КЛАСТЕРИЗАЦИЯ ТЕКСТОВ ФИНАНСОВЫХ СООБЩЕНИЙ

Егоркин А. А.[1]
(*ФГБОУ ВО Российский государственный социальный университет, Москва*)

Работа посвящена задаче кластеризации текстов финансовых сообщений алгоритмами машинного обучения. С помощью алгоритмов кластеризации можно выделить группы похожих финансовых сообщений, выделить среди них однотипные или подозрительные, а также использовать в дальнейшем анализе найденные кластеры, а не сами тексты сообщений. В работе используются такие алгоритмы кластеризации как K-means, DBSCAN и иерархический метод кластеризации. В качестве текстов финансовых сообщений в работе используется информация о банковских транзакциях. Ввиду того, что банковские проводки подчинены строгим правилам бухгалтерского учета, устанавливаемым Банком России, представляется возможным ввести метрику оценки качества кластеризации. Данная метрика позволяет ранжировать между собой качество кластеризации с помощью алгоритмов машинного обучения, а также подобрать параметры, используемые при обучении данных моделей. Отдельное внимание в статье уделено особенностям используемых данных и тому, каким образом эти особенности могут быть учтены в практической части. В практической части работы приводятся результаты применения моделей кластеризации с указанием оптимальных параметров данных алгоритмов. В заключении делается вывод о лучших алгоритмах кластеризации применительно к финансовым текстам.

Ключевые слова: метод k-средних, DBSCAN, иерархический метод кластеризации, кластеризация финансовых сообщений.

1. Введение

Кластеризация текстов финансовых сообщений является важной задачей в области обработки естественного языка и анализа данных. Она включает в себя группировку схожих текстов, что дает возможность выделять ключевые темы, паттерны и закономерности в большом объеме данных.

Кластеризация позволяет выявить основные тренды в финансовых новостях и сообщениях. Финансовые сообщения могут быть очень разнообразными и касаться различных аспектов,

[1] *Антон Александрович Егоркин, аспирант* (2-5@bk.ru).

таких как новости о компаниях, изменения в экономической политике, финансовая отчетность и др. Кластеризация помогает сегментировать информацию на более управляемые группы, что облегчает анализ и извлечение необходимой информации [11].

С помощью кластеризации можно оценивать общее настроение по определенным темам или компаниям. Например, если кластер содержит множество негативных сообщений о компании, это может указывать на снижение доверия к ней на рынке. Такой анализ настроений имеет значение для принятия инвестиционных решений [12].

При анализе финансовых сообщений кластеризация может помочь выявить необычные или аномальные тексты, которые могут указывать на мошенническую деятельность или другие риски [8]. Аномалии в кластерах могут стать индикаторами необходимости более глубокого анализа или расследования.

При кластеризации текстов финансовых сообщений, представляющих собой банковские проводки, в настоящей статье предлагается использовать правила бухгалтерского учета, которым подчинены указанные транзакции. Это позволяет учесть типизацию банковских операций при обучении моделей кластеризации, оптимизировать выбор параметров данных моделей и по-новому взглянуть на метрики качества кластеризации.

С учетом постоянно растущего потока финансовой информации обработка и анализ больших объемов данных становятся сложными задачами. Кластеризация помогает структурировать и упростить данные, что делает их более доступными для анализаторов и аналитиков.

Таким образом, кластеризация текстов финансовых сообщений представляет собой мощный инструмент для анализа данных, который может существенно повысить эффективность работы инвесторов, аналитиков и исследователей. С ее помощью возможно более глубокое понимание рыночных тенденций, уменьшение рисков и повышение качества принимаемых решений. В условиях быстроменяющегося финансового рынка умение извлекать смысл из объемных массивов информации становится ключевым фактором успеха.

1.1. АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

В числовых методах кластеризации каждое текстовое сообщение представляется в виде точки в пространстве $\mathbb{R}^n$ с помощью отображения $\varphi: T \to \mathbb{R}^n$, которое каждому объекту данных $t \in T$ сопоставляет его векторное представление $\varphi(t) \in \mathbb{R}^n$. Числовые методы кластеризации могут быть разделены на методы разбиения, иерархические, плотностные, сетевые и модельные [2].

В настоящей работе используются следующие алгоритмы: *K*-means (*k*-средних) [17], относящийся к методам разбиения, алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [18], относящийся к плотностным методам, и иерархический алгоритм кластеризации [10].

Алгоритм *K*-means оптимизирует расстояния точек кластеров от центров этих кластеров, в случае использования евклидового расстояния оптимизируемая функция выглядит следующим образом [20]:

(1) $arg_s min \sum_{i=1}^{k} \sum_{x \in S_i} (\overline{x_i} - x)^2$,

где k – количество кластеров; S_i – состав i-го кластера; $\overline{x_i}$ – центр i-го кластера.

