**УДК 930.2:94(3)**

***Кузнецов Алексей Валерьевич***

*Кандидат исторических наук,*

*научный сотрудник*

*Института всеобщей истории РАН,*

*г. Москва, Российская Федерация*

*e-mail:*

**Латентно-семантический анализ средневековых латинских текстов. «Истории готов, вандалов и свевов» Исидора Севильского**

***Abstract.*** *The article is devoted to the problem of applying of quantitative methods to medieval Latin texts. The author demonstrated the efficiency of using of the UDPipe package for automatic processing and statistical analysis of medieval Latin on the example of «Historia Wambae regis» by Julian of Toledo.*

***Key words****: quantitative methods, text mining, natural language processing, medieval Latin texts.*

Keywords: Text Analysis, Isidore of Seville, early Middle Age historiography, computational text analysis, topic modelling, Latent Semantic Analysis (LSA), singular value decomposition, clustering.

Using SVD, LSA can break down your TF-IDF term-document matrix into three simpler matrices. And they can be multiplied back together to produce the original matrix, without any changes. This is like factorization of a large integer. Big whoop. But these three simpler matrices from SVD reveal properties about the original TFIDF matrix that you can exploit to simplify it. You can truncate those matrices (ignore some rows and columns) before multiplying them back together, which reduces the number of dimensions you have to deal with in your vector space model. These truncated matrices don’t give the exact same TF-IDF matrix you started with—they give you a better one. Your new representation of the documents contains the essence, the “latent semantics” of those documents. That’s why SVD is used in other fields for things such as compression. It captures the essence of a dataset and ignores the noise. A JPEG image is ten times smaller than the original bitmap, but it still contains all the information of the original image. When you use SVD this way in natural language processing, you call it latent semantic analysis. LSA uncovers the semantics, or meaning, of words that is hidden and waiting to be uncovered. Latent semantic analysis is a mathematical technique for finding the “best” way to linearly transform (rotate and stretch) any set of NLP vectors, like your TF-IDF vectors or bag-of-words vectors. And the “best” way for many applications is to line up the axes (dimensions) in your new vectors with the greatest “spread” or variance in the word frequencies.13 You can then eliminate those dimensions in the new vector space that don’t contribute much to the variance in the vectors from document to document. Using SVD this way is called truncated singular value decomposition (truncated SVD). In the image processing and image compression world, you might have heard of this as principal component analysis (PCA). And we show you some tricks that help improve the accuracy of LSA vectors. These tricks are also useful when you’re doing PCA for machine learning and feature engineering problems in other areas.

LSA uses SVD to find the combinations of words that are responsible, together, for the biggest variation in the data. You can rotate your TF-IDF vectors so that the new dimensions (basis vectors) of your rotated vectors all align with these maximum variance directions. The “basis vectors” are the axes of your new vector space and are analogous to your topic vectors in the three 6-D topic vectors from your thought experiment at the beginning of this chapter. Each of your dimensions (axes) becomes a combination of word frequencies rather than a single word frequency. So you think of them as the weighted combinations of words that make up various “topics” used throughout your corpus. The machine doesn’t “understand” what the combinations of words means, just that they go together. When it sees words like “dog,” “cat,” and “love” together a lot, it puts them together in a topic. It doesn’t know that such a topic is likely about “pets.” It might include a lot of words like “domesticated” and “feral” in that same topic, words that mean the opposite of each other. If they occur together a lot in the same documents, LSA will give them high scores for the same topics together. It’s up to us humans to look at what words have a high weight in each topic and give them a name. But you don’t have to give the topics a name to make use of them. Just as you didn’t analyze all the 1,000s of dimensions in your stemmed bag-of-words vectors or TF-IDF vectors from previous chapters, you don’t have to know what all your topics “mean.” You can still do vector math with these new topic vectors, just like you did with TF-IDF vectors. You can add and subtract them and estimate the similarity between documents based on their topic vectors instead of just their word counts. LSA gives you another bit of useful information. Like the “IDF” part of TF-IDF, it tells you which dimensions in your vector are important to the semantics (meaning) of your documents. You can discard those dimensions (topics) that have the least amount of variance between documents. These low-variance topics are usually distractions, noise, for any machine learning algorithm. If every document has roughly the same amount of some topic and that topic doesn’t help you tell the documents apart, then you can get rid of it. And that will help generalize your vector representation so it will work better when you use it with documents your pipeline hasn’t yet seen, even documents from a different context. This generalization and compression that LSA performs accomplishes what you attempted in chapter 2 when you ignored stop words. But the LSA dimension reduction is much better, because it’s optimal. It retains as much information as possible, and it doesn’t discard any words, it only discards dimensions (topics). LSA compresses more meaning into fewer dimensions. We only have to retain the high-variance dimensions, the major topics that your corpus talks about in a variety of ways (with high variance). And each of these dimensions becomes your “topics,” with some weighted combination of all the words captured in each one.

