МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ПСИХОЛОГО-ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ»

СОГЛАСОВАНО:
Начальник отдела планирования и сопровождения научных проектов и мероприятий
Е.В. Аржаных _____
« »_____2019 г.

УТВЕРЖДЕНО:
Решением Научно-экспертного совета
(протокол №__)
от « »_____2019 г.
Врио ректора А.А. Марголис _____

ТЕХНИЧЕСКОЕ ЗАДАНИЕ

на выполнение научно-исследовательского проекта

«Разработка программных средств автоматической обработки научных текстов с использованием моделей компьютерной лингвистики для формирования базы результатов научных исследований»

Факультет «Информационные технологии»
Название структурного подразделения

ТЕХНИЧЕСКОЕ ЗАДАНИЕ
на выполнение научно-исследовательского проекта по теме:
«Разработка программных средств автоматической обработки научных текстов с использованием моделей компьютерной лингвистики для формирования базы результатов научных исследований»

1. **Научная новизна**
   Получение комплекса инструментальных средств для автоматизированной оценки результатов научных исследований, представленных в форме научных публикаций на английском языке, ключевым элементом новизны является применение маркеров дискурса для выделения ключевых результатов исследований и формирования пространства документов отражающего их взаимные отношения.

2. **На решение какой практической проблемы направлена работа**
   Снижение времени необходимого для анализа актуального состояния научных исследований в определённой предметной области. Результат достигается за счёт представления доступных публикаций по референтной теме в форме интерактивного графа в трёхмерном пространстве, спроецированном на плоскость. При этом узлы графа представляют выделенную в ходе автоматического анализа характеристику научного текста (результат исследования, методы исследования и др.), а рёбра отождествлены с отношениями между этими характеристиками.

3. **Сроки начала и окончания работ:** 01.03.2019-31.12.2019

4. **Научное направление исследования в соответствии с государственным рубрикатором научно-технической информации (ГРНТИ) (по третьему уровню иерархии)**
   20 Информатика
   16.31.21 Автоматическая обработка текста. Автоматический перевод.
   Автоматическое распознавание речи

5. **Руководитель проекта**

Юрьев Г.А. - кандидат физико-математических наук, доцент кафедры «Прикладная информатика и мультимедийные технологии» факультета «Информационные технологии» МГППУ;

6. **Исполнители проекта**

Верховская Е.А. – младший научный сотрудник Центра информационных технологий факультета «Информационные технологии» МГППУ;
Юрьева Н.Е. – кандидат технических наук, научный сотрудник центра информационных технологий факультета «Информационные технологии» МГППУ;
Константиновский А.А. – студент 4 курса факультета «Информационные технологии»,
Антипова С.Н. – преподаватель факультета «Информационные технологии».

1. **Предмет исследования**

Семантическое векторное пространство текстов научных публикаций

2. **Цель исследования**

Формирование подхода к автоматизированной обработке текстов научных публикаций, позволяющего выделить такие их содержательные части, как результаты, полученные автором, методы использованные при их получении, а также диспозицию результатов относительно близких по содержанию работ, находящихся в контексте их проблематики.

3. **Задачи исследования**

Проведение предварительной обработки корпуса текстов научных публикаций, с получением их векторного представления, отвечающего парадигме дистрибутивной семантики.

Выбор и разработка подходов к автоматическому извлечению смысловых единиц, обрабатываемых текстов и отношений между ними, с опорой на характерные для научного текста дискурсивные маркеры.

Автоматизированное построение сигнатур текстов, дающих наилучшие результаты в смысле дифференцирования известных из их метаданных категорий в векторном пространстве, сформированном на этапе предварительной обработки.

Создание алгоритмов и программных средств поиска отображений семантического векторного пространства текстов научных публикаций позволяющих визуализировать структуру отношений между результатами исследований соответствующих элементам множества (векторным представлениям текстов) принадлежащим симметричной окрестности определённой точки отображения с заданным радиусом.

Разработка эвристических критериев классификации основанных, в частности, на использовании элементов тезаурусного подхода и концепций латентно-семантического анализа.

Разработка базы метаданных исследований, отражающей их характеристики необходимые для представления неявной структуры связей в многомерном признаковом пространстве дескрипторов, извлечённых на этапах автоматизированной обработки.

Разработка технологии визуализации многомерного пространства корпуса текстов, отражающей их индивидуальные характеристики и неявную структуру связей между ними, включая диспозицию результатов исследований, лежащих в рамках единого контекста.

