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 в коллекции документов скрытые темы, аннотировать документы с помощью этих тем и организовывать большой объем неструктурированных данных. Такой подход к латентно-семантический анализ верен лишь отчасти, поскольку область применения этого метода значительно шире рамок 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, 22],[21, с. 19-20] 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].

Экспериментальная часть

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), 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 и сравнению их семантической близости включает следующие этапы [10, p. 931],[11, p.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. Отметим, 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 (леммам, как в нашем случае, 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. Делается это для того, чтобы в ходе анализа повысить значимость низкочастотных слов и одновременно понизить значимость высокочастотных слов. Практика показывает, что использование взвешивания 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 статистической меры 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 станет 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 (главным образом синонимию и полисемию слов) и раскрывая скрытые смыслы в документах и термах [41, p. 159-160].

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.

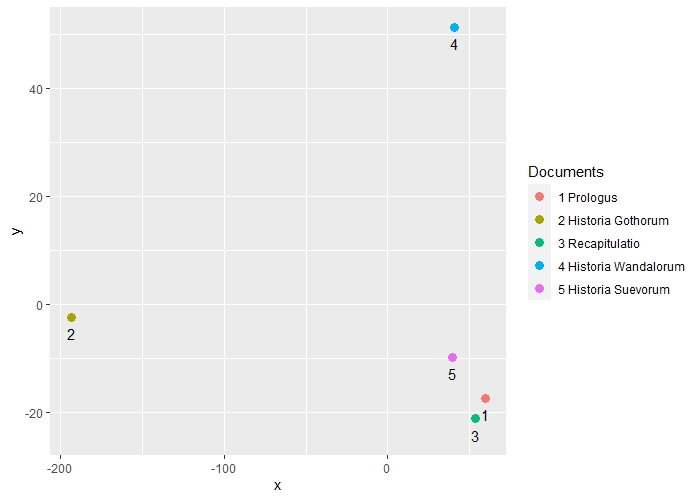


Рисунок 1. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» на плоскость.

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].

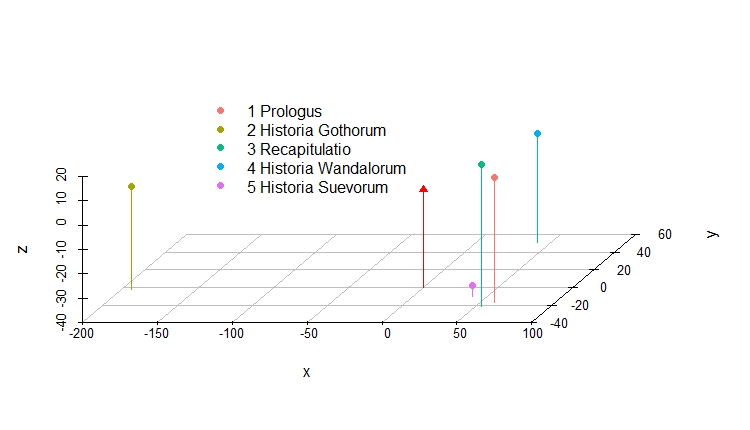


Рисунок 2. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» в трех измерениях.

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, 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[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].

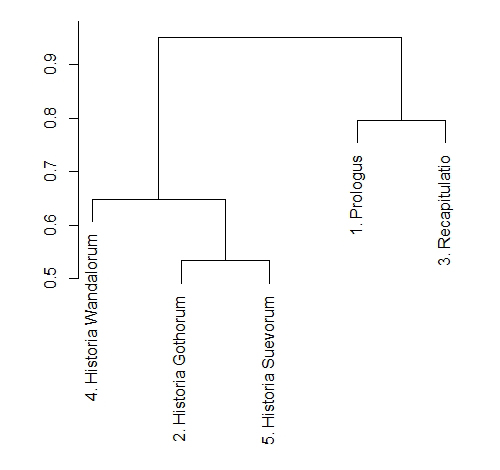


Рисунок 5. Иерархическая кластеризация фрагментов «Истории готов, вандалов и свевов» на основании косинусного расстояния между векторами документов в семантическом пространстве.

«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.

Выводы

Подводя итог, отметим, что проделанный латентно-семантический анализ латиноязычного текста хроники «История готов, вандалов и свевов» Исидора Севильского подтверждает гипотезу о выстраивании автором иерархии варварских народов. Расчет семантической близости разных разделов хроники демонстрирует наибольшее сходство между описаниями вестготов и свевов. Исидор Севильский излагает их историю, используя в описании похожие семантические структуры. Несколько особняком оказываются вандалы. Отметим, что к аналогичным результатам привело и проделанное нами ранее тематическое моделирование «История готов, вандалов и свевов» [43], показавшее совпадение преобладающих тем в разделах истории готов и свевов.