Hobson Lane

Используя SVD, LSA может разбить вашу матрицу терминологических документов TF-IDF на три более простые матрицы. И их можно умножить обратно вместе, чтобы получить исходную матрицу, без каких-либо изменений. Это как факторизация большого целого числа. Большой возглас Но эти три более простые матрицы из SVD раскрывают свойства исходной матрицы TFIDF, которые вы можете использовать для ее упрощения. Вы можете обрезать эти матрицы (игнорируя некоторые строки и столбцы) перед тем, как умножить их обратно вместе, что уменьшает количество измерений, с которыми вам приходится иметь дело в вашей модели векторного пространства. Эти усеченные матрицы не дают ту же самую матрицу TF-IDF, с которой вы начали - они дают вам лучшую. Ваше новое представление документов содержит суть, «скрытую семантику» этих документов. Вот почему SVD используется в других областях для таких вещей, как сжатие. Он отражает сущность набора данных и игнорирует шум. Изображение JPEG в десять раз меньше исходного растрового изображения, но оно по-прежнему содержит всю информацию об исходном изображении. Когда вы используете SVD таким способом в обработке естественного языка, вы называете это скрытым семантическим анализом. LSA раскрывает семантику или значение слов, которые скрыты и ожидают раскрытия. Скрытый семантический анализ - это математический метод для нахождения «наилучшего» способа линейного преобразования (поворота и растяжения) любого набора векторов НЛП, например, ваших векторов TF-IDF или векторов с набором слов. И «лучшим» способом для многих приложений является выравнивание осей (измерений) в ваших новых векторах с наибольшим «разбросом» или дисперсией частот слов13. Затем вы можете устранить те измерения в новом векторном пространстве, которые не t вносит большой вклад в дисперсию векторов от документа к документу. Использование SVD таким способом называется усеченным разложением по сингулярным значениям (усеченным SVD). В мире обработки изображений и сжатия изображений вы, возможно, слышали об этом как об анализе основных компонентов (PCA). И мы покажем вам некоторые приемы, которые помогут повысить точность векторов LSA. Эти приемы также полезны, когда вы делаете PCA для машинного обучения и решаете технические проблемы в других областях.

LSA использует SVD, чтобы найти комбинации слов, которые вместе несут ответственность за самые большие различия в данных. Вы можете вращать ваши векторы TF-IDF так, чтобы новые измерения (базисные векторы) ваших повернутых векторов совпали с этими направлениями максимальной дисперсии. «Базисные векторы» являются осями вашего нового векторного пространства и аналогичны вашим тематическим векторам в трех 6-D тематических векторах из вашего мысленного эксперимента в начале этой главы. Каждое из ваших измерений (осей) становится комбинацией частот слов, а не частотой одного слова. Таким образом, вы думаете о них как о взвешенных комбинациях слов, которые составляют различные «темы», используемые в вашем корпусе. Машина не «понимает», что означают комбинации слов, просто они соединяются. Когда он часто видит такие слова, как «собака», «кошка» и «любовь», он объединяет их в тему. Он не знает, что такая тема, скорее всего, о «домашних животных». В эту же тему может входить много слов, таких как «одомашненный» и «дикий», слова, которые означают противоположность друг другу. Если они встречаются вместе в одних и тех же документах, АЛП даст им высокие баллы по одним и тем же темам. Мы, люди, должны посмотреть, какие слова имеют большой вес в каждой теме, и дать им имя. Но вам не нужно давать названия темам, чтобы использовать их. Точно так же, как вы не проанализировали все тысячи измерений в ваших стеблевых векторах или векторах TF-IDF из предыдущих глав, вам не нужно знать, что означают все ваши темы. Вы все еще можете выполнять векторную математику с этими новыми тематическими векторами, так же, как вы делали с векторами TF-IDF. Вы можете добавлять и вычитать их и оценивать сходство между документами на основе их тематических векторов, а не только количества слов. LSA дает вам еще немного полезной информации. Как и часть «IDF» в TF-IDF, она говорит вам, какие измерения в вашем векторе важны для семантики (значения) ваших документов. Вы можете отбросить те измерения (темы), которые имеют наименьшую разницу между документами. Эти темы с малыми отклонениями, как правило, отвлекающие факторы, шум, для любого алгоритма машинного обучения. Если каждый документ содержит примерно одинаковое количество какой-либо темы, и эта тема не помогает вам разделить документы, то вы можете избавиться от них. И это поможет обобщить ваше векторное представление, чтобы оно работало лучше, когда вы используете его с документами, которые ваш конвейер еще не видел, даже с документами из другого контекста. Это обобщение и сжатие, которые выполняет LSA, завершают то, что вы пытались в главе 2, когда игнорировали стоп-слова. Но уменьшение размера LSA намного лучше, потому что оно оптимально. Он сохраняет столько информации, сколько возможно, и не отбрасывает слова, он только отбрасывает измерения (темы). LSA сжимает больше смысла в меньшее количество измерений. Нам нужно только сохранить измерения с высокой дисперсией, основные темы, о которых ваш корпус говорит различными способами (с высокой дисперсией). И каждое из этих измерений становится вашими «темами» с некоторой взвешенной комбинацией всех слов, записанных в каждом из них.

