

Исследование влияния даты сборки словаря на параметры quality_test.

По мотивам задачи:

<https://st.yandex-team.ru/FUNCTIONALITY-1670>

Последний ru словарь за 19 августа собирался по данным из MR за следующие даты: начальная дата 19 июля, последняя дата 17 августа.

Первый запуск был с sampling файлом, взятым из последнего ru словаря (sampling файл за 17 августа?). Вот таблица с основными параметрами:

date	recall	empty	saved_1	saved_4	saved_7	saved_10
2015-08-12	67.626800	34873	28.002300	38.977900	41.823100	43.163100
2015-08-13	67.716800	34862	28.192700	39.207800	42.029300	43.395800
2015-08-19	67.906800	34694	28.130300	39.133200	41.964100	43.208500

Тут приведены следующие значения:

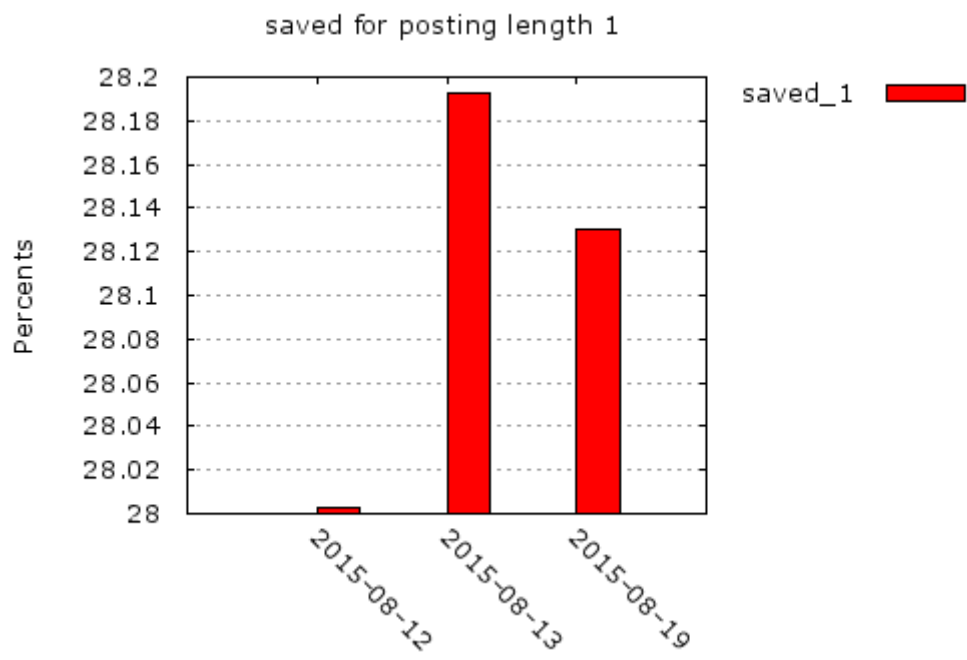
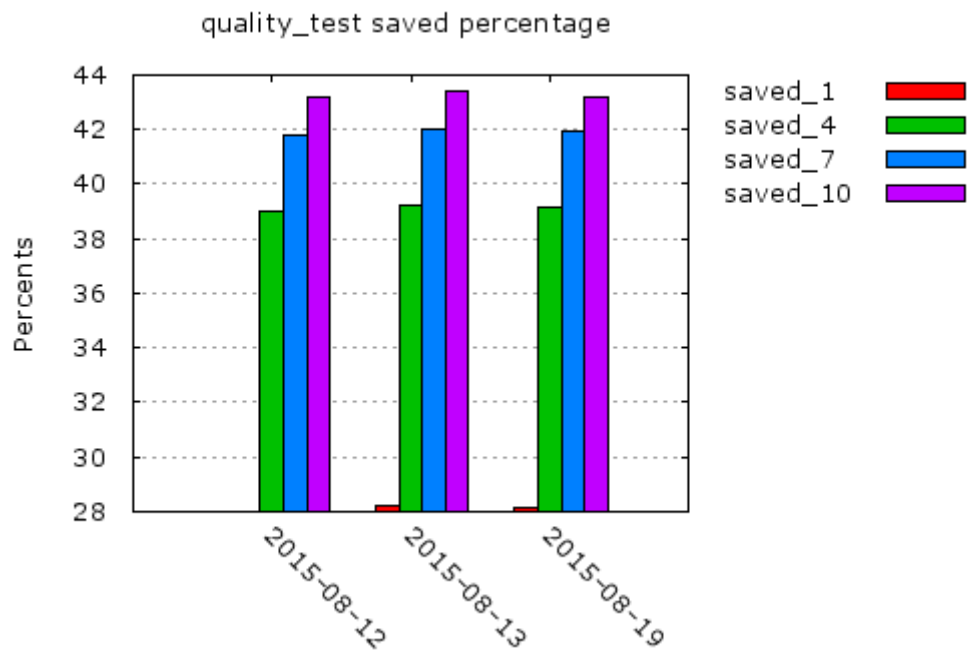
- date: дата словаря
- recall: значение параметра одноименного параметра из quality_test (проценты)
- empty: кол-во пустых ответов из quality_test
- saved_1: saved(проценты) при условии длины ввода 1
- saved_4: saved(проценты) при условии длины ввода 4
- saved_7: saved(проценты) при условии длины ввода 7
- saved_10: saved(проценты) при условии длины ввода 10

