Научная визуализация, 2019, том 11, номер 4, страницы 13 - 26, DOI: 10.26583/sv.11.4.02

Подходы к визуализации больших массивов текстовых данных на этапе их сбора и предобработки

Е. А. Макарова[1], Д.Г. Лагерев[2]
Брянский государственный технический университет

[1] ORCID: 0000-0002-5410-5890 , m4karova.e@yandex.ru
[2] ORCID: 0000-0002-2702-6492 , LagerevDG@mail.ru

Аннотация

В статье рассмотрен процесс анализа текстовых данных в процессе разработки управленческих решений. Наиболее подробно рассмотрен процесс сбора текстовых данных для дальнейшего анализа, а так же вопросы использования визуализации с целью увеличения эффективности использования человеческих ресурсов на этапах сбора и предобработки данных. Предложена модификация алгоритма для создания визуализации «облако n-gram», позволяющая сделать визуализацию доступной для людей с ограничениями по зрению. Так же предложены методы визуализации моделей векторного представления n-gram (word embedding). На основе проведенных исследований реализована часть программного комплекса, отвечающая за создание интерактивных визуализаций в браузере и взаимодействие с ними.

Ключевые слова: визуализация, обработка естественного языка, доступность веб-приложений.

Введение

В условиях ускорения научно-технического прогресса, как следствие, стремительно ускоряются темпы экономического развития как глобальных, так и локальных рынков. Согласно исследованию [10] количество сделок по поглощению и слиянию в 2017 году в России году увеличилось на 13%. Кроме того, растёт число выданных кредитов. Согласно данным Объединенного кредитного бюро, в годовом отношении количество выданных в России кредитов выросло на 22%, при этом объемы кредитования увеличились на 53%. Помимо ускорившегося оборота денежных средств, рост наблюдается так же в области трудового рынка. В исследовании рекрутинговой компании Antal 27% работодателей заявили, что текучесть персонала за прошедший год в их компаниях выросла. [9]

Возросшая скорость и количество сделок, заключаемых в различных сферах социально-экономической деятельности, выливается в рост нагрузки на управленцев разных уровней. Это требует либо увеличения кадрового состава лиц, принимающих управленческие решения, либо улучшения информационных систем для поддержки принятия управленческих решений с целью снижения нагрузки людей, занятых в процессе принятия решений. Помимо традиционных данных, используемых в подобных системах (таких как, например, кредитная история и капитал для скоринговых систем, используемых при одобрении получения займа) многие исследователи и производители технологических решений используют неструктурированные источники информации об юридических или физических лицах, являющихся участниками сделок. Примером такой информации могут служить данные из средств массовой информации, социальных сетей и т.д.

Кроме того, некоторые исследования показали, что добавление процесса анализа текстовых данных из социальных медиа в модели прогнозирования дают прирост точности. Так, например, помогают увеличить точность предсказания банкротства юридического лица [8]. Так же интересна работа, связанная с прогнозирование инцидентов, связанных с китайским энергетическим рынком. [17] Добавление текстовых данных в модель прогнозирования так же дало прирост точности прогнозирования.

Есть так же работы, которые указывают на то, что для ряда задач добавление текстовых данных в анализ не даст прироста точности прогнозирования, Так, например, текстовые данные добавленные в модель прогнозирования успеха в получении инвестиций для компаний не принесли прироста в точности модели. [18] Авторы другой работы [19] так же исследовали вопрос эффективности добавления текстовых данных в модели прогнозирование финансовых событий и пришли к выводу, что прирост достигается только при условии правильного сбора и предобработки текстовых данных.

Кроме того, открытые источники текстовых данных, как, например, новостные издания, обновляются по мере изменения ситуации, тогда как, например, финансовая отчетность по законам РФ предоставляется в контролирующие органы раз в год.

Наличие подобных регулярно обновляемых источников данных будет полезно как при разовом принятии решения, связанного с юридическим лицом (в случае, если это происходит задолго до подачи подачи отчетности в различные контролирующие органы), так и в случае процесса мониторинга деятельности организации или деятельности, связанной с организацией: например, лизинг крупной техники, долгосрочное кредитование юридических лиц и т.д. В таком случае, регулярный мониторинг открытых источников текстовой информации и их анализ позволит уменьшить возможные риски.

Соответственно, одним из этапов использования систем поддержки принятия управленческих решений является загрузка в них текстовой информации об объекте социально-экономической деятельности для дальнейшего использования.

Объекты социально-экономических отношений широко представлены в сети Интернет, как через «официальные страницы», так и в виде «цифровой репутации» - отзывов, новостей, того, что появляется в сети о них без их непосредственного вмешательства. Однако, количество подобных данных постоянно растёт (в том числе из-за дублирования данных, заимствования данных у одного источника от другого и т.д.), что требует оптимизации с точки зрения скорости и стоимости процесса их сбора и информации.

И, так как увеличивается количество лиц, с кем вынужден контактировать малый и средний бизнес в процессе своей деятельности, увеличивается риск сделки с юридическими или физическими лицами, неблагонадёжными с точки зрения налогового или иного права, что может повлечь за собой долгосрочные последствия, как, например, имидже вые потери и т.д., что может закончится банкротством юридического лица.

С одной стороны, решения нужно принимать всё быстрее, их становится всё больше, что может приводить к увеличению ошибок и рисков, решению этой проблемы может поспособствовать интеграция в СППР системы интеллектуального анализа данных, используя доступные для анализа большие объемы неструктурированных данных [18]. С другой стороны – процесс сбора и предобработки этих данных требует вовлечения серьезных человеческих и вычислительных ресурсов, что может нивелировать экономический эффект, полученный от добавления неструктурированных данных в процесс разработки управленческих решений.

В настоящий момент существуют различные аналитические системы, которые работают не только со структурированными, но и с неструктурированными, в т.ч. текстовыми данными, загруженными из социальных медиа [3]. Однако, сбор и предобработка данных для подобных систем всё ещё являются довольно трудоёмкими.

Задача извлечения информации из разноструктурированных источников.

Рассмотрим подробнее процесс сбора и анализа текстовой информации из различных источников, представленный на рис. 1.