LSI is an automatic indexing method that projects both documents and terms into a low dimensional space which, by intent, represents the semantic concepts in the document. By projecting documents into the semantic space, LSI enables the analysis of documents at a conceptual level, purportedly overcoming the drawbacks of purely term-based analysis. For example, in information retrieval, users may use many different queries to describe the same information need, and likewise, many of the relevant documents may not contain the exact terms used in the particular query. In this case, projecting documents into the semantic space enables the search engine to find documents containing the same concepts but different terms. The projection also helps to resolve terms that are associated with multiple concepts. In this sense, LSI overcomes the issues of synonymy and polysemy that plague term-based information retrieval. LSI was applied to text data in the 1980s and later used for indexing in information retrieval systems [23]. It has also been used for a variety of tasks, including assigning papers to reviewers [28] and cross-lingual retrieval. LSI is based on the singular value decomposition (SVD) of the termdocument matrix, which constructs a low rank approximation of the original matrix while preserving the similarity between the documents. LSI is meant to interpret the dimensions of the low-rank approximation as semantic concepts although it is surpassed in this regard by later improvements such as PLSI. We now describe the basic steps for performing LSI. Then, we will discuss the implementation issues and analyze the underlying mechanisms for LSI.

Mining text data 133-134

LSI - это метод автоматической индексации, который проецирует как документы, так и термины в низкоразмерное пространство, которое представляет семантические концепции в документе. Проецируя документы в семантическое пространство, LSI позволяет анализировать документы на концептуальном уровне, якобы преодолевая недостатки анализа, основанного исключительно на терминах. Например, при поиске информации пользователи могут использовать много разных запросов для описания одной и той же потребности в информации, и аналогичным образом многие из соответствующих документов могут не содержать точных терминов, используемых в конкретном запросе. В этом случае проецирование документов в семантическое пространство позволяет поисковой системе находить документы, содержащие те же понятия, но разные термины. Проекция также помогает разрешать термины, связанные с несколькими понятиями. В этом смысле LSI преодолевает проблемы синонимии и многозначности, которые мешают поиску информации на основе терминов. БИС применялась к текстовым данным в 1980-х годах, а затем использовалась для индексации в информационно-поисковых системах [23]. Он также использовался для решения различных задач, в том числе для предоставления рецензентов рефератов [28] и межъязыкового поиска. LSI основана на разложении по сингулярным значениям (SVD) матрицы терминов документов, которая создает низкое ранговое приближение исходной матрицы, сохраняя при этом сходство между документами. LSI предназначен для интерпретации измерений низкосортного приближения как семантических понятий, хотя в этом отношении он превосходит последующие улучшения, такие как PLSI. Теперь опишем основные этапы выполнения БИС. Затем мы обсудим вопросы реализации и проанализируем механизмы, лежащие в основе LSI.

Shah C., Jivani A. A Hybrid Approach of Text Summarization Using Latent Semantic Analysis and Deep Learning. Shah, Chintan and Anjali Jivani. // 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Bangalore, 2018. Pp. 2039-2044. 10.1109/ICACCI.2018.8554848.

Латентный семантический анализ (LSA) - это один из статистических методов для извлечения значения слов посредством статистического вычисления, применяемого к огромному текстовому корпусу. Основная цель состоит в том, чтобы информация обо всех контекстах слова, в которых появляется данное слово, обеспечивала набор взаимных ограничений, которые во многом определяют сходство значений слов и набор слов, связанных друг с другом. Компетентность репликации человеческих знаний LSA была доказана многими способами.

Латентный семантический анализ (LSA) - это алгебраический статистический метод, который извлекает значение слов и сходство предложений, используя информацию об использовании слов в контексте. Он хранит информацию о том, какие слова используются в предложении, сохраняя при этом информацию об общих словах среди предложений. Более общие слова между предложениями означают, что эти предложения более семантически связаны. Метод LSA может представлять значение слов и значение предложений одновременно. Он усредняет значение слов, содержащихся в предложении, чтобы выяснить значение этого предложения. Это представляет значение слов, усредняя значение предложений, которые содержат это слово. Метод LSA использует разложение по сингулярным значениям (SVD) для поиска семантически похожих слов и предложений. SVD - это метод, который моделирует отношения между словами и предложениями. Он имеет возможность снижения шума, что приводит к повышению точности.

ЛСА является мощным неконтролируемым аналитическим инструментом. Это один из самых выдающихся алгоритмов обучения в задачах поиска информации. Он обладает способностью раскрывать невидимую структуру слов среди слов, предложений или текста посредством разложения по сингулярным числам (SVD). Он также производит измерения отношения слово-слово, слово-документ и документ-документ, которые хорошо коррелируют с несколькими когнитивными феноменами человека, включающими ассоциацию или семантическое сходство. Производительность алгоритмов суммирования на основе LSA зависит от качества представления документа [Triantafillou, E.; Kiros, J.R.; Urtasun, R.; Zemel, R.: Towards generalizable sentence embeddings. In: Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, 2016, pp. 239–248] и алгоритма выбора предложений. Более ранние подходы суммирования документов на основе LSA имеют некоторые ограничения. В этой работе мы пытаемся преодолеть эти ограничения, улучшая представление документа и предлагая новый алгоритм выбора предложений.