Алгоритм DBSCAN основан на плотности данных. Он группирует в один кластер точки, которые расположены близко к друг-другу, иными словами, находятся в ε-окрестности. В случае, если в ε-окрестности точки нет ни одного другого элемента, данная точка классифицируется как шум [13].

Иерархическая кластеризация – способы упорядочивания данных, направленные на создание иерархии (дерева) вложенных кластеров. Различают агломеративные и дивизионные методы: в первом случае кластеры создаются путем объединения наиболее близких между собой мелких кластеров, во втором – путем разделения крупных кластеров на более мелкие. Мерой близости элементов в данном случае могут выступать различные метрики, например, среднее расстояние, максимальное или минимальное расстояние между кластерами, расстояние Уорда(Ward) и другие [19].

Ключевые (гипер)параметры перечисленных выше алгоритмов представлены в таблице 1.

Таблица 1. Основные параметры алгоритмов кластеризации

Алгоритм кластеризации	Параметры алгоритма
K-means	Количество кластеров
DBSCAN	• Минимальное количество элементов в одном кластере; • ε-окрестность.
Иерархический метод	• Вид расстояния между кластерами; • Количество кластеров.

Каждый из приведенных алгоритмов имеет свои достоинства и недостатки, в дальнейшем они будут рассмотрены через призму решаемой задачи.

2. Особенности используемых в работе данных

Текстовые финансовые сообщения, используемые в настоящей работе, представляют собой банковские проводки, в атрибутах которых есть текстовое поле – назначение платежа. Здесь и далее под термином финансовый текст/сообщение будет пониматься поле «назначение платежа» из атрибутов банковской проводки. Другими атрибутами банковских проводок являются:

- наименование плательщика и получателя;
- сумма и валюта операции;
- номера счетов и банки-корреспонденты плательщика и получателя;
- дата и время осуществление платежа;
- иные атрибуты.

Ниже приведены примеры банковских проводок (таблица 2).

В качестве характерных особенностей рассматриваемых текстовых сообщений можно выделить следующее:

- короткое содержание сообщений: более 99% текстов имеют длину менее 200 символов;
- наличие большого количества сокращений: дог., сч., опл., тов.;
- наличие большого количества служебных слов/аббревиатур и их сокращений: млн, млрд, рубль, руб, ндс, ооо;

Таблица 2. Примеры банковских проводок

Назначение платежа	Сумма платежа, руб.	Номер счета отправителя	Номер счета получателя	Отправитель	Получатель
Погашение основного долга по договору n 0-0-0-1 от 01.01.11 согласно распоряжению 0002 от 01.01.23	1000	40702810900010000002	45205810900010000001	Клиент	Банк
Оплата за услуги по договору n 0/0/0/3 от 02.02.2023г. сумма 2000-00 руб. ндс не облагается	2000	30102810900010000004	40802810900010000005	Банк-контрагент	Клиент
Перевод денежных средств по договору № 004 –аб/вг от 03.03.23 г. ндс не облагается	3000	40702810900010000005	30102810900010000012	Клиент	Банк-контрагент

- наличие в тексте непосредственного указания времени и даты проводки или договора: дата, число, год, январь, февраль и т.п.;
- специфические синонимы: вклад – депозит, а не вклад в общее дело;
- орфографические ошибки и опечатки;
- повторяющиеся *n*-граммы[1] в одном сообщении.

После предварительного анализа 20 тыс. финансовых текстов автором был сделан ряд выводов. Используемый словарь в финансовых сообщениях довольно скуден. Частота употребления лемм[2] крайне неравномерна, первые по частоте употребления пять лемм встречаются в 31% сообщений, а первые 62 (5 + 57) леммы употребляются в 58% (30,6% + 27,2%) сообщений, см рис. 1 и 2. Смысл сообщений не зависит от порядка слов и может быть понят по двум-трем ключевым словам в сообщении.