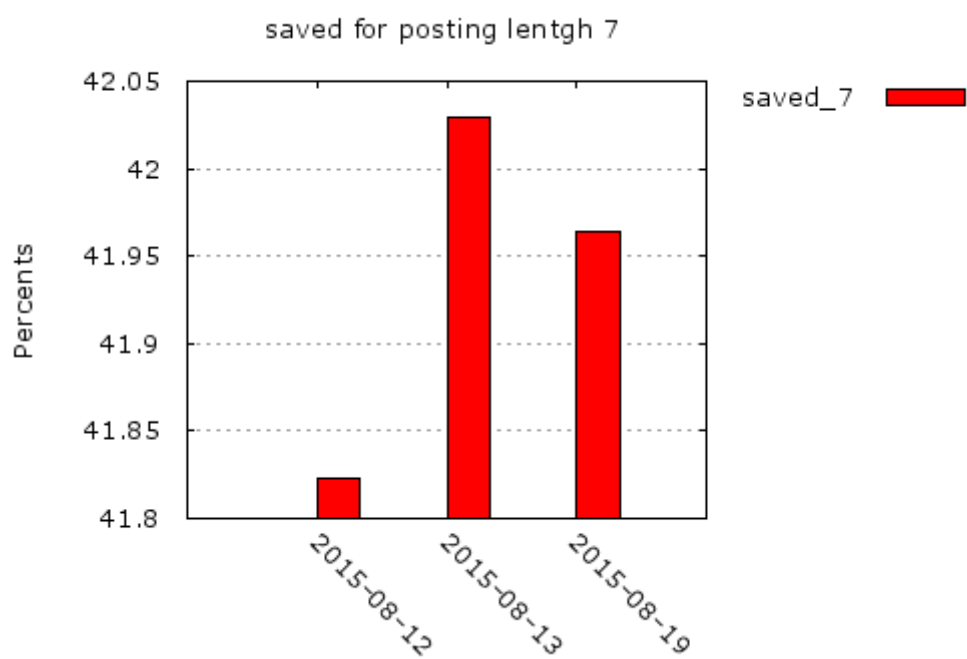
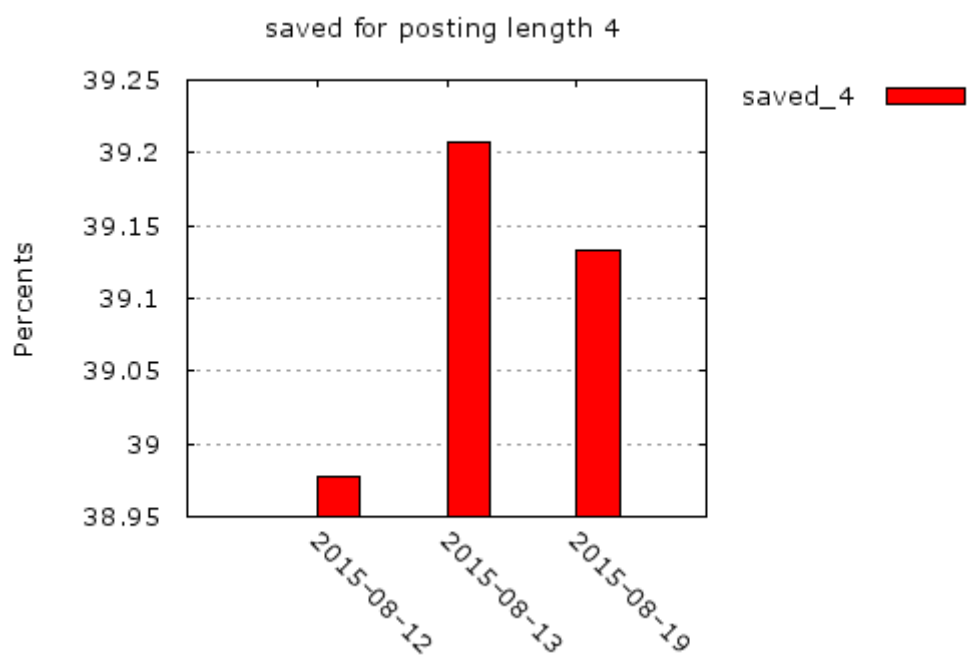
Ниже таблица разницы значений за последнюю даты и предыдущие даты.

