**УДК 930.2:94(3)**

***Кузнецов Алексей Валерьевич***

*Кандидат исторических наук,*

*научный сотрудник*

*Института всеобщей истории РАН,*

*г. Москва, Российская Федерация*

*e-mail:*

**Компьютерный анализ текстов на латинском языке: Латентно-семантический анализ «Истории готов, вандалов и свевов» Исидора Севильского**

**The computer analysis of Latin texts: Latent Semantic Analysis of «Historia de regibus Gothorum, Wandalorum et Suevorum» by Isidore of Seville**

Аннотация

В статье предпринимается попытка с использованием современных методов интеллектуального анализа текстов исследовать латиноязычный текст хроники «История готов, вандалов и свевов» выдающегося богослова и ученого VII века Исидора Севильского. Ставится цель проверить выдвинутую в историографии гипотезу о наличии у автора представлений об определенной иерархии варварских народов. Основное внимание направлено на раскрытие неявных семантических взаимоотношений между различными частями произведения с целью уточнения отношения автора к трем варварским народам.

Анализ текста проводился на языке программирования R. В качестве конкретного метода бы выбран метод латентно-семантического анализа, позволяющий проводить сравнение и кластеризацию текстов на основе семантического пространства, построенного путем сингулярного разложения терм-документной матрицы.

Актуальность исследования заключается в том, что впервые реализован полный цикл латентно-семантического анализа средневекового латиноязычного текста. Проведена предварительная подготовка, построено семантическое пространство текста памятника, осуществлено сравнение семантической схожести текстов на основе меры косинусного сходства. Результаты анализа позволяют утверждать, что Исидор Севильский действительно выстраивает иерархию из трех варварских народов, придавая большее сходство описанию вестготов и свевов и особняком ставя вандалов.

The article makes an attempt to study the Latin text of the chronicle «Historia de regibus Gothorum, Wandalorum et Suevorum» by the theologian and scientist of the 7th century Isidore of Seville by means of the methods of text mining. The main goal is to verify the hypothesis that the author has ideas about a hierarchy of barbarian kingdoms. The main focus is on clarifying of implicit semantic relationships between the different parts of the chronicle in order to find out the author’s attitude to the three barbaric kingdoms.

The analysis of the text was performed with the R programming language. As a method for analysis was chosen the method of latent-semantic analysis. It allows the computation of semantic similarity and clustering of texts, which are represented as vectors in the semantic space constructed by the singular decomposition of the term-document matrix.

The relevance of the study is that for the first time provides a full cycle of latent semantic analysis of medieval Latin text: text preprocessing, creation of semantic space, calculating of the semantic similarity of the texts on the basis of the cosine similarity measure. The results of the analysis suggest that Isidore of Seville really builds a hierarchy of three barbarian kingdoms, giving greater similarity to the description of the Visigoths and Suevs and putting Vandals apart.

**Ключевые слова**: Исидор Севильский, раннесредневековая историография, компьютерный анализ текстов, латентно-семантический анализ, сингулярное разложение, кластерный анализ.

**Keywords**: Isidore of Seville, early Middle Age historiography, computational text analysis, Latent Semantic Analysis, singular value decomposition, cluster analysis.

Введение

Исидор Севильский (ок. 560–636) известен как выдающийся ученый, богослов, церковный и политический деятель конца VI – начала VII веков, энциклопедист, внесший неоценимый вклад в сохранение и творческую переработку античного наследия. Высокий интерес к творчеству Исидора Севильского сохраняется с 1960-х гг., когда отмечался его 1400 летний юбилей [1, с. 205]. Среди множества проблем в поле зрения исследователей его трудов одной из наиболее часто рассматриваемых в настоящее время является, пожалуй, проблема отражения в них формирования национальной, политической и религиозной идентичности в государстве вестготов [2, 3, 4, 5]. В сравнительно недавно увидевшей свет монографии Дж. Вуд выдвинул гипотезу о наличии в исторических сочинениях Исидора единой программы, направленной на обоснование политического и религиозного господства вестготов в Испании [3, p.77, p. 159-260]. В «Истории готов, вандалов и свевов», по мнению Дж. Вуда, он выстраивает своеобразную иерархию варварских народов, на вершине которой располагаются вестготы [3, p. 153-161]. Другие варварские народы сознательно умаляются и делигитимизируются. Ярче всего это проявляется в используемых Исидором системах датирования исторических событий. В разделе, посвященном истории вестготов, он применяет двойную систему хронологических указателей – отмечает год в соответствии с Испанской эрой (отсчет лет с 1 января 38 года до н. э.) и соответствующий ему год правления римских императоров, а в разделах, посвященных истории вандалов и свевов, только по годам Испанской эры [3, p. 156.]. Особенно негативный образ Исидор создает вандалам, неоднократно указывая на их приверженность арианской вере. В противоположность им свевы, создавшие на северо-западе Испании собственное королевство и одновременно с вестготами перешедшие в никейскую веру, наделяются определенной религиозной и политической легитимностью. Исследователь выдвигает тезис, согласно которому Исидор таким образом подчеркивает, что свевы являются наиболее подходящим объектом для вестготского господства[3, p. 161]. Предположения Дж. Вуда были, впрочем, подвергнуты критике за недостаточную обоснованность его выводов текстами источников[6, с. 129]. В данной статье мы предполагаем попробовать значительно глубже понять отношение Исидора к трем варварским народам, раскрыть неявные идеи и смыслы, содержащиеся в «Истории готов, вандалов и свевов», используя метод латентно-семантического анализа – один из методов интеллектуального анализа текстов (англ. text mining). Интеллектуальный анализ текстов – это направление в искусственном интеллекте, цель которого получение информации из неструктурированных текстовых данных путём их преобразования в пригодный для анализа набор структурированных данных при помощи методов обработки естественного языка (англ. Natural Language Processing, NLP) и машинного обучения (англ. Machine Learning) [7, p. 1-15].

