Света Павлова
группа 131

Домашнее задание 1

Проверим гипотезу о том, что междометия и частицы более свойственны устной речи, чем письменной. Для этого возьмем междометия *ну*, *ой*, э, *ах* и частицы (?) *вот*, *ж*. В НКРЯ будем искать также э-э, э-э-э и т. д., *а-ах*, *а-а-ах* и т. д. Для э и *ж* будет омонимия с инициалами - пренебрежем ей. Сравним междометия и частицы со словами *говорить*, *сказать*, *человек* (общеупотребительные частотные), *кошка*, *кот* (общеупотребительные нечастотные). При этом дополнительной гипотезой будет то, что *котик* как обращение или как уменьшительно-ласкательное будет специфично для устной речи.

Specific corpus: устный подкорпус НКРЯ. Объем: 11 349 008 слов, 3 665 документов.

Reference corpus: НКРЯ. Объем: 265 401 717 слов, 109 028 документов.

w_i	Тип	$Count_{SpecC}$	ipm	R	$Count_{RefC}$	ipm	R	LogLike	R	Weird	R	$tf\text{-}idf_1$	R	doc N	tf_2	df_2	$tf\text{-}idf_2$	R	chi-sq	R
ну	спец.	116362	10253.05	2	242021	911.90	5	311771.03	1	11.24	3	89623.71	1	15821	6.07	0.84	5.08	8	726191.16	1
вот	спец.	132661	11689.22	1	433456	1633.21	4	267277.03	2	7.16	4	48191.95	2	32957	6.12	0.52	3.18	9	532032.69	2
ой	спец.	11166	983.87	7	13888	52.33	9	38055.27	4	18.80	2	16851.93	4	3274	5.05	1.52	7.69	1	104227.66	4
э	спец.	25678	2262.58	6	16271	61.31	8	109366.90	3	36.91	1	33794.59	3	5023	5.41	1.34	7.23	2	347171.57	3
ж	спец.	8428	742.62	8	69898	263.37	6	6199.31	7	2.82	5	8480.56	7	9783	4.93	1.05	5.16	7	8826.16	7
ах	спец.	2821	248.57	9	38849	146.38	7	635.17	9	1.70	7	3528.41	9	5795	4.45	1.27	5.67	5	754.65	9
говорить	общ.	49480	4359.85	3	553066	2083.88	3	20252.83	5	2.09	6	12443.60	6	39157	5.69	0.44	2.53	11	25784.31	5
сказать	общ.	47986	4228.21	4	663883	2501.43	2	10633.21	6	1.69	8	15394.11	5	35248	5.68	0.49	2.79	10	12564.23	6
человек	общ.	43202	3806.68	5	776244	2924.79	1	2630.06	8	1.30	9	4029.19	8	48683	5.64	0.35	1.97	12	2847.96	8
кошка	общ.	620	54.63	10	11907	44.86	10	21.56	11	1.22	11	906.17	10	3641	3.79	1.48	5.60	6	22.93	11
котик	спец.?	46	4.05	12	1159	4.37	12	0.25	12	0.93	12	116.06	12	326	2.66	2.52	6.72	3	0.25	12
кот	общ.	521	45.91	11	9686	36.50	11	24.32	10	1.26	10	863.89	11	2344	3.72	1.67	6.20	4	26.14	10

Кроме LogLikelihood посчитаем показатели weirdness, tf-idf по двум формулам и хи-квадрат.

1. Частотность в SpecC и RefC. Частоты распределяются ожидаемо. Заметим, что *ой*, *э*, *ж*, *ах* наименее частотны среди выбранных междометий и частиц в обоих корпусах (следуют за общеупотребительными частотными *говорить* и т. д.). Кошачьи всех видов находятся в конце списка.

2. LogLikelihood
   Ну → вот → э → ой → говорить → сказать → ж → человек → ах → кот → кошка → котик.
   Большинство междометий/частиц ожидаемо наверху списка, однако *говорить* и *сказать* обгоняют *ж* и *ах*. Тут начинает закрадываться подозрение, что *ж* и *ах* вполне себе свойственны письменной речи.

3. Weirdness
   $w = (Count_{SpecC} / Total_{SpecC}) / (Count_{RefC} / Total_{RefC})$ (лекции, статья Лукашевич и Логачева)
   Э → ой → ну → вот → ж → говорить → ах → сказать → человек → кот → кошка → котик.
   Этот показатель выделяет *э* и *ой*, скорее всего потому что частотность *ну* и *вот* в общем НКРЯ всё равно достаточно высокая. Кошачьи сравнимы с человеком.

4. $tf\text{-}idf_1 = tf * \log_{10}((N - df)/df)$ (статья Лукашевич и Логачева)
   Ну → вот → э → ой → сказать → говорить → ж → человек → ах → кот → кошка → котик.
   Для подсчета tf весь устный корпус считается одним документом, что странно. Используются абсолютные частоты. Результат ранжирования очень похож на LogLikelihood за исключением перестановки *сказать* и *говорить*.

5. $tf\text{-}idf_2 = (1 + \log_{10}tf) * \log_{10}(N/df)$ (из лекций Ионова)
   Ой → э → котик → кот → ах → кошка → ж → ну → вот → сказать → говорить → человек.
   Высокий вес придается редким словам, поэтому котики выбиваются в начало списка (ура!). Но вообще это ранжирование больше говорит о редкости, чем о специфичности, что не соответствует нашей задаче.

6. Хи-квадрат
   Ну → вот → э → ой → говорить → сказать → ж → человек → ах → кот → кошка → котик.
   Нет значимой разницы между корпусами только у котика. Ранжирование в точности совпадает с LogLikelihood, но есть две проблемы:
   - величины слишком большие → какая бы маленькая разница ни была у корпусов, она будет значима.
   - ни величины хи-квадрат, ни соответствующие вероятности p нельзя использовать для сравнения разных слов (то есть фактически ранжирование не имеет смысла), потому что этот критерий говорит именно о значимости разницы, а не о её величине.

Междометия и частицы, по-видимому, действительно более свойственны устной речи, чем письменной (кроме *ж* и *ах*). Вот это открытие! Гипотеза о разговорности котика не подтвердилась (но надо учитывать, что для него мало данных).