1. Изучение wikidata относительно предметной области география, а также сбор классов и проставление коэффициента значимости.
2. Поиск пересечений между классами 1 и 2 уровня значимости.
3. По сформированному файлу Entities.xlsx с классами, для которых были проставлены оценки (0,1,2) проводился сбор представителей (файлы типа entities\_...) и поиск популярностей (файлы типа entities\_with\_popularity). Затем проводился сбор предикатов для каждого класса в целом (файлы типа …aggregated) и для топ 5% популярных (файлы типа …aggregated\_with\_popularity).
4. Файлы с предикатами сначала были очищены скриптом от технических, затем отданы на обработку первокурсникам.
5. По полученным файлам был сформирован файл Analytics.xlsx с данными по встречаемости предикатов среди всех и среди популярных. А также по всем сформулированным на тот момент интентам.
6. Создали файл со всеми предикатами из Wikidata.
7. Файл Analytics.xlsx был модифицирован в файл Entities\_from\_questions\_analytics\_в3.xlsx , в котором статистика дополнена данными с файлов Никиты Блохина с поиском QID по вопросам толокеров. Также добавлен файл Not\_found\_in\_our\_pool.xlsx c представителями, которые не были обнаружены среди наших 78 файлов.
8. Аналитика вопросов на покрываемость класса предикатами.Выявление интентов для возможного обогащения из внешних источников.
9. Создание файла Rules.xlsx с правилами по выбору предикатов.
10. Был создан файл Qualifiers.xlsx со всеми используемыми единицами измерений по нашим интентам.
11. Сбор файлов .ttl для дампа.
12. Формирование файла Искусственные\_классы.xlsx - объединение 78 классов + популярных из ненайденных в синтетические классы.
13. Формирование Sankey диаграмм (проход вверх по иерархии и нахождение общего надкласса).
14. Классы\_из\_вопросов.xlsx – объединение классов из вопросов в синтетические.
15. Очистка подклассов синтетических классов (пустые, нерусские и логические несоответствующих)
16. Формирование сводной таблицы для синтетических предикатов, а также создание файлов для высоты и расположения с оценкой процентов для возможных предикатов.