Valdez, Danny & Pickett, Andrew & Goodson, Patricia. (2018). Topic Modeling: Latent Semantic Analysis for the Social Sciences\*. Social Science Quarterly. 10.1111/ssqu.12528.

Нашей первой техникой является скрытое семантическое индексирование (LSI), которое было разработано в 1970-х годах как статистический метод для корреляции семантически связанных терминов от корпусов. БИС используется не только для суммирования текста, но и для поиска и поиска информации. LSI использует очень популярный метод разложения по сингулярным значениям (SVD), который мы подробно обсудили в разделе «Важные понятия». Основным принципом, лежащим в основе LSI, является то, что подобные термины, как правило, используются в одном и том же контексте и, следовательно, имеют тенденцию встречаться чаще. Термин LSI происходит от того факта, что этот метод способен обнаруживать скрытые скрытые термины, которые семантически коррелируют для формирования тем.

Our first technique is Latent Semantic Indexing (LSI), which was developed in the 1970s as a statistical technique to correlate semantically linked terms from corpora. LSI is not just used for text summarization, but also in information retrieval and search. LSI uses the very popular Singular Value Decomposition (SVD) technique, which we discussed in detail in the “Important Concepts” section. The main principle behind LSI is that similar terms tend to be used in the same context and hence tend to co-occur more. The term LSI comes from the fact that this technique has the ability to uncover latent hidden terms that correlate semantically to form topics.

Sarkar D. Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data. The bag-of-words representation commonly used in text analysis can be analyzed very efficiently and retains a great deal of useful information, but it is also troublesome because the same thought can be expressed using many different terms or one term can have very different meanings. Dimension reduction can collapse together terms that have the same semantics, to identify and disambiguate terms with multiple meanings and to provide a lower-dimensional representation of documents that reflects concepts instead of raw terms. In this chapter, we survey two influential forms of dimension reduction. Latent semantic indexing uses spectral decomposition to identify a lower-dimensional representation that maintains semantic properties of the documents. Topic modeling, including probabilistic latent semantic indexing and latent Dirichlet allocation, is a form of dimension reduction that uses a probabilistic model to find the co-occurrence patterns of terms that correspond to semantic topics in a collection of documents. We describe the basic technologies in detail and expose the underlying mechanism. We also discuss recent advances that have made it possible to apply these techniques to very large and evolving text collections and to incorporate network structure or other contextual information.

Представление "мешок слов", обычно используемое в текстовом анализе, может быть проанализировано очень эффективно и сохранить большое количество полезной информации, но это также проблематично, потому что одна и та же мысль может быть выражена с помощью многих различных терминов или один термин может иметь очень разные значения. 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 этой главе мы рассмотрим две влиятельные формы сокращения размеров. Латентная семантическая индексация использует спектральную декомпозицию для идентификации низкомасштабного представления, которое поддерживает семантические свойства документов. Тематическое моделирование, включая вероятностное латентное семантическое индексирование и латентное выделение Дирихле, представляет собой форму сокращения размерности, которая использует вероятностную модель для поиска паттернов совместного появления терминов, соответствующих семантическим темам в коллекции документов. Мы подробно описываем основные технологии и раскрываем лежащий в их основе механизм. Мы также обсуждаем последние достижения, которые позволили применить эти методы к очень большим и развивающимся текстовым коллекциям и включить сетевую структуру или другую контекстуальную информацию.

Latent semantic analysis (LSA) is an algorithm applied to approximate the meaning of texts, thereby exposing semantic structure to computation. LSA combines the classical vector-space model − well known in computational linguistics − with a singular value decomposition (SVD), a two-mode factor analysis. Thus, bag-of-words representations of texts can be mapped into a modified vector space that is assumed to reflect semantic structure. In this contribution the author describes the lsa package for the statistical language and environment R and illustrates its proper use through an example from the area of automated essay scoring. I. CONCEPTS USED IN LATENT SEMANTIC ANALYSIS HEN applying a latent semantic analysis (Deerwester et al., 1990), a process is executed that typically involves several (optional) steps and involves various data-types created as an output of these steps. To clarify which entities and processes are involved when performing an LSA, the following concepts shall be defined. Term: The 'word' as it is written down in a document. Corpus: The collection of documents containing texts that consist out of terms separated by punctuation marks. Textmatrix: A representation of the document collection in matrix format: the cells contain the frequency, how often a particular term appears in a specific document. Terms are the rows, documents the columns. By transforming a corpus to this representation format, documents are treated as so-called bag of words, where the term order is neglected. Latent-Semantic Space: When applying a singular-value decomposition (SVD) to a textmatrix, the matrix is resolved into the term-vector matrix T (constituting the left singular vectors), the document-vector matrix D (constituting the right

*It is a challenge task to discover major topics from text, which provide a better understanding of the whole corpus and can be regarded as a text categorization problem. The goal of this paper is to apply latent semantic analysis (LSA) approach to extract common factors that representing concepts hidden in a large group of text. LSA involves three steps: the ﬁrst step is to set up a term-document matrix; the second step is to transform the term frequencies into a term-document matrix using various weighting schemes; the third step performs singular value decomposition (SVD) on the matrix to reduce the dimensionality. The reduced-order SVD is the best k-dimensional approximation to the original matrix. The experiment uses more than ﬁfteen hundreds research paper abstracts from a speciﬁc ﬁeld. Because diﬀerent factor solutions of the LSA suggest diﬀerent levels of aggregation, this work examines thirteen solutions in the experiment. The results show that LSA is able to identify not only principle categories, but also major themes contained in the text.*

