Лексическая кластеризация продуктового справочника методами машинного обучения без учителя для анализа данных. Фейзуллин К.М.

Научный руководитель – доцент , к.ф-м.н. Платонов Е.Н.

МАИ, Москва

Данная работа нацелена помочь разделить множество разнородных наименований в единые кластеры для создания новых признаков при решении задачи оттока клиентов в ретейл компании. Для более детального анализа покупок клиентов можно использовать справочник товаров – выделить определенные категории, бренды и назначения (типы продуктов). В нашем случае – это справочник одной из известных торговых сетей по продаже косметики и парфюмерии. Если категории и бренды в справочнике указаны верно, то тип продукта имеет неоднозначное название, которое может меняться при едином фактическом типе товара.

Для решения данной задачи был выбран высокоуровневый язык программирования Python версии 3.7, так как он имеет множество библиотек для удобной работы с данными разного рода.

Для лексического анализа был взять список типов продуктов, которые покупались за последний год. Данный список включает чуть более 500 уникальных наименований. Перед тем как заняться реализацией решения задачи, было решено изучить имеющиеся данные с помощью облака слов, где были ярко выражены наиболее часто встречающиеся слова или словосочетания. Таким образом удалось выделить общие понятия, которые будут иметь шумовую составляющую при дальнейшей кластеризации и будут расширять признаковое пространство – это значит, что в дальнейших шагах эти слова будут исключены.

Следующей проблемой стало разнообразие окончаний в русском языке. К счастью, с помощью библиотеки NLTK с поддержкой русского языка был произведен перевод всех слов в их начальную форму, что определенно повысит качество кластеризации.

Так как для кластеризации нужно придать наименованиям типов продуктов численное значение, то был выбран метод преобразования текстовой коллекции в матрицу вхождений токенов – слов. То есть каждому наименованию ставится в соответствии вектор размерности N, где N – количество уникальных токенов, а элементы вектора $X = \{x_1, \dots, x_n\}$ – количество вхождений слова в строку наименования типа продукта. Далее вектор нормализуется с помощью $l2$ – нормы.

Так как нет четкого представления, на какое количество кластеров мы можем разбить справочник, было решено воспользоваться методом распространения близости (Affinity Propogation)[1]. Помимо того, что алгоритм AP имеет меньшую ошибку, по сравнению с алгоритмом k – средних[1], одной из главных особенностей алгоритма является отсутствие количества кластеров как входного параметра. Количество групп - центров будет определено после исполнения всех итераций алгоритма.

В результате работы алгоритма удалось выделить 56 лексических кластеров, содержащие в каждом от 1 до 32 корректных наименований в 55 – и группах и одна группа была отведена для всех наименований, для которых не смог образоваться кластер, содержащий 131 наименование.

Целью следующего шага исследований будет задача обработки уже всех наименований типов продуктов справочника и дальнейшая автоматизация алгоритма, чтобы определять для новых типов продуктов соответствующий им кластер в базе данных.

Список использованных источников:

1. Frey, Brendan J. and Delbert Dueck. Clustering by Passing Messages Between Data Points // Science 315 (2007): 972 - 976.