Приложение 2
к протоколу заседания
ученого совета НИУ ВШЭ
от 28.05.2021 № 5

Концепция
лаборатории естественного языка ВШЭ-Яндекс
факультета Санкт-Петербургская школа физико-математических и компьютерных наук НИУ ВШЭ - Санкт-Петербург

1. Обоснование актуальности создания совместно с ООО «Яндекс» лаборатории естественного языка ВШЭ-Яндекс факультета Санкт-Петербургская школа физико-математических и компьютерных наук НИУ ВШЭ - Санкт-Петербург (далее – лаборатория)

Колоссальный рост объема разнообразной информации в современном обществе (30% в год), называемый информационным взрывом, настоятельно требует как новых решений в области искусственного интеллекта и анализа данных, так и новых, высококвалифицированных кадров в этой области. Спрос на таких специалистов растет экспоненциально. Особенно остро ощущается потребность в научно-педагогических кадрах — практически в каждом зарубежном университете ежегодно открываются вакансии на специалистов в этих областях. При этом количество таких специалистов в мире значительно меньше, чем количество открываемых вакансий.

Похожая ситуация наблюдается и в Санкт-Петербурге. Ведущие вузы города (СПбГУ, ИТМО, НИУ ВШЭ СПб, Политех) открывают новые бакалаврские и магистерские программы в области искусственного интеллекта и анализа данных, к ним обращаются ведущие компании (Яндекс, JetBrains, Газпромнефть, Сбербанк и др.) с предложениями о проведении научных исследований в этой области. Однако количество высококвалифицированных специалистов в этой области в городе критически мало.

2. Предпосылки создания лаборатории

ООО «Яндекс» готово помогать организационно и финансово в привлечении ведущих специалистов (как наших бывших соотечественников, так и иностранцев) в области искусственного интеллекта и анализа данных в Санкт-Петербург. При этом ближайшая цель такого проекта - создание точки развития наук в области искусственного интеллекта и анализа данных, которая бы со временем превратилась в полноценную научную лабораторию в этой области, способную как решать современные прикладные и фундаментальные задачи, так и способствовать подготовке кадров высшей квалификации (магистров и аспирантов).

С точки зрения ООО «Яндекс», факультет Санкт-Петербургская школа физико-математических и компьютерных наук НИУ ВШЭ – Санкт-Петербург является идеальной площадкой для создания такой лаборатории. ООО «Яндекс» активно участвует в подготовке бакалавров и магистров, предоставляя студентам бакалаврских и магистерских программ научно-исследовательские проекты, практики, приглашая их на стажировки, являясь одной из основных компаний, куда по окончании бакалавриата и магистратуры уходят на работу выпускники. При этом компания ООО «Яндекс» хотела бы углубить и расширить это взаимодействие, участвуя в подготовке аспирантов этой школы, привлекая ведущих ученых и участвуя в совместных научно-исследовательских проектах. Для этого компания готова предложить паритетное (50 на 50) софинансирование для привлечения ведущих ученых на

долгосрочной основе в лабораторию под названием «лаборатория естественного языка ВШЭ-Яндекс».

Со своей стороны, факультет Санкт-Петербургская школа физико-математических и компьютерных наук НИУ ВШЭ – Санкт-Петербург давно и успешно работает в области искусственного интеллекта и анализа данных. Так, на факультете еженедельно проходят 4 содержательных научно-исследовательских семинара в этой области («Агентные системы и обучение с подкреплением», «Прикладное машинное обучение и глубокое обучение», «Машинное обучение в программной инженерии», «Машинное обучение в биологии и медицине»), на которых выступают как известные в своих областях специалисты, так и студенты бакалаврских и магистерских программ факультета. Появление на факультете новой лаборатории естественного языка, проведение в ней содержательных научных исследований в области искусственного интеллекта и анализа данных существенно усилит научную и проектную составляющие работы со студентами и аспирантами факультета.

Кроме того, создаваемая лаборатория планирует наладить тесное сотрудничество в научной сфере с недавно созданной научно-учебной лабораторией компании Яндекс (далее – НУЛ) факультета компьютерных наук НИУ ВШЭ https://cs.hse.ru/big-data/yandexlab/. Тематики НУЛ перекликаются с планируемыми темами исследований создаваемой лаборатории естественного языка ВШЭ-Яндекс. В частности, планируется совместная работа по направлениям анализа текстов на естественном языке и машинного перевода, а также по направлению методов машинного обучения.