Метод латентно-семантического анализа.

Применительно к анализу текстов на естественном языке латентно-семантический анализ (англ. Latent Semantic Analysis, LSA) понимается как метод, раскрывающий взаимосвязь между набором документов и термами (словами, n-граммами), представленными в виде векторов в многомерном латентно-семантическом пространстве [8, p. 111-113]. Под вектором документа понимается вектор, координатами которого являются частоты вхождений термов словаря в этот документ. Теоретическим основанием латентно-семантического анализа является так называемая дистрибутивная гипотеза, согласно которой, слова со схожим значением имеют тенденцию встречаться в схожих контекстах[9]. В латентно-семантическом пространстве представленные в виде векторов слова и документы, сходные по значению, будут стремиться находиться в близких областях пространства, что дает возможность сравнивать семантические значения слов и документов. Сравнение в большинстве случаев реализуется путем вычисления косинуса угла между векторами соответствующих слов и документов [10, p. 931, 11, p.84-87].

Часто латентно-семантический анализ рассматривают в контексте развития методов тематического моделирования[12; 13, p. 368-389], то есть способов построения моделей коллекции текстовых документов на основе анализа совместной встречаемости термов, которые определяют, к каким те мам относится каждый из документов, и какие термы образуют эти темы. Тематическое моделирование помогает обнаружить а коллекции документов скрытые темы, аннотировать документы с помощью этих тем и организовывать большой объем неструктурированных данных. Это верно лишь отчасти, поскольку применения латентно-семантического анализа уже давно перешагнуло рамки тематического моделирования.

Впервые латентно-семантический анализ был описан и запатентован в конце 1980-х годов под наименованием латентно-семантическое индексирование (англ. Latent Semantic Indexing) [14, 15] как метод автоматического индексирования текстов и информационного поиска [16]. Использование латентно-семантического анализа позволило повысить эффективность работы информационно-поисковых систем, преодолев две основные сложности в информационном поиске: полисемию (многозначность терминов) и синонимию (сходство значения различных слов) [17, с. 411-412]. Этот метод также успешно используется для решения других задач обработки естественного языка, в том числе для автоматического поиска рецензентов [18], реферирования текстов [19], тематического моделирования [12, с. 219-222]. Несколько позднее метод латентно-семантического анализа стал успешно использоваться в психологии для построения когнитивных моделей понимания и формирования знания [20, 21, с. 19-20, 22] моделей долговременной и кратковременной памяти у детей разного школьного возраста на базе детских текстов [23], оценки понимания текстов [24]. Отдельно стоит отметить, что латентно-семантический анализ нашел применение в системах тестирования знаний, в случаях когда надо проверить правильность ответов в свободной форме [20]. Для нас наиболее значимым является то, что латентно-семантический анализ был и остается одним из наиболее эффективных методов классификации и индексации текстов, оценки их семантической близости [25, 26, 27].

РЕАЛИЗАЦИЯ

Латентно-семантический анализ проводился с использованием латинского текста «Истории готов, вандалов и свевов» из первого издания в Patrologia Latina Жак-Поль Миня [28]. Электронный текст памятника доступен в «Patrologia Latina Database»[29], а также на сайте проекта «The Latin Library»[30].