The LSA analysis can be summarized in three main steps. The ﬁrst step is to set up a term- document matrix in which each row stands for a key word or term and each column stands for a document or context in which the key word appears. An entry in the matrix is the frequency of a key word in the corresponding document. The second step is to transform the term frequencies in a term-document matrix using various weighting schemes. The third step is to perform SVD on the matrix to reduce the dimensionality, which is the key feature of the LSA method. In this step only the k largest singular values are retained. The reduced-order SVD is the best k-dimensional approximation to the original matrix[Dumais, S. T. (2004). Latent Semantic Analysis, Annual Review of Information Science and Technology, 38: 189-230.]

Abstract

The recent influx in generation, storage and availability of textual data presents researchers with the challenge of developing suitable methods for their analysis. Latent Semantic Analysis (LSA), a member of a family of methodological approaches that offers an opportunity to address this gap by describing the semantic content in textual data as a set of vectors, was pioneered by researchers in psychology, information retrieval, and bibliometrics. LSA involves a matrix operation called singular value decomposition, an extension of principal component analysis. LSA generates latent semantic dimensions that are either interpreted, if the researcher’s primary interest lies with the understanding of the thematic structure in the textual data, or used for purposes of clustering, categorisation and predictive modelling, if the interest lies with the conversion of raw text into numerical data, as a precursor to subsequent analysis. This paper reviews five methodological issues that need to be addressed by the researcher who will embark on LSA. We examine the dilemmas, present the choices, and discuss the considerations under which good methodological decisions are made. We illustrate these issues with the help of four small studies, involving the analysis of abstracts for papers published in the European Journal of Information Systems.

Keywords: text mining, analysis of textual data, singular value decomposition, clustering, factor analysis.

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 60-х годов XX века, когда отмечался его1400 летний юбилей [Уколова В.И. Античное наследие и культура раннего средневековья (конец V - начало VII века). М., 1989. 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 [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.; Wood J. The Politics of Identity in Visigothic Spain. Religion and Power in the Histories of Isidore of Seville. Leiden, Boston: Brill, 2012.; 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).; Сямтомов И. В. Понятие «Gens» и королевская власть в системе вестготского права (IV–VIII вв.) // Vox medii aevi. 2015. №2-3. С. 13-14. URL: https://cyberleninka.ru/article/n/ponyatie-gens-i-korolevskaya-vlast-v-sisteme-vestgotskogo-prava-iv-viii-vv (дата обращения: 23.03.2020).]. В сравнительно недавно увидевшей свет монографии Дж. Вуд выдвинул гипотезу о наличии в исторических сочинениях Исидора единой ~~исторической~~ программы, направленной на обоснование политического и религиозного господства вестготов в Испании [Wood J. The Politics of Identity in Visigothic Spain. Religion and Power in the Histories of Isidore of Seville. Leiden, Boston: Brill, 2012. P.77, 159-260.]. В «Истории готов, вандалов и свевов», по мнению Дж. Вуда, он выстраивает своеобразную иерархию варварских народов, на вершине которой располагаются вестготы [Wood J. The Politics of Identity in Visigothic Spain. Religion and Power in the Histories of Isidore of Seville. Leiden, Boston: Brill, 2012. P. 153-161.]. Другие варварские народы сознательно умаляются и делигитимизируются. Особенно ярко 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 [Wood J. The Politics of Identity in Visigothic Spain. Religion and Power in the Histories of Isidore of Seville. Leiden, Boston: Brill, 2012. 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[Wood J. The Politics of Identity in Visigothic Spain. Religion and Power in the Histories of Isidore of Seville. Leiden, Boston: Brill, 2012. P. 161.]. Предположения Дж. Вуда были, впрочем, подвергнуты критике за недостаточную обоснованность его выводов текстами источников[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. 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 – одного из методов интеллектуального анализа текстов (англ. text mining). Интеллектуальный анализ текстов – это одно из направлений в искусственном интеллекте, цель которого получение информации из неструктурированных текстовых данных путём их преобразования в пригодный для анализа набор структурированных данных на основе методов обработки естественного языка (англ. Natural Language Processing, NLP) и машинного обучения (англ. Machine Learning) [Kwartler T. Text mining in practice with R. NJ: John Wiley & Sons, 2017. Pp. 1-15.].

