Анастасия Костяницына
БКЛ-151

Домашнее задание №1
Отчет

Задание 1.1.

В качестве текстов для коллекции были выбраны отрывки из книг Д.К. Роулинг о Гарри Поттере. Отрывок для анализа ключевых слов был выбран случайным образом.

Выбранные ключевые слова:

Лексема	Частота по частотному словарю	Частота в коллекции	Частота в тексте
Сириус	-	0.000557	0.015337
Гарри	-	0.020475	0.018881
наложить	16.4	7.6e-05	0.000699
заклятие	-	0.00027	0.001399
Думбльдор	-	0.001841	0.002098
Уизли	-	0.001098	0.00979
палочка	21.1	0.00185	0.002797
волшебный	28.2	0.000802	0.001399
Гермиона	-	0.0031	0.001399
рыжий	43.2	0.000101	0.001399
Вольдеморт	-	0.000693	0.000699
Хогварц	-	0.000422	0.000699
смерть	284.1	0.000633	0.000699
заколдовать	-	3.4e-05	0.000699

1) Есть ли среди выбранных вами ключевых слов редкие слова?

> Если рассматривать с точки зрения частотного словаря, то некоторые ключевые слова были попросту не найдены, так как многие из ключевых слов являются именами персонажей: Сириус, Гарри и тд. Самыми редкими словами и для одного текста, и для целой коллекции стали: заколдовать и наложить. Выбор на них пал из-за связи с тематикой книги.

2) Есть ли среди выбранных вами слов слова, вошедшие в топ 500 по частоте?

По частотному словарю - нет.

3) К каким частям речи относятся выбранные вами слова, слов какой части речи больше?

В основном это существительные, имена собственные. Так же есть по два прилагательных и глагола.

4) Какие слова встретились во всех или в большинстве документов? Каковы их грамматические характеристики.

Из списка практически во всех документах встретились слова Гарри, Гермиона и палочка (от самого встречающегося к менее). Гарри и Гермиона - существительные, имена собственные. Гарри - м. род, одушевлённое, неизменяется. Гермиона - ж.род, одушевлённое, изменяется.

Задание 1.2.

Составьте матрицу term x document для всей коллекции и еще для трех высокочастотных слов (по частотному словарю для русского языка).

Была составлена матрица term x document для каждого слова и для каждого текста в коллекции. Также была составлена матрица, каждая ячейка в такой таблице рассчитывалась по формуле tf * idf, где tf = Кол-во вхождений слова в тесте/ Кол-во слов в тексте, а idf = Логарифм от Кол-ва документов в коллекции/ Кол-во документов с данным словом. Столбики в такой таблице - леммы слов, строчки - документ. Всего 58 документов в коллекции.

Term x document для всей коллекции:
Так как таблица для всех слов слишком большая, прилагаю только первые 5 строк.

```
1  X.head()
```

	adalbert	all	alohomora	argus	arsenic	bagshot	bathilda	beat	beater	blot	...	ясный	ястреб	ястребтетеревятник	ято	ящик	яя	ёкнуть	ёлка	ёл
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	

5 rows × 12177 columns

tf * idf для всей коллекции:

В качестве трех высокочастотных слов для русского языка были выбраны слова: *быть, что, не*.

```
1  X.head()
```

	adalbert	all	alohomora	argus	arsenic	bagshot	bathilda	beat	beater	blot	...	ясный	ястреб	ястребтетеревятник	ято	ящик	яя	ёкнуть	ёлка	ёл
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	

5 rows × 12177 columns

Матрица term x document для шести слов из списка ключевых слов и трех высокочастотных слов:

