ПРИЛОЖЕНИЕ А

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет компьютерных наук
Департамент программной инженерии

СОГЛАСОВАНО
Доцент базовой кафедры «Системное программирование» ИСП РАН

_____ Д.Ю. Турдаков
«___» _____ 2017 г.

УТВЕРЖДАЮ
Академический руководитель образовательной программы «Программная инженерия»

_____ В.В. Шилов
«___» _____ 2017 г.

Подп. и дата	
Инв. № дубл.	
Взам. инв. №	
Подп. и дата	
Инв. № подл.	

ИНСТРУМЕНТ КЛАСТЕРИЗАЦИИ И ВИЗУАЛИЗАЦИИ КЛАСТЕРОВ НАУЧНЫХ СТАТЕЙ

Техническое задание

ЛИСТ УТВЕРЖДЕНИЯ

RU. 17701729.503390-01 ТЗ 01-1-ЛУ

Исполнитель:
студент группы БПИ133
_____ /Григорьев А.А. /
«____»_____ 2017 г.

2017

УТВЕРЖДЕНО

RU. 17701729.503390-01 ТЗ 01-1-ЛУ

Инструмент кластеризации и визуализации кластеров научных статей

Техническое задание

Листов 19

2017

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.506100 —01 ТЗ				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

АННОТАЦИЯ

Техническое задание – это основной документ, оговаривающий набор требований и порядок создания программного продукта, в соответствии с которым производится разработка программы, ее тестирование и приемка.

Настоящее Техническое задание на разработку проекта «Инструмент кластеризации и визуализации кластеров научных статей» содержит следующие разделы: «Введение», «Основание для разработки», «Назначение разработки», «Требования к программе», «Требования к программной документации», «Технико-экономические показатели», «Стадии и этапы разработки», «Порядок контроля и приемки» и приложения.

В разделе «Введение» указано наименование и краткая характеристика области применения программы для «Инструмент кластеризации и визуализации кластеров научных статей».

В разделе «Основания для разработки» указан документ на основании, которого ведется разработка и наименование темы разработки.

В разделе «Назначение разработки» указано функциональное и эксплуатационное назначение программного продукта.

Раздел «Требования к программе» содержит основные требования к функциональным характеристикам, к надежности, к условиям эксплуатации, к составу и параметрам технических средств, к информационной и программной совместимости, к маркировке и упаковке, к транспортировке и хранению, а также специальные требования.

Раздел «Требования к программным документам» содержит предварительный состав программной документации и специальные требования к ней.

Раздел «Технико-экономические показатели» содержит ориентировочную экономическую эффективность, предполагаемую годовую потребность, экономические преимущества разработки программы «Инструмент кластеризации и визуализации кластеров научных статей».

Раздел «Стадии и этапы разработки» содержит стадии разработки, этапы и содержание работ.

В разделе «Порядок контроля и приемки» указаны общие требования к приемке работы.

Настоящий документ разработан в соответствии с требованиями:

1) ГОСТ 19.101-77 Виды программ и программных документов [1];
2) ГОСТ 19.102-77 Стадии разработки [2];
3) ГОСТ 19.103-77 Обозначения программ и программных документов [3];
4) ГОСТ 19.104-78 Основные надписи [4];
5) ГОСТ 19.105-78 Общие требования к программным документам [5];
6) ГОСТ 19.106-78 Требования к программным документам, выполненным печатным способом [6];
7) ГОСТ 19.201-78 Техническое задание. Требования к содержанию и оформлению [7].

Изменения к данному Техническому заданию оформляются согласно ГОСТ 19.603-78 [8], ГОСТ 19.604-78 [9].

Содержание

ВВЕДЕНИЕ

Наименование программы

Полное наименование программы – «Инструмент кластеризации и визуализации кластеров научных статей».

Краткая характеристика области применения

Инструмент создаётся как компонент системы под названием Research Supporter. Данная система представляет собой интерактивный граф цитирования **научных статей**, где каждая статья является узлом графа, а рёбрами представлены факты наличия одной статьи в **списке источников** другой. Данный граф расположен на карте, где его узлы можно произвольно перемещать.

Разрабатываемая программа позволяет, разместив на карте некоторое количество научных статей, автоматически разделить их на группы семантически близких документов, а также визуализировать полученные группы статей.

Основная цель разрабатываемой программы -- облегчить работу и взаимодействие **исследователей** с научными статьями.