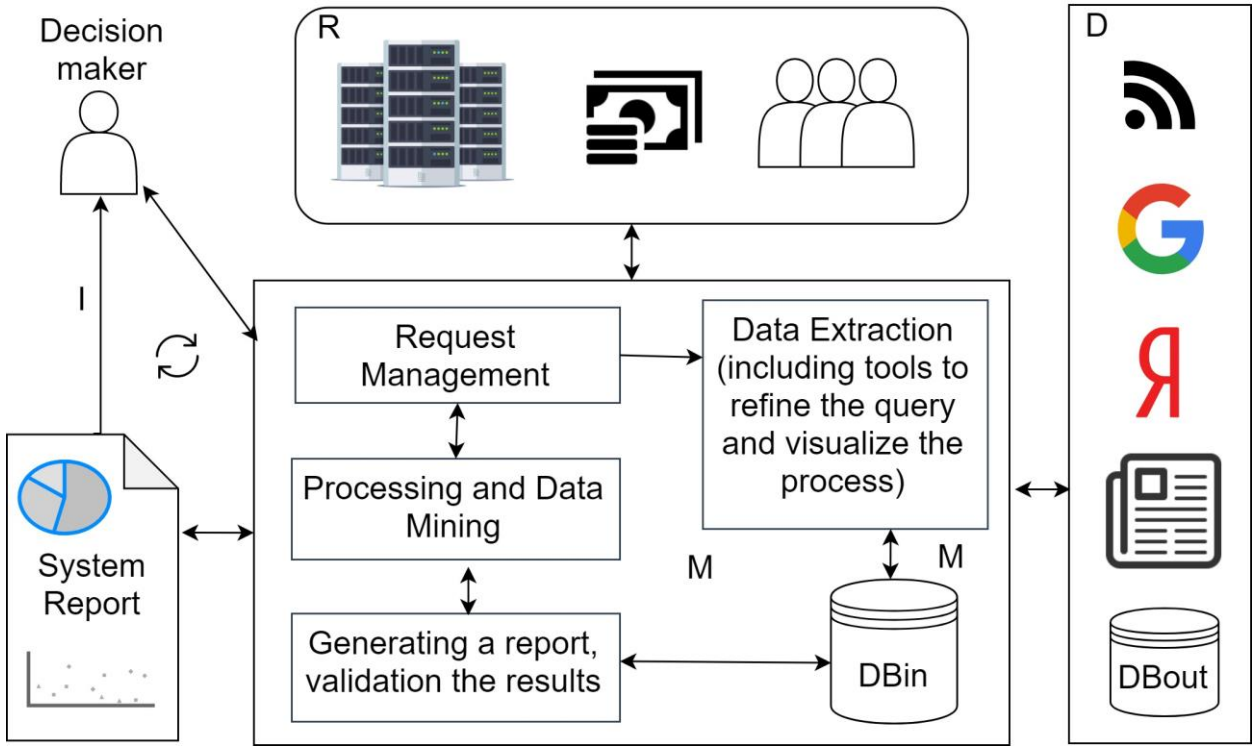

Рис 1. Общая схема процесса сбора и обработки данных

Все процессы, описанные на представленной схеме, важны для эффективного использования неструктурированных текстовых данных для принятия управленческих решений. Однако, в данной работе наибольший акцент сделан именно на процессе сбора информации, т. к. от её качества зависит точность и ресурсозатратность дальнейшего анализа. В том числе, особое внимание уделено процессу «мониторинга» - т.е. регулярного сбора, предобработки и анализа информации, связанной с исследуемой сущностью (объектом). Данные, которые будут получены в ходе работы экспертов с программным комплексом для сбора и анализа информации, будут хранится в базе данных от одного сеанса мониторинга к другому, что позволит минимизировать экспертное время , затрачиваемого для решения задач, которые не представляется возможным или не целесообразно решать автоматически на данном этапе.

На схеме, DB_{in} – внутренняя база данных, содержащая в себе обученные модели для сбора и анализа информации, а так же накопленную информацию об объектах анализа. DB_{out} – базы данных (структурированные источники), подключаемые пользователем.

Концептуальная модель сбора текстовой информации:

$$S = < R, M, D; I >,$$

где**: R –** ресурсы (временные, материальные, человеческие)

M — информация о предыдущих генерациях.

D – данные, отправляемых на анализ;

I – количество релевантной задаче информации, которой располагает система;

$$R = < R_m, R_h, R_p, T >$$

где**:** $\mathbf{R_m}$ – деньги, затрачиваемые на платные сервисы (различные API и сервисы);

$\boldsymbol{R_p}$ – количество доступных экспертов в предметной области.

R_h – ограничения используемых аппаратных ресурсов. Для некоторых задач ограничение по ресурсам будет означать лишь скорость вычислений и, соответственно, рассматриваться совместно с параметром T, но, для некоторых задач обработки языка, например, при использовании векторного представления слов, объем доступной оперативной памяти будет иметь ключевое значение для возможности применения этих методов);

T – время, потраченное на сбор информации, которое, в своё очередь, можно разложить на следующие компоненты:

$$T = T_u + (T_e + T_d) + T_a$$

T_u – время, затраченное основным пользователем системы;

T_e – время, затраченное экспертом, который будет проверять и производить ручное разрешение различных сложных для машинной обработки ситуаций;

T_d – время задержки между ответом эксперта и продолжением обработки (погрешность для учета некруглосуточной доступности эксперта);

T_a – время, потраченное на автоматическую обработку;

Задача оптимизации сбора данных состоит в том, чтобы уменьшать параметры R, H, D, увеличивая при этом параметр I.

Так же предполагается, что ряд параметров должен будет уменьшаться при каждом последующем использовании системы за счёт обучения пользователей и моделей, накоплении полезных знаний об объектах исследования.

В вопросах повышения эффективности сбора информации, существуют две крайности: сделать всю работу полностью автоматической, тем самым сэкономив на человеческих ресурсах, или сделать контроль за процессов полностью «ручным». В данной работе рассматривается «промежуточный» вариант, когда используется привлечение эксперта для оценки эффективности процесса сбора, но с помощью различных инструментов, таких как, например, визуализация, время его работы существенно сокращается [17].

Кроме того, в разрабатываемом программном комплексе применяются следующие подходы для оптимизации сбора информации перед анализом:

1) уточнение поисковых запросов;
2) игнорирование дублирующийся информации;
3) предварительный анализ данных и т.д.

Визуализация больших массивов текстовых данных для решения задачи оптимизации извлечения данных.

Рассмотрим некоторые функции, для более эффективной работы которых требуется человеческое вмешательство и для которых в рамках работы над системой были изучены и доработаны различные методы визуализации [4]. В качестве источника данных в данном примере мы будет использовать Интернет-СМИ, но разрабатываемые методы применимы ко всем схожим по структуре источником с некоторой настройкой.

При настройке выгрузки текстовых документов из некоторого источника, по которому возможен поиск, пользователей системы может столкнуться с тем, что запрос не соответствует необходимому результату. Например, если запрос оказался слишком «общим» или в тех же источниках присутствует информация об одноименных объектах. Выходом из данной ситуации может служить просмотр части собранных текстовых документов, их кратких содержаний или каких-либо метаданных. Это потребует от пользователя (эксперта в предметной области или наемного работника) траты значительного времени. Другой способ ознакомить пользователя с загружаемыми данными – визуализировать их. В предыдущей работе [9] уже было продемонстрировано, что разница в содержании документов заметна на визуализации типа «облако n-gram», и было отмечено, что данный метод требует дальнейшей доработки. В текущей реализации визуализация претерпела ряд изменений, как, например, объединение весов слов,

которые имеют семантическую близость ниже определенного порога, исключение из визуализации «стоп-слов» и слов с маленькими весами.