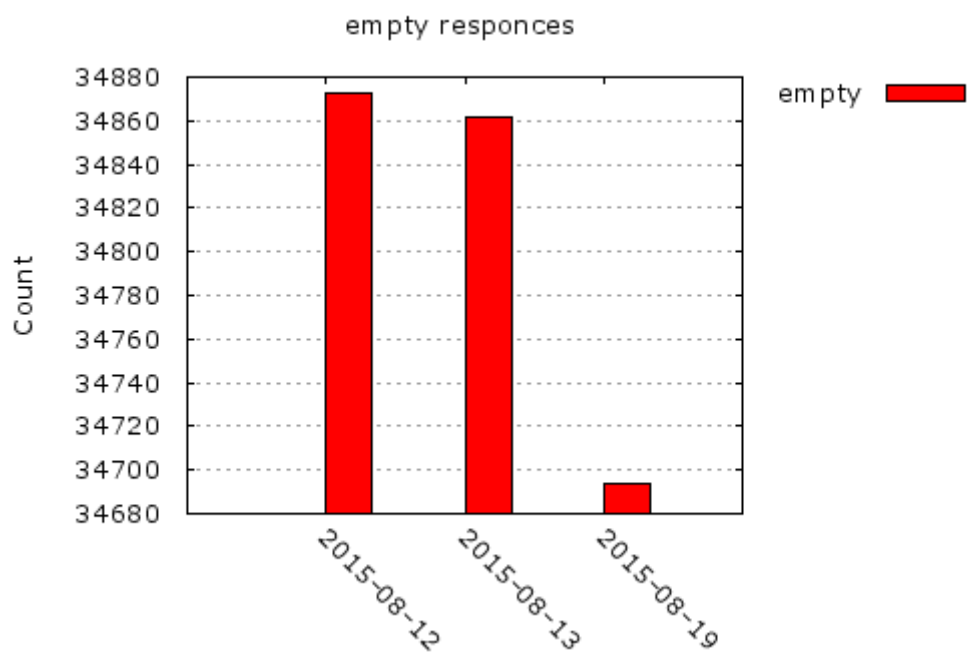
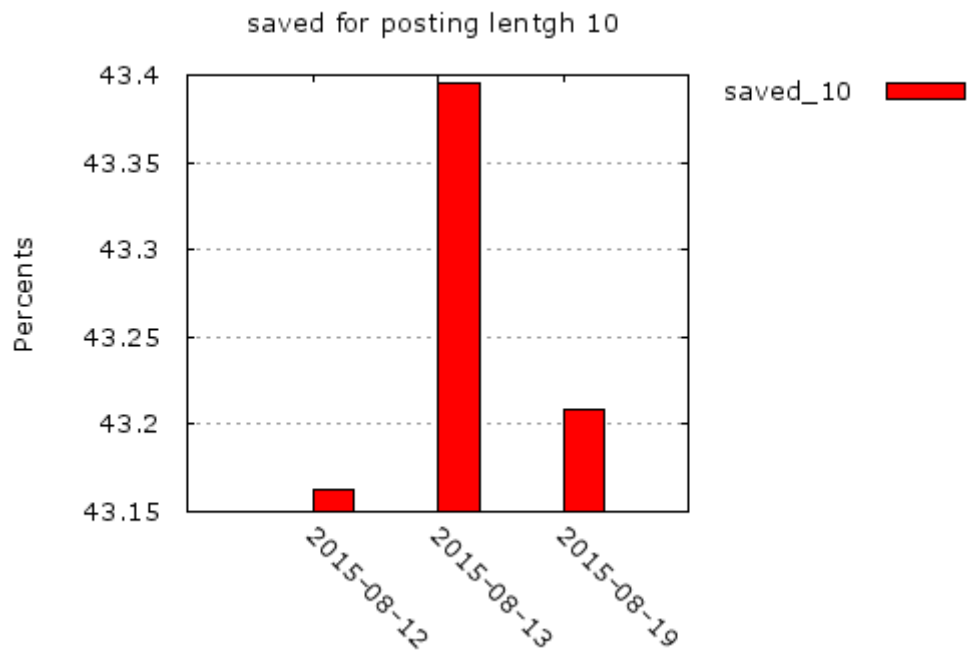
date	d_recall	d_empty	d_saved_1	d_saved_4	d_saved_7	d_saved_10
2015-08-12	-0.28	179.00	-0.13	-0.16	-0.14	-0.05
2015-08-13	-0.19	168.00	0.06	0.07	0.07	0.19
2015-08-19	0.00	0.00	0.00	0.00	0.00	0.00