Для проведения латентно-семантического анализа части произведения Исидора Севильского были представлены в виде отдельных документов. Обычно в сочинении выделяются четыре части «Пролог» (Prologus), известный также под названием «Похвала Испании» (Laus Spaniae / De laude Spaniae), и три части посвященные соответственно истории готов, вандалов и свевов. В первой части (главы 1-70) речь идет о происхождении готов, переселении их на земли Римской империи, создании готских королевств, правлению вестготских королей в Испании и объединении всей Испании под их властью. В конце первой части (главы 66-70) расположены «Выводы» (Item recapitulatio ejusdem Isidori in Gothorum laudem) краткое изложение истории готов. «Выводы» мы также выделили в отдельный документ. Как и «Пролог» они стилистически отличающиеся от всего текста и по аналогии с «Похвалой Испании» представляет собой своеобразную «Похвалу готам». Во второй части (главы 71-84) рассказывается о нашествии вандалов, создании их государства в Испании и Африке и о падении их государства. В третьей части – об образовании королевства свевов в Испании и их присоединении к королевству вестготов. Таким образом мы будем анализировать всего пять документов: «Пролог», «Историю готов», «Выводы», «Историю вандалов», «Историю свевов».

Анализ текста проводился при помощи среды программирования R. Текст скрипта на языке R и все ресурсы доступны в репозитории сервиса GitHub [31].

В обобщенном виде латентно семантический анализ текста применительно к классификации документов включает следующие этапы [10, p. 931, 11, Pp.80-87, 32, p. 456-460]:

1. предварительная обработка текста;
2. создание терм-документной матрицы из набора документов;
3. взвешивание термов в терм-документной матрице;
4. создание семантического пространства набора документов
5. сравнение векторов термов или документов в латентном семантическом пространстве, путем вычисления косинуса угла между векторами или другим методом.

Предварительная обработка текстов является важнейшим этапом для любых методов и приемов интеллектуального анализа текстов, от неё во многом зависят полученные результаты. Цель предварительной обработки – преобразовать неструктурированные тексты в пригодный для анализа формат данных ~~при помощи методов обработки естественного языка~~. В зависимости от стоящих задач предварительная обработка текста может включать в различном сочетании следующие операции[7, p. 37-44, 11, p. 45-59, 33]: 1. Токенизация – разбиение текста на фрагменты (абзацы, предложения, слова, N-граммы). 2. Очистка текста – удаление лишних пробелов и пустых строк, типографских знаков, чисел, знаков препинания, перевод всех букв в нижний регистр. 3. Удаление стоп-слов – малозначимых и низкоинформативных (как правило, служебные части речи, местоимения, числительные). 4. Лемматизация – приведение слова к словарной форме (в латинском языке словарная форма глаголов соответствует форме первого лица единственного числа, существительных и прилагательных – именительному падежу единственного числа) или стемминг – выделение основы слова. Очевидно, что стемминг для латиноязычных текстов не применим, поскольку приведет к потере или искажению значимой информации. 5. Частеречная разметка – определение части речи и морфологической формы слов в тексте. 6. Синтаксический парсинг – определение синтаксических зависимостей слов в предложении.

Предварительная обработка латиноязычных текстов имеет свою специфику, обусловленную развитой системой словоформ[33], без приведения которых к словарной форме адекватный анализ текста провести невозможно. Обработка текста «Истории готов, вандалов и свевов» Исидора Севильского была проделана с использованием пакетов tm [34] и UDPipe [35]. При помощи пакета tm буквы были переведены в нижний регистр, удалены числа и знаки препинания. Для удаления стоп-слов был составлен собственный их список. За основу взят список латинских стоп-слов проекта Perseus Digital Library [36], в который дополнительно были добавлены римские цифры во множестве присутствующие в тексте, а также некоторые часто встречающиеся, но малоинформативные слова. В их числе такие как «annus», «aera» используемые Исидором для обозначения дат. Далее с помощью пакета UDPipe текст был разбит на предложения, проведена лемматизация Как результат – сформирована таблица данных (data frame), пригодная для дальнейшего статистического анализа.

