Непарадоксальные парадоксы канона. «Дальнее чтение» Франко Моретти

https://sysblok.ru/permhse/neparadoksalnye-paradoksy-kanona-dalnee-chtenie-franko-moretti/

«Впереди сущий сахарный голод»: что говорит о дневниках 1917 года анализ данных

https://sysblok.ru/linguistics/vperedi-sushhij-saharnyj-golod-chto-govorit-o-dnevnikah-1917-goda-analiz-dannyh/

+ STM

https://bookdown.org/joone/ComputationalMethods/topicmodeling.html

https://www.theanalyticslab.nl/nlpblogs\_5\_state\_of\_the\_art\_in\_nlp\_transformers\_and\_bert/

https://www.theanalyticslab.nl/nlpblogs\_0\_preparing\_restaurant\_review\_data\_for\_nlp\_and\_predictive\_modeling/

https://www.theanalyticslab.nl/nlpblogs\_1\_identifying\_topics\_using\_topic\_-modeling\_lda/

**2.3 Filter tokens**

Topic model is fitted, so, let's explore!

The fitted lda object contains a number of matrices:

* phi: matrix with distribution of tokens (in rows) over topics (in columns)
* theta: matrix with distribution of documents (hence: reviews, in rows) over topics (in columns)

Both for phi and for theta, the sum over all columns is equal to 1, meaning:

* For phi, the sum of all token scores within a topic is 1 - higher scores meaning higher importance of that token within the topic.
* For theta, the sum of all topic scores within a document is 1 - higher scores meaning the topic is more present in that document.

ТЕМЫ В ДНЕВНИКЕ

Марта Баллард, акушерка из штата Мэн, вела дневник с 1785 по 1812 год

За это время она приняла 816 родов и сделала около 10000 записей в дневнике

ТЕМЫ В ДНЕВНИКЕ

MIDWIFERY: birth deld safe morn receivd calld left cleverly pm labour fine reward arivd infant expected recd shee born patient

CHURCH: meeting attended afternoon reverend worship foren mr famely performd vers attend public supper st service lecture discoarst administred supt

DEATH: day yesterday informd morn years death ye hear expired expird weak dead las past heard days drowned departed evinn

GARDENING: gardin sett worked clear beens corn warm planted matters cucumbers gatherd potatoes plants ou sowd door squash wed seeds

SHOPPING: lb made brot bot tea butter sugar carried oz chees pork candles wheat store pr beef spirit churnd flower

ILLNESS: unwell mr sick gave dr rainy easier care head neighbor feet relief made throat poorly takeing medisin ts stomach

**Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama**

http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

Blevins C. Topic Modeling Martha Ballard’s Diary, April 1, 2010 [Электронный ресурс]. URL: http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/ (дата обращения 25.10.2021).

