КЛАСТЕРИЗАЦИЯ ЖАЛОБ ПАЦИЕНТОВ ИЗ ДОКУМЕНТА «ОСМОТР ЛЕЧАЩИМ ВРАЧОМ»

Е.В. Кащеева
Научный руководитель: доцент, к.т.н. С.В. Аксёнов
Национальный исследовательский Томский политехнический университет,
Россия, г. Томск, пр. Ленина, 30, 634050
E-mail: ev.kashcheeva@mail.ru

CLUSTERING OF PATIENT COMPLAINTS FROM THE DOCUMENT «EXAMINATION BY THE ATTENDING PHYSICIAN»

E.V. Kashcheeva
Scientific Supervisor: Assistant Professor, Candidate of Technical Sciences S.V. Axyonov
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050
E-mail: ev.kashcheeva@mail.ru

Abstract. *This article describes the stages of preparing text data for analysis and building a clusterer of patient complaints from the document "Examination by the attending physician". The article discusses the processes of tokenization, stemming, and deleting stop words. Also, the determination of the optimal number of clusters and the results of clustering using the K Means method are described.*

Введение. Многие сферы деятельности человека претерпевают различного рода изменения, это связано с совершенствованием, оптимизацией и автоматизацией определенных процессов. Для медицинских учреждений разрабатываются информационные системы, позволяющие 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. 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 состоит из 11 блоков и содержит подробную информацию о состоянии пациента при поступлении в стационар.

Целью данной работы является проведение кластеризации жалоб пациентов из блока 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 38 пациентов.

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

Таблица 1

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.

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

Затем строится модель кластеризации. Было решено использовать метод k-средних, основная идея которого 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», который 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 слева изображено графическое представление метода «Локтя». Можем увидеть, что точке, начиная с которой значения искажения перестают значительно уменьшаться, соответствует количество кластеров равное 36. Проанализировав данные вручную, были найдены слова, которые пишутся по-разному, однако смысл имеют один и тот же, так называемые синонимы. После замены синонимов повторно определили оптимальное количество кластеров, число которых сократилось до 35 (Рис. 1 справа).

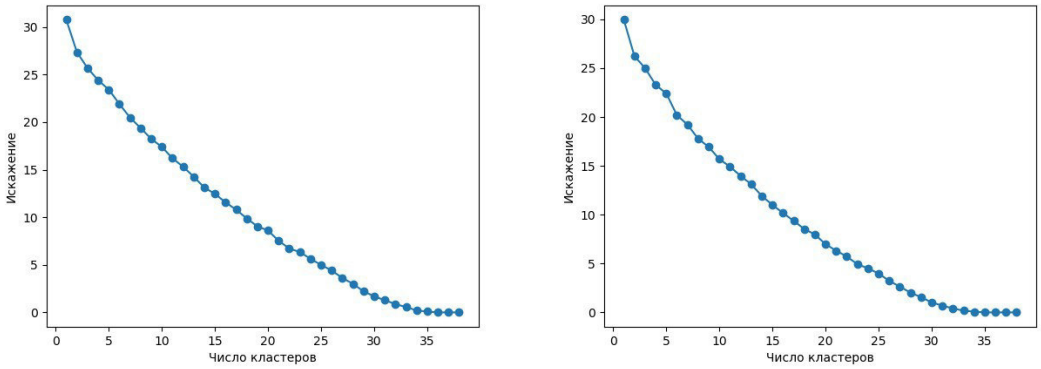

Рис. 1. Графическое представление метода «Локтя» до и после работы с синонимами

На рисунке 2 представлено распределение жалоб пациентов по кластерам. К нулевому кластеру относятся жалобы трех пациентов, к первому кластеру относятся жалобы двух пациентов, ко второму по тридцать четвертый кластер относятся по одному элементу из списка жалоб.

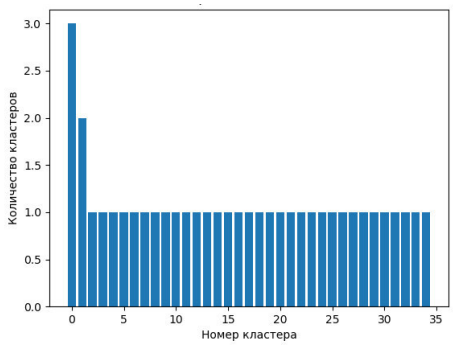

Рис. 2. Распределение жалоб по кластерам

Заключение. Таким образом, в рамках данной работы была проведена подготовка текстовых данных для анализа, после чего было определено оптимальное количество кластеров и построен кластеризатор. По результатам работы кластеризатора можно сделать вывод, что среди жалоб пациентов большинство составляют уникальные жалобы, однако некоторые из жалоб всё-таки были объединены.

СПИСОК ЛИТЕРАТУРЫ

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/Kha shin/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).