	гарри	заколдовать	волшебный	гермиона	палочка	думбльдор	не	быть	что
0	0.000139	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000288	0.000422
1	0.001041	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000426	0.000393
2	0.000722	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000304	0.000323
3	0.000567	0.000000	0.000495	0.000000	0.000280	0.000000	0.0	0.000281	0.000445
4	0.000898	0.000000	0.001867	0.000000	0.002848	0.000000	0.0	0.000297	0.000356
5	0.000011	0.000000	0.000520	0.000000	0.000737	0.001850	0.0	0.000113	0.000339
6	0.000032	0.000000	0.000767	0.000000	0.001448	0.009083	0.0	0.000164	0.000259
7	0.000703	0.000000	0.001360	0.000000	0.001387	0.026101	0.0	0.000135	0.000292
8	0.000860	0.000000	0.001965	0.000188	0.002784	0.043651	0.0	0.000142	0.000264
9	0.001129	0.000000	0.001042	0.000799	0.000787	0.000000	0.0	0.000136	0.000287
10	0.000980	0.000000	0.000610	0.000117	0.000173	0.000000	0.0	0.000120	0.000246
11	0.001114	0.000000	0.001273	0.000122	0.001442	0.000000	0.0	0.000184	0.000289
12	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
13	0.000053	0.000000	0.002954	0.000000	0.003349	0.000000	0.0	0.000096	0.000157
14	0.000533	0.000000	0.001966	0.000251	0.000743	0.024459	0.0	0.000135	0.000291

15	0.001235	0.000000	0.000380	0.000000	0.000215	0.001350	0.0	0.000133	0.000330
16	0.001091	0.000000	0.002286	0.001267	0.001296	0.000903	0.0	0.000110	0.000199
17	0.001115	0.000000	0.000343	0.001578	0.000777	0.000609	0.0	0.000149	0.000262
18	0.000872	0.000000	0.000615	0.000472	0.002091	0.000364	0.0	0.000144	0.000267
19	0.000803	0.000000	0.000935	0.000072	0.001484	0.006981	0.0	0.000143	0.000371
20	0.001013	0.000000	0.002521	0.000264	0.002728	0.000407	0.0	0.000147	0.000246
21	0.000845	0.000000	0.000836	0.003044	0.000592	0.007424	0.0	0.000168	0.000432
22	0.000682	0.001299	0.000817	0.000784	0.000927	0.003632	0.0	0.000093	0.000228
23	0.001006	0.000000	0.000618	0.000000	0.000350	0.000000	0.0	0.000179	0.000192
24	0.000828	0.000000	0.000602	0.000000	0.000682	0.000000	0.0	0.000273	0.000298
25	0.000672	0.000000	0.000000	0.000951	0.001756	0.000000	0.0	0.000231	0.000256
26	0.000385	0.000000	0.000000	0.001417	0.000000	0.000000	0.0	0.000254	0.000280
27	0.000588	0.002132	0.000000	0.000772	0.000000	0.000000	0.0	0.000208	0.000277
28	0.000704	0.000000	0.000000	0.000240	0.000709	0.000000	0.0	0.000207	0.000207
29	0.000453	0.000000	0.000639	0.000490	0.000724	0.000000	0.0	0.000251	0.000264
30	0.000688	0.000000	0.000611	0.001874	0.000692	0.000000	0.0	0.000240	0.000316
31	0.000699	0.000000	0.000000	0.000714	0.000000	0.000000	0.0	0.000257	0.000385
32	0.000648	0.000000	0.000000	0.000238	0.000000	0.000000	0.0	0.000218	0.000218
33	0.000614	0.000000	0.000000	0.002590	0.001739	0.000000	0.0	0.000228	0.000292
34	0.000782	0.000000	0.000000	0.001677	0.000354	0.000000	0.0	0.000232	0.000387
35	0.000734	0.000000	0.000000	0.001862	0.001375	0.000000	0.0	0.000263	0.000238
36	0.000694	0.000000	0.000000	0.003074	0.003494	0.000000	0.0	0.000217	0.000242

