**Нейросетевое моделирование для автоматической настройки задач в системах обработки больших данных** Бобряков А.С.

Научный руководитель — доцент, к.ф-м.н. Осипова В.А.

МАИ, Москва

В настоящее время с увеличением объема хранимых данных возникает потребность эти данные обрабатывать для получения статистик и новой информации, необходимой для дальнейшего анализа. Для этого применяют системы обработки больших данных: Hadoop, Spark и другие.

Основной проблемой использования таких систем является их предварительная настройка. Для каждой выполняемой задачи необходимо минимизировать целевые затраты: время выполнения, используемую память, другие целевые критерии, присущие конкретной области. Для этого необходимо подобрать оптимальные значения параметров, которых предоставляется несколько сотен [1].

Большинство таких параметров настраиваются по предоставленному руководству один раз администратором, но рекомендованные значения могут негативно влиять на конкретную выполняемую задачу. Также обычно не учитывается топология архитектуры кластеров, текущая нагрузка, обязательные параллельные задачи, которые запускаются самой системой обработки больших данных (например, удаление мусора, перераспределение данных между узлами, очистка буферов и др.). Важно понимать, что большинство параметров непредсказуемо взаимосвязаны, например, при увеличении значения одного параметра можно получить и положительное, и отрицательное влияние на систему другого параметра. Чтобы не задумываться обо всех смежных процессах, система рассмотрена в виде “черного ящика”, который моделировался нейронной сетью.

В качестве архитектуры нейронной сети использовали три полносвязных слоя с оптимизатором Adam [2] на объеме данных в 1млн записей, которые отражают статистики выполнения одной map-reduce задачи с варьированием параметров конфигурации кластера, параметров системы Hadoop и присущих конкретной задачи характеристик. Это позволило получить модель для определения длительности выполнения задачи. Такую обученную нейросеть предполагается встроить в процессы запуска с целью предварительной оптимизации желаемого времени выполнения, что позволит целесообразно использовать доступные ресурсы в условиях высокой нагрузки на кластер. Список использованных источников:

1. Muhammad Bilal and Marco Canini. 2017. Towards automatic parameter tuning of stream processing systems. In Proceedings of the 8th ACM Symposium on Cloud Computing

(SoCC’17). ACM, 189–200.

1. Diederik P. Kingma, Jimmy Ba Adam: A Method for Stochastic Optimizatioт

[Электронный ресурс] // arXiv.org. 2015. URL: https://arxiv.org/abs/1412.6980

**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 в создании новых признаков для решения задачи оттока клиентов в ретейл компании. Для более детального анализа покупок клиентов можно использовать справочник товаров – выделить определенные категории, бренды и назначение (тип продукта). 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 стала особенность русского языка – склонения. К счастью, с помощью библиотеки NLTK с поддержкой русского языка был произведен перевод всех слов в их начальную форму.

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.