На основе созданной с помощью пакета UDPipe таблицы данных можно сформировать матрицу терм-документ (term-document matrix) – базу латентно-семантического анализа, математическую матрицу, описывающую частоту терминов в коллекции документов [11, p.77, 32, p. 458-459]. В такой матрице каждый столбец соответствуют документу, а строки – термам (словам, фразам, N-граммам). Количество столбцов равно количеству документов, а количество строк – размеру словаря, числа в ячейках обозначают количество встречаемости слов в каждом из документов [См. табл. 1].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **accipio** | 0 | 1 | 0 | 1 | 2 |
| **acies** | 0 | 3 | 0 | 0 | 0 |
| **adeo** | 0 | 11 | 1 | 0 | 0 |
| **africus** | 0 | 3 | 0 | 7 | 1 |
| **agilanis** | 0 | 3 | 0 | 0 | 0 |
| **alanus** | 0 | 0 | 1 | 2 | 0 |
| **alaricus** | 0 | 6 | 0 | 0 | 0 |
| **aliquandiu** | 0 | 3 | 0 | 0 | 0 |
| **anteus** | 0 | 3 | 0 | 0 | 0 |
| **antiquus** | 0 | 2 | 1 | 0 | 0 |
| **apostolus** | 0 | 3 | 0 | 0 | 0 |

Таблица 1. Матрица терм-документ (фрагмент).

Перед формированием латентно-семантического пространства, как правило, рассчитывают значимость или иначе вес терма (в нашем случае слова) в документе относительно всего корпуса документов ~~[37]~~. Делается это для того, чтобы в ходе анализа повысить значимость низкочастотных слов и одновременно понизить значимость высокочастотных слов. Практика показывает, что использование взвешивания слов в латентно-семантическом анализе дает лучший результат, чем без взвешивания, поскольку именно низкочастотные слова точнее отражают содержание документа[37]. Кроме того применения взвешивания позволяет снизить влияние размера текстов при их анализе [38] Сейчас разработано большое число алгоритмов определяющих значимость слов [8, p. 93-95.], но чаще всего применяется статистическая мера *частотность терминов-обратная частотность документов* или TF-IDF (от англ. term frequency – inverse document frequency).

Частотность терминов (term frequency, TF) – это величина, показывающая насколько часто слово встречается в документе. Она дает возможность оценить важность данного слова в пределах конкретного документа. Рассчитывается как частное от деления количества раз, которое слово встречается в тексте, и общего количества слов в тексте. Обратная частотность документов (inverse document frequency, IDF) – это инверсия частотности, с которой определенное слово встречается в коллекции документов. Она рассчитывается как логарифм от общего количества документов, делённого на количество всех документов, в которых встречается конкретное слово. Статистическая мера TF-IDF рассчитывается как произведение TF и IDF [17, с. 134-137].

При применении статистической меры TF-IDF к терм-документной матрице значения в её ячейках меняются [См. табл. 2]. Вес терма увеличивается, если он встречается множество раз в небольшом количестве документов, тем самым усиливая их уникальность. Вес терма уменьшается, если он встречается небольшое количество раз в одном документе или во множестве документов, снижая их уникальность. Вес терма будет минимальным, если он во множестве встречается в большинстве документов [17, с. 136.].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **accipio** | 0.000000 | 1.736966 | 0.000000 | 1.736966 | 3.473931 |
| **acies** | 0.000000 | 9.965784 | 0.000000 | 0.000000 | 0.000000 |
| **adeo** | 0.000000 | 25.541209 | 2.321928 | 0.000000 | 0.000000 |
| **africus** | 0.000000 | 5.210897 | 0.000000 | 12.158759 | 1.736966 |
| **agilanis** | 0.000000 | 9.965784 | 0.000000 | 0.000000 | 0.000000 |
| **alanus** | 0.000000 | 0.000000 | 2.321928 | 4.643856 | 0.000000 |
| **alaricus** | 0.000000 | 19.931569 | 0.000000 | 0.000000 | 0.000000 |
| **aliquandiu** | 0.000000 | 9.965784 | 0.000000 | 0.000000 | 0.000000 |
| **anteus** | 0.000000 | 9.965784 | 0.000000 | 0.000000 | 0.000000 |
| **antiquus** | 0.000000 | 4.643856 | 2.321928 | 0.000000 | 0.000000 |
| **apostolus** | 0.000000 | 9.965784 | 0.000000 | 0.000000 | 0.000000 |

Таблица 2. Матрица терм-документ (фрагмент) после применения ~~алгоритма~~ TF-IDF.

На основе подготовленной таким образом терм-документной матрицы мы построили семантическое пространство при помощи пакета lsa [39]. Пакет lsa наиболее популярный инструмент для латентно-семантического анализа в среде R. Он хорошо документирован, имеются подробные руководства для работы с ним [32, p. 450-496].