37	0.000869	0.000000	0.000000	0.003632	0.000000	0.000000	0.0	0.000248	0.000404
38	0.000842	0.000000	0.000000	0.002582	0.001041	0.000000	0.0	0.000164	0.000228
39	0.000568	0.001966	0.000000	0.002136	0.000351	0.000000	0.0	0.000217	0.000460
40	0.000999	0.000000	0.001843	0.000471	0.000000	0.000000	0.0	0.000279	0.000241
41	0.000840	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000331	0.000269
42	0.000971	0.000000	0.000000	0.002349	0.000000	0.000000	0.0	0.000266	0.000443
43	0.000802	0.000000	0.000000	0.004523	0.000703	0.000000	0.0	0.000257	0.000321
44	0.000588	0.000000	0.000000	0.002351	0.000347	0.000000	0.0	0.000279	0.000329
45	0.000529	0.000000	0.000000	0.003405	0.000000	0.000000	0.0	0.000275	0.000301
46	0.000719	0.000000	0.000000	0.002833	0.000000	0.000000	0.0	0.000331	0.000394
47	0.000681	0.000000	0.000000	0.003132	0.000331	0.000000	0.0	0.000313	0.000337
48	0.000644	0.000000	0.000000	0.002050	0.000000	0.000000	0.0	0.000258	0.000331
49	0.000831	0.000000	0.000000	0.002315	0.000000	0.000000	0.0	0.000237	0.000512
50	0.000743	0.000000	0.000000	0.003772	0.000348	0.000000	0.0	0.000165	0.000419
51	0.000728	0.000000	0.000000	0.004849	0.000682	0.000000	0.0	0.000199	0.000261
52	0.001065	0.002029	0.000000	0.006122	0.000000	0.000000	0.0	0.000238	0.000198
53	0.000656	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000182	0.000312
54	0.000696	0.000000	0.000000	0.000229	0.000000	0.000000	0.0	0.000246	0.000443
55	0.000711	0.000000	0.000000	0.002421	0.000000	0.000000	0.0	0.000261	0.000326
56	0.000446	0.000000	0.000000	0.001094	0.000000	0.000000	0.0	0.000177	0.000192
57	0.000000	0.000000	0.000966	0.000000	0.000000	0.000000	0.0	0.000200	0.000140

Матрица tf * ifd для шести слов из списка ключевых слов и трех высокочастотных слов:

	гарри	заколдовать	волшебный	гермиона	палочка	дамблдор	не	быть	что
0	0.000139	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000288	0.000422
1	0.001041	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000426	0.000393
2	0.000722	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000304	0.000323
3	0.000567	0.000000	0.000495	0.000000	0.000280	0.000000	0.0	0.000281	0.000445
4	0.000898	0.000000	0.001867	0.000000	0.002848	0.000000	0.0	0.000297	0.000356
5	0.000011	0.000000	0.000520	0.000000	0.000737	0.001850	0.0	0.000113	0.000339
6	0.000032	0.000000	0.000767	0.000000	0.001448	0.009083	0.0	0.000164	0.000259
7	0.000703	0.000000	0.001360	0.000000	0.001387	0.026101	0.0	0.000135	0.000292
8	0.000860	0.000000	0.001965	0.000188	0.002784	0.043651	0.0	0.000142	0.000264
9	0.001129	0.000000	0.001042	0.000799	0.000787	0.000000	0.0	0.000136	0.000287
10	0.000980	0.000000	0.000610	0.000117	0.000173	0.000000	0.0	0.000120	0.000246
11	0.001114	0.000000	0.001273	0.000122	0.001442	0.000000	0.0	0.000184	0.000289
12	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
13	0.000053	0.000000	0.002954	0.000000	0.003349	0.000000	0.0	0.000096	0.000157
14	0.000533	0.000000	0.001966	0.000251	0.000743	0.024459	0.0	0.000135	0.000291
15	0.001235	0.000000	0.000380	0.000000	0.000215	0.001350	0.0	0.000133	0.000330
16	0.001091	0.000000	0.002286	0.001267	0.001296	0.000903	0.0	0.000110	0.000199
17	0.001115	0.000000	0.000343	0.001578	0.000777	0.000609	0.0	0.000149	0.000262
18	0.000872	0.000000	0.000615	0.000472	0.002091	0.000364	0.0	0.000144	0.000267
19	0.000803	0.000000	0.000935	0.000072	0.001484	0.006981	0.0	0.000143	0.000371
20	0.001013	0.000000	0.002521	0.000264	0.002728	0.000407	0.0	0.000147	0.000246