Векторное представление слов при анализе естественного языка сейчас активно используется в различных исследованиях и прикладных задачах. Например, модели word2vec используются для решения таких задач, как поиск слов с наименьшим семантическим расстоянием (семантически близких).

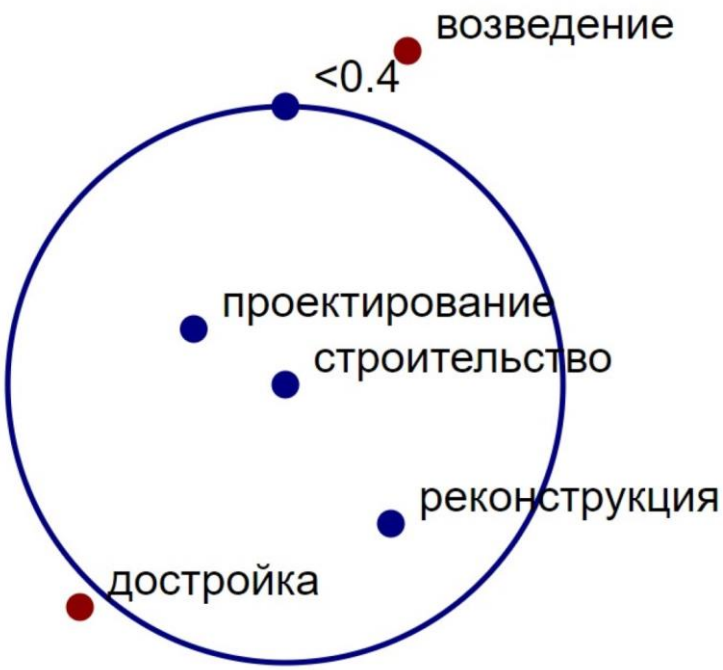

Рис 2 «Круговая» визуализация семантически близких слов к слову «строительство», которые будут объединены при создании визуализации вида «облако слов».

В данном примере (рис. 2) по поиску информации о юридическом лице, деятельность которого связана со сферой строительства, такие слова как «проектирование» и «реконструкция» объединены со словом «строительство», которое наиболее часто встречалось в исследуемой выборке новостей о юридическом лице. На данном изображении круг – это граница семантической близости, равная косинусному расстоянию между словами, равному 0,48, внутрь которого попадают слова, имеющие семантическую близость со словом «строительство» меньшую, чем обозначенная граница. Это позволило не только уменьшить количество слов, которые попали на анализ эксперту (и, соответственно, уменьшить время его работы), но и увеличить акцент на часто встречающихся темах.

В таблице 1 Приведены по три ближайших слова к словам, имевший большой вес в списках ключевых слов, выделенных из коллекции документов по исследуемым предприятиям, а конкретнее – “ **машиностроение** ”, “ **строительство** ”, “ **банкротство** ”. Используются модель, обученные на Национальном корпусе русского языка(www.ruscorpora.ru) и модель, обученная на коллекции новостей на русском языке с 2013го по 2016ый годы (NEWS).

Русскоязычные модели, построенные на методах word2vec, чувствительны к части речи конкретных слов, для работы с ними использовался предварительный POS-tagging с помощью библиотеки pymorphy.

Таблица 1 . Примеры наиболее семантически близких слов на разным датасетам.

word	**машиностроение**	**строительство**	**банкротство**
Top-1 НКРЯ	промышленность	возведение	кредитор
Top -2 НКРЯ	приборостроение	достройка	банкрот
Top -3 НКРЯ	метталургия	проектирование	приватизация
Top -1 NEWS	станкостроение	проектирование	самобанкротство
Top -2 NEWS	автомобилестроение	реконструкция	неплатежеспособность
Top -3 NEWS	приборостроение	достройка	разорение

Одним из направлений доработки используемого метода визуализации будет его адаптация для использования различными группами людей, в том числе, с ограниченными возможностями. При разработке визуализаций важно учитывать возможности взаимодействия с веб-страницей всех групп пользователей, в т.ч. имеющих ограничения различного характера. Это необходимо не только с точки зрения соответствия международным стандартам, но и с точки зрения увеличения числа потенциальных пользователей. Так, например, различными нарушениями цветовосприятия, которые могут помешать пользователю полноценно взаимодействовать с визуализацией, страдают более чем 5% населения [7]

Тема доступной визуализации в последние годы приобретает большой интерес у исследователей и поставщиков услуг по разработке программного обеспечения [13]. Например, руководство по созданию доступных визуализаций данных опубликовала в открытом доступе компания Square [12]. Среди предложенных ими визуализаций присутствуют различные типы диаграмм и графиков.

Сравнительно малоисследованными с этой стороны являются визуализации, связанные с анализом текстовой информации. Далее мы рассмотрим два примера подобных визуализаций, важных для сбора текстовых данных в описываемом программном комплексе.

В классических работах, посвященных построению облаку n-gram (или же «облаку тегов», «облаку слов»), [14, 10], в которых описаны алгоритмы, на которые опираются библиотеки, реализующие данные визуализации, не могли быть учтены рекомендации WCAG по адаптации приложений для людей с нарушениями зрения, т. к. появились до разработки этих рекомендаций.

В качестве элемента программного комплекса был реализован модуль клиент-серверной архитектуры, реализующий интерактивные визуализации и применение пользовательского анализа к процессу сбора данных. Так, например, разработанная визуализация «облако n-gram» учитывает рекомендации WCAG 2.0, поэтому в алгоритм внесены следующие ограничения и дополнения:

1) ограничение на контрастность используемых цветов
2) исключение вертикальных направлений текста [11]
3) установка минимальных и максимальных размеров текста
4) добавление расширенных настроек пользователя.

Т.к. интерфейс существующей системы разрабатывался как веб-приложение, будет разумно при разработке опираться на алгоритмы, используемые для создания и отображения облаков тегов [14], адаптировав их под решаемую задачу и рекомендации WCAG 2.1.

Многие готовые инструменты для визуализации не учитывают контрастность для разных групп населения, включая людей, страдающих нарушениями зрения и цветовосприятия. Однако, стоит понимать, что часто целью создания облака тегов является эффектная иллюстрация массива информации, а не детальный анализ [10].

Контрастность цветов по WCAG 2.1:

$$(L_1 + 0.05) / (L_2 + 0.05) > C_{min}$$

L_1, L_2- относительная яркость сравниваемых цветов.

Так как все слова в визуализации будут являются интерактивными, то необходимую контрастность для них следует рассчитывать как для элементов управления, т. е. $\mathbf{C_{min} = 3}$ для n-gram, располагающихся отдельно. Кроме того, контраст для каждого отдельного цвета по сравнению с фоном должен быть равен $\mathbf{C_{min} = 4{,}5}$ [1]. Произведенные расчёты показали, что возможно найти только **2 цвета**, которые будут контрастны одновременно и с фоном и между собой.