[1] *n-грамма – последовательность, состоящая из n элементов, которые в рамках настоящей статьи могут быть словами или буквами, разделенные знаком пробел.*

[2] *Лемма – начальная форма слова, в русском языке для существительных и прилагательных это форма именительного падежа единственного числа, для глаголов – форма инфинитива.*

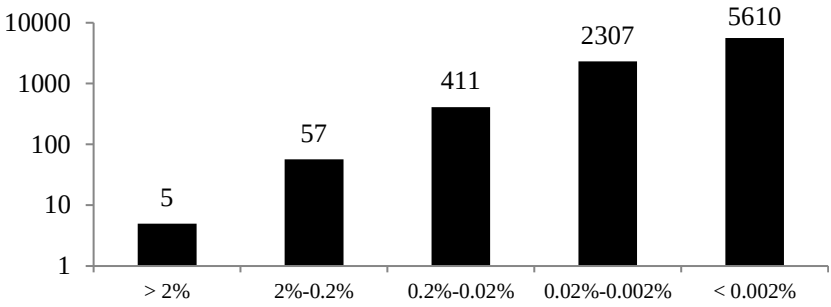

Рис. 1. Количество лемм, сгруппированных по частоте их употребления, логарифмическая шкала

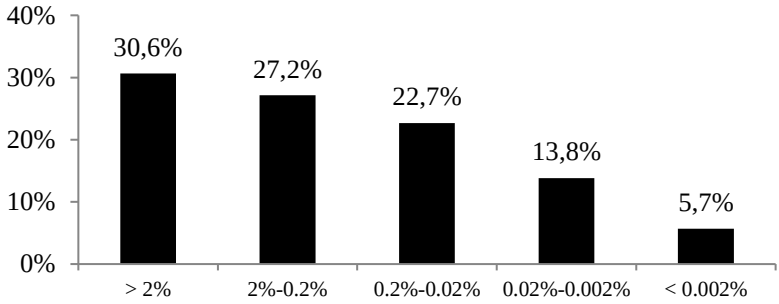

Рис. 2. Частота употребления лемм, сгруппированная по частоте их употребления

2.1 ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

Как уже отмечалось выше, исходные данные содержат достаточный объем «шума», дубликатов и других служебных символов, которые не несут в себе смысловой нагрузки, но могут ухудшать качество итоговой кластеризации. Для того чтобы отчистить исходные данные, а также сократить количество переменных проводится предварительная обработка данных.

В настоящей работе предварительная обработка данных была разделена на следующие этапы (таблица 3).

Таблица 3. Этапы предварительной обработки данных

№	Название этапа	Описание этапа
1	Лемматизация[1] и токенизация[2] с длиной токена ≥ 2 символов	Предварительное создание лемм токенов для дальнейшей обработки.
2	Замена сокращений и синонимов	Этап 2 идет перед этапом удаления стоп-слов[3], так как часть стоп-слов может использоваться в первоначальных данных в виде сокращений (пн, дек).
3	Удаление стоп-слов и пунктуации	В стандартный перечень стоп-слов добавляются служебные слова и аббревиатуры (тыс, руб, млн, ооо) и слова, описывающих дату и время (январь, понедельник, год …).
4	Токенизация с длиной токена ≥ 3 символов;	Токенизация с длиной токена, равной 3 и более символа, так как все токены с меньшей длиной были изменены или удалены на предыдущих этапах.
5	Создание массива *n*-грамм, состоящего из 1 токена	Промежуточная операция для дальнейшего использования. Массив упорядочивается по частоте употребления токена во всех сообщениях.
6	Упорядочивание токенов внутри сообщения, по частоте употребления токенов	На основании массива токенов, полученного на предыдущем этапе, осуществляется сортировка токенов внутри каждого сообщения. Так как смысл сообщения не зависит от порядка слов, то сортировка осуществляется по частоте употребления токена, так чтобы на первом месте всегда шли наиболее часто используемые токены.

[1] *Лемматизация— это процесс приведения всех встречающихся форм слова к одной, начальной форме слова- лемме.*

[2] *Токенизация — это процесс разбиения текста на минимальные значимые единицы, называемые токенами. Здесь токенами являются слова* (*леммы*), *предлоги, сокращения.*

[3] *Стоп-слова — это общеупотребительные слова в языке, которые, как правило, несут мало смысловой нагрузки* (*предлоги, частицы и т.п.*). *Также к ним могу относиться знаки препинания и числа. Удаление стоп-слов позволяет сократить объем данных для анализа.*

Таблица 4 (*продолжение*)

7	Перевод текста в векторное представление	Перевод всех текстовых сообщений в вектора. Элементы вектора представляют *n*-грамму из двух или трех токенов. При этом отсеваются *n*-граммы, частота появления которых в тексте ниже заданной границы. В дальнейшем эта граница будет являться одной из переменных при обучении моделей кластеризации.
8	Замена похожих между собой *n*-грамм	Замена в массиве векторов, полученном на предыдущем шаге, *n*-грамм, имеющих высокую схожесть между собой. При этом в итоговом массиве остаются наиболее часто употребляемые *n*-граммы. В качестве методов сравнения *n*-грамм между собой используются алгоритмы нечеткого сравнения строк: редакционное расстояние Левенштейна, расстояние Jaro, методы ratio [5].

