# Формулировка задачи

Тематическое моделирование интересов реальных пользователей на основе их поисковых запросов

## Описание

Проанализировать пользовательские запросы к движку поиска в заданном формате (описание формата будет доступно) с целью:

1. выделить группы похожих запросов у разных пользователей (по UserId);
2. для выделенных групп запросов по "облакам слов" выработать метрику, максимизирующую качество маппинга некоторых из выделенных облаков на предоставленные тематики (облака слов для тематик из СКАН) - т.е. смаппить в каком-то виде полученные облака слов на тематики из СКАН (тоже облака слов);
3. для тех облаков слов из запросов, которые не удалось смаппить на тематики СКАНа сформировать названия потенциальных "тематик", покрывающих семантику этих запросов (тематическое моделирование).

## Требования

Дашборд. Можно использовать Jupyter Notebook.

**Метод 1**: по тексту поискового запроса получить список идентификаторов пользователей (userId) с похожими запросами и значениями метрики похожести.

**Метод 2**: по тексту поискового запроса получить название соответствующих запросу тематик из списка и значения метрики соответствия для каждой тематики.

Для п.2: в случае, если соответствия тематикам не найдены - возвращать потенциальные названия новых тематик. Если никаких предложений нет - то возвращать служебное сообщение о том, что ничего не найдено.

*\*Вместо дашборда может быть web-сервис, возможно даже без собственного UI - только REST-API (swagger), без авторизации.*

*\*\*Чем проще запустить для проверки, тем лучше. Идеально – исходный код, который собирается в контейнер и просто работает.*

# Продукты разработки

1. Отчёт, содержащий методику расчета, метрику/дерево метрик, позволяющие определять кластеры пользователей на основе схожести их поисковых запросов:
   1. разбиение пользователей на кластеры
   2. методику расчета, метрику/дерево метрик, позволяющие определять соответствие поисковых запросов тематикам из предоставленного списка
   3. вычисленные значения метрик соответствия запросов тематикам
   4. методика определения потенциальных названий тематик для запросов, которые не удалось поставить в соответствие тематикам из списка
   5. названия потенциальных тематик для некоторого набора групп поисковых запросов

## Планируемое назначение решения

Решение будет использовано для выдачи рекомендаций по составлению поисковых запросов на основе метрики их похожести, а также предлагать тематики для дальнейшей проработки и улучшения алгоритмов поиска.

## Какие проблемы решает разрабатываемый продукт

Решение должно:

* позволять выявлять группы пользователей с похожими интересами для проработки рекомендаций по составлению поисковых запросов;
* также выявлять потенциальные темы, которые интересуют наших клиентов с целью их дальнейшей проработки и улучшения алгоритмов поиска.

## Бизнес-функция продукта

* Оптимизация затрат
* Повышение конверсии и вовлеченности в продукт

## Целевая аудитория продукта

Клиенты-пользователи продукта Интерфакс СКАН.

## Под капотом

1. Пользовательские идентификаторы (UserId), разбитые на группы в соответствии с некоторой метрикой сходства их поисковых запросов
2. Облака слов и соответствующие им названия тематик из предоставленного списка тематик, сгруппированные по пользователям (для каждой тематики количество и список пользователей которые делали эти запросы)
   1. значение метрики качества (степень соответствия запроса тематике)
3. Сортировка по убыванию значения метрики качества и по количеству пользователей
4. Для запросов, которым не удалось поставить в соответствие тематики из списка, названия потенциальных "тематик", покрывающих семантику этих запросов

## Референсы

### **Сервисы заказчика**

* Веб-сервис <https://scan-interfax.ru/>
* Мобильное приложение “СКАН Мониторинг медиа” <https://clck.ru/34pYhj>

### **Примеры работ**

* Работы Воронцова <https://github.com/bigartm/bigartm>
* Исследование на Твиттере <https://trec.nist.gov/pubs/trec20/papers/Morpheus.microblog.update.pdf>
* Описание от Яндекса <https://yandex.ru/company/researches/2017/classifier>

# Оценивание

Оценка результатов проводится экспертно:

У заказчика будут формулы и бейзлайн, по которым он оценит предложенные разбиения пользователей на кластеры и значения метрик схожести запросов и соответствия тематикам.

# Что предоставляет заказчик

“Для решения нашей задачи никакие специфические доступы (в том числе к нашей системе) не требуются, т.к. все необходимые данные присутствуют в датасете. На выходе мы ожидаем какое-нибудь техническое решение (готовое приложение/сервис/контейнер/или даже просто код, который можно запускать) и какой-нибудь читабельный отчет с результатами. Техническую и отчетную часть можно совместить, например, в виде **Jupyter Notebook** или чего-угодно аналогичного. Особых требований у нас нет.”

## Датасеты

<https://drive.google.com/drive/folders/16rC5nqJ0V4iUQsrPOp7O68-QIp28UP5w>