Так же появляются ограничения на размер шрифта. С одной стороны, минимальный размер n-gramы не должен быть меньше 16pt [1]. С другой — тот же стандарт накладывает условие, что все тексты на странице должны увеличиваться до 200%, сохраняя читабельность, что накладывает границу на максимально возможный размер шрифта при отображении страницы в размере 100%. Для того, чтобы сохранять примерное положение контейнеров, в которых будет находится текст при увеличении, при разработке интерфейса использовалась технология CSS Grid [15] и алгоритм Slicing floorplan [14].

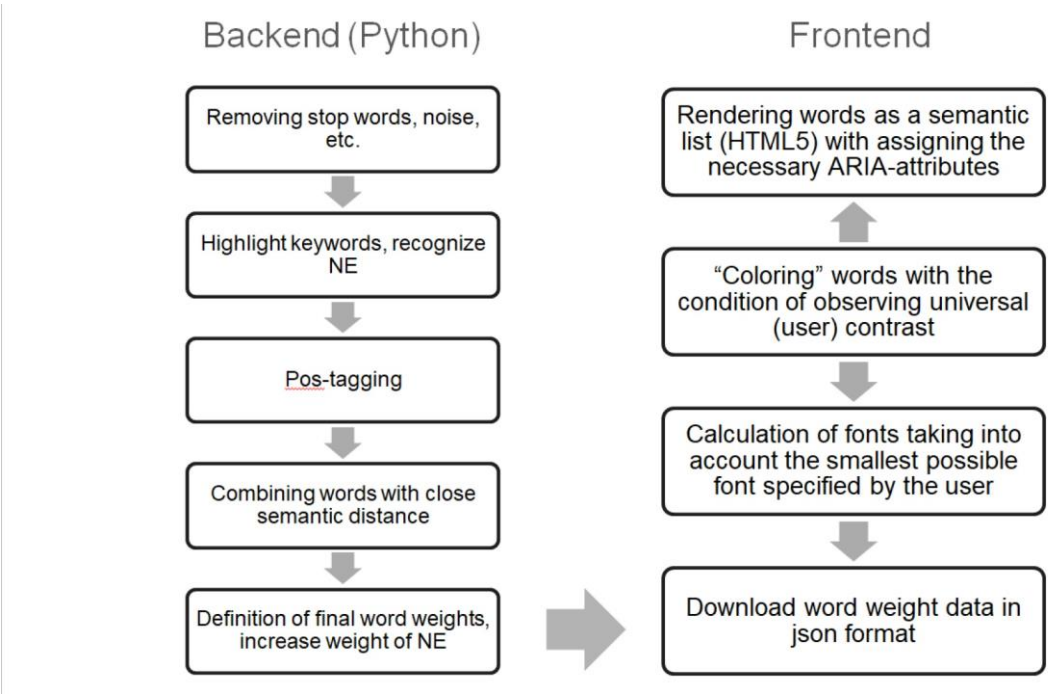

Рис. 3. Создание визуализации

Кроме того, в разделе «расширенные настройки», пользователю даётся возможность поменять минимальный шрифт (при этом отображение будут пересчитано и перерисовано с учетом изменившихся пропорций), задать удобную для него цветовую схему и поменять порядок слов с произвольного (как на приведенной иллюстрации) на «по убыванию», когда n-gramы будут располагаться от имеющий наибольших вес к имеющим наименьший вес.

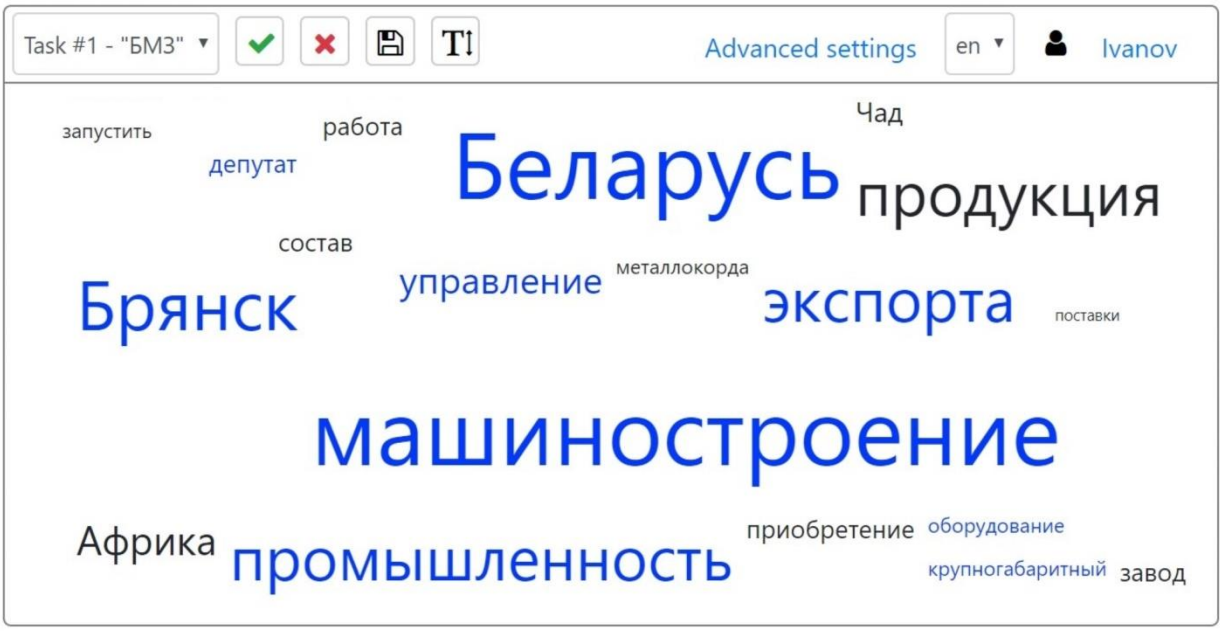

Рис 4. Визуализация коллекции текстовых документов «облако n-gram»

