ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к проекту DEDUPLICATE для компании ФЕРТОИНГ

В Фертоинг, как и во многих компаниях, в связи с отсутствием стандарта присвоения имен карточкам учета, номенклатура замусорена многочисленными дубликатами, текстуально разными, но семантически идентичными. Часть заведены по небрежности, часть — умышленно, это своеобразное хитрое "резервирование" товара в новой, экзотически названной карточке с целью укрытия остатков от коллег из смежного отдела, которые в своей работе применяют те же расходные материалы. В итоге остатки одного и того же — по своим физическим/эксплуатационным характеристикам — распределяются в системе по нескольким виртуальным наименованиям. Поиск затруднен, агрегирование не происходит. Идут годы, сотрудники, авторы этих дубликатов, уходят, и про некоторые такие товарные "схроны" просто забывают. Осмысленное управление товарными запасами в такой системе невозможно: компания банально плохо понимает, чем в действительности располагает.

Назначение данной группы python-модулей — поиск таких семантических дубликатов. Номенклатура, отобранная по ключевым словам, разбирается парсером на нормированные атрибуты, описывающие потребительские свойства выборки. По результатам лексико-статистического анализа и попарного сравнения атрибутов, в той или иной логической строгости, потенциальным дубликатам присваивается оценка, характеризующая степень сходства.

Порядок работы итеративный: выбор наименований из источника по ключевым словам, разбор, формирование нового источника и ранжированного списка его ключевых слов и т.д. По соображениям конфиденциальности данных компании, некоторые записываемые кодом отчеты из данной демонстрации изъяты. Работа кода показана на примере ключевых слов *болт*, *винт*, *гайка*, *шайба*, *шуруп*.

Парсер командной строки снабжен справкой:

```
PS D:\YandexDisk\courses\Projects\dedupl> python main.py -h
usage: main.py [-h] [-e [EXCLUDE ...]] [-t THRESHOLD]
               source_file {any,all} keywords [keywords ...]
Argument parser for Fertoing 'Deduplicate' project

positional arguments:
  source_file           Input file with inventory items to parse.
  {any,all}             Argument manages items pick basing on presence of KEYWORDS.
                        Use corresponds to 'any' and 'all' builtins.
  keywords              List of words to pick items by.
                        Lower case matches any case, upper case matches exact input.

options:
  -h, --help            show this help message and exit
  -e [EXCLUDE ...], --exclude [EXCLUDE ...]
                        List of words to filter items out. Any word excludes item.
                        Lower case matches any case, upper case matches exact input.
  -t THRESHOLD, --threshold THRESHOLD
                        Min ratio of similarity of items in report. Defaults to 0.01.

Meredelin Evgeny, meredelin@pm.me, 2022
```