По итогам предварительной обработки данных каждому текстовому сообщению ставится в соответствие вектор, состоящий из *n*-грамм. При этом используются *n*-граммы, состоящие из двух или трех токенов, которые наиболее часто употребляются во всех текстовых сообщениях.

Пример перевода текста в векторное представление, в соответствии с описанным выше алгоритмом (этапы 1–7), приведен ниже в таблице 4.

Первоначальное векторное представление данного текста будет иметь следующий вид:

$$\varphi_0 = \begin{bmatrix} \text{"договор облагаться"} \\ \text{"облагаться средство"} \\ \text{"средство денежный"} \\ \text{"денежный перевод"} \\ \text{"договор облагаться средство"} \\ \text{"облагаться средство денежный"} \\ \text{"средство денежный перевод"} \end{bmatrix}.$$

На этапе 8 из вектора φ_0 удаляются наиболее похожие между собой элементы. Для этого *n*-граммы сравниваются между собой, и среди схожих элементов отдаётся предпочтение наиболее часто встречающимся *n*-граммам. В качестве метрики

сходства используется максимальное значение из расстояния Jaro и метода set_ratio [5].

Таблица 5. Пример перевода текста в векторное представление

Исходный текст	перевод денежных средств по дог. № 004 от 03.03.13 г. ндс не облагается				
Стоп-слова и знаки препинания	по от ндс не № . 0 1 3 4				
Леммы токенов с длиной ≥2 символов без стоп-слов	перевод	денежный	средство	дог	облагаться
Замена синонимов и сокращений				договор	
Частота появления токена (порядковый номер)	41	11	9	1	5
Токены, упорядоченные по частоте	договор	облагаться	средство	денежный	перевод
Создание массива 2/3 грамм					
2 *n*-грамм	договор облагаться				
		облагаться средство			
			средство денежный		
				денежный перевод	
3 *n*-грамм	договор облагаться средство				
		облагаться средство денежный			
			средство денежный перевод		

Частота употребления *n*-граммы (*df*) – это доля финансовых сообщений, в которых встречалась соответствующая *n*-грамма во всем используемом наборе данных. В таблице 5 *n*-граммы упорядочены по частоте их употребления.

Таблица 6. Описание алгоритма удаления схожих n-грамм

df	*max (Jaro; Set_ratio)*	средство денежный	договор облагаться	облагаться средство	~~облагаться средство денежный~~	денежный перевод	~~договор облагаться средство~~	~~средство денежный перевод~~
5,3%	средство денежный		0,17	0,64	1	0,79	0,64	1
4,8%	договор облагаться	0,17		0,76	0,71	0,18	1	0,23
1,2%	облагаться средство	0,64	0,76		1	0,29	1	0,59
0,7%	~~облагаться средство денежный~~	1	0,71	1		0,67	0,83	0,81
0,7%	денежный перевод	0,79	0,18	0,29	0,67		0,28	1
0,6%	~~договор облагаться средство~~	0,64	1	1	0,83	0,28		0,48
0,6%	~~средство денежный перевод~~	1	0,23	0,59	0,81	1	0,48	

В рассматриваемом примере на первой итерации рассматривается *n*-грамма «средство денежный», наиболее близкие ей *n*-граммы: «облагаться средство денежный» и «средство денежный перевод». Так как последние две *n*-граммы реже употребляются, то они исключаются из дальнейшего анализа. На втором шаге рассматривается *n*-грамма «договор облагаться», наиболее близкая ей *n*-грамма – это «договор облагаться средство», последняя также исключается из итогового векторного представления. После того как были рассмотрены все *n*-граммы (с учетом удаления на предыдущих итерациях), итоговое векторное представление принимает следующий вид:

$$\varphi = \begin{bmatrix} \text{"средство денежный"} \\ \text{"договор облагаться"} \\ \text{"облагаться средство"} \\ \text{"денежный перевод"} \end{bmatrix}.$$

Таким образом, результатом работы алгоритма предварительной подготовки данных, описанного выше, является представление финансового сообщения «перевод денежных средств

по дог. № 004 от 03.03.13 г. ндс не облагается» виде вектора φ, компонентами которого являются следующие n-граммы: «средство денежный», «договор облагаться», «облагаться средство», «денежный перевод».

3. Требования к кластеризации финансовых сообщений

Задачи кластеризации относятся к задачам обучения без учителя [7]. Обычно неизвестно правильное разделение выборки на кластеры. За счет этого метрики качества кластеризации стремятся минимизировать расстояние между элементами одного кластера и максимизировать расстояние между кластерами. К описанной выше метрике можно отнести коэффициент силуэта [4].

В случае финансовых текстов некоторая информация о возможной классификации сообщения содержится в номерах банковских счетов.