ОСНОВАНИЯ ДЛЯ РАЗРАБОТКИ

Задание на выпускную квалификационную работу. Тема работы: «Инструмент кластеризации и визуализации кластеров научных статей». Национальный исследовательский университет – Высшая школа экономики, факультет компьютерных наук, департамент программной инженерии, в соответствии с Приказом НИУ ВШЭ № 2.3-02/0812-04 от 08.12.2016

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.506100 —01 ТЗ				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

НАЗНАЧЕНИЕ РАЗРАБОТКИ

Функциональное назначение

Программа предоставляет возможность разделения набора научных статей на кластеры, статьи в которых семантически более близки друг к другу, чем к статьям из других кластеров. Кроме того, программа визуализирует сгенерированные кластеры разделяя их на карте на группы, обозначенные непрерывными контурами, обхватывающими каждый кластер статей.

Эксплуатационное назначение

Программа является компонентом системы для работы с научными статьями, позволяющей облегчить процесс исследовательского поиска для научных работников. Каждый пользователь данной программы может хранить на интерактивной карте выбранные научные статьи, видеть отношения между ними в виде графа цитирования и разделения статей на кластеры по семантической близости.

ТРЕБОВАНИЯ К ПРОГРАММЕ

1. Требования к функциональным характеристикам

Программа состоит из двух основных компонент: клиентской и серверной частей, между которыми должно быть налажено взаимодействие

1.1. Требования к серверной части

На серверной части должен быть реализован алгоритм кластеризации статей, разделяющий находящиеся в базе данных статьи, принадлежащие определённой **исследовательской карте** (research map), на группы таким образом, чтобы статьи в одной группе были семантически ближе по отношению друг к другу, чем по отношению к статьям из других групп.
Семантическая близость между статьями определяется как семантическая близость между текстами их заголовков и аннотаций.
Для определения семантической близости используется косинусное расстояние между векторными представлениями текстов с помощью алгоритма paragraph2vec [10].

Также должно быть реализовано взаимодействие с базой данных для получения статей и сохранения сгенерированных кластеров.
Каждый сгенерированный кластер должен быть представлен как структура, состоящая из собственного уникального по отношению ко всем сущностям в базе данных идентификатора и списка идентификаторов статей, относящихся к этому кластеру.

Должна быть возможность задавать количество кластеров, на которые будут разделены статьи.

1.2. Требование к взаимодействию клиентской и серверной частей

Взаимодействие между клиентской и серверной частями должно осуществляться посредством HTTP-запросов.
При получении GET-запроса от клиента, сервер должен ответить сообщением в формате JSON, содержащим список сгенерированных кластеров с их уникальными идентификаторами и идентификаторами статей, относящихся к ним.

1.3. Требования к клиентской части

Клиентская часть должна быть реализована в виде веб-приложения, запускающегося в браузере и представлена в виде интерактивной карты с расположенным на ней графом цитирования. В графе цитирования статьи являются узлами графа, а ребрами представляется факт наличия одной статьи в списке источников другой. На карте же узлы графа цитирования отображаются в виде прямоугольных элементов с текстовой информацией о статье, а рёбра -- стрелками от цитирующей статьи к цитируемой.

Веб-приложение должно предоставлять следующие возможности:
- разделить на кластеры все статьи на карте;
- разделить на кластеры выбранные статьи;
- разделить на новые кластеры статьи, помещённые в выбранные кластеры;
- удалить выбранные кластеры;
- удалить все кластеры;

Также при разделении статей на кластеры пользователь должен иметь возможность задать количество получаемых кластеров.

Каждый кластер должен быть представлен в виде замкнутого контура, внутри которого располагаются все узлы, представляющие статьи, относящиеся к данному кластеру.
Данные контуры должны автоматически перерисовываться в результате перемещения узлов статей по исследовательской карте.
Кроме того, визуальные отображения кластеров должны взаимодействовать между собой таким образом, чтобы минимизировать площадь пересечения их внутреннего пространства.

2. Требования к надежности

2.1. Требования к обеспечению надежного (устойчивого) функционирования программы

Пользователю, работающему с программой через веб-браузер должен быть предоставлен непрерывный доступ к веб-приложению, расположенному по определённому url-адресу. Веб-сервис не должен непредвиденно прерывать свою работу.

2.2. Время восстановления после отказа

В случае отказа работы серверной части и последующей недоступности веб-приложения, время восстановления не должно превышать одних рабочих суток.

2.3. Отказы из-за некорректных действий оператора

После запуска программы на сервере, отказ программы вследствие некорректных действий оператора должен быть исключён. В том числе, должна быть исключена возможность

непреднамеренного выключения программы не связанного с техническими неполадками сервера.