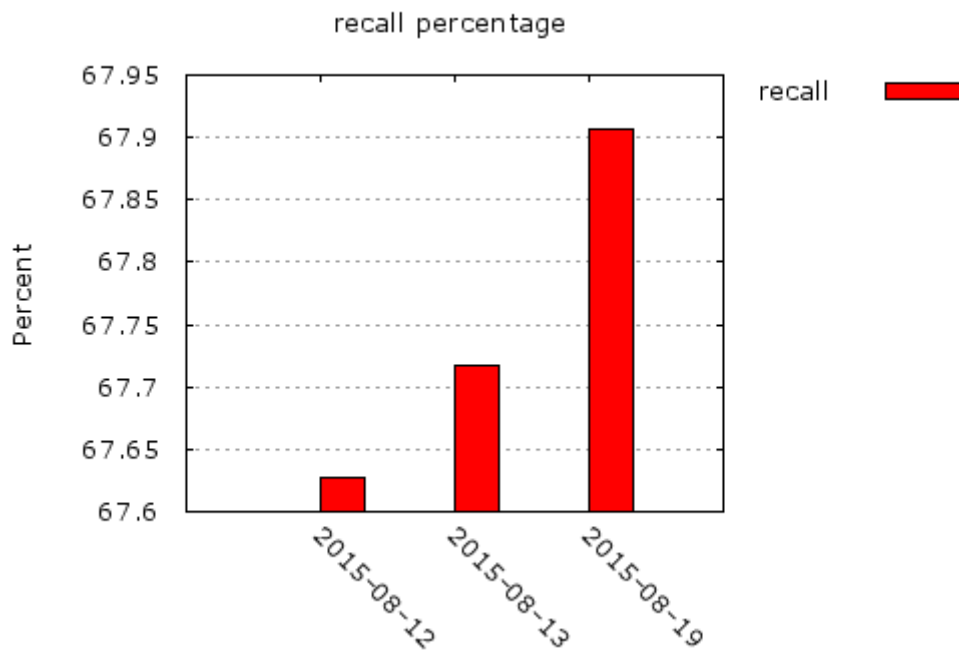
Видно, что тут нет такого очевидного и явного падения параметров quality_test. Например, saved_* параметры quality_test за 13 августа даже чуть лучше, чем за более позднюю дату – 19 августа. Но параметры recall и пустые ответы все-таки хуже для более старых словарей для всех дат.

Ниже графики основных параметров.









Я решил проверить параметр `quality_test` с `sampling` файлом, собранным за какое-то число в будущем. Ниже таблицы значений `qualit_test`.

Запуск `quality_test` с `sampling` файлом за **18 августа**. `Sampling` файл получил скриптом Алисы: `suggest_framework/suggest_scripts/suggest_tester/bin/load_sampling.sh`.

date	recall	empty	saved_1	saved_4	saved_7	saved_10
2015-08-12	67.130000	36824	27.759100	38.057700	40.618000	42.011000
2015-08-13	67.130000	36832	28.043200	38.252400	40.862300	42.162700
2015-08-19	67.260000	36617	27.863900	38.133800	40.706600	42.025600

Изменение параметров относительно последней даты:

date	d_recall	d_empty	d_saved_1	d_saved_4	d_saved_7	d_saved_10
2015-08-12	-0.13	207.00	-0.10	-0.08	-0.09	-0.01
2015-08-13	-0.13	215.00	0.18	0.12	0.16	0.14
2015-08-19	0.00	0.00	0.00	0.00	0.00	0.00

Как и в случае с запуском с `sampling` файлом, взятым из последнего `ru` словаря, тут не видно явного ухудшения параметров. Параметры, `saved_*` даже лучше для более раннего словаря от 13 августа, чем для словаря от 19 августа. Опять же, как и в предыдущем запуске, параметры `recall` и `empty responses` хуже для более старых словарей. Но незначительно.

Запуск `qualit_test` с `sampling` файлом от **19 августа**.

date	recall	empty	saved_1	saved_4	saved_7	saved_10
2015-08-12	66.740000	38543	26.490700	37.166100	39.738600	41.091000
2015-08-13	66.840000	38544	26.871900	37.386700	39.948300	41.321800
2015-08-19	67.040000	38402	26.364500	37.034900	39.611600	41.010200

Изменение параметров относительно последней даты:

date	d_recall	d_empty	d_saved_1	d_saved_4	d_saved_7	d_saved_10
2015-08-12	-0.30	141.00	0.13	0.13	0.13	0.08
2015-08-13	-0.20	142.00	0.51	0.35	0.34	0.31
2015-08-19	0.00	0.00	0.00	0.00	0.00	0.00

Картина аналогичная предыдущему запуску от 18 августа.