Банковский счет, приведенный в таблице 2 и состоящий, из 20 цифр, имеет определённую структуру [9]:

- первые три цифры обозначают счет первого порядка;
- первые пять цифр – счет второго порядка;
- цифры с шестой по восьмую – код валюты (810 – российский рубль, 840 – доллар США и др.);
- девятая цифра – это контрольная сумма;
- цифры с десятой по тринадцатую обозначают код подразделения банка;
- последние семь цифр – это внутренний номер клиента в банке.

В соответствии с порядком бухгалтерского учета номер счета второго порядка характеризует вид клиента, который совершает операцию, а также и тип отношений между клиентом и банком. Например, счет второго порядка 45205 указывает на то, что клиент-юридическое лицо имеет кредит в банке, а счет 40702 – то, что клиент-юридическое лицо открыл текущий счет в банке. Таким образом, перевод денежных средств со счета 40702 на счет 45202 означает погашение юридическим лицом кредита, ранее полученного в банке. Связка счетов второго по-

рядка отправителя и получателя называется корреспонденцией счетов, в данном случае «40702-45202».

Логично предположить, что проводки с одинаковой корреспонденцией счетов второго порядка должны относиться к одному кластеру. Данное условие можно формализовать минимумом энтропии Шеннона [21] внутри одной корреспонденции счетов

(2) $H_{micro} = \frac{-\sum_j^M \sum_{i=1}^N C_j\, p_{ji}\, log_N p_{ji}}{\sum_j^M C_j},\ H \in [0,1],$

где p_{ij} – доля проводок, относящихся к кластеру *i, в j*-й корреспонденции счетов; *N* – количество кластеров; *M* – количество корреспонденций счетов; C_j – количество транзакций в *j*- корреспонденции счетов.

В то же время необходимо, чтобы между кластерами были максимальные отличия и при этом не было доминирующего кластера, в который были бы классифицированы большинство сообщений. Данное условие можно формализовать максимизацией энтропии Шеннона между кластерами:

(3) $H_{macro} = -\sum_{i=1}^N p_i\, log_N p_i,$

p_i – доля проводок, относящихся к кластеру *i*.

Итоговые требования к кластеризации финансовых текстов выглядят следующим образом:

(4) $\begin{cases} H_{micro} \to min, \\ H_{macro} \to max; \end{cases}$ или $\begin{cases} H_{micro} \to min, \\ 1 - H_{macro} \to min. \end{cases}$

Исходя из поставленного условия, необходимо соблюсти баланс между энтропией на микроуровне и на макроуровне. Эмпирически было получено, что наиболее простой функцией, которая позволяет обеспечить необходимый баланс является полином:

(5) $f(H_{micro}, H_{macro}) = [(H_{micro}^n + (1 - H_{macro})^n)/2]^{\frac{1}{n}}.$

При степени *n* = 1 целевой функции безразлично, что будет достигнуто минимум H_{micro} или максимум H_{macro}.

При *n* < 1 устойчивое равновесие будет находиться в крайних точках: либо при минимуме H_{micro}, либо при максимуме H_{macro}.

Устойчивое равновесие между H_{micro} и H_{macro} достигается при *n* > 1. Аналогичная функция при *n* = 2 используется

в [14, 15] для оценки стабильности работы алгоритма нахождения центральности в многослойных сетях.

Нормируем целевую функцию таким образом, чтобы она могла принимать значения в диапазоне [0, 1].

4. Результаты кластеризации на реальных данных

В качестве информации о транзакциях используются данные, используемые в статье [3]. Для анализа использовалось 20 тыс. штук финансовых транзакций.

Оптимальные переменные моделей кластеризации определялись посредством метода перебора параметров grid search [15].

В качестве глобального параметра для всех алгоритмов кластеризации использовалась минимальная частота употребления *n*-граммы – df_{min}, упомянутая в разделе 2.1 (таблица 6).

Таблица 7. Количество n-грамм, используемых в анализе

df_{min} (частота употребления *n*-граммы)	Первоначальное количество *n*-грамм	Количество *n*-грамм после сокращения	% сокращения *n*-грамм	Доля сообщений без *n*-грамм
0,002	435	244	–44%	6,7%
0,003	263	149	–43%	9,2%
0,004	206	117	–43%	11,2%
0,005	168	92	–45%	11,8%
0,006	135	74	–45%	13,1%
0,007	116	63	–46%	13,8%
0,008	101	55	–46%	14,5%
0,009	88	47	–47%	15,8%

Указанная величина показывает, какая минимальная частота употребления *n*-граммы в общем количестве сообщений необходима для того, чтобы включить данную *n*-грамму в качестве переменной алгоритма кластеризации. При этом чем будет выше данный показатель, тем меньше будет переменных для анализа и тем быстрее и устойчивей будет осуществляться расчет. Однако одновременно будет расти доля сообщений (редких

по контексту), у которых не будет выделено ни одной *n*-граммы, что будет приводить к ухудшению качества кластеризации.