МЕТОД

Применительно к анализу текстов на естественном языке под латентно-семантическим анализом (англ. Latent semantic analysis, LSA) понимается метод, анализирующий взаимосвязь между набором документов и термами (словами, n-граммами), представленными в виде векторов в многомерном латентно-семантическом пространстве [Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W. (2007). Handbook of Latent Semantic Analysis. Mahwah, NJ : Erlbaum.]. Под вектором документа понимается вектор, координатами которого являются частоты вхождений термов словаря в этот документ. 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 [Günther F., Dudschig C., Kaup B. LSAfun - An R package for computations based on Latent Semantic Analysis // Behavior Research Methods. Volume 47. 2015. P. 931. doi:10.3758/s13428-014-0529-0; 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. Pp.84-87.]. Теоретическим основанием латентно-семантического анализа является так называемая дистрибутивная гипотеза, согласно которой, слова со схожим значением имеют тенденцию встречаться в схожих контекстах[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.; Harris Z. Distributional Structure // Word. Vol. 10, 1954. Issue 2-3. Pp. 146-162, DOI: 10.1080/00437956.1954.11659520].

Часто латентно-семантический анализ рассматривают в контексте развития методов тематического моделирования[Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. Том 23, 2012. С. 215-244.; Crain S.P., Crain S.P., Zhou K., Yang Sh.-H., Zha H. Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond // Mining Text Data. / Eds.: Aggarwal C., Zhai C. Springer, 2012. Pp. 129-161.; Sarkar D. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Bangalore, 2019. Pp. 368-389. ], то есть способов построения моделей коллекции текстовых документов на основе анализа совместной встречаемости термов, которые определяют, к каким темам 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.

Латентно-семантический анализ был описан и запатентован в конце 1980-х годов под наименованием латентно-семантическое индексирование (англ. Latent Semantic Indexing) [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.; 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.] как метод автоматического индексирования текстов и информационного поиска [Dumais S.T. Latent Semantic Analysis // Annual Review of Information Science and Technology, 38. 2004. Pp. 189-230; 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, NJ: Erlbaum. 2007. Pp. 293-322.; 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. С. 411 – 417.; Стенин А. А., Тимошин Ю. А., Мелкумян Е. Ю., Курбанов В. В. Латентно-семантический метод извлечения информации из Интернет ресурсов // Восточно-Европейский журнал передовых технологий. Автоматика. Вычислительная техника. Вып. 9 (64). Том 4. 2013. С. 19-22.]. 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) [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. C. 411-412. ]. Затем этот метод стал успешно использоваться в психологии для построения когнитивных моделей понимания и формирования знания [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.; Величковский Б.М. Когнитивная наука. Основы психологии познания. Том II. Москва, 2006. C. 19-20.; 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, NJ: Erlbaum. 2007. Pp. 3-32.] моделей долговременной и кратковременной памяти у детей разного школьного возраста на базе детских текстов [Denhière G., Lemaire B., Bellissens C., Jhean-Larose S. Psychologie cognitive et compréhension de texte: une démarche théorique et expérimentale // S. Porhiel, D. Klinger (еds.). L'unité texte. Pleyben: Perspectives, 2004. Pp. 74–95.; Lemaire B., Denhière G. Cognitive Models based on Latent Semantic Analysis // Tutorial given at the 5th International Conference on Cognitive Modeling (ICCM'2003), Bamberg, Germany, April 9 2003. Pp. 23–25.; 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, NJ: Erlbaum, 2007. P. 143-167.], оценки понимания текстов [Воронин В. М., Курицин С. В., Наседкина З. А., Ицкович М. М. Использование латентного семантического анализа как альтернативы пропозиционального анализа в исследованиях понимания текста // Гуманизация образования. 2017. №2. С. 11-19 . URL: https://cyberleninka.ru/article/n/ispolzovanie-latentnogo-semanticheskogo-analiza-kak-alternativy-propozitsionalnogo-analiza-v-issledovaniyah-ponimaniya-teksta (дата обращения: 01.04.2020).; Курицин С.В., Воронин В.М. Исследование оценки понимания нарративных и экспозиторных текстов с применением латентного семантического анализа // Сибирский психологический журнал, 2009. № 33. С. 25-30.; Воронин В. М., Курицын С.В. Латентный семантический анализ и понимание текста // Психологический вестник Уральского государственного университета. Вып. 9. Екатеринбург, 2010. С. 15-27.]. Отдельно стоит отметить, что латентно-семантический анализ нашел применение в системах проверки знаний, в случаях когда надо проверить правильность ответов в свободной форме [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. 1997. 104. — P. 211–240]. 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 [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.; Kuralenok I., Nekrest'yanov I. Automatic document classification based on latent semantic analysis // Programming and Computer Software, Vol.26, No. 4, 2000. Pp. 199-206.; Бондарчук Д.В. Использование латентно-семантического анализа в задачах классификации текстов по эмоциональной окраске // Бюллетень результатов научных исследований. 2012. № 2(3). С. 146–152.; 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.; 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.].

РЕАЛИЗАЦИЯ

Латентно-семантический анализ проводился с использованием латинского текста «Истории готов, вандалов и свевов» из первого издания в Patrologia Latina Жак-Поль Миня [Isidorus Hispalensis. Historia de regibus Gothorum, Wandalorum et Suevorum // Patrologiae Cursus Completus. Series Latina. Vol. 83. Paris: 1850. Col. 1057-1082.]. Электронный текст памятника доступен в «Patrologia Latina Database»[Patrologia Latina Database [Электронный ресурс]. URL: http://pld.chadwyck.co.uk/ (дата обращения 12.01.2020).], а также на сайте проекта «The Latin Library»[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).].