Запуск с sampling файлом за **20 августа**.

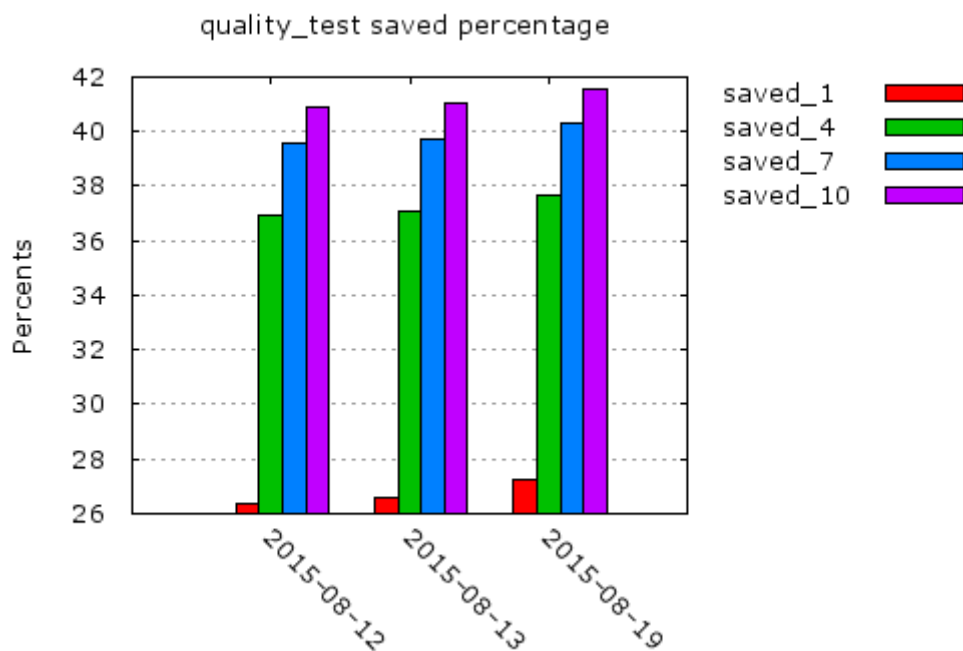
Получается такая таблица.

date	recall	empty	saved_1	saved_4	saved_7	saved_10
2015-08-12	66.320000	36658	26.380100	36.962100	39.608000	40.877500
2015-08-13	66.380000	36603	26.569500	37.096400	39.751100	41.024200
2015-08-19	66.700000	36406	27.261300	37.689300	40.290600	41.562900