Семантическое пространство формируется путем применения к матрице терм-документ сингулярного разложения (англ. Singular Value Decomposition, SVD). Математические основы сингулярного разложения применительно к латентно-семантическому анализу неоднократно подробно описаны [11, p.79-80, 40]. Подготовленная нами матрица первоначально раскладывается на три матрицы, согласно формуле: M = U\*Vt\*S, где M – исходная матрица, U – ортогональная матрица документов, Vt – ортогональная транспонированная матрица термов, а S – диагональная матрица. Если в диагональной матрице S оставить только k наибольших сингулярных значений, а в матрицах U и Vt только соответствующие этим значениям столбцы и строки, то мы получим три новые уменьшенные матрицы Uk, Vtk и Sk. Произведение трех новых уменьшенных матриц даст нам новую матрицу Mk, которая максимально приближена к исходной матрице M, но имеет меньший ранг, равный k. Иными словами матрица Mk будет наилучшей малоранговой аппроксимацией матрицы M. В этом случае строки матрицы Uk будут образами термов, а столбцы матрицы Vtk – образами документов в k – мерном семантическом пространстве (Mk).

Матрица Mk будет целиком отображать структуру скрытых зависимостей документов и термов матрицы M, но одновременно обладать рядом преимуществ. Она будет меньше по размеру, поэтому её анализ потребует меньше ресурсов, у неё будет ниже разреженность (значительно меньше пустых ячеек), она будет содержать только самые сущностные семантические отношения документов и слов, игнорируя несущественный шум и раскрывая скрытые смыслы в документах и термах [41, p. 159-160]. Сокращение размерности может сводить вместе термины, имеющие одинаковую семантику, для идентификации и устранения неоднозначности терминов с несколькими значениями и для обеспечения более низкого размерного представления документов, отражающих понятия, а не необработанные термины.

В качестве эксперимента построим семантическое пространство как на основе терм-документной матрицы, к которой была применена статистическая мера TF-IDF, так и на основе первоначальной матрицы без взвешивания.

Для наглядности покажем положение документов в семантическом пространстве сделав их проекцию на плоскости [См. рис. 1] и в трехмерном пространстве [См. рис. 2]. В первом случае мы видим, что из пяти, анализируемых нами частей «Истории готов, вандалов и свевов» Исидора Севильского, «Пролог», «Выводы» к истории готов и «История свевов» составляют плотную группу. Однако следует учитывать, что проекция векторов многомерного пространства на плоскость не дает нам возможности на этом основании судить о схожести документов.

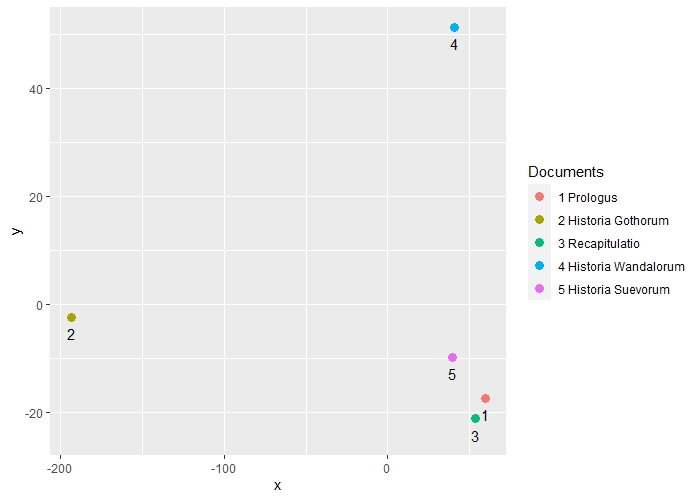


Рисунок 1. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» на плоскость.

Проекция документов в семантическом пространстве в трех измерениях также не дает возможности судить о семантической схожести документов, поскольку и здесь есть доля условности из-за сокращения количества измерений. Как мы уже отмечали выше, наиболее часто используемым способом оценить схожесть документов в векторном семантическом пространстве является расчет меры косинусного сходства (англ. cosinus similarity) – косинуса угла между векторами. На примере проекции документов в семантическом пространстве в трех измерениях будет рассчитываться косинус угла между векторами, начинающимися в точке с координатами [0,0,0], помеченной красным треугольником, и заканчивающимися в точках помеченных кружочками [См. рис. 2].

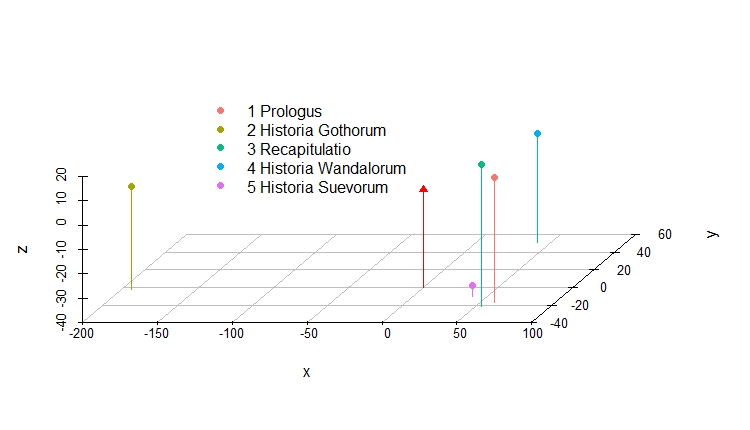


Рисунок 2. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» в трех измерениях.