**Interpretable Outputs: Criteria for Machine Learning in the Humanities**

http://digitalhumanities.org:8081/dhq/vol/15/2/000555/000555.html

Те дисциплины и области, которые считают себя в первую очередь гуманистическими, уже давно проявляют особый интерес к изучению нюансов, неопределенности и двусмысленности в наших объектах. Хотя часто можно предположить, что исследования, основанные на данных, противоречат таким концепциям, это не так. Фактически, данные могут способствовать обнаружению и выявлению неоднозначности. Тем не менее, по мере того как вычислительные методы, особенно те, которые предназначены для работы с крупномасштабными текстовыми архивами, все чаще используются такими гуманистами, существует риск того, что доступ к этим неоднозначным измерениям данных будет утерян. Этот риск существует как для исследователя, использующего эти инструменты для интерпретации своих объектов, так и для скептически настроенного читателя или рецензента таких интерпретаций. Визуально привлекательные графические изображения данных, высокая точность классификации и оценки достоверности, а также впечатляющая сводная статистика могут быть легко структурированы риторически, чтобы апеллировать к предвзятым представлениям и общепринятым представлениям. Эффективная и академически ответственная аргументация и интерпретация в гуманитарных науках требует не только открытости, то есть неограниченного доступа к методам и данным, используемым для создания визуализаций и демонстрации точности, но и того, что выбранные наборы данных и модели поддаются интерпретации. Когда признаки/особенности/свойства выводятся из текста, как в случае с машинным обучением, имеющим место в гуманитарных науках, эти особенности должны быть представлены вместе с данными, которые помогают читателям понять их значение. Гуманистическая работа, которая в первую очередь аргументирована - следует признать, что, конечно, не все работы в области цифровых гуманитарных наук являются аргументированными - зависит от определенных рамок, с помощью которых читатели могут оценивать утверждения. Эти структуры обычно действуют в рамках устоявшихся интерпретирующих сообществ ученых, работающих над аналогичными проблемами или использующих аналогичные подходы [Fish 1980]. В то время как данные и статистика имели минимальное присутствие в гуманитарных науках до двадцать первого века, современные вычислительные подходы, если они представлены должным образом, полностью совместимы с существующими гуманистическими аргументационными системами [Dobson 2019]. В то время как некоторые вычислительные области переживают кризис воспроизводимости, настоящее обсуждение не является в первую очередь аргументом о доступе, аргументом, который критикует использование так называемых алгоритмов черного ящика или выступает за то, чтобы цифровые гуманитарные науки присоединились к движению к открытой науке. , а скорее это эссе аргументирует важность слушания, а не сокрытия шума, выявления сложностей и двусмысленностей и интерпретации нюансов в данных, полученных от наших объектов. Это будет сделано путем объяснения роли герменевтики в цифровых гуманитарных науках, прежде чем обратиться к трем важнейшим участкам вычислительных рабочих процессов, которые будут служить в качестве тематических исследований: создание объектов данных, методы тематического моделирования и алгоритмы классификации. Эти три стороны, с их возрастающей сложностью, позволяют нам увидеть основные общие ограничения интерпретируемости всего программного «стека» инструментов, обычно используемых сегодня в цифровых гуманитарных науках.

**Тематические Модели**

[20](http://digitalhumanities.org:8081/dhq/vol/15/2/000555/000555.html#p20)

Тематическое моделирование, одна из первых вычислительных моделей, используемых для исследовательского анализа данных в гуманитарных науках, продолжает оставаться доминирующим методом. Тематические слова или списки характеристик - гораздо чаще, чем частота, с которой эти темы связаны с документами - привлекают внимание гуманистов. В результате неконтролируемой модели «темы», созданные как один из выходных данных алгоритма тематического моделирования, не имеют имени и представляются интерпретатору как интерпретируемый объект. Выходные данные тематического моделирования могут быть очень разнообразными, в зависимости от построения рабочего процесса, включая выбранные алгоритмы, используемые параметры и выполненную предварительную обработку. Нан Да утверждает, что эта изменчивость делает тематическое моделирование «чрезвычайно чувствительным к параметризации, склонным к переобучению и довольно нестабильным» [Da, Nan Z. “The Computational Case against Computational Literary Studies.” Critical Inquiry 45, no. 3 (March 2019): 601–639., 625]. Представление отдельных тем из тематической модели с помощью Wordcloud делает тематические модели особенно трудными для интерпретации. Зритель должен сначала сделать вывод о взаимосвязи между словами, обычно представленными с разными размерами шрифта для изображенных слов или реже с цветовым кодированием, чтобы понять, как читать эти слова по отношению друг к другу. Часто критики, использующие тематические модели, обозначают свою модель тем, что они считают смыслом темы.

[21](http://digitalhumanities.org:8081/dhq/vol/15/2/000555/000555.html#p21)

В обзорной статье Дэвида М. Блея, популяризировавшей тематическое моделирование для гуманитарных наук, Блей продемонстрировал частичный вывод модели скрытого распределения Дирихле (LDA) из 100 тем из 17 000 статей, опубликованных в Science. Этот метод является одним из нескольких, используемых в тематическом моделировании, и остается самым популярным из-за его широкой доступности и простоты. Представление Блея этих тем и его объяснение его выводов по темам помогли установить образец того, что мы могли бы думать как отображение данных без данных. Термины, конечно, являются данными, но значения, придаваемые этим терминам, и их относительная важность для тем скрыты. Способ, которым Блей присваивает ярлыки своим темам, также приводит к некоторой дополнительной путанице, вводя еще один явно субъективный уровень интерпретации. Мы понимаем, что слова, перечисленные первыми в каждом столбце таблицы с пометкой «15 самых частых слов из наиболее часто встречающихся тем» [Blei, David M. “Probabilistic Topic Models.” Communications of the ACM 55, no. 4 (April 1, 2012): 77-84, 79], могут быть наиболее важными для определения значения списка, но насколько на самом деле важны эти слова в тему? Значение важного в такой модели, как LDA, означает статистически значимое или относительно высокое значение для статистически рассчитанного результата. Статистическую важность этих терминов трудно интерпретировать без числовых значений, по которым мы можем определить отношения между элементами в этих списках. В то время как «Эволюция», «Болезни» и «Компьютеры» - это относительно простые ярлыки тем с наиболее часто встречающимися словами, для темы «Генетика» необходимо было перейти к четвертому по частоте термину, чтобы найти правильное обозначение. Дело здесь не в том, что эти темы неправильно помечены, а в том, что метод, с помощью которого определяются метки, непрозрачен - они назначаются специальным образом оператором - и информация, необходимая для определения отношений между этими словами, - которая должна говорят, что кроме ранжированного порядка числовые значения не представлены - недоступно.

