УДК 004.8

СЕМАНТИЧЕСКИЙ АНАЛИЗ КОРПУСОВ ТЕКСТОВ С ПОМОЩЬЮ МЕТОДОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Шайкина А.А. (Университет ИТМО),
Функнер А.А. (Университет ИТМО),
Научный руководитель – к.т.н., Ковальчук С.В.
(Университет ИТМО)

Аннотация: Данная работа посвящена сравнительному анализу корпусов медицинских текстов с помощью моделей тематического моделирования и сегментации медицинских данных. Для семантического сравнения используются несколько корпусов неструктурированных медицинских текстов, в том числе из ФГБУ «НМИЦ им. В. А. Алмазова» Минздрава России и ФГБНУ "Научный центр здоровья детей" (НЦЗД).

Введение. Большой объем полезной медицинской информации хранится в неструктурированном виде на естественном языке в формате электронных медицинских карт (ЭМК). В лечебно-профилактических учреждениях используются разные медицинские информационные системы, клинические протоколы и форматы ЭМК. Всё это, а также профиль отделения и медицинского учреждения, где накапливаются данные, влияют на содержание и структуры хранимых медицинских текстов. Целью данного исследования является провести сравнительный семантический анализ с помощью средств тематического моделирования и сегментации для определения различий в корпусах медицинских текстов, полученных из разных источников. Результаты данной работы могут быть использованы для моделирования цифровой клиники и цифрового профиля врача.

Основная часть. Для анализа предлагается использовать сегментацию медицинских записей. Она позволяет разделять исследуемый текст на предложения, каждое из которых принадлежит конкретной теме. Темы определяются с помощью тематического моделирования, которое определяет скрытые темы и соответствующие термины для каждой темы. Алгоритм построения тематической модели получает на вход наборы текстовых документов. Предварительно данные необходимо предобработать: соединить данные из разных медицинских систем, очистить данные от зашумлений, привести все слова к их нормальной форме.

После этого на основе «очищенных» данных строится тематическая модель с помощью библиотеки BigARTM - библиотека с открытым кодом для тематического моделирования больших коллекций текстовых документов и массивов данных. BigARTM по скорости вычислений опережает другие доступные библиотеки и имеет встроенную библиотеку регуляризаторов и метрик качества, и позволяет добавлять свои. С помощью регуляризаторов можно задавать желаемые свойства тематической модели. Для данной работы использовалось два типа регуляризаторов: сглаживающий регуляризатор для выделения слов фоновой темы и разреживающий регуляризатор тематических областей. С помощью построенной тематической модели по наибольшему совпадению слов из тем для каждого предложения определяется одна из полученных тем. После этого проводится сравнение тем, терминов и их частот для исследуемых корпусов.

Выводы. Семантический анализ текстов позволяет сравнивать разные наборы данных и определять особенности работы медицинских учреждений и медицинских информационных систем на основе предоставленных корпусов клинических записей. Предложенный подход основан на интеллектуальных методах обработки данных, не требует для начала работы размеченных текстов и может быть использован для любого нового корпуса записей.

Шайкина А.А. (автор)
Функнер А.А. (автор)
Ковальчук С.В. (научный руководитель)