* 1. Для рекраулинга в БД git200\_crawl.sn\_activity фиксируется:
     + last\_date - дата последней активности группы / дата последней активности поста группы
     + upd\_date - дата последнего рекраулинга группы / дата последнего рекраулинга поста

Запись посл. активности группы, отличается от записи посл. активности поста тем, что в первой нет id поста, т.е. эта запись относится к группе в целом.

Автоматически созданный замещающий текст:
souces 
ec773e72.D53d-433D-sscc-s 
ff10d1ff.e8aS-47fD-S4fc-1dB8 
e3D78883-3ed3-481d.a275-7 
fc8Efd7f.e4S8-4fSf.b70c.acd 
ffd2594e.87D1-4734-8DBE-s,7 
1 531 
16 758 516 
16 758 516 
16 758 516 
1€ 758 
16 758 516 
16 758 516 
1€ 758 
114 70€ 21 08 
осчз 
1138SD 2D2D-11-1E 2351 
114 2D2D-1D-31 15:14 
114 2D2D-1D-23 2239 
114 422 2D2D-1D-23 2247:0043 
114 Зн 
ира «ate 
2020-11-17 
2020-11-17 2221Ђ2*аз 
2020-11-17 2221Ђ2*аз 
2020-11-17 2221Ђ2*аз 
2020-11-17 2221Ђ2*аз 
2020-11-17 2221Ђ2*аз 
2020-11-17 2221Ђ2*аз 

* 1. Параметры проектов git000\_cfg.projects для рекраулинга:
     + date\_deep - глубина скачивания. git000\_cfg.projects
     + recrawl\_days\_post - количество дней рекраулинга поста, от даты последней активности поста
     + recrawl\_days\_reply - количество дней рекраулинга ответов на сообщения, от даты сообщения

* 1. Принцип рекраулинга постов

Исходная постановка: при повторном краулинге посты дублировать не нужно, но новые обсуждения пользователей по данному посту нужно скраулить. Наличие новых сообщений пользователей для постов должны учитывать дату поста, дату последней активности поста, дату последнего рекраулинга поста. Т.е. очень старые посты не должны рекраулиться.

Первое ограничение - дата последнего рекраулинга группы upd\_date. Если дата поста меньше данной upd\_date группы и upd\_date группы не пустая, то считается что данный пост уже скраулен и не рекраулится. Как только попался первый ранее скрауленный пост, фетчинг страниц прекращается.

* 1. Принцип рекраулинга сообщений 1-го уровня (reply to post)

В алгоритм краулинга группы требуется передавать список постов, сообщения по которым нужно рекраулить. Для определения данного списка постов учитываются две даты - дата последней активности поста и дата последнего рекраулинга поста.

Пост включается в список рекраулинга до тех пор, пока дата его последнего рекраулинга не выйдет за период от даты последней активности, до даты последней активности + кол. дней рекраулинга поста. Т.е. условие включения post в рекраулинг сообщений следующее: post upd\_date in (last\_date, last\_date + recrawl\_days\_post)

Сообщения новых постов краулятся в полном объеме.

Сообщения старых постов рекраулятся если дата сообщения >= post upd\_date. Данное условие более «узкое» чем условие из следующего пункта, поэтому оно не используется в алгоритме.

* 1. Принцип рекраулинга сообщений 2-го уровня (reply to reply)

Сообщения с датой ранее даты последнего рекраулинга поста уже загружены, но не могут не рекраулится, т.к. на них могут быть даны ответы "reply to reply". Т.е. они до определенного момента должны рекраулиться, для проверки появления ответов на них. Для этой цели введен параметр recrawl\_days\_reply означающий количество дней, в течении которых система считает, что на сообщение может быть дан ответ.

Для вычисления берется дата посл. рекраулинга поста, из нее вычитается кол. дней ожидания: и все сообщения с датой позднее полученной - рекраулятся. Т.е. условия включения сообщения в рекраулинг следующее: reply date > post upd\_date - recrawl\_days\_reply

В алгоритм требуется подать словарь: (id\_post : upd\_date - recrawl\_days\_reply)