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 [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).].

В обобщенном виде латентно семантический анализ текста применительно к классификации документов включает следующие этапы [Wild F. An LSA package for R // Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL’07). Heerlen, 2007. Pp. 11-12.; 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. 456 – 460.; Günther F., Dudschig C., Kaup B. LSAfun - An R package for computations based on Latent Semantic Analysis // Behavior Research Methods. Volume 47. 2015. Pp. 931.; 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. Pp.80-87.; Wild F. Learning Analytics in R with SNA, LSA, and MPIA. Springer, 2016. P. 77-79.]:

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[Ингерсолл Г. С., Мортон Т. С., Фэррис Э. Л. Обработка неструктурированных текстов. Поиск, организация и манипулирование. / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 414 с.: ил. С. 44-72.; Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. СПб.: Питер, 2019. С. 66-72; 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; 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. Pp. 45-59; Kwartler T. Text mining in practice with R. Hoboken, NJ : John Wiley & Sons, 2017. С. 37-44.]: 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 (служебные части речи, местоимения, числительные, а также слов встречающихся очень редко или, напротив, очень часто). 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 [Кузнецов А.В. Применения инструментов text mining для анализа средневековых латиноязычных текстов: предварительная обработка текстов // Научные ис-следования и разработки. Сборник научных работ 57й Международной научной конференции Евразий-ского Научного Объединения (г. Москва, ноябрь 2019). — Москва: ЕНО, 2019. C. 68-70.; ~~Kuznetsov A.V. Text Preprocessing of the «History of the Goths, Vandals and Suevi» by Isidore of Seville. [Электронный ресурс]. URL: https://github.com/alexeyvkuznetsov/Latin\_Text\_Preprocessing/ (дата обращения 25.11.2019).~~]. Обработка текста «Истории готов, вандалов и свевов» Исидора Севильского была проделана при помощи пакетов tm [tm: Text Mining Package [Электронный ресурс]. URL: https://CRAN.R-project.org/package=tm (дата обращения 12.01.2020).] и UDPipe [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).]. 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 [Perseus Stop Words. [Электронный ресурс]. URL: http://www.perseus.tufts.edu/hopper/stopwords/ (дата обращения 12.01.2020).], 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[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).] текст был разбит на предложения, проведена лемматизация, ~~частеречная разметка и синтаксический парсинг~~. 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 [ 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. 458 – 459.; 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. Pp.77.]. 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 [Dumais S. T. (1991). Improving the retrieval of information from external sources. Behavior Research Methods, Instrumentation, and Computers, 23, 229–236.; Martin D. I., Berry M. W. (2007). Mathematical foundations behind Latent Semantic Analysis // T.K. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis (pp. 35–56). Mahwah, NJ: Erlbaum.]. 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[Dumais S. T. Improving the retrieval of information from external sources. // Behavior Research Methods, Instrumentation, and Computers, 23(2). 1991. Pp. 229–236.]. Кроме того применения взвешивания позволяет снизить влияние размера текстов при их анализе [Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing and Management. 1988. Volume 24, Issue No. 5. P. 513 – 523.] Сейчас разработано большое число алгоритмов определяющих значимость слов [Lane H., Howard C., Hapke H. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Manning Publications Co., 2019. Pp. 93-95.], но чаще всего применяется статистическая мера *частотность терминов-обратная частотность документов* или TF-IDF (от англ. term frequency – inverse document frequency) [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].

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.

~~Статистическую меру TF-IDF широко используют в поисковых алгоритмах, чтобы точнее определить релевантность текста пользовательскому запросу, а также близость различных документов при их кластеризации.~~

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 [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. 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | V1 | V2 | V3 |
| 1 Prologus | -0.009659629 | 0.004349679 | -0.01505734 |
| 2 Historia Gothorum | -0.988460934 | -0.130194319 | -0.06795501 |
| 3 Recapitulatio | -0.036141000 | 0.003984056 | -0.00724269 |
| 4 Historia Wandalorum | -0.107332095 | 0.957605576 | -0.26700821 |
| 5 Historia Suevorum | -0.100126534 | 0.256916733 | 0.96115012 |

Таблица ???. Матрица документов в трехмерном латентно-семантическом пространстве.

На основе подготовленной таким образом терм-документной матрицы мы построили семантическое пространство при помощи пакета lsa [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).]. Пакет lsa наиболее популярный инструмент для латентно-семантического анализа в среде R. Он хорошо документирован, имеются подробные руководства для работы с ним [Wild F. An LSA package for R // Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL’07). Heerlen, 2007. Pp. 11-12.; 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.; Wild F. Learning Analytics in R with SNA, LSA, and MPIA. Springer, 2016. **Pp. 71 – 106**.; Ashish K., Avinash P. Mastering Text Mining with R. Master text-taming techniques and build effective text-processing applications with R. Birmingham, 2016. Pp. 131-134.; Wild F., Stahl C. Investigating Unstructured Texts with Latent Semantic Analysis // Advances in Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Eds.: Decker R., Lenz H.J. Springer, Berlin, Heidelberg. 2007. Pp. 383-390.].