3. Цель создания лаборатории

Целью создания лаборатории является проведение исследований фундаментальных теоретико-информационных и статистических свойств дискретных последовательностей. Профильными прикладными направлениями работы лаборатории будут генерация текстов и компьютерная лингвистика, а также разработка программных средств анализа текстов и компонент интеллектуальных информационных систем. Теоретические исследования лаборатории будут связаны с теорией информации и Representation Learning. Кроме того, в рамках работы над генеративными языковыми моделями лаборатория будет неизбежно сталкиваться с дефицитом математического аппарата для описания генерации текстов. Создание математического аппарата для строгого определения таких понятий, как «семантическая информация», «новизна последовательности» или «сюжетная линия» является вторым направлением теоретических исследований лаборатории. Прикладные и теоретические направления работы лаборатории будут взаимно дополнять друг друга.

4. Задачи лаборатории

1. Осуществление научно-исследовательской, экспертно-аналитической деятельности в области анализа текстов и компьютерной лингвистики.
2. Реализация теоретических и прикладных исследовательских проектов по направлению фундаментальных теоретико-информационных и статистических свойств дискретных последовательностей.
3. Разработка программных средств анализа текстов и компонент интеллектуальных информационных систем.

5. Описание научно-исследовательской деятельности лаборатории

Планируемые направления исследований:

1. Анализ статистических свойств текстов на естественном языке (русский и английский языки).
2. Разработка методов проектирования глубинных нейронных сетей и методов глубинного обучения (deep learning) для генерации текстов.
3. Интерпретация сложных моделей машинного обучения.
4. Анализ данных в компьютерной лингвистике.
5. Исследования теоретико-информационных свойств текстов на естественном языке.
6. Исследования методов представления текстовой информации.
7. Разработка математического аппарата для создания теории семантической информации.

6. Финансирование деятельности лаборатории

На старте предполагается софинансирование лаборатории компанией ООО «Яндекс» в размере пяти миллионов рублей в год. Конкретный объем средств в дальнейшем будет зависеть от количества и качества привлекаемых сотрудников, и будет уточнён в 2023 году по итогам оценки эффективности работы лаборатории.

7. Кадровое обеспечение лаборатории

Планируемый руководитель создаваемой лаборатории Иван Ямщиков с отличием защитил диссертацию в Бранденбургском Техническом Университете, является приглашённым профессором университета Лиссабона. Иван работал аналитиком в компании Яндекс, был пост-доком в Институте Макса Планка в Лейпциге. Иван Ямщиков — автор ряда научных публикаций на ведущих конференциях по обработке естественного языка (EMNLP, AAAI), активно популяризирует машинное обучение и выступает на отраслевых конференциях по машинному обучению.

Предполагается привлекать в лабораторию в год от 3 до 5 студентов, аспирантов и научных сотрудников с тем, чтобы в перспективе за три года выйти на следующий кадровый состав лаборатории: 1-2 ведущих ученых, 3-5 научных сотрудников, 3-5 аспирантов, а также 8-10 студентов бакалавриата и магистратуры.

На данный момент к научной работе формируемого в настоящее время коллектива уже привлечен внешний научный сотрудник Шарвин Резаголи — перспективный специалист в области theoretical computer science. В ближайшие несколько лет он планирует работать над исследованием клеточных автоматов в контексте обработки потоковой информации динамическими обучающимися системами. Его результаты могут быть востребованы в рамках работы лаборатории для исследований вопросов обработки и генерации естественного языка. В данный момент господин Резаголи руководит исследовательской группой в частной компании, но он заинтересован в публикации научных результатов, не связанных с прикладными задачами компании. Уже достигнута договоренность, что свои научные результаты господин Резаголи будет публиковать с аффилиацией Высшей школы экономики (кампус в Санкт-Петербурге) без привлечения дополнительного финансирования со стороны ВШЭ. Он также будет выступать в качестве соруководителя ряда теоретических исследований сотрудников лаборатории.