Благодаря описанному выше алгоритму предварительной обработки данных удается сократить количество переменных почти в два раза.

На примере алгоритма *K*-means в таблице 7 иллюстрируется выбор оптимальных параметров кластеризации.

Таблица 8. Определение оптимальных параметров кластеризации при[1] $n = 2$

df_{min}	Количество кластеров						
	4	5	6	7	8	9	10
0,002	0,457	0,420	0,469	0,467	0,455	0,432	0,424
0,003	0,457	0,430	0,422	0,385	0,429	0,431	0,407
0,004	0,445	0,422	0,438	0,432	0,439	0,431	0,418
0,005	0,472	0,422	0,403	0,436	0,452	0,437	0,425
0,006	0,459	0,420	0,459	0,463	0,432	0,444	0,418
0,007	0,459	0,436	0,470	0,466	0,432	0,436	0,416
0,008	0,460	0,413	0,413	0,398	0,437	0,442	0,434
0,009	0,461	0,426	0,398	0,432	0,430	0,429	0,432

В данном случае оптимальными представляются следующие параметры: $df_{min} = 0{,}003$, количество кластеров – 7.

Итоговые результаты работы алгоритмов кластеризации приведены в таблице 8.

Таблица 9. Результаты кластеризации при[1] $n = 2$

Модель	*K*-means	Hierarchy	DBSCAN
Минимум целевой функции	0,385	0,392	0,506
$1 - H_{macro}$	0,321	0,259	0,147
H_{micro}	0,44	0,491	0,704
df_{min}	0,003	0,004	0,009
Количество кластеров	7	6	29
Вид расстояния	–	Ward	–
ε-окрестность	–	–	0,6
Минимальное количество элементов в кластере	–	–	145

[1] *Переменная n задается в формуле* (*5*).

5. Анализ результатов

Для наилучшей интерпретации полученных результатов с помощью алгоритма понижения размерности (T-distributed Stochastic Neighbor Embedding) представим результаты кластеризации методом *K*-means в двумерном пространстве (рис. 3).

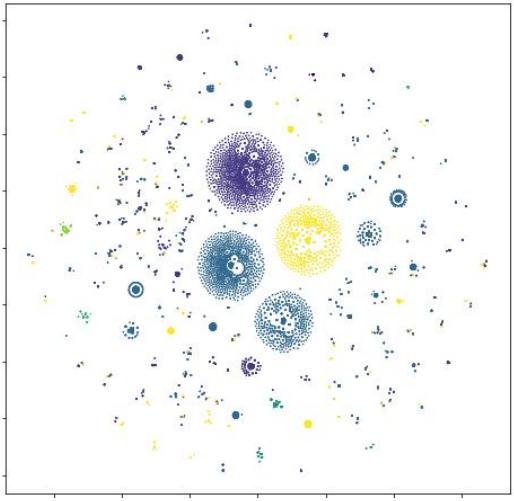

Рис. 3. Результаты кластеризации

Кластеры имеют сферическую форму. Это обуславливает хорошие результаты иерархического алгоритма и *K*-means. Алгоритм DBSCAN показал наихудший результат по причине того, что кластеры имеют различную плотность [6].

Отдельно необходимо отметить, что результаты, показанные иерархическим алгоритмом и *K*-means, близки между собой. Пересечение между кластерами составило около 92%. Учитывая, что согласно некоторым исследованиям, алгоритм *K*-means наилучшим образом подходит для работы с русским языком [1], далее ограничимся анализом результатов алгоритма *K*-means.

В таблице 9 приведены *n*-граммы – центры, вокруг которых формируются кластеры алгоритмом *K*-means.

