 91 0
0 3 0 0 0 0 0 89
80
60
40
20
0