**УДК 930.2:94(3)**

***Кузнецов Алексей Валерьевич***

*Кандидат исторических наук,*

*научный сотрудник*

*Института всеобщей истории РАН,*

*г. Москва, Российская Федерация*

*e-mail:*

**: латентный семантический анализ средневековых латинских текстов.**

***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, Latent Semantic Analysis (LSA), IS Research Methods, Measurement, Metrics, SVD, Sparse SVD.

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. Apress/Springer, 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. 1997. 104. 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 [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.].

Типичный процесс LSA с использованием пакета R выглядит аналогично показанному на рисунке 1. Во-первых, текстовая матрица строится с помощью textmatrix () из входного корпуса. Текстовая матрица может (но не обязательно) взвешиваться с использованием одной из различных предоставленных схем взвешивания (более подробно см. Wild (2005)).

Затем выполняется разложение по сингулярному значению над этой текстовой матрицей, а полученные частичные матрицы усекаются и возвращаются lsa (). Количество сохраняемых измерений может быть установлено с помощью различных процедур рекомендации (например, dimcalc\_kaiser ()). Полученное в результате скрытое семантическое пространство может быть преобразовано обратно в формат текстовой матрицы с помощью as.textmatrix ().

Наиболее часто используется пакет lsa [lsa: Latent Semantic Analysis [Электронный ресурс]. URL: https://cran.r-project.org/package=lsa (02.01.2020)]

В R пакет lsa (Wild, 2011 ) может быть использован для создания семантических пространств пространств, основанных на алгоритме LSA. Этот пакет также включает в себя основные функциональные возможности для сходства и вычисления окрестностей.

Стандартный ЛСА не предусматривает никакой предварительной работы с исходной матрицей, однако ее преобразование может значительно повысить эффективность данного метода. Предлагается произвести с матрицей следующие действия, которые позволят существенно уменьшить ее размерность:

- удалить строки, соответствующие так называемым стоп-словам (кто, куда, ли, лучше, между и т. д.), а также словам, содержащимся почти в каждом тексте и не несущим никакой смысловой нагрузки (следует помнить, что для каждой предметной области может быть свой список стоп-слов);

- удалить строки, соответствующие редким словам, не встречающимся ни в одном тексте из выборки более одного раза;

- привести все словоформы к исходной форме, например с помощью операции стемминга (процесс нахождения основы слова);

- из текстов некоторых тематик полезно удалить имена собственные, которые, так же как стоп-символы, не несут в себе никакой смысловой нагрузки;

- из текстов некоторых тематик имеет смысл удалить всю цифровую информацию (числительные, цифры).

После того как вся возможная шумовая информация удалена, можно приступать к следующему шагу ЛСА - сингулярному разложению исходной матрицы [2].

Сингулярное разложение - это математическая операция, раскладывающая матрицу на три составляющие. Сингулярное разложение можно представить в виде следующей формулы:

А = USVT, (1)

где А - исходная матрица; U и Vt - ортогональные матрицы; S - диагональная матрица, значения на диагонали которой называются сингулярными коэффициентами матрицы А. Сингулярное разложение позволяет выделить ключевые составляющие исходной матрицы.

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[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).]. Сначала 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).].

МАТЕМАТИЧЕСКИЕ ОСНОВЫ

Для сравнения результатов мы построили 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.

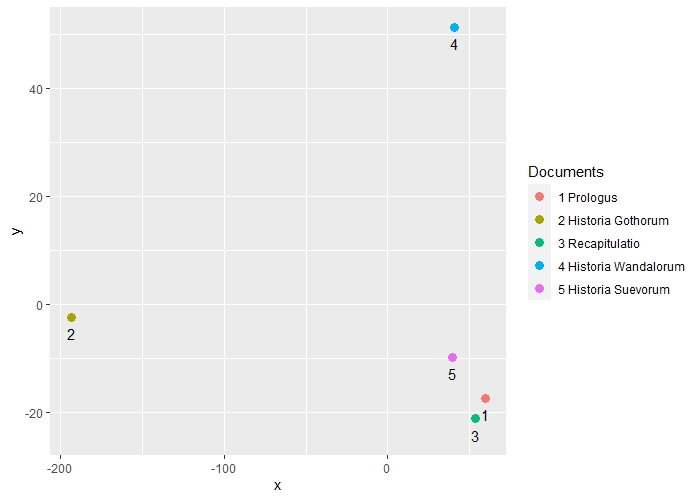


Рисунок 1. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» на плоскость.

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].

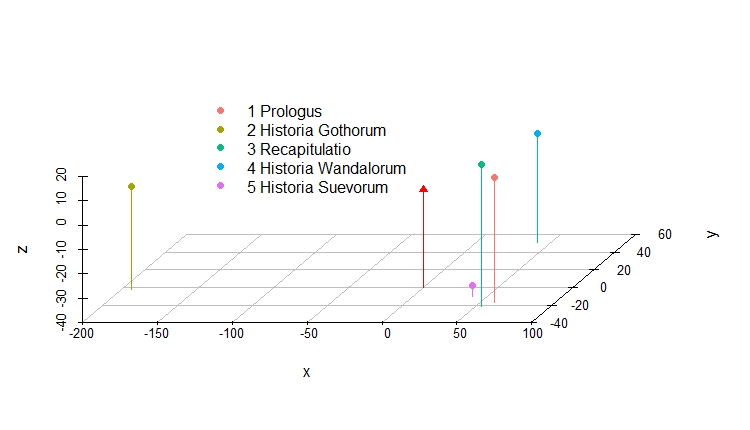


Рисунок 2. Проекция документов в семантическом пространстве «Истории готов, вандалов и свевов» в трех измерениях.

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].

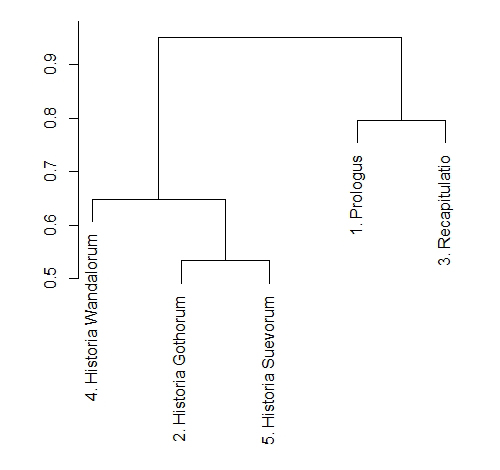


Рисунок 5. Иерархическая кластеризация фрагментов «Истории готов, вандалов и свевов» на основании косинусного расстояния между векторами документов в семантическом пространстве.

«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.