## Анализ частотности поисковых запросов

Работая с СПС Консультант Плюс пользователи задают поисковые запросы. Понимание того, как со временем изменяется интерес пользователей к определенным запросам, может быть полезен при решении различных внутренних задач, таких как анализ качества поиска, написание контента и др.

**Данные**

Файл с поисковыми запросами в формате: “userId \t day \t query”, где

* userId – уникальный идентификатор пользователя;
* day – порядковый номер дня предоставляемого периода времени;
* query – текст поискового запроса;

Кодировка файла – utf-8.

**Задачи**

1. **Поиск всплесков пользовательского интереса**

Интересы пользователей изменяются со временем: появляются новые потребности или «старые» потребности по каким-то причинам становятся более популярными.

**Требуется:**

* научиться определять периоды времени, в течение которых для запроса наблюдается всплеск пользовательского интереса (рост частотности запроса);
* на конкретный момент времени(день) ранжировать запросы по важности(силе) выявленного всплеска;

Важность(сила) всплеска должна вычисляться из тех соображений, что для дальнейшего использования (например, написания под этот запрос статьи на сайте) более интересны всплески по достаточно частотным запросам.

Более формально: для каждой пары (запрос, номер дня) требуется вычислить значение, характеризующее наличие или отсутствие, а также силу всплеска данного запроса на конкретный день.

При этом важно понимать, что на момент конкретного дня НЕ будут известны данные из будущего, поэтому при вычислении всплеска для пары (запрос, номер дня) могут использоваться только данные, известные на момент окончания этого дня.

1. **Выявление периодических потребностей(\*)**

**Требуется:**

* на основании всех предоставленных данных выявить периодические(сезонные) запросы (длительности периодов у разных запросов могут отличаться и быть от одного месяца до года).

Более формально: для каждого запроса требуется определить набор периодов, которые ему соответствуют.

1. **Прогнозирование частотности поисковых запросов(\*\*)**

В продолжение задач 1 и 2, можно поставить задачу прогнозирования частотности запроса в будущем.

**Требуется:**

* на основании всех предоставленных данных спрогнозировать частотность запросов на каждую из 4 недель, следующих после предоставляемого периода времени (частотность запроса на конкретную неделю – суммарная частотность данного запроса по всем дням этой недели).

Для оценки качества прогнозов можно использовать метрику [RMSLE](https://www.kaggle.com/wiki/RootMeanSquaredLogarithmicError).

**Оценка результата**

1. По каждой из сделанных задач необходимо предоставить воспроизводимый код на любом из следующих языков программирования (языки даны в порядке приоритета):

* Python (плюсом будет iPython notebook);
* Java;
* R;
* C++;
* C#.

1. Т.к. число уникальных запросов достаточно велико (бОльшая часть запросов имеет низкую суммарную частоту):

* для задачи 1. в качестве результата должен быть сформирован текстовый файл, содержащий информацию только о тех запросах, которые на протяжении всего предоставляемого периода времени хотя бы в один из дней попали в ТОП-100 по важности(силе) всплеска.
* для задачи 2. в качестве результата должен быть сформирован файл только с периодическими запросами (по ним должен определиться хотя бы один период).
* для задачи 3. в качестве результата должен быть сформирован файл с прогнозами для 1.000.000 наиболее частотных запросов за весь предоставляемый период времени.

1. Желательно в каком-то виде визуализировать получаемые результаты по каждой из сделанных задач (график частотности запроса от времени + результаты по задачам для данного запроса).

**Замечания**

1. Задачи со звездочками являются дополнительными.
2. В формулировках задачах специально не задаются жесткие критерии оценки качества получаемых результатов для возможности творчески подойти к решению каждой из задач.