Разработка программного обеспечения для автоматизированного анализа корпуса текстов с возможностью его динамического пополнения и переоценки текущей структуры связей.

Разработка программного обеспечения для визуализации многомерного пространства корпуса текстов в соответствии с предложенной технологией.

4. Гипотезы исследования

Текстовое описание научного исследования представленного в форме статьи содержит доступную для автоматизированного извлечения информацию о его содержательных особенностях, которая может быть использована для размещения в многомерном признаковом пространстве, дающем возможность представить неявные связи между схожими по направлению исследованиями, представленными в форме научных статей.

11. Методы исследования

Методы математической статистики, системного анализа, машинного обучения, латентно- семантического анализа, дистрибутивной семантики, дискурс анализа.

12.Календарный план выполнения работ

№	Содержание выполняемых работ	Сроки выполнения	Результаты выполнения работ	Исполнители
1.	Разработка архитектурных решений, разработка структуры хранения данных на всех этапах реализации проекта, определение и согласование инструментальных средств решения подзадач и выработка методологии интеграции разработанных подсистем, реализация алгоритмов предобработки и хранения данных, разработка интерактивных форм представления данных с использованием веб-технологий, разработка и внедрение свёрточных алгоритмов построения стилистических профилей документов и системы вывода на их основе.	март-сентябрь 2019	Построение семантической интерпретации слов и конструкций; установление "содержательных" семантических отношений между элементами текста; внедрение свёрточных алгоритмов построения стилистических профилей документов и системы вывода на их основе.	Юрьев Григорий Александрович
2	Представление результатов выполнения первого этапа работ, включая демонстрацию	октябрь 2019		Юрьев Григорий Александрович

	функционирующих интерфейсов взаимодействия с базой результатов исследований построенной на ограниченном корпусе статей. Выбор и согласование статей для отработки интеллектуальных алгоритмов компьютерного анализа научных текстов осуществляется в рабочем порядке, в ходе выполнения первого этапа работ.			
3.	Расчёт моделей для латентно-семантического анализа, расчёт моделей векторизации документов для реализации поиска семантических соответствий, основанного на концепциях дистрибутивной семантики.	март-октябрь 2019	Внедрение инструментов анализа в структуру прикладного ПО.	Верховская Екатерина Андреевна
4.	Координация деятельности рабочих групп, контроль за соблюдением порядка исполнения работ, модульное и комплексное тестирование программного обеспечения, участие в разработке и реализации интерактивных форм представления	март-ноябрь 2019	Реализации интерактивных форм представления информации пользователю.	Юрьева Наталия Евгеньевна

	информации пользователю.			
5.	Реализация нейросетевых алгоритмов для решения задач вывода характеристик исследования из данных документов, анализа и прогнозирования объёмов исследований по отраслям и направлениям на основе текущих трендов. Использование нейросетевых алгоритмов в задачах коллаборативной фильтрации. Участие в разработке моделей для латентно-семантического анализа и моделей векторизации документов.	март-ноябрь 2019	Модели для латентно-семантического анализа и векторизации документов.	Константиновский Александр Александрович
6.	Представление промежуточных результатов выполнения работ на научно-экспертном совете.	декабрь 2019	Промежуточный вариант отчета.	Юрьев Григорий Александрович

13. Результаты выполнения проекта

Инструмент для автоматизированного анализа текстов научных публикаций, с выделением его основной тематики, результатов, методов их получения в контексте выявленной предметной области.

Интерактивные средства визуализации результатов анализа в форме веб-приложения.

14. Способ реализации результатов проекта (предполагаемое использование результатов проекта)

Программные инструменты для автоматизированного анализа структуры результатов научных исследований, выполненных в рамках заданных предметных областей с использованием дополнительных классифицирующих признаков.

15. Список планируемых научных статей по теме исследования в российских и зарубежных рецензируемых изданиях (рекомендованных для публикации результатов исследований сотрудников МГППУ)

№	Автор/соавтор	Название статьи	Журнал	Импакт-фактор	Срок подготовки статьи
Раздел 1. Российские журналы из списка рекомендованных					
1	Юрьев Г.А., Верховская Е.А., Константиновский А.А.	Использование дискурсивных маркеров для решения задачи сопоставления результатов научных исследований, представленных в форме научных публикаций.	«Экспериментальная психология»		
Раздел 2. Зарубежные журналы из списка рекомендованных					
1					
2					
3					