21	0.000845	0.000000	0.000836	0.003044	0.000592	0.007424	0.0	0.000168	0.000432
22	0.000682	0.001299	0.000817	0.000784	0.000927	0.003632	0.0	0.000093	0.000228
23	0.001006	0.000000	0.000618	0.000000	0.000350	0.000000	0.0	0.000179	0.000192
24	0.000828	0.000000	0.000602	0.000000	0.000682	0.000000	0.0	0.000273	0.000298
25	0.000672	0.000000	0.000000	0.000951	0.001756	0.000000	0.0	0.000231	0.000256
26	0.000385	0.000000	0.000000	0.001417	0.000000	0.000000	0.0	0.000254	0.000280
27	0.000588	0.002132	0.000000	0.000772	0.000000	0.000000	0.0	0.000208	0.000277
28	0.000704	0.000000	0.000000	0.000240	0.000709	0.000000	0.0	0.000207	0.000207
29	0.000453	0.000000	0.000639	0.000490	0.000724	0.000000	0.0	0.000251	0.000264
30	0.000688	0.000000	0.000611	0.001874	0.000692	0.000000	0.0	0.000240	0.000316
31	0.000699	0.000000	0.000000	0.000714	0.000000	0.000000	0.0	0.000257	0.000385
32	0.000648	0.000000	0.000000	0.000238	0.000000	0.000000	0.0	0.000218	0.000218
33	0.000614	0.000000	0.000000	0.002590	0.001739	0.000000	0.0	0.000228	0.000292
34	0.000782	0.000000	0.000000	0.001677	0.000354	0.000000	0.0	0.000232	0.000387
35	0.000734	0.000000	0.000000	0.001862	0.001375	0.000000	0.0	0.000263	0.000238
36	0.000694	0.000000	0.000000	0.003074	0.003494	0.000000	0.0	0.000217	0.000242
37	0.000869	0.000000	0.000000	0.003632	0.000000	0.000000	0.0	0.000248	0.000404
38	0.000842	0.000000	0.000000	0.002582	0.001041	0.000000	0.0	0.000164	0.000228
39	0.000568	0.001966	0.000000	0.002136	0.000351	0.000000	0.0	0.000217	0.000460
40	0.000999	0.000000	0.001843	0.000471	0.000000	0.000000	0.0	0.000279	0.000241
41	0.000840	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000331	0.000269
42	0.000971	0.000000	0.000000	0.002349	0.000000	0.000000	0.0	0.000266	0.000443

43	0.000802	0.000000	0.000000	0.004523	0.000703	0.000000	0.0	0.000257	0.000321
44	0.000588	0.000000	0.000000	0.002351	0.000347	0.000000	0.0	0.000279	0.000329
45	0.000529	0.000000	0.000000	0.003405	0.000000	0.000000	0.0	0.000275	0.000301
46	0.000719	0.000000	0.000000	0.002833	0.000000	0.000000	0.0	0.000331	0.000394
47	0.000681	0.000000	0.000000	0.003132	0.000331	0.000000	0.0	0.000313	0.000337
48	0.000644	0.000000	0.000000	0.002050	0.000000	0.000000	0.0	0.000258	0.000331
49	0.000831	0.000000	0.000000	0.002315	0.000000	0.000000	0.0	0.000237	0.000512
50	0.000743	0.000000	0.000000	0.003772	0.000348	0.000000	0.0	0.000165	0.000419
51	0.000728	0.000000	0.000000	0.004849	0.000682	0.000000	0.0	0.000199	0.000261
52	0.001065	0.002029	0.000000	0.006122	0.000000	0.000000	0.0	0.000238	0.000198
53	0.000656	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000182	0.000312
54	0.000696	0.000000	0.000000	0.000229	0.000000	0.000000	0.0	0.000246	0.000443
55	0.000711	0.000000	0.000000	0.002421	0.000000	0.000000	0.0	0.000261	0.000326
56	0.000446	0.000000	0.000000	0.001094	0.000000	0.000000	0.0	0.000177	0.000192
57	0.000000	0.000000	0.000966	0.000000	0.000000	0.000000	0.0	0.000200	0.000140

1) Назовите те слова, у которых мощность обратного индекса (количество документов, в которых слово встречается) равна количеству документов в коллекции

Во всех документах встретилось слова *не*.