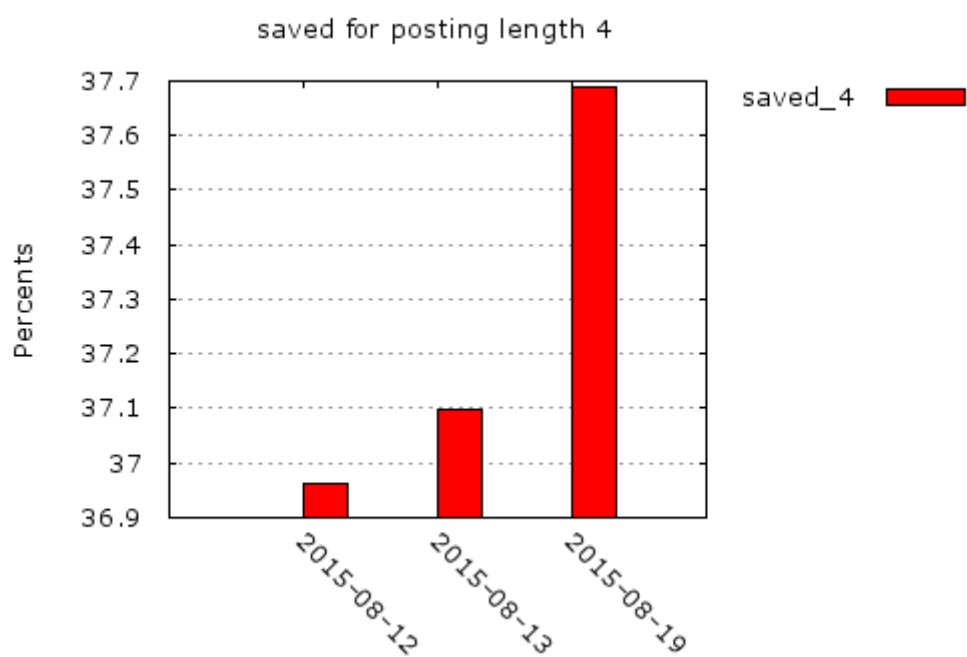
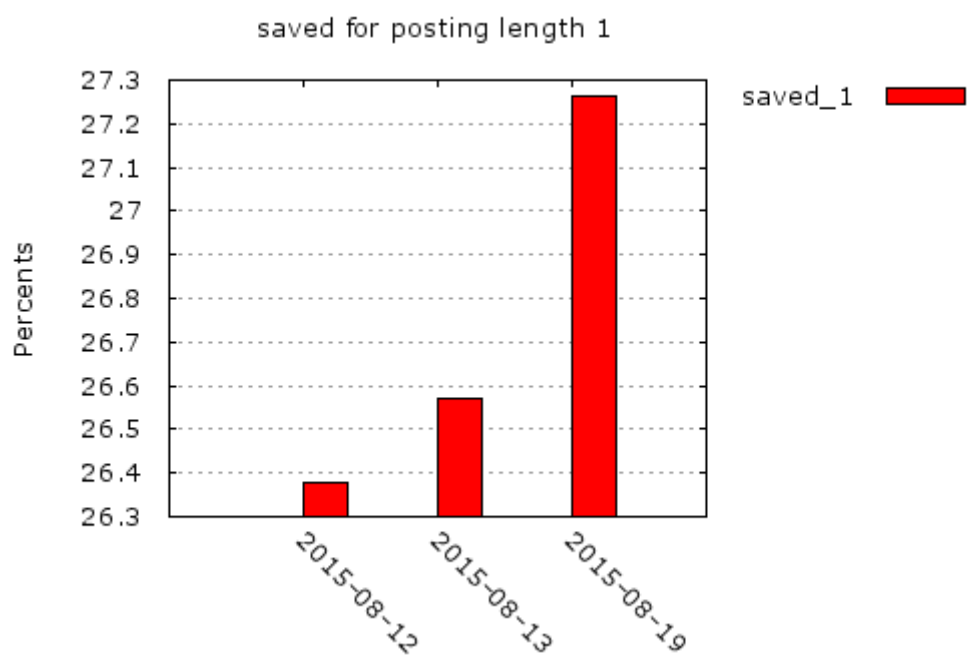
Ниже таблица разницы значений за последнюю даты и предыдущие даты.