Теоретически, значение косинуса угла между векторами может варьироваться между -1 (полная оппозиция) до 1 (полная идентичность). В случае с векторным представлением текстов в семантическом пространстве косинус угла между векторами может принимать значения от 0 (соответствует углу 90°) до 1, поскольку частота терма или статистическая мера TF-IDF не могут принимать отрицательное значение. Угол между двумя векторами документов или слов в семантическом пространстве не может быть больше, чем 90°. Чем выше будет значение косинуса угла между векторами документов ~~в семантическом пространстве~~, тем о большем семантическом сходстве документов можно говорить.

На основе сформированного семантического пространства была построена матрица косинусного сходства. Результаты визуализированны в виде тепловой карты [См. рис. 3]. Полученные данные свидетельствуют, что среди выделенных нами частей «Истории готов, вандалов и свевов» Исидора Севильского наибольшей семантической схожестью обладают «История готов» и «История свевов», степень сходства между остальными частями значительно меньше.



Рисунок 3. Визуализация матрицы косинусного сходства для частей «Истории готов, вандалов и свевов» в виде тепловой карты. С использованием статистической меры TF-IDF.

Матрица косинусного сходства, построенная на основе семантического пространства без использования статистической меры TF-IDF, дает другие значения, но тенденции остаются прежние: наибольшей семантической схожестью обладают «История готов» и «История свевов» [См. рис. 4].



Рисунок 4. Визуализация матрицы косинусного сходства для частей «Истории готов, вандалов и свевов» в виде тепловой карты. Без использования статистической меры TF-IDF.

Для ещё большей наглядности проведем кластерный анализ документов на основании косинусного расстояния между векторами документов в латентно-семантическом пространстве. Кластеризация – это один из методов анализа, группирующий объекты на основании избранной меры сходства/различия в группы (кластеры) таким образом, чтобы объекты (в нашем случае документы - части «Истории готов, вандалов и свевов») внутри каждой группы были похожи друг на друга, а объекты из разных групп явно отличались ~~друг от друга~~ [42, p. 385]. Для кластеризации предлагаем использовать один из наиболее распространенных методов – иерархическую кластеризацию, а конкретнее одну из её разновидностей – агломеративную кластеризацию. Она более всего подходит для небольшого числа анализируемых объектов и получаемых кластеров. Иерархическая кластеризация строит иерархию кластеров в виде перевернутого дерева – дендрограммы. В случае агломеративной кластеризации построение дендрограммы начинается снизу, с «листьев», соответствующих объектам в анализируемом наборе данных. На следующем шаге наиболее схожие между собой «листья» объединяются в кластеры и формируют «ветви» дендрограммы. Алгоритм повторяется до тех пор, пока все объекты не будут объединены в одном кластере – «стволе» дендрограммы [11, p. 67-68, 42, p. 390-391, 394]. Для кластеризации матрица косинусного сходства была преобразована в матрицу косинусного расстояния между документами, которая и была проанализирована с помощью функции hclust.

В результате была построена дендрограмма [См. рис. 4].

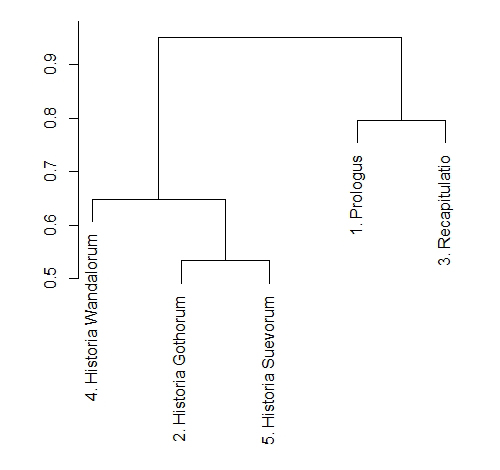


Рисунок 5. Иерархическая кластеризация фрагментов «Истории готов, вандалов и свевов» на основании косинусного расстояния между векторами документов в семантическом пространстве.