2) Найдите тексты, удовлетворяющие запросу Word1&Word2&¬Word3

Насколько я понимаю, требуется найти тексты, в которых встречается слова 1 и 2, но не встречается слово 3. Так как в задании не уточнено, что именно подразумевается под словами 1, 2, 3, я рассмотрела комбинации слов из колонок таблиц. При составлении комбинаций не учитывался порядок. Далее привожу таблицу, где пересечение множеств не дало пустое множество.

Word1&Word2&¬Word3	Document
('гарри', 'заколдовать', 'волшебный')	{27, 52, 39}
('гарри', 'заколдовать', 'палочка')	{27, 52}
('гарри', 'заколдовать', 'думбльдор')	{27, 52, 39}
('гарри', 'волшебный', 'гермиона')	{3, 4, 5, 6, 7, 13, 15, 23, 24}
('гарри', 'волшебный', 'палочка')	{40}
('гарри', 'волшебный', 'думбльдор')	{3, 4, 40, 9, 10, 11, 13, 23, 24, 29, 30}
('гарри', 'гермиона', 'палочка')	{32, 37, 40, 42, 45, 46, 48, 49, 52, 54, 55, 56, 26, 27, 31}
('гарри', 'гермиона', 'думбльдор')	{9, 10, 11, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 54, 55, 56}
('гарри', 'палочка', 'думбльдор')	{3, 4, 9, 10, 11, 13, 23, 24, 25, 28, 29, 30, 33, 34, 35, 36, 38, 39, 43, 44, 47, 50, 51}
('заколдовать', 'гермиона', 'палочка')	{27, 52}
('заколдовать', 'гермиона', 'думбльдор')	{27, 52, 39}
('заколдовать', 'палочка', 'думбльдор')	{39}
('волшебный', 'гермиона', 'палочка')	{40}
('волшебный', 'гермиона', 'думбльдор')	{40, 9, 10, 11, 29, 30}
('волшебный', 'палочка', 'думбльдор')	{3, 4, 9, 10, 11, 13, 23, 24, 29, 30}
('гермиона', 'палочка', 'думбльдор')	{33, 34, 35, 36, 38, 39, 9, 10, 11, 43, 44, 47, 50, 51, 25, 28, 29, 30}

Задание 1.3.

Составьте список из 10 лексем/словоформ одного из документов коллекции; в список должны войти:

шесть ключевых слов,

два частотных слова из топ-100 по частотному словарю (см. Новый частотный словарь русского языка http://dict.ruslang.ru/freq.php или другой частотный список);

два редких слова (встретившихся во всей Вашей коллекции не больше трех раз, а в Вашем тексте - не больше одного).

Так как список ключевых слов составлялся на основе отрывка одного из документов, то наличие 6 слов из этого списка и два частотных слова могут задать множество, состоящее только из этого же документа. Поэтому именно я перебрала возможные комбинации из списка ключевых слов, для того, чтобы выбрать какой-то другой текст.

Так как tf * ifd был рассчитан ранее, дополнительных вычислений на этом шаге не потребовалось.

В итоге были выбран 20 текст и следующие слова:

Слово (Лексема)	Count(wi)	Fr(Coll) Countcoll(wi)/ N(Coll)	tf
быть	37	0.012036	0.008224
гарри	103	0.020707	0.022894
думбльдор	21	0.001899	0.004668
кошачий	1	0.000017	0.000222
наложить	2	0.000077	0.000445
не	112	0.023765	0.024894
палочка	14	0.001899	0.003112
троица	1	0.000017	0.000222
уизли	3	0.001124	0.000667
хогварц	6	0.000409	0.001334