Кроме того, у потенциального руководителя лаборатории — Ивана Ямщикова, есть кадровый резерв из молодых специалистов, которые могли бы быть привлечены к работе лаборатории в качестве научных сотрудников, стажеров и лаборантов. Это студенты бакалаврской программы «Прикладная математика и информатика» и магистерской

программы «Программирование и анализ данных», с которыми Иван Ямщиков уже ведет научную работу, а также аспиранты других российских и зарубежных университетов, планирующие прийти в лабораторию в качестве научных сотрудников. Планируется, что в работе в создаваемой лаборатории активное участие примут студенты четвертого курса бакалавриата НИУ ВШЭ – Санкт-Петербург Владислав Мосин и Максим Сурков, которые в данный момент выполняют бакалаврские дипломы под руководством Ивана Ямщикова. В научную команду сотрудников лаборатории планируют войти Илларион Дмитриев, студент первого курса магистратуры НИУ ВШЭ, и Эмилия Шаймуратова — студентка третьего курса бакалавриата НИУ ВШЭ. Также под научным руководством Ивана Ямщикова кандидатскую диссертацию выполняет аспирант УрФУ Вячеслав Шибаев, который является автором публикаций на конференциях списка CORE A*. Планируется, что Вячеслав Шибаев присоединиться к работе в лаборатории в качестве младшего научного сотрудника с последующим развитием научного трека в рамках лаборатории и НИУ ВШЭ – Санкт-Петербург. В прохождении стажировки в создаваемой лаборатории с последующем поступлением в аспирантуру НИУ ВШЭ – Санкт-Петербурге заинтересована Тинатин Осмонова — выпускница магистратуры Академии ОБСЕ. Кроме того, к работе планируется привлечь несколько молодых специалистов из российских технологических компаний, заинтересованных в исследовательской деятельности и возможной смене своей карьерной траектории с отраслевой на научную.

8. Требуемые помещения и оборудование

Планируется, что лаборатория будет располагаться на факультете Санкт-Петербургская школа физико-математических и компьютерных наук по адресу Кантемировская ул., 3, корп. 1. Дополнительных помещений и оборудования на данный момент для создаваемой лаборатории не требуется.

Примерные количественные планы функционирования лаборатории

	Наименование показателя	**2021 г.**	**2022 г.**	**2023 г.**
1.	Количество работников подразделения, включая руководителя	4	7	9
1.1.	Студентов/аспирантов НИУ ВШЭ в штате подразделения	2	4	6
1.2.	Пост-доков в штате подразделения	0	1	2
2.	Количество статей работников подразделения с аффилиацией НИУ ВШЭ, в международных журналах, индексируемых WoS (ед.), а также с докладами на конференциях	1	6	9
2.1.	Из них статей в журналах уровня Q1/Q2 и с докладами на конференциях уровня A* (ед.)	1	2	4

Целевые конференции лаборатории

В современных компьютерных науках приоритетными являются не журнальные публикации, а участие в крупных тематических конференциях. В контексте исследований лаборатории следующие конференции класса A* по версии ВШЭ являются для неё целевыми: NeurIPS, ICML, AAAI, KDD, ICLR.

Компьютерная лингвистика и обработка естественного языка — более узкое направление исследований в рамках компьютерных наук, поэтому, помимо указанных конференций, целесообразно ориентироваться на наиболее цитируемые конференции по

данным платформы Google Scholar (https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics), в особенности на конференции ACL, EMNLP, HLT-NAACL, EACL, COLING, CoNLL.

Профильные журналы из первого квантиля Scopus или WebOfScience также являются целевыми направлениями для публикаций лаборатории, однако не являются первым приоритетом.

Объем финансирования лаборатории из различных источников, млн.руб.

Источник финансирования	2021	2022	2023
1. ООО «Яндекс»	5	5	5
2. Субсидия на выполнение государственного задания по науке	5	5	5
3. НИУ ВШЭ – Санкт-Петербург	-	5	5
4. Гранты РФФИ, РНФ	-	1	7
Итого	**10**	**16**	**22**

Таким образом для создаваемой лаборатории в таблице выше определена схема ее финансирования без необходимости дальнейшего привлечения дополнительных средств со стороны НИУ ВШЭ помимо запланированных. Размер софинансирования лаборатории со стороны ВШЭ на 2022-2023 гг. может быть скорректирован в зависимости от результативности работы лаборатории. Основные расходы лаборатории будут состоять в выплате заработной платы работникам лаборатории. Кроме того, в расходной части планируются командировки, связанные с поездками на конференции, а также расходы, связанные с публикацией результатов исследований.

Примерный объем расходов лаборатории

	2021	**2022**	**2023**
Количество ставок, шт.	5	7	9
Средняя заработная плата сотрудника лаборатории, тыс. руб.	128	133	180
Суммарные расходы на заработную плату, включая ЕСН, тыс. руб.	10 000	16 690	25 310