«Листья» в нижней части дендрограммы соответствую, анализируемым нами фрагментам «Истории готов, вандалов и свевов» Исидора Севильского. Учитывая особенности алгоритма иерархической кластеризации, можно утверждать, что чем ниже в структуре дендрограммы происходит слияние «листьев» и «ветвей», тем выше степень сходства между анализируемыми объектами. Если же объединение в один кластер происходит сравнительно высоко, вблизи от «ствола» дендрограммы, то в таком кластере могут быть объединены довольно непохожие друг на друга объекты [11, p. 99-102, 42, p. 391-392]. Мы видим, что на самом нижнем уровне в один кластер объединяются разделы, посвященный истории готов и свевов. Это свидетельствует об их максимальной схожести. Чуть выше с ними объединяется раздел, посвященный истории вандалов. Только значительно выше в один кластер объединяются «Пролог» и «Выводы» к истории готов.

1. Уколова В.И. Античное наследие и культура раннего средневековья (конец V - начало VII века). М., 1989.

2. Velázquez I. Pro patriae gentisqve Gothorvm statv (4th council of Toledo, Canon 75, A. 633) // Regna and Gentes: The Relationship between Late Antique and Early Medieval Peoples and Kingdoms in the Transformation of the Roman World / ed. by Goetz H.-W., Jarnut J., Pohl W. Leiden, Boston: Brill, 2003. P. 161-217.

3. Wood J. The Politics of Identity in Visigothic Spain. Religion and Power in the Histories of Isidore of Seville. Leiden, Boston: Brill, 2012.

4. Марей Е. С. Феномен «вестготской симфонии» в 75-ом каноне IV-го Толедского собора 633 г. (к проблеме перехода к средневековой государственности) // Электронный научно-образовательный журнал «История». 2012. T. 3. Выпуск 3 (11) [Электронный ресурс]. Доступ для зарегистрированных пользователей. URL: https://history.jes.su/s207987840000372-8-1/ (дата обращения: 24.03.2020).

5. Сямтомов И. В. Понятие «Gens» и королевская власть в системе вестготского права (IV–VIII вв.) // Vox medii aevi. 2015. №2-3. С. 90-110. URL: https://cyberleninka.ru/article/n/ponyatie-gens-i-korolevskaya-vlast-v-sisteme-vestgotskogo-prava-iv-viii-vv (дата обращения: 23.03.2020).

6. Воронцов С. А. Wood J. The politics of identity in Visigothic Spain. Religion and power in the histories of Isidore of Seville. Brill, 2012 // Вестник ПСТГУ. Серия 1: Богословие. Философия. 2012. №42 (4). С. 125-131.

7. Kwartler T. Text mining in practice with R. New Jersey: John Wiley & Sons, 2017.

8. Lane H., Howard C., Hapke H. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Manning Publications Co., 2019.

9. Sahlgren M. The Distributional Hypothesis. From context to meaning // Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica : журнал. – 2008. – Vol. 20, no. 1. – P. 33–53.

10. Günther F., Dudschig C., Kaup B. LSAfun - An R package for computations based on Latent Semantic Analysis // Behavior Research Methods. Volume 47. 2015. Pp. 930–944 doi:10.3758/s13428-014-0529-0

11. Anandarajan M., Hill C., Nolan T. Practical Text Analytics. Maximizing the Value of Text Data. (Advances in Analytics and Data Science. Vol. 2.) Springer, 2019.

12. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. Том 23, 2012. С. 215-244.

13 Sarkar D. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Bangalore, 2019.

14. Deerwester S., Dumais S.T., Landauer T.K., Furnas G., Beck L. Improving Information Retrieval with Latent Semantic Indexing // Proceedings of the 51st Annual Meeting of the American Society for Information Science, vol. 25. 1988. Pp. 36–40.

15. Deerwester S., Dumais S. T., Furnas, G. W., Landauer, T. K., Harshman, R. Indexing by Latent Semantic Analysis // Journal of the American Society for Information Science, 41. 1990. Pp. 391-407.

16. Dumais S.T. LSA and Information Retrieval: Getting Back to Basics // Handbook of Latent Semantic Analysis. Eds.: Landauer T. K., McNamara D. S., Dennis S., Kintsch, W. Mahwah, New Jersey: Erlbaum. 2007. Pp. 293-322.

17. Маннинг, К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце; пер. с англ. под ред. П. И. Браславского, Д. А. Клюшина, И. В. Сегаловича. Москва: И. Д. Вильямс, 2011.

18. Dumais S., Nielsen J. Automating the assignment of submitted manuscripts to reviewers // SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. Copenhagen, Denmark, June 21-24, 1992. Pp. 233–244. https://doi.org/10.1145/133160.133205

19. Ozsoy M., Cicekli I., Alpaslan F. Text Summarization of Turkish Texts using Latent Semantic Analysis // COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China. Vol. 2. Pp. 869-876.