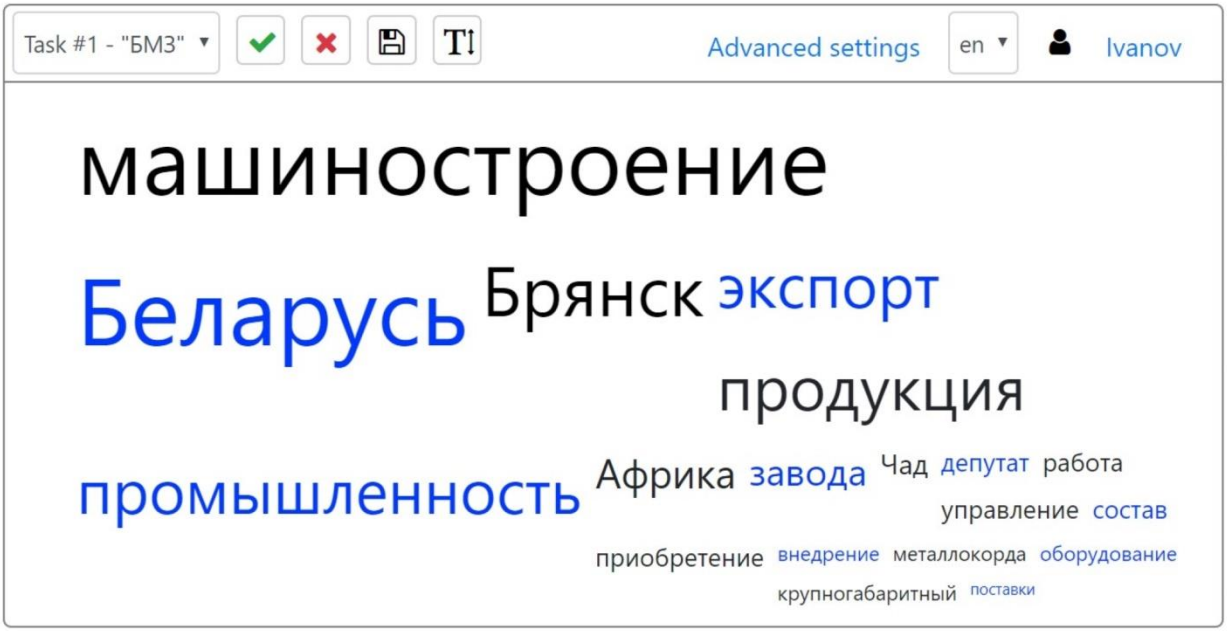

Рис 5. «Упорядоченная» визуализация коллекции текстовых документов «облако n-gram»

Кроме того, у пользователя должна быть возможность добавить пользовательские настройки цветов и размеров визуализации. Рассмотрим конкретный пример. В предыдущей работе [9] подробно описывалось, как визуальный анализ части текстовых документов по поисковому запросу позволяет понять, нужно ли добавить или исключить из запроса различные поисковые сущности. На рис. 2 представлена реализация визуализации по настройке сбора данных по объекту «БМЗ». Цель данной визуализации — проследить, правильно ли был передан контекст запроса, который подразумевал поиск предприятия, находящегося в Брянской области. Как может видеть пользователь из визуализации, настройки поиска были некорректны, что привело к наличию в собранных данных множества документов, связанных с деятельностью аналогичного предприятия в республике Беларусь.

Кроме того, т. к. речь идёт об отображении в браузере, все элементы должны будут иметь атрибут «tabindex» по возрастанию по мере убывания значимости в выборке и атрибут «aria-label» с «весом» данного элемента, чтобы облегчить понимание людям, которые имеют проблемы со зрением и пользуются специальными программами для чтения с экрана.

Обычно для решения одной задачи настройки сбора данных требуется просмотр (или визуализация) 20-30 случайных документов из коллекции документов, в зависимости от количества имеющихся данных. Был проведен эксперимент по влиянию способа решения задачи настройки сбора данных в котором принимало участие пять групп пользователей: пользователи групп 1 и 2 для решения задачи настройки сбора данных использовали визуализацию документов со стандартными настройками, пользователи групп 3 и 4 - беглый просмотр документов, пользователи группы 5 – визуализацию с пользовательской настройкой, время настройки так же включено в итоговый расчёт.

Результаты тестирования на некоторых задачах представлены в таблице. До начала работы с представленными задачами все пользователи проходили обучение на тестовой задаче. Некоторые группы пользователей выполняли только одну группу задач (например, анализа сущностей, связанных с «БМЗ») в один момент времени, другие после решения одной задачи, сразу приступали к следующей.

Таблица 2. Среднее время, затраченное пользователем на один документ на один документ (в секундах).

Задача Группа	«БМЗ»	«БМЗ + Брянск»	«БМЗ + Брянск + производство»	«Экофрио»	«Экофрио + картофель»
Группа 1 (одна задача)	12,5	13	11,5	13	12
Группа 2 (три задачи)	12	13,5	13	14	13
Группа 3 (одна задача)	17,5	19	18,5	16	14
Группа 4 (три задачи)	17	15	14,5	17,5	15
Группа 5 (три задачи)	11	10,5	11	12	11,5

В среднем, визуализация даёт прирост от 18 до 42 процентов. Если пользователем была осуществлена предварительная настройка визуализации под особенности собственного восприятия, скорость обработки задач увеличивалась.

В предыдущей работу по данной теме [9] было продемонстрировано, как модели с использованием word embedding[2], обученные на разных коллекциях текстовых документах, по-разному группируют слова с точки зрения их семантической близости.

Кроме того, в моделях, построенных на word embedding встречаются ошибки, связанные с содержанием исходных данных.

Так же, помимо слов, которые могут быть определены как «соседние» при помощи моделей word2vec, среди списка ключевых слов, которые были выделены в процессе апробации разработанного метода, было так же много слов, которые относились к той же тематике, но имели далёкое косинусное расстояние от исследуемого слова. Ручная разметка слов, подходящих под одну тематику, существенно уменьшает размер визуализации вида «облака слов», однако является достаточно трудоёмкой, что будет подробнее рассмотрено ниже. Однако, в первую очередь при настройке пользователю предоставляются слова с наименьшим семантическим расстоянием, что позволяет быстрее настраивать группировку слов.

Существуют различные способы визуализации векторных пространств, например:

1) t-SNE [23]
2) 3D-визуализация векторного пространства [16]

Приведенные выше методы так же эффективны при решении определенных задач, однако, при разработке данного программного комплекса мы не ставим своей целью визуализацию всего векторного пространства, а только некоторых его «разрезов». Кроме того, важной частью разработки данных визуализаций является их читаемость и доступность, что ограничивает метафоры визуализации, которые мы можем здесь использовать.

Для удобства просмотра слов, семантически близких данному, а так же добавления новых слов в «окружение» данного из семантического пространства, была разработана «круговая» и «вертикальные» визуализации (см рис. 6 и 7)

По предварительным оценкам пользователей, принимавших участие в тестировании интерфейса, круговая визуализация позволяет вмещать слова с большим удобством и компактностью. Однако, данный вид визуализации позволяет наложения текстов один на другой, что противоречит требования WCAG по параметрам доступности веб-приложений. В текущей версии интерфейса для работы с визуализациями оставлена возможность переключиться на удобный пользователю вариант в пользовательском интерфейсе (расширенные настройки). В центре визуализации — слово, чьё положение в модели word embedding исследуется. Расстояния от n-gram определяются таким образом, чтобы двумерный вектор был равен показателю similarity этой n-gramы с исследуемой. Далее алгоритм подбирает положения для n-gram таким образом, чтобы обеспечить читаемость n-gram, включая рекомендации, описанные выше (отсутствие пересечений с другими элементами, горизонтальный текст приемлемого размера). Пример данной визуализации для n-gram, имеющих максимальную семантическую близость со словом «строительство» представлен на рис. 3.

