УДК 676.22.017

О. Л. Панченко, магистрант
(БГТУ, г. Минск)

КЛАСТЕРИАЗЦИЯ И ВИЗУАЛИЗАЦИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

Кластеризация документов – это процесс обнаружения естественных групп в коллекции документов. Задачу кластеризации – дано множество объектов, в которой необходимо найти группы похожих объектов. Есть две основные проблемы: не известно количество кластеров и не известны истинные кластеры, которые нужно выделять. Кластеризация автоматически выявляет группы семантически похожих документов. Группы формируются только на основе попарной схожести описаний документов, и никакие характеристики этих групп не задаются заранее, в отличие от классификации документов, где категории задаются заранее.

В общем случае задача кластеризации текста распадается на две: техническая задача преобразования в некоторую матричную, векторную или любую другу модель и математическая задача кластеризации.

Сначала необходимо выполнить предварительную обработку документов. Она включает в себя следующие этапы: фильтрация; токенизация; стемминг; удаление стоп-слов; сокращение; создание взвешенной матрицы терм-документ – переход к векторной форме документа. Для этого используется преобразование TD-IDF.

Одним из главных моментов обработки является отбор признаков. Это делается по следующим причинам. Во-первых, признаков может быть слишком много, больше чем нужно. Во-вторых, существуют признаки, из-за которых при решении задачи возникает много проблем. В-третьих, ускорение модели.

Задача визуализации данных — это частный случай нелинейного понижения размерности, когда данные проецируются на плоскость или в трёхмерное пространство так, чтобы изображение наглядно показывало структуру объектов. Иерархические методы кластеризации позволяют строить довольно удобную визуализацию их работы, так называемые дендрограммы.

Кластеризация является единственным решением задачи, когда нет точного представления о составе и структуры данных, а ручной отбор сложен, либо не соответствует временным и человеческим ресурсам.