**Задание #2:**

Нужно предложить и обосновать подход для группировки идентичных наименований товаров.

 Пример входа:

[1] "Кисломолочный напиток Актимель ежевично-черничный 100 г"

[2] "Напиток Данон Актимель черника/ежевика 2,5% 6\*0,1л Россия"

[3] "Гель для душа Роскошная мягкость Черная орхидея Palmolive, 250 мл"

[4] "Гель Palmolive Naturals д/душа Роскошная мягкость Черная орхидея 250мл Италия"

[5] "Гель для душа Palmolive "Черная орхидея" 250 мл"

[6] "Корм Whiskas говядина/кролик канапе новогодний 85г Россия"

Пример выхода: [1 2] [3 4 5] 6 (см. вложенный файл)

**Файл: test\_task\_NLP.json**

****Ответ:****

**1. Данные необходимо привести к единой табличной форме. Необходимо произвести их очистку. Удалить все незначащие символы, знаки препинания и шумовые слова. Также необходимо произвести леммализацию. Данный процесс нужно контролировать, потому-что можно потерять смысл в названиях. Данные лучше разделить на несколько групп с разной степенью очистки.**

**2. Выполнить генерацию признаков (фичей). Можно применить классические nlp-методы: length, counts, алгоритмы Levenshtein, Jellyfish, SequenceMatcher, fuzzy, word2vec и другие. Можно применить рекурентные нейронные сети, они хорошо выявляют признаки из последовательностей.**

**3. Применить методы выявляющие структуру данных: матрица смежности, частотный анализ, графовые методы.**

**4. Модели обучать с использованием стекинг-подходов. Рекурентные сети для выявления признаков в данных, затем модели на деревьях решений, например XGBoost.**

**5. Тестировать полученные модели для группировки однотипных наименований товаров.**

****Обоснование**: п.1 и п.2 стандартный подход к работе с текстовыми данными. На этом этапе нельзя точно быть уверенным, какая комбинация признаков и какая степень очистки будет влиять на точность. Необходимо исследовать получаемые результаты. Плюс немаловажную роль играет заданная точность к разрабатываемому подходу.**

**Если изучить данные, то в глаза бросается некая структура, по которой составлены наименования товаров. Можно выразить в виде формулы:**

***наименование = название + марка товара + параметры товара***

**Этот тривиальный фактор может быть использован при обучении модели, потому-что все названия более-менее имеют стандартное название. Поэтому, в п.3 предлагается использование методов для выявления структуры данных, хотя они напрямую не совсем относятся к nlp.**