|  |  |  |  |
| --- | --- | --- | --- |
| “Генетика” | “Эволюция” | “Болезнь” | “Компьютеры” |
| человек | эволюция | болезнь | компьютер |
| геном | эволюционный | хозяин | Модели |
| ДНК | вид | бактерия | Информация |
| генетический | организмы | болезни | данные |
| гены | жизнь | сопротивление | компьютеры |
| последовательность | происхождение | бактериальные | система |
| геном | эволюционный | хозяин | Модели |
| ген | биология | новое | сеть |
| молекулярный | Группы | деформации | системы |
| последовательность действий | филогенетическая | контроль | Модель |
| Карта | живущий | инфекционный | параллель |
| Информация | разнообразие | малярия | методы |
| генетика | Группа | паразит | сети |
| отображение | новое | паразиты | Программное обеспечение |
| проект | два | объединенный | новое |
| последовательности | Обычный | туберкулез | моделирование |

**Таблица 1.**

Модель темы, отображающая слова без данных и с заголовками “тема”, назначенными оператором [[Blei 2012](http://digitalhumanities.org:8081/dhq/vol/15/2/000555/000555.html" \l "blei2012)].

[22](http://digitalhumanities.org:8081/dhq/vol/15/2/000555/000555.html#p22)

Scikit-learn, который предоставляет несколько алгоритмов, используемых для создания тематических моделей, предоставляет методы для доступа к данным, используемым для присвоения слов тематическим группам с моделями LDA. Эти значения можно запросить с помощью атрибута «n\_components» реализации Scikit-learn модели LDA. Эти «компоненты» или «вариационные параметры для распределения тематических слов» могут предоставлять либо то, что пакет называет псевдосчетом, то есть количество раз, когда каждый из терминов словаря был назначен теме, либо число, которое представляет нормализованное распределение. слов в каждой теме: «Поскольку полное условное распределение слов по теме - это Дирихле, компоненты\_ [i, j] можно рассматривать как псевдосчет, который представляет количество раз, когда слово j было присвоено теме i» [Scikit-Learn 2020]. Псевдосчет особенно полезен при определении относительной значимости каждого термина в списке тем или контейнере.

[23](http://digitalhumanities.org:8081/dhq/vol/15/2/000555/000555.html#p23)

Существует несколько методов, которые используют выходные данные стандартных моделей LDA для предоставления статистики и визуализаций более высокого уровня, которые обеспечивают дополнительную степень проверки как тем, так и отдельных функций, включенных в темы. Несмотря на неоднократные призывы к большей ясности при представлении тематических моделей в гуманитарных науках, эти методы, позволяющие более тщательно изучить, не нашли широкого применения гуманистами [Schmidt, Benjamin. “Words Alone: Dismantling Topic Models in the Humanities.” Journal of Digital Humanities 2, no. 1 (Winter 2012).]. Пакет LDAvis для R и Python предоставляет несколько показателей для изучения и понимания вклада отдельных терминов в темы. Одна из метрик, используемых LDAvis, известна как термит и использует сериацию для вставки извлеченных терминов обратно в некоторый более широкий текстовый контекст, то есть появление интересующего слова до векторизации [Chuang, Jason, Christopher D. Manning, and Jeffrey Heer, “Termite: Visualization Techniques for Assessing Textual Topic Models.” In Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI ’12. The International Working Conference. Capri Island, Italy: ACM Press, 2012, 74-78]. Этот восстановленный контекст сохраняет порядок слов и улучшает понимание тем и слов, но он доступен только в том случае, если доступны предварительно векторизованные источники, другими словами, текстовые источники с неизменным исходным порядком слов. Выбор интерпретируемых объектов данных, в данном случае наличие полных текстовых источников со связанной библиографической информацией, а не просто векторизованных источников, которые могут быть распространены учеными, работающими с копирайтерскими или иными ограниченными наборами текстовых данных, напрямую информирует об интерпретируемости более высоких - уровневые модели.