Расчет tf.idf

Слово (Лексема)	Count(doc)	Count(wi)	DocLength	N	idf	tf	tf.idf
быть	57	37	4499	58	0.017392	0.008224	0.000143
гарри	56	103	4499	58	0.035091	0.022894	0.000803
думбльдор	13	21	4499	58	1.495494	0.004668	0.006981
кошачий	2	1	4499	58	3.367296	0.000222	0.000748
наложить	7	2	4499	58	2.114533	0.000445	0.000940
не	58	112	4499	58	0.000000	0.024894	0.000000
палочка	36	14	4499	58	0.476924	0.003112	0.001484

троица	2	1	4499	58	3.367296	0.000222	0.000748
уизли	29	3	4499	58	0.693147	0.000667	0.000462
хогварц	13	6	4499	58	1.495494	0.001334	0.001994

1) Соответствуют ли те слова, которые попали вверх списка, упорядоченного по убыванию tf.idf, Вашей интуиции?

Да, в топ списка попали все слова из списка ключевых слов

2) Все ли ключевые слова попали в верхнюю часть списка (в первые шесть слов), ранжированного по tf.idf?

Все, кроме одного (уизли).

3) Какие слова попали вниз ранжированного списка? Каковы их характеристики с точки зрения грамматических характеристик, семантики;

В конце списка оказались слова, выбранные из топа частотных слов для русского языка: быть и не.
Быть - глагол, несовершенный вид, непереходный, изолированное спряжение; не - частица; неизменяемое.

4) Как, по-вашему, должен быть устроен список «стоп»-слов, данные о которых нет смысла включать в таблицу?

В этот список должны попасть слова, для которых tf.idf либо 0, либо очень близкий к 0. Нам не нужны слова, которые встречаются практически во всей коллекции текстов, например, предлоги, междометия, цифры, частицы.

Усложненный вариант:

Определите тематический вес слов в одном из документов коллекции; отсортируйте слова из текста по tf.idf (выделить топ 20 и 20 с минимальным tf.idf).

Лексема	tf.idf largest	Лексема	tf.idf smallest
вернона	0.017175	люба	0.000061
петуния	0.015291	необходимый	0.000061
дудлить	0.015238	бы	0.000058
дядя	0.014883	забыть	0.000057
фигга	0.010351	остаться	0.000055

тётя	0.008807	кроме	0.000052
дементор	0.008596	тем	0.000048
думбльдор	0.006981	который	0.000043
миссис	0.004970	свой	0.000043
мундугнус	0.004538	один	0.000039
министерство	0.004092	момент	0.000037
дурслей	0.003852	только	0.000035
колдовство	0.003695	наш	0.000035
сын	0.003607	через	0.000031
сова	0.003554	чтобы	0.000027
кухня	0.003475	другой	0.000024
конверт	0.003313	без	0.000023
слушание	0.003292	для	0.000012
авоська	0.002994	не	0.000000
уладить	0.002708	себя	0.000000

Дополнительные вопросы:

1) какие слова из списка тематически значимых слов, составленного вручную, вошли в список топ 20 слов по tf.idf, а какие не вошли;

   Вошло только слово *думбльдор, остальные (гарри, наложить, уизли, палочка, хогварц, быть, не, троица, кошачий) не вошли.*

2) задайте пороговое значение по tf.idf для ключевых слов;

   Наверное, 0.006

3) какие слова, на ваш взгляд, имеют высокий tf.df (выше порогового значения), но не являются ключевыми;

   Кажется, что слова дядя и тетя не совсем подходят для того, чтобы быть ключевыми, так как могут много где встречаться. Однако, зная сами рассказы, я бы предложила в качестве ключевых использовать словосочетания дядя Вернон и тетя Петуния, которые так же встретились в топе списка.

Задание 1.4.

