УДК 004.89

# Сравнительный анализ продуктовых предпочтений в России и Китае с применением машинного обучения

Чжан Цзяи (Университет ИТМО)

Научный руководитель: Добренко Наталья Викторовна (к.т.н., доцент ИКТ) (Университет ИТМО)

**Введение.** В ходе работы с помощью библиотеки python BeautifulSoup4 собраны данные из нескольких сайтов еды, которые имеют рецепты блюд России и Китая. Эти данные используются для создания тематического моделирования. После этого было произведено сравнение результатов полученных тематических моделей, созданных с помощью методов LDA и NMF. С помощью изучения тем, которые представляют особенность блюд китайских и российских, получены различия продуктовых предпочтений в России и Китае. Реализована разработка системы для рекомендации блюд по названиям ингредиентов с применениям встраивания слов.

**Основная часть.** В датасете о российских рецептах есть 38843 рецепта, а в датасете с китайскими рецептами 8689 рецептов. Основные исследования проводятся на этапах приготовления блюд и ингредиентах приготовления.

Для создания тематического моделирования использованы методы LDA и NMF [1, 2]. Чтобы оценить качество модели использовалась модель согласованности (Coherence Model). CV основана на скользящем окне, сегментации одним набором верхних слов и косинусной мере подтверждения, которая использует нормализованную поточечную взаимную информацию (NPMI) и косинусное сходство. В результатах метод NMF получает больше согласованности.

Затем созданы тематическая модель горячего блюда в России и тематическая модель горячего блюда в Китае. Произведено сравнение тем двух моделей, при этом было выявлено, что в основной состав в российских горячих блюдах входят курица, картофель и рыба и их разницы преставились на гарнирах и заправках. А в состав китайских горячих блюд входит больше различных видов продуктов и есть несколько тем, в которых не входит мясо.

После тематического моделирования разработана система для рекомендации блюд по названиям ингредиентов с применениям встраивания слов. Встраивание слов (Word embedding) – это собирательное название для набора языкового моделирования и изучения функций методы обработки естественного языка (NLP), при которых слова или фразы из словаря отображаются в векторы из

действительных чисел. Нет готовой обученной модели для еды, поэтому в работе для создания модели использовалась модель Word2Vec.

Чтобы оценить модели, проверка осуществлялась по синонимам. Для получения хорошего результата в работе использовалось косинусное сходство для поиска названия блюда, которое похоже на входящие ингредиентов.

В работе разработано 5 моделей для рекомендации блюд (горячее блюдо, салат, суп, закуска, выпечка). В результате пользователь получает название блюда и его энергетическую ценность.

**Вывод.** Согласно сравнительному анализу тем горячих блюд в России и в Китае можно выделить разницу между рецептами двумя странами, а именно это: в Китае обычно не едят горячие блюда в качестве основного продукта питания и в России в качестве основных блюд обычно подают горячие блюда.

При создании моделей тем по российским и китайским горячим блюдам в работе анализированы и изучены характеристики, различия и предпочтения этих двух стран. С помощью встраивания слов создана система для рекомендации блюд. Для вычисления сходства между входящими ингредиентами и блюдами в моделях использовалось косинусное сходство. Модели выдают нормальные результаты.

**Список использованных источников:**

1. Свердлов С. А. «Метод LDA в психологической оценке понимания текстов»// Казанский педагогический журнал №3, 2021.

2. Плешакова Е. С., Гатауллин С.Т., Осипов А. В., Романова Е. В., Марунько А.С. «Применение методов тематического моделирования в задачах распознавания темы текста для обнаружения телефонного мошенничества» // Программныесистемы и вычислительные методы. – 2022. – № 3. DOI: 10.7256/2454–0714.2022.3.38770.