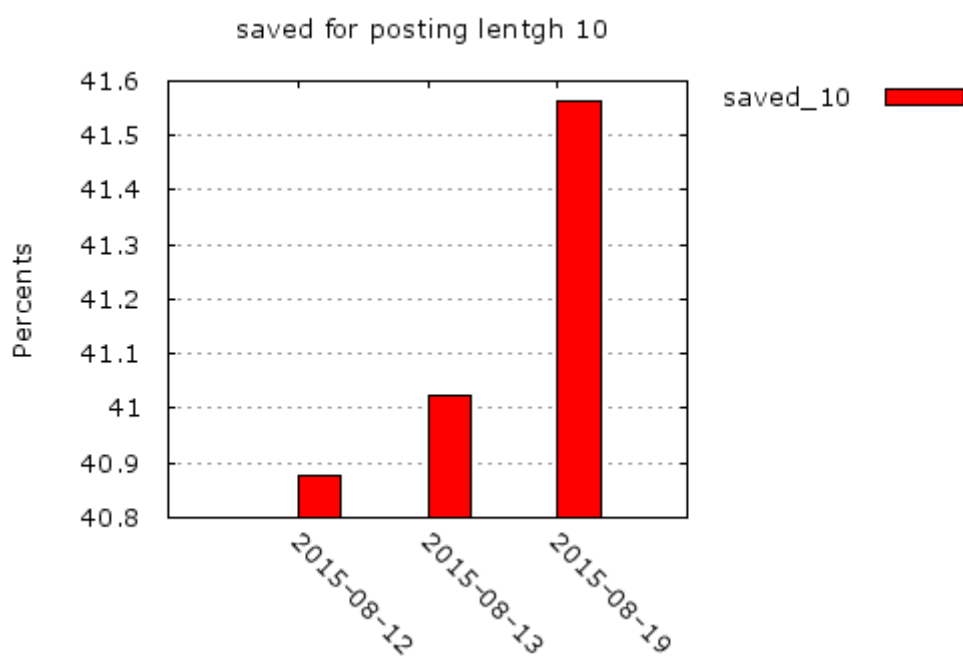
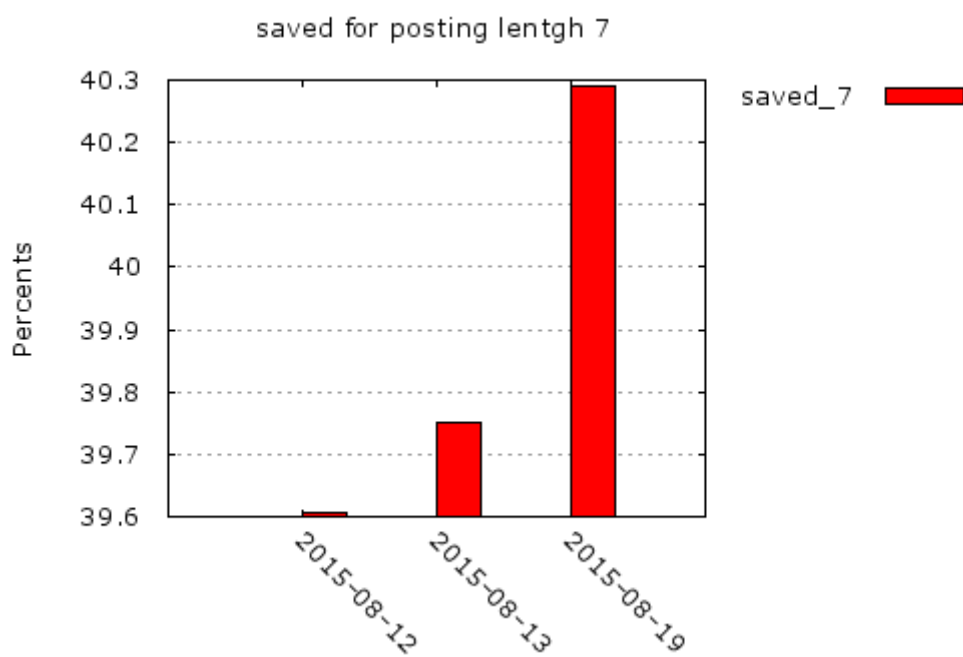
date	d_recall	d_empty	d_saved_1	d_saved_4	d_saved_7	d_saved_10
2015-08-12	-0.38	252.00	-0.88	-0.73	-0.68	-0.69
2015-08-13	-0.32	197.00	-0.69	-0.59	-0.54	-0.54
2015-08-19	0.00	0.00	0.00	0.00	0.00	0.00

А вот графики по датам. Ниже график изменения saved_* по датам.

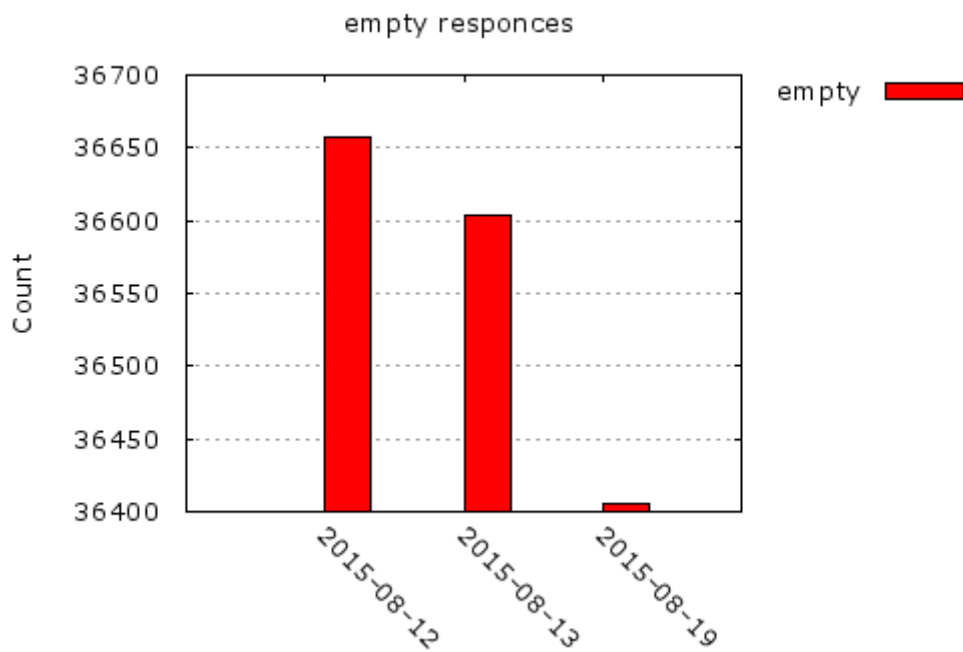


Вот график изменения каждого отдельного saved (так как на общем графике изменения меньших значений видны не очень хорошо).