Был проведен следующий эксперимент: было подсчитано количество ключевых слов из собранных документов, прошедших предварительную обработку (удаление стоп-слов, выделение именованных сущностей и т.д.). На следующем шаге были объединены слова с «безопасным» порогом семантической близости, и подсчитано, насколько уменьшилась размер списка слов для визуализации «облако тегов» (первая колонка). После этого была произведена ручная разметка «объединения слов» с использованием разработанной визуализации и подсчитан процент, на сколько уменьшилось пространство слов для визуализации от изначального (второй столбец). Затем, было произведено несколько запусков с загрузкой документов, опубликованных в разные промежутки времени, т.е. симуляция «мониторинга» и подсчитано, как часто приходилось прибегать к ручному редактированию при повторных загрузках, сколько времени на это было потрачено (третий столбец) и на сколько удалось при этом уменьшить размер списка слов. При повторных сборах ручное редактирование происходило только в том случае, если в списке слов с наибольшими весами появились новые сущности. Результаты по ряду объектов представлены в таблице 3

Таблица 3. Уменьшение размер списка при стандартной выборке / Уменьшение выборки с настройкой / время на редактирование (в секундах на один документ)

Объект	Запуск 1			Запуск 2			Запуск 3			Запуск 4			Запуск 5		
БМЗ	7%	23%	12	8%	25%	0	8%	23%	0	7%	24%	7	6%	22%	0
Изотерм	10%	17%	15	9%	15%	8	11%	16%	0	11%	17%	0	7%	14%	0
Экофрио	6%	27%	14	6%	26%	0	7%	25%	3	6%	26%	0	5%	28%	6

На основе проведенных экспериментов был сделан вывод, что настройка слов, входящих в одно «понятие» занимает довольно много времени. Если сравнить с таблицей 2, то становится очевидно, что настройка «объединения слов» при первом запуске соизмеримо со временем обработки всей визуализации. Однако, если мы рассматриваем процесс продолжительного мониторинга (или повторных загрузок при редактировании запроса, что так же является важной частью работы системы), то подобная настройка даёт существенное, примерно 6-23%, уменьшение списка исследуемых слов при соизмеримых временных затратах.

В описываемой системе данные об «объединении слов» используются не только для упрощения визуализации, но и для удаления дублирующихся документов при дальнейшей их обработке.

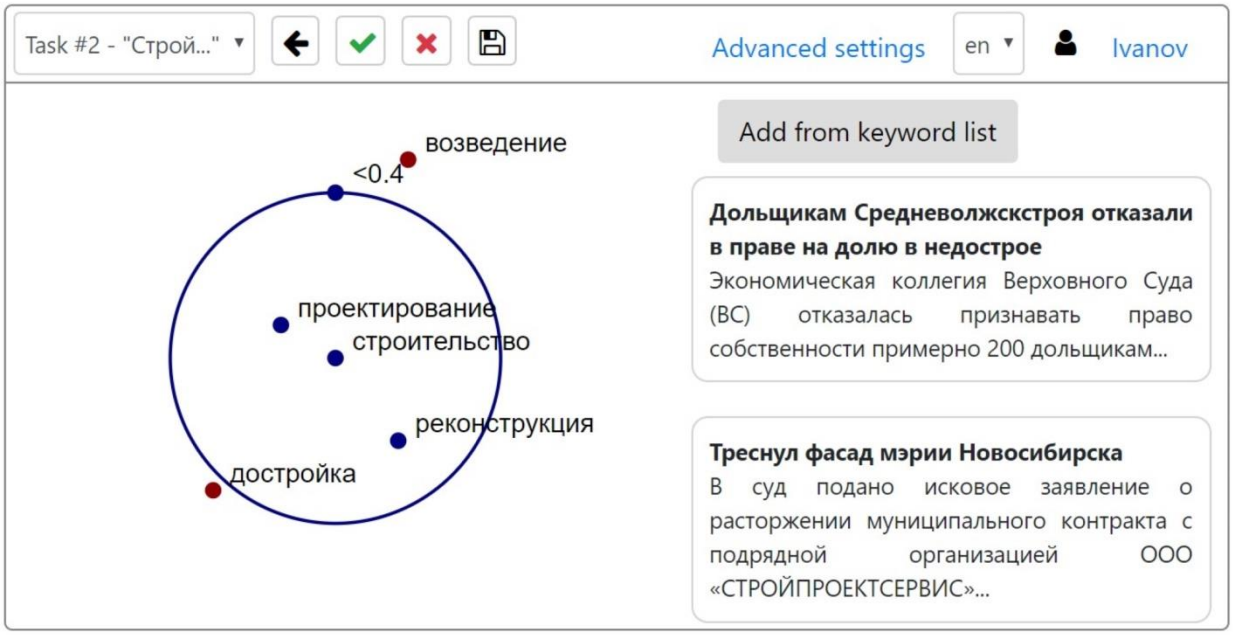

Рис 6. «Круговая» визуализация ближайших соседей n-граты в модели word embedding

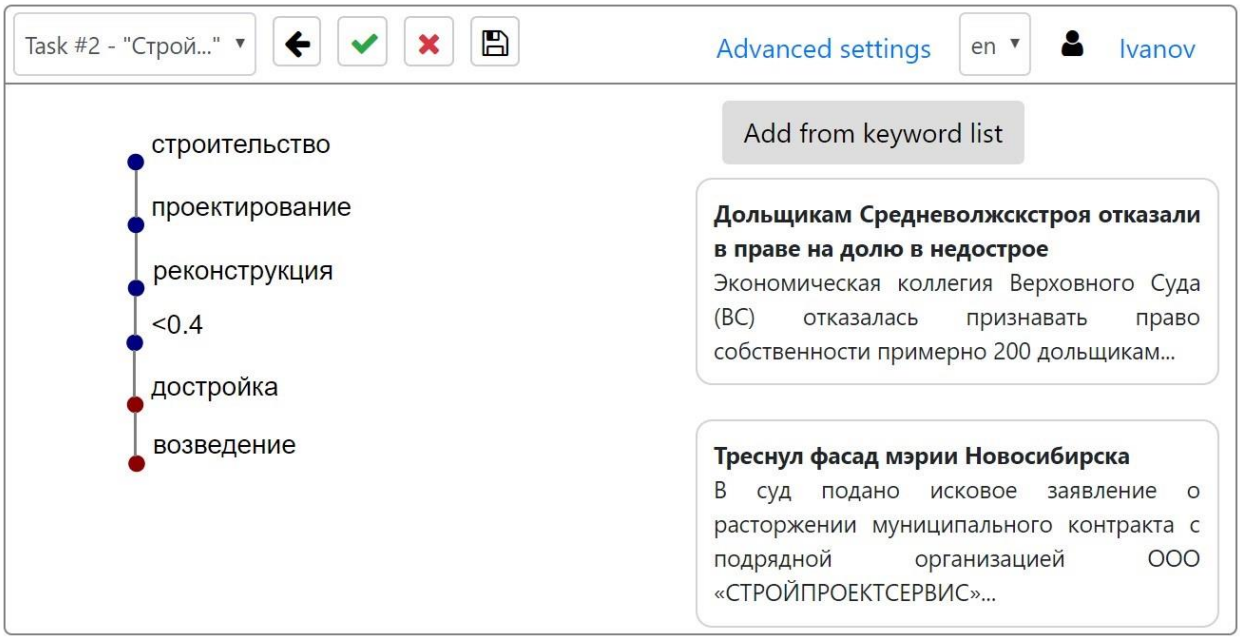

Рис 7. «Круговая» визуализация ближайших соседей n-граты в модели word embedding

В центре визуализации находится слово, положение которого в модели встраивания слова изучается. Расстояния от n-граммы определены так, чтобы двумерный вектор был равен показателю подобия этого n-грамма исследуемому (по умолчанию это значение равно 0,4). Кроме того, алгоритм выбирает позиции для n-грамм таким образом, чтобы обеспечить читаемость n-грамм, включая рекомендации, описанные выше (нет пересечений с другими элементами, горизонтальный текст приемлемого размера). Пример этой визуализации для n-граммов, имеющих максимальную семантическую близость со словом «строительство», представлен на рисунке 6.

