КЛАСТЕРИЗАЦИЯ ЖАЛОБ ПАЦИЕНТОВ ИЗ БЛОКА ДОКУМЕНТА «ОСМОТР ЛЕЧАЩИМ ВРАЧОМ»

Е.В. Кащеева
Научный руководитель: С.В. Аксёнов, к.т.н., доцент ОИТ ИШИТР
Томский политехнический университет
E-mail: ev.kashcheeva@mail.ru

Введение

Многие сферы деятельности человека претерпевают различного рода изменения, это связано с совершенствованием, оптимизацией и автоматизацией определенных процессов. Для медицинских учреждений разрабатываются информационные системы, которые позволяют ускорить ввод информации и формирование на их основе различных отчетов и документов. В виду широкого развития аналитики данных и машинного обучения, появляется возможность извлекать полезную информацию из собранных данных, выявлять определенные закономерности.

Данные могут быть представлены в различных форматах: таблицы, изображения, аудио, видео, текст и т.д. Для каждого типа представления данных существует свой подход в обработке и анализе данных. Чаще встречающейся формой представления медицинских данных являются изображения и текст на естественном языке.

Постановка задачи

Отделением инфекционных заболеваний Сибирского государственного медицинского университета были предоставлены деперсонализированные истории болезни пациентов, страдающих рожистыми воспалениями. История болезни включает в себя документ «Осмотр пациента лечащим врачом», который содержит подробную информацию о состоянии пациента при поступлении в стационар.

Документ «Осмотр пациента лечащим врачом» включает в себя следующие 11 блоков: номер пациента, пол и возраст; дата и время осмотра; жалобы (описание беспокоящих пациента факторов, записанных со слов пациента); анамнез болезни (хронология развития симптомов заболевания со слов пациента до поступления под наблюдение врача); анамнез жизни (описание условий жизни и труда пациента, ранее перенесенных заболеваний); анамнез врачебно-трудовой экспертизы (ВТЭ, информация о листе нетрудоспособности); объективный статус (указание наличия или отсутствия патологий по каждому из рассматриваемых органов и систем организма); локальный статус (максимально подробные данные исследования поражённой системы); диагноз при поступлении; обоснование диагноза (обосновывается диагноз с указанием данных, которые подтверждают его); диагноз.

Целью данной работы является проведение кластеризации жалоб пациентов из блока документа «Осмотр пациента лечащим врачом».

Объектом исследования является история болезни пациента. Предметом исследования является блок «жалобы» документа «Осмотр пациента лечащим врачом».

Описание работы программы

Код программы, выполняющей кластеризацию жалоб пациентов из одноименного блока документа «Осмотр лечащим врачом», был написан на языке программирования Python. Данные хранятся в текстовом файле с расширением «.txt». Файл содержит в себе документы «Осмотр лечащим врачом» для 21 пациента.

Первым этапом работы программы является формирование списка жалоб по каждому пациенту. В найденном блоке удаляются все знаки препинания, кроме разделителей целой и дробной частей значений величины температуры. Также все знаки переводятся в нижний регистр.

Затем сформированные элементы списка жалоб подвергаются токенизации. Под токенизацией понимают разбиение текста на более мелкие части, токены [1]. К ним относят слова и знаки пунктуации. В нашем случае, в качестве отдельных токенов выступают слова.

Одной из особенностей работы с данными, представленными на естественном языке, является приведение слов к начальной форме. Данный процесс необходим, чтобы исключить принятие за разные слова различные формы одного слова. Процесс нахождения лексической основы для заданного исходного слова называется стеммингом (лемматизацией), наиболее известным алгоритмом стемминга является «Стеммер Портера» [2]. Принцип работы данного алгоритма заключается в отбрасывании суффиксов и окончаний, используя основные морфологические правила языка. Данный алгоритм реализован на языке Python и находится в пакете библиотек и программ для символьной и статистической обработки естественного языка «NLTK».

В таблице представлен пример жалоб пациента, полученных токенов, а также результата стемминга.