Таблица 10. Центры кластеров алгоритма K-means

№	*n*-грамма	Номер кластера						
		1	2	3	4	5	6	7
1	налог платёж	0	0	0,01	0	1	0	0
2	соглашение факторинговый	0	0	0,01	0	1	0	0
3	платёж соглашение	0	0	0	0	1	0	0
4	факторинговый финансирование	0	0	0	0	0,97	0	0
5	средство денежный	0,01	0,1	0,08	0	0	0	1
6	денежный перечисление	0	0,01	0,02	0	0	0	1
7	валюта конвертация	0	0	0,03	0	0	0	1
8	денежный перечисление поручение	0	0	0	0	0	0	1
9	перечисление поручение	0	0	0	0	0	0	1
10	перечисление поручение валюта	0	0	0	0	0	0	1
11	распоряжение погашение	0	0	0	1	0	0,71	0
12	договор согласно	0	0	0,02	1	0	0	0
13	согласно распоряжение	0	0	0	1	0	0	0
14	погашение ксв	0	0	0,01	0,97	0	0	0
15	долг основный	0	0	0,01	0,77	0	0	0
16	договор облагаться	0,04	0	0,12	0	0	1	0
17	кредит кредитный	0	0	0	0	0	1	0
18	облагаться распоряжение	0	0	0	0	0	1	0
19	предоставление кредит	0	0	0	0	0	1	0
20	кредитный заёмщик	0	0	0	0	0	0,87	0
21	договор оплата	1	0	0	0	0	0	0
22	оплата счёт	0,38	1	0	0	0	0	0
23–149	прочие *n*-граммы							
	Количество кластеров	8880	3990	5602	677	319	297	156
	Доля	44,6%	20,0%	28,1%	3,4%	1,6%	1,5%	0,8%

Анализ полученных кластеров позволяет группировать их по экономическому содержанию операций. Так, кластеры №1–3 описывают платежные операции клиентов (переводы, обслуживание счетов, оплата по договорам), кластер №4 описывает опе-

рации по погашению ранее выданных банком кредитов, кластер №5 – факторинговые операции, кластер №6 – операции по выдаче кредитов, кластер №7 – конверсионные операции.

Также необходимо отметить, что несмотря на проделанную работу по предобработке данных, не удалось полностью отчистить *n*-граммы от технических сокращений и аббревиатур. Так, один из токенов *n*-граммы №14 представляет собой техническое обозначение кредитного договора – «ксв».

Основной объем финансовых операций классифицируется в первые три кластера. Это обусловлено тем, что данный вид операций наиболее часто встречается в банках. Среди первых десяти наиболее частых корреспонденций счетов все относятся к платежным операциям. Доля первых десяти корреспонденций операций составляет около 80% от всех рассматриваемых транзакций. Таким образом, неудивительно, что именно кластеры, описывающие данные операции, имеют наибольшее количество элементов.

6. Заключение

В работе были изучены особенности финансовых сообщений. Исходные текстовые данные содержат много шума и неточностей. Перед анализом необходимо провести большую работу по подготовке данных, так как использование данных в «сыром» виде приведет к неудовлетворительным результатам работы алгоритмов кластеризации. Дальнейшее сокращение количества переменных позволяет улучшить вычисления, не потеряв при этом в качестве.

В настоящем исследовании был предложен вариант целевой функции, базирующейся на энтропии внутри одной корреспонденции счетов и межкластерной энтропии. Использование целевой функции, основанной на корреспонденции банковский счетов, позволяет обучить модель кластеризации финансовых сообщений, определив лучшие параметры моделей кластеризации.

Во многих практических задачах для типизации банковских операций достаточно сгруппировать сделки по их корреспонденции счетов. В данном подходе этот принцип усовершенствуется, что позволяет методами машинного обучения определять

кластеры, учитывая бухгалтерские особенности данных операций, и одновременно позволяя определить экономический смысл сделок исходя из содержания текстовых сообщений.

Наилучший результат в кластеризации финансовых сообщений показали иерархический метод и *K*-means, наихудший – DBSCAN. Такие результаты определяются структурой данных: кластеры имеют сферическую форму и разную плотность.

Результаты иерархического метода и *K*-means совпадают более чем на 90%. Оптимальным количеством кластеров является 7 шт. Бо́льшую долю операций занимают кластеры, описывающие платежи и денежные переводы, – это связано со структурой исходных данных.

Литература

1. АЛЬ ДАУД Д. *Применение алгоритма кластеризации k-means для анализа вариативности языковой картины мира носителей арабского и русского языков: корпусный подход* // Успехи гуманитарных наук. – 2024. – №4. – С. 84–90.
2. ВИШНЯКОВ И.Э. *Выявление и кластеризация шаблонных текстов в больших массивах сообщений* // Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия Приборостроение. – 2022. – №4(141). – С. 20–35.
3. ЕГОРКИН А.А. *Определение центральности графа алгоритмом PageRank с учетом весов связей* // Управление большими системами. – 2024. – Вып. 111. – С. 81–96.
4. ЕГОРКИН А.А. *Особенности использования алгоритма классификации k-means для данных, подчиненных степенному закону распределения* // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. – 2023. – №9.– С. 65–69.
5. ЕФРЕМОВА А.Н. *Сравнительный анализ алгоритмов нечеткого сравнения строк* // Труды Братского государственного университета. Серия: Естественные и инженерные науки. – 2023. – Т. 1. – С. 45–50.