Заключение

С гуманистической точки зрения мы могли бы думать о моделях данных, созданных из текстовых источников, как об альтернативных представлениях предоставленных текстов, и их преобразования могут быть, как утверждает Кэтрин Боде, перформативными материализациями текстовых источников [Bode 2020]. Это могут быть творческие перформансы в соответствии с понятием вычислительной трансформации как деформации или аргументативные действия. Даже в качестве действия или перформанса гуманисты лучше всего будут служить своей аудитории, выбрав в качестве основных строительных блоков своей работы модели данных и алгоритмы, обеспечивающие максимальную степень интерпретируемости. Понимание значения конкретного экземпляра исполнения или исполнения требует внимательного действия со стороны аудитории. Обращать внимание - значит участвовать в совместном создании смысла с критической / творческой работой. Эта совместная деятельность по созданию смысла лицензируется доступом к общему словарю и потенциальному пространству для игровых манипуляций. Если выдвигаются аргументированные утверждения, они затем оцениваются и подтверждаются общими предположениями и возможностью проверить и подтвердить, что данные действительно понятны в соответствии с нормами разделяемого интерпретирующего сообщества. Таким образом, проверка и интерпретация функционируют вместе, чтобы оживить отношения между исследователем и читателем. Обеспечение интерпретируемости вычислительной работы имеет важное значение для сохранения двух различных направлений в цифровых гуманитарных науках: поддержание стандартов ответственной науки, сформулированных в движении к открытым и воспроизводимым рабочим процессам, и обеспечение совместной деятельности по созданию смысла между критиком и читателем, что характеризует многие гуманистические подходы. интерпретация.

На протяжении всего этого эссе мы видели ограничения интерпретируемости, обнаруженные при выборе и использовании общих векторных моделей данных, алгоритмов и параметров тематического моделирования, а также в алгоритмах классификации. Вычислительные рабочие процессы бывают составными и модульными. Альтернативные процедуры и варианты существуют почти на каждом уровне рабочего процесса. Это и актив, и пассив. Эта гибкость в качестве метода проведения исследования позволяет цифровым гуманистам выбирать и связывать воедино лучшие модели и алгоритмы для своих целей, но по мере увеличения абстракции и сложности также увеличиваются риски для интерпретируемости и доступа к базовому разнообразию данных. Способность гуманистов, в частности, исследовать и интерпретировать тексты и методы, используемые для создания модели, так же важна, если не более важна, чем сообщение общих лучших результатов модели - мы видели это в тематическом исследовании моделирования - или точность конкретной классификационной модели. Именно по этим причинам гуманисты, использующие вычислительные методы для проведения своих исследований, должны выбирать и отдавать предпочтение тем моделям и методам, которые в наибольшей степени обеспечивают возможность проверки, исследования и интерпретации разнообразных данных и параметров.

**Da 2019** Da, Nan Z. “The Computational Case against Computational Literary Studies.” Critical Inquiry 45, no. 3 (March 2019): 601–39.

Johannes Ledolter, Lea VanderVelde, A case study in text mining: Textual analysis of the Territorial Papers, Digital Scholarship in the Humanities, Volume 35, Issue 1, April 2020, Pages 101–126, <https://doi.org/10.1093/llc/fqz007>

https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3343549

A Case Study in Text Mining: Textual Analysis of the Territorial Papers

Digital Scholarship in the Humanities, 2019

U Iowa Legal Studies Research Paper No. 2019-04

26 Pages Posted: 1 Mar 2019 Last revised: 22 Feb 2021

Topic Modelling of Historical Languages in R

by Thomas Koentges | Feb 5, 2016 | Open source | 0 comments

http://www.dh.uni-leipzig.de/wo/topic-modelling-of-historical-languages-in-r/

https://www.researchgate.net/publication/261703677\_Historizing\_topic\_models\_A\_distant\_reading\_of\_topic\_modeling\_texts\_within\_historical\_studies