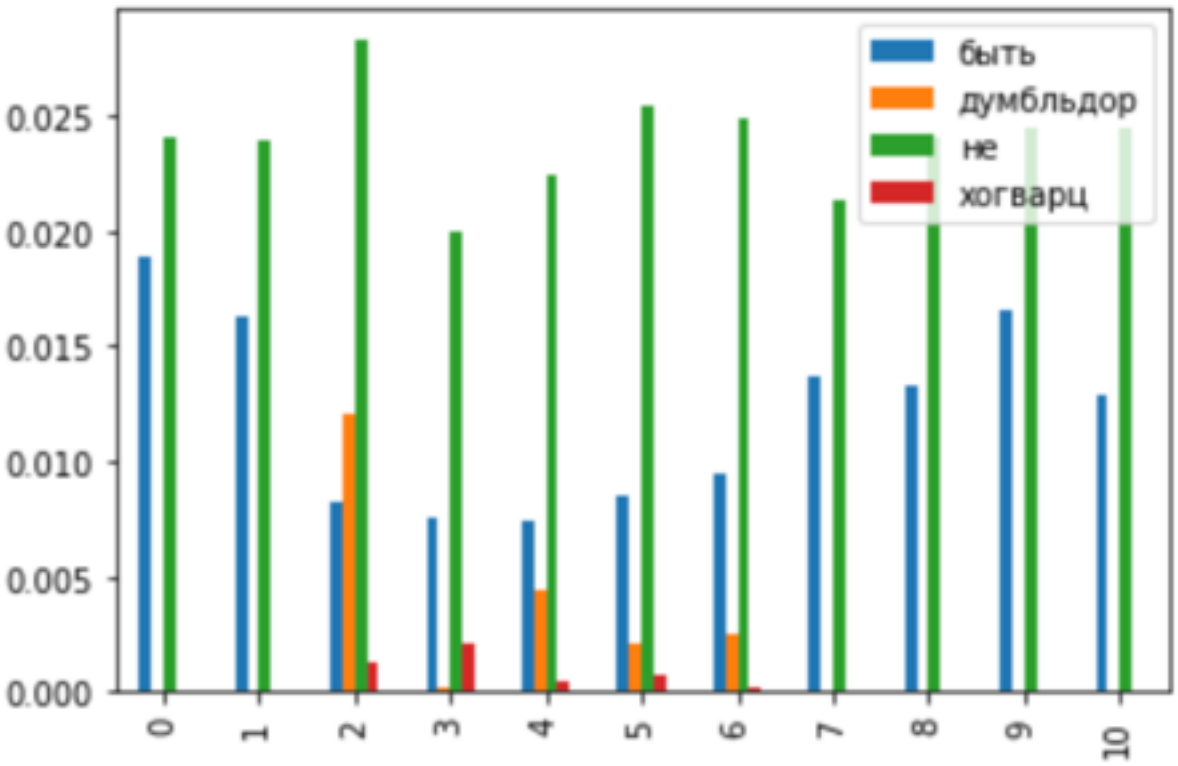
4 слова: два высокочастотных в языке - не, быть

два – с самым высоким tf.idf. - думбльдор и хогварц

ipm

	быть	думбльдор	не	хогварц
0	18823.082042	0.000000	24146.782013	0.000000
1	16332.020367	0.000000	23921.606302	0.000000
2	8168.693009	12063.069909	28305.471125	1234.802432
3	7553.303375	191.222870	19982.789942	2103.451573
4	7461.258848	4400.229577	22479.433710	478.285824
5	8452.844525	2089.467186	25453.509355	664.830468
6	9491.268033	2467.729689	24867.122248	94.912680
7	13736.263736	0.000000	21314.892005	0.000000
8	13230.013230	0.000000	24097.524098	0.000000
9	16539.923954	0.000000	24524.714829	0.000000
10	12889.773481	0.000000	24547.436262	0.000000

Диаграмма относительной частоты слова по подкорпусам



1) Отличаются ли диаграммы для самых частотных в языке слов и для слов с высоким tf.idf в Вашем списке, если отличаются, то чем?

> Как видно из диаграммы, частотные слова для русского языка присутствуют во всех частях коллекции. Практически для всех частей их частотность превышает частотности слов из нашего списка (за исключением 2 текста, в котором частотность слова думбудьдор превышает частотность слова быть).