6. ЖИЛОВ Р.А. *Интеллектуальные методы кластеризации данных* // Известия Кабардино-Балкарского научного центра РАН. – 2023. – №6(116). – С. 152–159.
7. КУРОЧКИН С.В., ПАВЛОВ Н.А., ТКАЧЕНКО М.В. и др. *Методы машинного обучения в риск-профилировании инвестора* // AlterEconomics. – 2024. – Т. 21, №3. – С. 527–552.
8. ПИМЕНОВ В.И. *Использование искусственного интеллекта в противодействии экстремистской деятельности* // Интеллектуальные информационные системы: теория и практика: Сборник научных статей по материалам V Международной конференции. – 2024. – С. 114–118.
9. *Положение Банка России от 24.11.2022 N 809-П. О плане счетов бухгалтерского учета для кредитных организаций и порядке его применения.*
10. ПОНОМАРЕВ Д.С. *Иерархическая кластеризация на языке R для производственно-экономических показателей пенитенциарной системы* // Экономика. Информатика. – 2023. – Т.50, №3. – С. 655–668.
11. САВЕЛЬЕВА М.Ю. *Выявление направления искажения финансовых результатов в бухгалтерской отчетности компаний в регионах Сибирского федерального округа* // Вестник Самарского государственного экономического университета. – 2020. – №4(186). – С. 109–123.
12. ФЕДОРОВА Е.А., АФАНАСЬЕВ Д.О., ДЕМИН И.С. *Разработка тонально-тематического словаря EcSentiThemeLex для анализа экономических текстов на русском языке* // Прикладная информатика. – 2020. – Т. 15, №6(90). – С. 58–77.
13. ACTKINSON B., GRIFFIN R.J. *Detecting plumes in mobile air quality monitoring time series with density-based spatial clustering of applications with noise* // Atmos. Meas. Tech. – 2023 – Vol. 16 – P. 3547–3559,
14. BAPTISTA A., BRIÈRE G., BAUDOT A. *Random walk with restart on multilayer networks: from node prioritisation to supervised link prediction and beyond* // BMC Bioinformatics. – 2024. – 19 p.
15. BAPTISTA A., GONZALEZ A., BAUDOT A. *Universal Multilayer Network Exploration by Random Walk with Restart* // Commun Phys. – 2022. – Vol. 5. – 10 p.

16. BUDIMAN F. *Parameters Testing Optimization Using Cross Validation and Grid Search to Improve Multiclass Classification* // Scientific Visualization. – 2019. – P. 80–90
17. IKOTUN A.M., EZUGWU A.E., ABUALIGAH L. et al. *K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data* // Kean Publications. – 2023 – P. 178–210.
18. TAN P.N., STEINBACH M.M., KUMAR V. *Introduction to data mining* // Pearson Education. – 2023. – 175 p.
19. WARD J.H. *Hierarchical grouping to optimize an objective function* // American Statistical Association. – 1963 – 236 p.
20. YUBO ZHUANG, XIAOHUI CHEN, YUN YANG et al. *Statistically Optimal K-means Clustering via Nonnegative Low-rank Semidefinite Programming* // Published as a conference paper at ICLR. – 2024 – 14 p.
21. WENTAO YE1, JIAJU ZHANG *Shannon entropy in quasiparticle states of quantum chains* // Journal of Physics A Mathematical General. – 2024 – 41 p.

CLUSTERING OF FINANCIAL MESSAGE TEXTS

Anton Egorkin, Russian State Social University, Moscow, Postgraduate Student (2-5@bk.ru).

Abstract: The paper is devoted to the problem of clustering financial message texts by machine learning algorithms. Clustering algorithms can be used to identify groups of similar financial messages, identify the same type or suspicious ones, and use the clusters found rather than the message texts themselves in further analysis. Clustering algorithms such as K-means, DBSCAN and the Hierarchical Clustering method are used in the work. Information about bank transactions is used as texts of financial messages in the work. Due to the fact that bank transactions are subject to strict accounting rules established by the Bank of Russia, it is possible to introduce a metric for assessing the quality of clusterization. This metric allows you to rank the quality of clustering using machine learning algorithms, as well as select the parameters used in training these models. Special attention in the article is paid to the specifics of the data used, and how these features can be taken into account in the practical part. In the practical part of the paper, the results of using clustering models are presented, indicating the optimal parameters of these algorithms. In conclusion, it is concluded that the best clustering algorithms are applied to financial texts.

Keywords: K-means, DBSCAN, Hierarchical clustering method, clustering of financial messages.

УДК 519.8
ББК 22.18

Статья представлена к публикации
членом редакционной коллегии П.В. Сараевым.

Поступила в редакцию 14.02.2025.
Опубликована 31.07.2025.