20. Landauer T. K., Dumais S. T. A solution to Plato’s problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge // Psychological Review. 104(2), 1997. Pp. 211–240.

21. Величковский Б.М. Когнитивная наука. Основы психологии познания. Том II. Москва, 2006.

22. Landauer T.K. LSA as a Theory of Meaning // Handbook of Latent Semantic Analysis. Eds.: Landauer T. K., McNamara D. S., Dennis S., Kintsch W. Mahwah, New Jersey: Erlbaum. 2007. Pp. 3-32.

23. Denhière G., Lemaire B., Bellissens C., Jhean-Larose S. A semantic space modeling children’s semantic memory // Handbook of Latent Semantic Analysis. Eds.: Landauer T. K., McNamara D. S., Dennis S., Kintsch . Mahwah, New Jersey: Erlbaum, 2007. P. 143-167.

24. Воронин В. М., Курицын С.В. Латентный семантический анализ и понимание текста // Психологический вестник Уральского государственного университета. Вып. 9. Екатеринбург, 2010. С. 15-27.

25. Кураленок И.Е., Некрестьянов И.С. Автоматическая классификация документов на основе латентно-семантического анализа // Труды первой всероссийской научно-методической конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». СПб, 1999. C. 89–96.

26. Kou G., Peng Y. An Application of Latent Semantic Analysis for Text Categorization // International Journal of Computers Communications & Control. 10(3). June, 2015. Pp. 357-369.

27. Краснов С.А., Илатовский А.С., Хомоненко А.Д., Арсеньев В.Н. Оценка семантической близости документов на основе латентно-семантического анализа с автоматическим выбором ранговых значений // Труды СПИИРАН. 2017. № 54 (5). C. 185-204.

28. Isidorus Hispalensis. Historia de regibus Gothorum, Wandalorum et Suevorum // Patrologiae Cursus Completus. Series Latina. Vol. 83. Paris: 1850. Col. 1057-1082.

29. Patrologia Latina Database [Электронный ресурс]. URL: http://pld.chadwyck.co.uk/ (дата обращения 12.01.2020).

30. Sancti Isidori Hispalensis Episcopi Historia de regibus Gothorum, Wandalorum et Suevorum [Электронный ресурс]. URL: https://www.thelatinlibrary.com/isidore/historia.shtml (дата обращения 12.01.2020).

31. Kuznetsov A.V. The computer analysis of Latin texts: Latent Semantic Analysis of «Historia de regibus Gothorum, Wandalorum et Suevorum» by Isidore of Seville [Электронный ресурс]. URL: https://github.com/alexeyvkuznetsov/Latin\_Text\_LSA/ (дата обращения 12.01.2020).

32. Gefen D., Endicott J. E., Fresneda J. E., Miller J., Larsen K. R. A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community // Communications of the Association for Information Systems. Vol. 41, Article 21. November 2017. Pp. 450-496.

33. Кузнецов А.В. Применения инструментов text mining для анализа средневековых латиноязычных текстов: предварительная обработка текстов // Научные исследования и разработки. Сборник научных работ 57й Международной научной конференции Евразийского Научного Объединения (г. Москва, ноябрь 2019). Москва: ЕНО, 2019. C. 68-70.

34. tm: Text Mining Package [Электронный ресурс]. URL: https://CRAN.R-project.org/package=tm (дата обращения 12.01.2020).

35. Natural Language Processing with R and UDPipe. Tokenization, Parts of Speech Tagging, Lemmatization, Dependency Parsing and NLP flows [Электронный ресурс]. URL: https://bnosac.github.io/udpipe/en/ (дата обращения 12.01.2020).

36. Perseus Stop Words. [Электронный ресурс]. URL: http://www.perseus.tufts.edu/hopper/stopwords/ (дата обращения 12.01.2020).

37. Dumais S. T. Improving the retrieval of information from external sources. // Behavior Research Methods, Instrumentation, and Computers, 23(2). 1991. Pp. 229–236.

38. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing and Management. 1988. Volume 24, Issue No. 5. P. 513-523.

39. Wild F. lsa: Latent Semantic Analysis. (R package version 0.73.2). [Электронный ресурс]. URL: https://CRAN.R-project.org/package=lsa (дата обращения 12.01.2020).

40. Martin D. I., Berry M. W. Mathematical foundations behind Latent Semantic Analysis Handbook of Latent Semantic Analysis. Eds.: Landauer T. K., McNamara D. S., Dennis S., Kintsch W. Mahwah, New Jersey: Erlbaum, 2007. Pp. 35-56.

41. Turney P. D., Pantel P. From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research, 37. March 2010. Pp. 141–188.

42. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer, 2015.