Таблица. Исходное предложение, результаты токенизации и стемминга

Жалобы пациента	повышение температуры с ознобом затем появление гиперемии на коже левой голени болезненность в левой паховой области
Результат токенизации	'повышение', 'температуры', 'с', 'ознобом', 'затем', 'появление', 'гиперемии', 'на', 'коже', 'левой', 'голени', 'болезненность', 'в', 'левой', 'паховой', 'области'
Результат стемминга	'повышен', 'температур', 'с', 'озноб', 'зат', 'появлен', 'гиперем', 'на', 'кож', 'лев', 'голен', 'болезнен', 'в', 'лев', 'пахов', 'област'

При работе с данными, представленными на естественном языке, помимо приведения слов к начальной форме также необходимо исключить слова, которые не несут никакой смысловой нагрузки. К ним относятся союзы, предлоги, местоимения, частицы и т.д. Такие слова называют стоп-словами. В библиотеке «NLTK» Python есть списки стоп-слов для различных языков, в том числе и для русского.

После того, как данные подготовлены для построения моделей, создается матрица весов TF-IDF. TF-IDF (term frequency – inverse document frequency) – статистическая мера, которая используется для оценки важности слова в контексте документа, который является частью коллекции документов. Вес слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции [3]. В данном случае, в качестве слов выступают токены, подвергшиеся стеммингу. А в качестве документов выступают жалобы отдельных пациентов, т.е. элементы сформированного списка жалоб. Для каждого слова рассчитывается вес, оценивается важность слова в пределах отдельного документа.

Затем мы приступаем к построению модели кластеризации. Было решено использовать метод k-средних. Основная идея алгоритма k-средних заключается в том, что данные произвольно разбиваются на кластеры, после чего итеративно перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике [4].

Прежде чем приступить к разбиению данных на кластеры, необходимо выяснить оптимальное количество кластеров. Для этого используется метод локтя. Данный метод подразумевает многократное циклическое исполнение алгоритма с увеличением количества выбираемых кластеров, а также последующим откладыванием на графике балла кластеризации, вычисленного как функция от количества кластеров. Балл является мерой входных данных по целевой функции, т.е. формой отношения внутрикластерного расстояния к межкластерному расстоянию. На рисунке 1 изображено графическое представление метода локтя. Мы можем увидеть, что точке, начиная с которой значения искажения перестают значительно уменьшаться, соответствует количество кластеров равное 19.

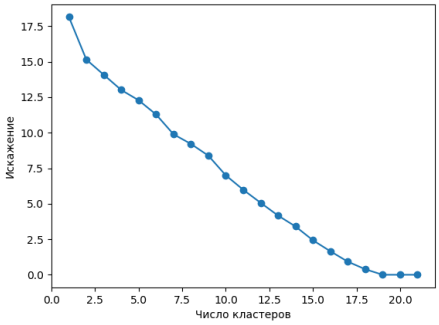

Рис. 1. Графическое представление метода локтя

Значение 19 является оптимальным количеством кластеров. Результат кластеризации на 19 кластеров представлен на рисунке 2. В левом столбце указан номер кластера, к которому относится жалоба из правого столбца.

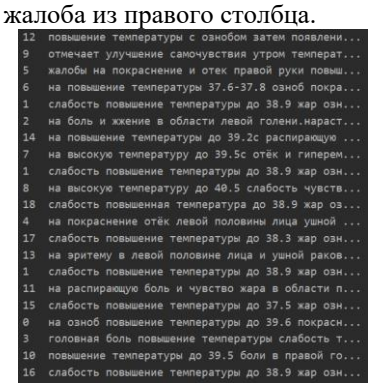

Рис. 2. Результат кластеризации

Заключение

Таким образом, при кластеризации жалоб 21 пациента, было выделено 19 кластеров. К первому кластеру относятся жалобы трех пациентов, к нулевому и со второго по восемнадцатый кластер относятся по одному элементу из списка жалоб отдельных пациентов.

Список использованных источников:

1. Забайкин, А.В. Функция токенизации текста на python [Электронный ресурс] / Заметки, идеи и скрипты. – URL: http://zabaykin.ru/ (дата обращения: 24.01.2020).
2. Хашин, С.И. Стеммер Портера [Электронный ресурс] / Полезные функции на C++. – URL: http://math.ivanovo.ac.ru/dalgebra/Khashin/cutil/porter.html (дата обращения: 24.01.2020).
3. TF-IDF [Электронный ресурс] / Википедия. – URL: https://ru.wikipedia.org/wiki/TF-IDF (дата обращения: 24.01.2020).
4. Алгоритм k-средних (k-means) [Электронный ресурс] / AlgoWiki. – URL: https://algowiki-project.org/ru/Алгоритм_k_средних _(k-means) (дата обращения: 24.01.2020).