В таблице 4 продемонстрировано, как использование визуализации «облако n-gram» и применение результатов анализа к параметрам поискового запроса увеличивают количество релевантных документов, полученных при сборе данных.

Таблица 4. Влияние ручной корректировки запроса на количество релевантных задаче поиска документов

Количество релевантных документов	Объект 1 «БМЗ»	Объект 2 «Изотерм»	Объект 3 «Экофрио»	Объект 4 «Спецстрой»
До корректировки пользователем	20%	30%	85%	10%
После корректировки	85%	45%	90%	20%

В среднем, был отмечен рост количества релевантных документов в среднем **на 24%.** Количество релевантных документов в рамках эксперимента определялось методом экспертного просмотра случайных 20 документов из поисковой выборки. Время, сэкономленное при использовании визуализации по сравнению с беглым просмотром текстов, различно в зависимости от опыта работы пользователя с системой и колеблется от 18 до 42%.

Заключение

Визуализация больших массивов текстовой информации позволяет существенно сократить время, затрачиваемое на их обработку человеком. Кроме того, была разработана часть программного комплекса, реализующая визуализации текстовых данных и моделей векторного представления слов. При разработке алгоритмов визуализации учитывались международные стандартов для создания веб-приложений для людей с ограниченными возможностями, делая их, таким образом, доступными широкому кругу пользователей.

Предоставление интерактивной визуализации для редактирования объединения слов позволило уменьшить размер списка слов для визуализации вида «облако слов», однако, по предварительной оценке, выигрыш по времени при данной ручной настройке возможен только при процессе регулярного мониторинга или многократном повторении процесса загрузки и визуализации загруженных данных с сохранением данных о настройках задачи.

В дальнейшем планируется продолжить исследование методов эффективного сбора данных для использования их в анализе для поддержки принятия управленческих решений. В том числе, планируется более подробное изучение векторного представления n-gram и его использования для определения и удаления дублирующихся данных.

Список источников

1. Web Content Accessibility Guidelines (WCAG) 2.1 https://www.w3.org/TR/WCAG21/
2. Kutuzov, Andrey and Andreev, Igor. Texts in, meaning out: neural language models in semantic similarity task for Russian. Proceedings of the Dialog 2015 Conference, Moscow, Russia (2015)
3. Prangnawarat, Narumol; Hulpus, Ioana; Hayes, Conor / Event Analysisin Social Media using Clustering of Heterogeneous Information Networks // The 28th International FLAIRS Conference (AAAI Publications) (AAAI)
4. Zhao, Jian & Gou, Liang & Wang, Fei & Zhou, Michelle. (2014). PEARL: An Interactive Visual Analytic Tool for Understanding Personal Emotion Style Derived from Social Media. 2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings. 10.1109/VAST.2014.7042496.
5. За прошедший год текучесть персонала в компаниях увеличилась [Электронный ресурс] Режим доступа: https://antalrussia.ru/news/staff-turnover-2018/ Дата обращения: 14.04.19
6. Рынок поглощений и слияний в России в 2017 году. [Электронный ресурс] Режим доступа: https://home.kpmg/content/dam/kpmg/ru/pdf/2018/03/ru-ru-ma-survey-2017.pdf Дата обращения: 14.04.19
7. Causes of Colour Blindness http://www.colourblindawareness.org/colour-blindness/causes-of-colour-blindness/
8. Mai, Feng & Tian, Shaonan & Lee, Chihoon & Ma, Ling. (2018). Deep Learning Models for Bankruptcy Prediction using Textual Disclosures. European Journal of Operational Research. DOI 10.1016/j.ejor.2018.10.024.
9. Zakharova A.A., Lagerev D.G., Makarova E.A. Evaluation of the semantic value of textual information for the development of management decisions. // CPT2019 The Conference Proceedings – May 13 -17, 2019, TzarGrad, Moscow region, Russia
10. Viégas, Fernanda B., Martin Wattenberg, and Jonathan Feinberg. 2009. “Participatory visualization with Wordle.” IEEE Transactions on Visualization and Computer Graphics 15, no. 6 (Nov/Dec 2009): 1137–1144. doi:10.1109/TVCG.2009.17
11. Make your information more accessible. National Disability Authority. http://nda.ie/Resources/Accessibility-toolkit/Make-your-information-more-accessible/
12. Accessible Colors for Data Visualization https://medium.com/@zachgrosser/accessible-colors-for-data-visualization-2ad64ac4ee7e
13. The Future of Data Visualization: Predictions for 2019 and Beyond A https://depictdatastudio.com/the-future-of-data-visualization-predictions-for-2019-and-beyond/
14. Kaser, O., & Lemire, D. (2007). Tag-Cloud Drawing: Algorithms for Cloud Visualization. Tagging and Metadata for Social Information Organization; a workshop at WWW2007, Banff, Alberta, Canada Retrieved December 5, 2007
15. CSS Grid – Table layout is back. Be there and be square. https://developers.google.com/web/updates/2017/01/css-grid
16. Exploring word2vec embeddings as a graph of nearest neighbors. Available by link: https://github.com/anvaka/word2vec-graph
17. Podvesovskii A.G., Isaev R.A. (2018) Visualization Metaphors for Fuzzy Cognitive Maps. Scientific Visualization, vol. 10, no. 4, pp. 13-29. doi: 10.26583/sv.10.4.02
18. Podvesovskii A.G., Gulakov K.V., Dergachyov K.V., Korostelyov D.A., Lagerev D.G. (2015) The choice of parameters of welding materials on the basis of fuzzy cognitive model with neural network identification of nonlinear dependence. Proceedings of the 2015 International Conference on Mechanical Engineering, Automation and Control

