Женя Миллер

HW1. Отчёт

1.1.

Я работала с корпусом рассказов Антона Чехова объёмом 192 документа и 403149 словоформ, для анализа мною был выбран его рассказ «Супруга» объёмом 2001 словоформ. В нём я выделила такие ключевые слова:

word	ipm	ipm_collection	ipm_text
николай	164.2	106.7	2998.5
евграфыч	0	14.9	2998.5
телеграмма	30.8	74.4	3498.3
ольга	116.7	488.7	4497.8
дмитриевна	8.8	39.7	4497.8
жена	376.8	1381.6	4997.5
теща	14.3	111.6	2498.8
ницца	3.4	9.9	1499.3
ножка	31.9	84.3	3498.3
развод	22.1	32.2	3498.3

Ответы на вопросы:

Есть ли среди выбранных вами ключевых слов редкие слова?

Да, имена собственные (*Ницца, Дмитриевна, Евграфыч)* и, например, *телеграмма* и *теща.*

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

Да, *жена* – топ-255.

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

Все выбранные слова – существительные.

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

Во всех текстах, понятно, встретились одни и те же предлоги, союзы, местоимения частотные глаголы (*говорить, мочь, видеть*).

1.2.

Шесть ключевых слов – *ольга, дмитриевна, ножка, жена, развод, телеграмма.*

Ещё три частотных слова (в художественных текстах) – *рука, вдруг, дверь.*

Пять текстов – рассказы Чехова «Анна на шее», «Ариадна», «Супруга», «Три года» и «Убийство».

	Супруга	Ариадна	Анна на шее	Три года	Убийство
ольга	9	0	0	0	0
дмитриевна	9	0	0	0	0
ножка	7	1	0	0	0
жена	9	13	6	40	4
телеграмма	7	1	0	5	0
развод	7	0	0	0	0
рука	2	13	15	78	16
дверь	0	1	0	17	7
вдруг	1	8	6	23	12
всего слов	2001	8290	3961	26993	9197

Найдите тексты, удовлетворяющие запросу Word1&Word2&¬Word3
– Слова *ножка* и *телеграмма.*

1.3.

10 лексем:

- шесть ключевых слов – *ольга, дмитриевна, ножка, жена, развод, телеграмма*
- два частотных слова из топ-100 по частотному словарю – *человек, время*
- два редких слова – *адюльтер, хирург*

DocLength = 2001 – количество словоформ в тексте

N(coll) = 403149 – количество словоформ в коллекции

N = 192 – количество текстов в коллекции

Лексема	Fr (L)	Count	Fr(Coll)	Count(w)	tf	Count(doc)	idf	tf*idf
дмитриевна	8,8	16	0,00003969	9	0,004497751	3	1,806179974	0,008124
ольга	116,7	197	0,00048865	9	0,004497751	7	1,438203189	0,006469
развод	22,1	13	0,00003225	7	0,003498251	6	1,505149978	0,005265
телеграмма	30,8	30	0,00007441	7	0,003498251	11	1,241908544	0,004345
ножка	31,9	34	0,00008434	7	0,003498251	21	0,961081934	0,003362
адюльтер	1,2	1	0,00000248	1	0,00049975	1	2,283301229	0,001141
жена	376,8	557	0,00138162	9	0,004497751	109	0,245874731	0,001106
хирург	22,1	2	0,00000496	1	0,00049975	2	1,982271233	0,000991
время	2015,7	559	0,00138658	5	0,002498751	131	0,166029933	0,000415
человек	2723	1354	0,00335856	8	0,003998001	173	0,045255126	0,000181

Fr (L) – частотность по частотному словарю.
Count – количество употреблений в коллекции.
Fr(Coll) – частотность в коллекции.
Count(w) – количество употреблений в тексте.
Count(doc) – количество текстов, в которых встретилось слово.

Ответы на вопросы:

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

Да, соответствуют.

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

Да.

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

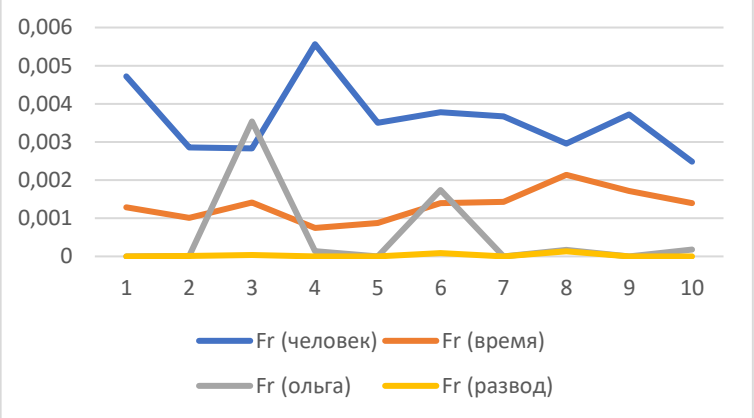
Вниз попали слова *человек* и *время* – одни из самых частотных слов в русском языке, семантически и стилистически нейтральные.

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

Возможно, слова служебных частей речи (предлоги, союзы), местоимения. Слова, частотность которых не меняется от текста к тексту.

1.4.

	Fr (человек)	Fr (время)	Fr (ольга)	Fr (развод)
1	0,0047197	0,0013	0	0
2	0,0028544	0,001	0,0000110	0,0000110
3	0,0028307	0,0014	0,00353832	0,0000354
4	0,0055631	0,0007	0,00013568	0
5	0,0035049	0,0009	0	0
6	0,0037771	0,0014	0,00173834	0,0000858
7	0,0036712	0,0014	0	0
8	0,00296	0,0021	0,00017187	0,000133677
9	0,0037217	0,0017	0	0
10	0,0024832	0,0014	0,00018059	0



На графике частотных слов из нашего текста видно пики, но в основном они очень редко встречаются. Частотные слова языка (*человек, время*) примерно одинаково распределены по текстам.