3. Условия эксплуатации

3.1. Климатические условия эксплуатации

Требований к климатическим условиям эксплуатации не предъявляется

3.2. Требования к видам обслуживания

Обслуживание не требуется.

3.3. Требования к численности и квалификации персонала

Для управления системой достаточно одного человека, способного запустить на сервере систему управления базами

4. Требования к составу и параметрам технических средств

Для использования системы необходим веб-браузер, запущенный на компьютере, у которого есть доступ к сети Интернет.

5. Требования к информационной и программной совместимости

5.1. Требования к информационным структурам и методам решения

Кластеризация статей на серверной части должна быть разделена на два этапа:

- Отображение статей в векторное пространство (векторизация)
- Кластеризация полученных векторных представления статей

Для векторизации статей необходимо использовать метод paragraph2vec в то время как кластеризация полученных векторов должна осуществляться с помощью алгоритма kMeans.

Для визуализации сгенерированных кластеров должен использоваться алгоритм bubble sets.

5.2. Требования к используемым программным средствам.

Должен быть гарантирован доступ к веб-приложению через веб-браузер (Google Chrome/Mozilla Firefox/Opera).

5.3. Требования к исходным кодам и языкам программирования

Разрабатываемая программа должна являться компонентой инструмента Research Supporter. Поэтому серверная часть программы реализуется на языке Scala, а клиентская -- на языке TypeScript.

5.4. Требования к защите информации и программы

Требований к защите информации и программы не предъявляется

6. Требования к маркировке и упаковке

Требований к маркировке и упаковке не предъявляется

7. Требования к транспортировке и хранению

Требований к транспортировке и хранению не предъявляется

8. Требования к хранению и транспортировке носителя

8.1. Требования к хранению и транспортировке носителя

Специальных требований к транспортировке не предъявляется.

8.2. Требования к хранению и транспортировке программных документов, предоставляемых в печатном виде

Требования к транспортировке и хранению программных документов являются стандартными и должны соответствовать общим требованиям хранения и транспортировки печатной продукции:

В помещении для хранения печатной продукции допустимы температура воздуха от 10°C до 30°C и относительная влажность воздуха от 30% до 60%.

Документацию хранят и используют на расстоянии не менее 0.5 от источников тепла и влаги.

Не допускается хранение печатной продукции в помещениях, где находятся агрессивные агенты – растворители, спирт, бензин.

Не допускается попадание на документацию агрессивных агентов.

Транспортировка производится в специальных контейнерах с применением мер по предотвращению деформации документов внутри контейнеров, а также проникновения влаги, вредных газов, пыли, солнечных лучей и образованию конденсата внутри контейнеров.

Программные документы, предоставляемые в печатном виде должны соответствовать общим правилам учета и хранения программных документов, предусмотренных стандартами Единой системы программной документации и соответствовать требованиям ГОСТ 19.602-78 [17].

9. Специальные требования

Специальные требования к данной системе не предъявляются.

ТРЕБОВАНИЯ К ПРОГРАММНОЙ ДОКУМЕНТАЦИИ

Состав программной документации

1) «Инструмент кластеризации и визуализации кластеров научных статей». Техническое задание (ГОСТ 19.201-78);
2) «Инструмент кластеризации и визуализации кластеров научных статей». Программа и методика испытаний (ГОСТ 19.301-78);
3) «Инструмент кластеризации и визуализации кластеров научных статей». Руководство оператора (ГОСТ 19.505-79);
4) «Инструмент кластеризации и визуализации кластеров научных статей». Код программы (ГОСТ 19.401-78).

Специальные требования к программной документации

1) Все документы к программе должны быть выполнены в соответствии с ГОСТ 19.106-78 [6] и ГОСТ к этому виду документа (см. п. 5.1.).

ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ

Ориентировочная экономическая эффективность

Ориентировочная экономическая эффективность обусловлена следующими фактами.

Использование разрабатываемого инструмента сократит время затрачиваемое на поиск нужных научных статей.

Также, разрабатываемый инструмент сократит время на ознакомление с предметной областью и облегчит процесс взаимодействия и навигации по научным статьям, что позволит повысить эффективность работы научных сотрудников.

Предполагаемая потребность

Предполагаемая потребность обуславливается тем фактом, что на данный момент не существует инструмента, позволяющего облегчить процесс исследовательского поиска для научных сотрудников, который занимает значительную часть их работы.