Systems (MEACS) (Tomsk, Russia, December 1-4, 2015), IEEE Catalog Number: CFP1561Y-ART, pp. 02-38-NSAP. doi: 10.1109/MEACS.2015.741490
19. Zakharova A.A., Vekhter E.V., Shklyar A.V. (2017) Methods of Solving Problems of Data Analysis Using Analytical Visual Models. Scientific Visualization, vol. 9, no. 4, pp. 78-88. doi: 10.26583/sv.9.4.08
20. Xu, Wei & Pan, Yuchen & Chen, Wenting & Fu, Hongyong. (2019). Forecasting Corporate Failure in the Chinese Energy Sector: A Novel Integrated Model of Deep Learning and Support Vector Machine. Energies. 12. 2251. 10.3390/en12122251.
21. Dorfleitner, Gregor & Priberny, Christopher & Schuster, Stephanie & Stoiber, Johannes & Weber, Martina & Castro, Ivan & Kammler, Julia. (2016). Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. Journal of Banking & Finance. 64. 169-187. 10.1016/j.jbankfin.2015.11.009.
22. Guo, Li & Shi, Feng & Tu, Jun. (2017). Textual Analysis and Machine Leaning: Crack Unstructured Data in Finance and Accounting. The Journal of Finance and Data Science. 2. 10.1016/j.jfds.2017.02.001
23. Visualizing Tweets with Word2Vec and t-SNE, in Python. Available by link: https://leightley.com/visualizing-tweets-with-word2vec-and-t-sne-in-python/

Scientific Visualization, 2019, volume 11, number 4, pages 13 - 26, DOI: 10.26583/sv.11.4.02

Approaches to visualizing big text data at the stage of collection and pre-processing

E. A. Makarova[1], D. G. Lagerev[2]
Bryansk State Technical University

[1] ORCID: 0000-0002-5410-5890 , m4karova.e@yandex.ru
[2] ORCID: 0000-0002-2702-6492 , LagerevDG@mail.ru

Abstract

This paper describes the text data analysis in the course of management decision making. We examine in detail the process of collection of text data for further analysis and the use of imaging to increase the efficiency of human resources during collection and data pre-processing. A modification of the algorithm for creating an “n-gram cloud” visualization is proposed, which makes visualization accessible to people with visual impairments. Also, a method of visualization of n-gram vector representation models (word embedding) is proposed. On the basis of the conducted research, a part of a software package was implemented, which is responsible for creating interactive visualizations in a browser and interoperating with them.

Keywords: visualization, natural language processing, web application accessibility.

References

1. Accessible Colors for Data Visualization. Available by link: https://medium.com/@zachgrosser/accessible-colors-for-data-visualization-2ad64ac4ee7e
2. Causes of Colour Blindness. Available by link: http://www.colourblindawareness.org/colour-blindness/causes-of-colour-blindness/
3. CSS Grid – Table layout is back. Be there and be square. Available by link: https://developers.google.com/web/updates/2017/01/css-grid
4. Kaser O., Lemire D. (2007). Tag-Cloud Drawing: Algorithms for Cloud Visualization. Tagging and Metadata for Social Information Organization. A workshop at WWW2007, pp 1086-1087.
5. KPMG presents the results of a survey of Russia's mergers and acquisitions market in 2017. Available by link: https://home.kpmg/ru/en/home/media/press-releases/2018/03/ma-survey-2017.html
6. Kutuzov A, Kutuzov I. (2015) Texts in, meaning out: neural language models in semantic similarity task for Russian. Proceedings of the Dialog 2015 Conference, Moscow, Russia
7. Mai F., Mai T., Ling C., Ling M. (2018). Deep Learning Models for Bankruptcy Prediction using Textual Disclosures. European Journal of Operational Research. doi: 10.1016/j.ejor.2018.10.024.
8. Make your information more accessible. National Disability Authority. Available by link:http://nda.ie/Resources/Accessibility-toolkit/Make-your-information-more-accessible/
9. Podvesovskii A.G., Isaev R.A. (2018) Visualization Metaphors for Fuzzy Cognitive Maps. Scientific Visualization, vol. 10, no. 4, pp. 13-29. doi: 10.26583/sv.10.4.02
10. Podvesovskii A.G., Gulakov K.V., Dergachyov K.V., Korostelyov D.A., Lagerev D.G. (2015) The choice of parameters of welding materials on the basis of fuzzy cognitive model with neural network identification of nonlinear dependence. Proceedings of the

2015 International Conference on Mechanical Engineering, Automation and Control Systems (MEACS) (Tomsk, Russia, December 1-4, 2015), IEEE Catalog Number: CFP1561Y-ART, pp. 02-38-NSAP. doi: 10.1109/MEACS.2015.741490
11. Prangnawarat N., Hulpus I., Hayes C. (2015) Event Analysisin Social Media using Clustering of Heterogeneous Information Networks. The 28th International FLAIRS Conference (AAAI Publications) (AAAI)
12. Staff turnover has started to grow. Available by link: https://www.antalrussia.com/news/staff-turnover-has-started-to-grow/
13. Exploring word2vec embeddings as a graph of nearest neighbors. Available by link: https://github.com/anvaka/word2vec-graph
14. The Future of Data Visualization: Predictions for 2019 and Beyond A. Available by link: https://depictdatastudio.com/the-future-of-data-visualization-predictions-for-2019-and-beyond/
15. Viégas B., Wattenberg M., Feinberg J. (2009) Participatory visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics 15, no. 6, pp. 1137–1144. doi:10.1109/TVCG.2009.17
16. Web Content Accessibility Guidelines (WCAG) 2.1. Available by link: https://www.w3.org/TR/WCAG21/
17. Zakharova A.A., Lagerev D.G., Makarova E.A. (2019) Evaluation of the semantic value of textual information for the development of management decisions. CPT2019 The Conference Proceedings, TzarGrad, Moscow region, Russia
18. Zakharova A.A., Vekhter E.V., Shklyar A.V. (2017) Methods of Solving Problems of Data Analysis Using Analytical Visual Models. Scientific Visualization, vol. 9, no. 4, pp. 78-88. doi: 10.26583/sv.9.4.08
19. Zhao J., Zhao G., Zhao L., Zhao W., (2014). PEARL: An Interactive Visual Analytic Tool for Understanding Personal Emotion Style Derived from Social Media. IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings. doi: 10.1109/VAST.2014.7042496.
20. Xu, Wei & Pan, Yuchen & Chen, Wenting & Fu, Hongyong. (2019). Forecasting Corporate Failure in the Chinese Energy Sector: A Novel Integrated Model of Deep Learning and Support Vector Machine. Energies. 12. 2251. 10.3390/en12122251.
21. Dorfleitner, Gregor & Priberny, Christopher & Schuster, Stephanie & Stoiber, Johannes & Weber, Martina & Castro, Ivan & Kammler, Julia. (2016). Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. Journal of Banking & Finance. 64. 169-187. 10.1016/j.jbankfin.2015.11.009.
22. Guo, Li & Shi, Feng & Tu, Jun. (2017). Textual Analysis and Machine Leaning: Crack Unstructured Data in Finance and Accounting. The Journal of Finance and Data Science. 2. 10.1016/j.jfds.2017.02.001
23. Visualizing Tweets with Word2Vec and t-SNE, in Python. Available by link: https://leightley.com/visualizing-tweets-with-word2vec-and-t-sne-in-python/