Часть 1. Работа с данными

Входные данные для тестового задания можно найти здесь.
Ваша задача - подготовить и обработать исходные данных так, чтобы их можно было использовать во второй части задания.

Требования к выходным данным:

1) В выходной таблице должны остаться только следующие колонки: `area, cluster, cluster_name, keyword, x, y, count, color`, где

- area - область,
- cluster - номер кластера,
- cluster_name - название кластера,
- keyword - словосочетание,
- count - показатель,
- x и y - координаты для диаграммы рассеяния,
- color - цвет точки на карте для данного словосочетания

2) Колонку `color` нужно добавить самостоятельно - цвета вы можете взять из цветовых палеток Tableu или по своему усмотрению.

3) Цвет задается каждому словосочетанию согласно следующими правилам:

- внутри одной области цвета словосочетаний в одном кластере должны быть одинаковые, в разных - отличаться (например, у "Кластер 1" все слова будут окрашены в красный, у "Кластер 2" - в зеленый и т.д.)
- цвета кластеров в разных областях могут повторяться
- цвета кластеров в разных областях с разным номером не имеют никакой связи (у одной области [`area`] слова из "Кластер 1" могут быть красного цвета, в другой области у слов из "Кластер 1" может быть другой цвет)

3) Не должно быть дубликатов слов в одной и той же области (`area`), но словосочетание может повторяться из area в area

4) Колонки должны называться именно так, как указано в п.1

4) Сортировка должна происходить по колонкам `area, cluster, cluster_name, count` (по `count` значения сортируются в убывающем порядке, в остальных - по возрастающему).

5) Количество переданных в исходных ключевых слов должно совпадать с количество слов в выходных данных (за исключением дублированных строк или строк с пустыми\неформатными значениями по ключевым показателям [перечислены в п. 1], если такие имеются).

6) Никакие другие особенности оформления не должны учитываться при обработке данных (заливка и пр.)

7) Выходные данные должны быть аккуратно оформлены (заголовки закреплены, включен фильтр)

Формат представления выходных данных: google spreadsheet-таблица.

Выполнение данной работы желательно с помощью библиотеки pandas (Python)

Часть 2. Построение графиков

На основании обработанных данных постройте по одной диаграмме рассеяния для каждой области (area) (пример внешнего вида см. в приложенном svg-файле).

Строгие требования к визуализации:

- Наличие Footer-подписи на изображении.
- Наличие легенды цветов и кластеров.
- Минимизация наложения (слепливания) подписей к друг на друга (постарайтесь сделать так, чтобы наложение было минимальным).

Желательные требования к визуализации:

- Перенос слишком длинных словосочетаний (например, слова длиннее 15 символов, можно разбить на "solar\n cell").
- Обводка точек.

Формат представления выходных данных:
Png-файлы размером не менее 1500x1500 пикселей с визуализациями ***для каждой области (area)***.

Выполнение данной работы желательно с помощью одной из библиотек:

- Matploptlib (Python)
- plotly (Python) и т.п.

Строгие требования к результатам:

- Код для первой части задания (с комментариями и приложенным README)
- Код для второй части задания (с комментариями и приложенным README)
- Таблица с трансформированными данными (открыть доступ по ссылке)
- Набор визуализаций (выложить на облачное хранилище)

Формат представления кода:

- GitHub
- Визуализаций и таблиц - Google Docs