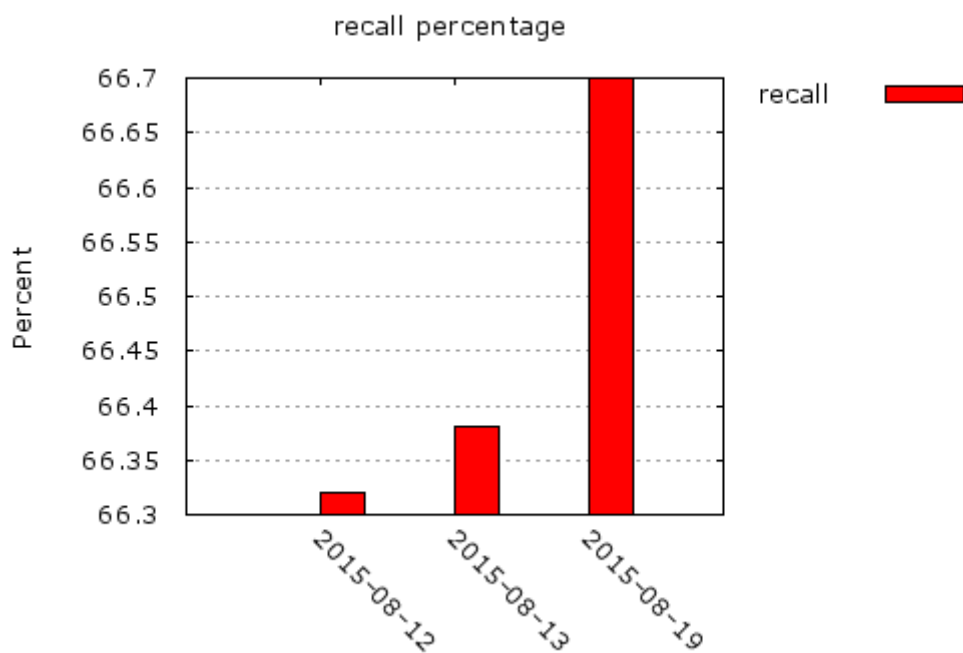




Вот график количества пустых ответов по датам. Тут, чем больше пустых ответов, тем хуже. В отличие от предыдущих графиков.



И, наконец, график recall:



Из графиков и таблиц видим “идеальную” картину ухудшения параметров теста `quality_test`. То есть, все основные параметры `quality_test` хуже для старых словарей, или, напротив, лучше для новых словарей) Насколько серьезно они ухудшились мне, честно говоря, тяжело судить. Ни один параметр, кроме количества пустых ответов, не вылез за границу в 1%.

Наконец, итоговый запуск с `sampling` файлом от **21 августа**. Получаю такие таблицы:

date	recall	empty	saved_1	saved_4	saved_7	saved_10
2015-08-12	66.810000	35228	27.494300	38.050100	40.515300	41.764700
2015-08-13	66.840000	35173	27.695100	38.195900	40.624500	41.906900
2015-08-19	67.480000	34983	28.306800	38.797100	41.318900	42.637500

date	d_recall	d_empty	d_saved_1	d_saved_4	d_saved_7	d_saved_10
2015-08-12	-0.67	245.00	-0.81	-0.75	-0.80	-0.87
2015-08-13	-0.64	190.00	-0.61	-0.60	-0.69	-0.73
2015-08-19	0.00	0.00	0.00	0.00	0.00	0.00

Тут мы также видим ухудшение основных параметров quality_test для старых словарей.

Итог. Запуск quality_test, к сожалению, не показывает каких-то явных и очень сильных изменений при двух разных подходах в задаче <https://st.yandex-team.ru/FUNCTIONALITY-1670>: при использовании sampling файла от последнего словаря и при явной сборке sampling файла за последующие даты. Для старых словарей всегда немного ухудшаются параметры recall (десятые доли процента) и количество пустых ответов (empty, ухудшается на 1.5-2%). Но параметры saved* не имеют ярко выраженную тенденцию ухудшаться даже на словарях, которые были собраны с разницей во времени в неделю(!). Иногда словарь собранный за более раннюю дату имеет даже лучшие параметры saved*, чем словарь собранный за более позднюю дату. В моих тестах получилось, что иногда(!) прогон с sampling файлом, который имеет более позднюю дату, чем сам словарь, дает более явную картину ухудшения параметром quality_test. Например, прогон тестов с sampling файлом за 18, 19 августа дал картину, аналогичную запуску с sampling файлом от последнего ru словаря (крайняя дата MR для словаря от 19 августа – 17 августа). А прогон с sampling файлом за 20, 21 августа дал явную картину ухудшения параметров quality_test для более старых словарей.