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 [Martin D. I., Berry M. W. (2007). Mathematical foundations behind Latent Semantic Analysis // Handbook of Latent Semantic Analysis. Eds.: Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W. Mahwah, NJ: Erlbaum. 2007. Pp. 35–56.; Wild F. Learning Analytics in R with SNA, LSA, and MPIA. Springer, 2016. Pp. 73 – 76.; 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. Pp.79-80.]. 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), уменьшен шум, сохранена только самая значимая информация, раскрывающая скрытые смыслы в документах и термах [Eckart C., Young G. The Approximation of One Matrix by Another of LowerRank // Psychometrika. 1936. №3(1). Pp. 211–218.; 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. Pp. 159-160].

Согласно теореме Эккарта-Янга, если в матрице W оставить только наибольшие сингулярные значения σ , а в матрицах U и V — соответствующие этим значениям столбцы, то матрицы Uσ и Vσ будут лучшими их приближениями, отражающими ассоциативные зависимости представления термов и документов в пространстве размерности σ [17, 20].

Бубнов В.П. и др. Модели информационных систем: учеб. пособие // М.: ФГБОУ «Учебно-методический центр по образованию на железнодорожном транспорте». 2015. 188 с.

Хомоненко А.Д., Дашонок В.Л., Краснов С.А. Выявление противоречий в семантически близкой информации на основе латентно-семантического анализа // Проблемы информационной безопасности. Компьютерные системы. 2014. № 2. С. 73–84.

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.

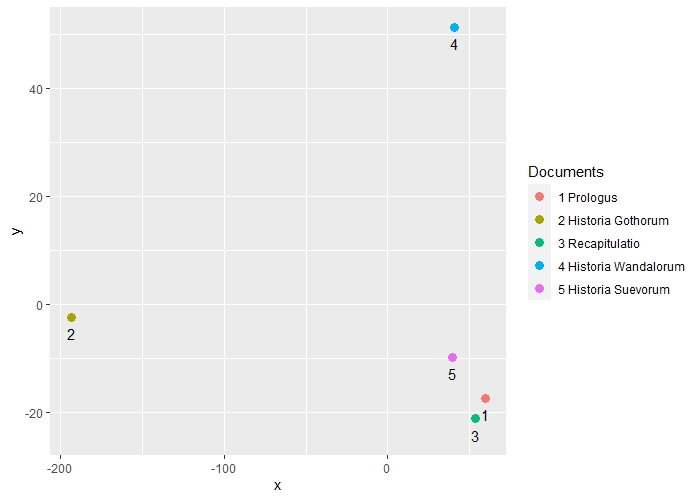


Рисунок 1. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» на плоскость.

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].

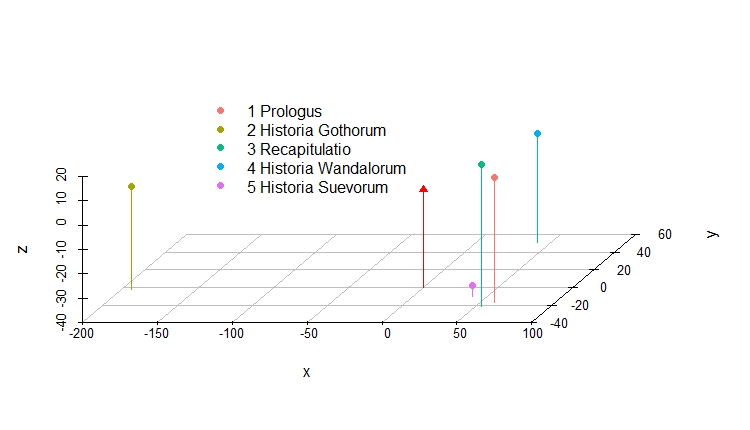


Рисунок 2. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» в трех измерениях.

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. Полученные 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 между собой 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~~ [James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer, 2015. 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 [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. P. 98.; Kassambara A. Practical Guide to Cluster Analysis in R. Unsupervised Machine Learning. STHDA, 2017. P. 67-68.; James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer, 2015. 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 из стандартного пакета stats [].

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

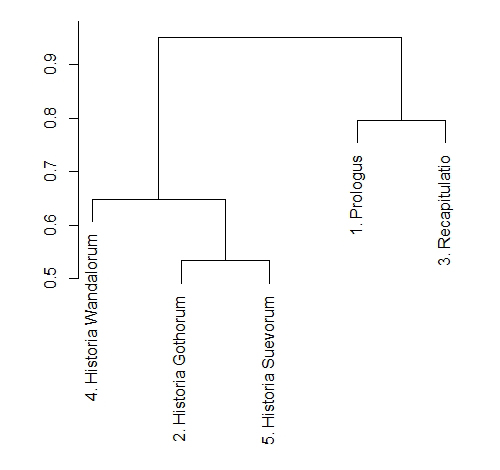


Рисунок 5. Иерархическая кластеризация фрагментов «Истории готов, вандалов и свевов» на основании косинусного расстояния между векторами документов в семантическом пространстве.

«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 [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. P. 99-102.; James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer, 2015. 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.