Экономические преимущества разработки по сравнению с отечественными и зарубежными образцами или аналогами

На данный момент не существует прямых аналогов разрабатываемого инструмента. Наиболее широко для работы с научными статьями используется сервис Google Scholar. Данный сервис предоставляет возможности поиска научных статей по их заголовкам, авторам, и метаинформации, а также находить для каждой статьи те научные работы, которые её процитировали.

Research Supporter с инструментом для кластеризации же позволяет сохранять выбранные статьи, добавлять к ним комментарии и разделять на кластеры по семантической близости.

СТАДИИ И ЭТАПЫ РАЗРАБОТКИ

Стадии и этапы разработки были выявлены с учетом ГОСТ 19.102-77 [2]:

Стадии разработки	Этапы работ	Содержание работ
1. Исследование методов кластеризации текстов	Ознакомление с предметной областью	Поиск научных статей по теме
		Выбор методов кластеризации для сравнения
		Выбор наборов научных статей
		Выбор метрик сравнение методов кластеризации
	Проведение экспериментов	Написание системы для проведения экспериментов
		Реализация методов кластеризации
		Запуск методов кластеризации
	Сравнительный анализ	Анализ результатов экспериментов
		Написание статьи о проведённых экспериментах
2. Техническое задание	Обоснование необходимости разработки программы	Постановка задачи
		Выделение сценариев использования
	Разработка и утверждение технического задания	Определение требований к программе.
		Определение стадий, этапов и сроков разработки программы и документации на нее.
		Согласование и утверждение технического задания.
3. Технический проект	Разработка технического проекта	Выбор технических средств
		Разработка архитектуры программы
	Утверждение технического проекта	Разработка плана разработки программы.
		Написание пояснительной записки.

4. Рабочий проект	Разработка программы	Реализация серверной части
		Реализация клиентской части
	Разработка программной документации	Написание программных документов в соответствии с требованиями ГОСТ 19.101-77 [1].
	Испытания программы	Базовое тестирование работоспособности программы
5. Внедрение	Подготовка и защита программного продукта.	Подготовка программы и программной документации для презентации и защиты.
		Презентация программного продукта.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.506100 —01 ТЗ				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПОРЯДОК КОНТРОЛЯ И ПРИЕМКИ

Виды испытаний

Проводится контроль функциональных требований, представленных в техническом задании.

Тестирование делится на несколько этапов:

- Проверка работы сервера посредством отправки на него GET-запросов для получении информации о кластерах
- Проверка работы клиента и визуализации кластеров в веб-приложении

Общие требования к приемке работы

Прием программного продукта происходит при полной работоспособности программы при различных входных данных, при выполнении указанных в пункте 4.1.1 настоящего документа функций, при выполнении требований указанных в пункте 4.2. настоящего документа и при наличии полной документации к программе, указанной в пункте 5.1, выполненной в соответствии со специальными требования указанными в пункте 5.2 настоящего технического задания.

ПРИЛОЖЕНИЕ 1

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. ГОСТ 19.101-77 Виды программ и программных документов. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
2. ГОСТ 19.102-77 Стадии разработки. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
3. ГОСТ 19.103-77 Обозначения программ и программных документов. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
4. ГОСТ 19.104-78 Основные надписи. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
5. ГОСТ 19.105-78 Общие требования к программным документам. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
6. ГОСТ 19.106-78 Требования к программным документам, выполненным печатным способом. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
7. ГОСТ 19.404-79 Пояснительная записка. Требования к содержанию и оформлению. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
8. ГОСТ 19.603-78 Общие правила внесения изменений. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
9. ГОСТ 19.604-78 Правила внесения изменений в программные документы, выполненные печатным способом. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
10. Le Q., Mikolov T. Distributed representations of sentences and documents //Proceedings of the 31st International Conference on Machine Learning (ICML-14). – 2014. – С. 1188-1196.

ПРИЛОЖЕНИЕ 2

ГЛОССАРИЙ

Понятие	**Определение**
Исследовательская карта	Интерактивное svg-изображение, на котором отображается граф цитирования и кластеры статей.
Научная статья	Статья опубликованная в научном журнале.
Исследователь	Научный работник проводящий исследования.
Список источников	Список ресурсов, на которые ссылается данная научная статья, расположенный, как правило, в конце статьи.
Контур	Непрерывная замкнутая кривая линия.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.506100 —01 ТЗ				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

Лист регистрации изменений									
Номера листов (страниц)					Всего листов (страниц в докум.)	№ документа	Входящий № сопроводительного докум. и дата	Подл.	Дата
Изм.	Измененных	Замененных	новых	аннулированных					

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.506100 —01 ТЗ				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата