Кластеризация запрашиваемых на рынке труда компетенций IT-специалистов и обзор области возможного применения обработки данных рынка труда для корректировки повышения качества подготовки IT-специалистов

1. Проблематика

Сфера подготовки кадров высшей квалификации на протяжении всего своего существования всегда была существенно интегрирована в вопрос актуализации предоставляемых навыков и компетенций. Для достижения этих целей высшие учебные заведения расширяют зоны сотрудничества с конечными бенефициарами учебного продукта - промышленными предприятиями и организациями. К таким средствам относятся:

1. Подбор профессорско-преподавательского состава в том числе из профессионалов реального сектора экономики, от предприятий, заинтересованных в молодых специалистах для обеспечения непрерывности деятельности и поддержания квалифицированного кадрового резерва.
2. Формирование базовых кафедр в рамках действующих академических структур учебных заведений, обеспечивающих плотную интеграцию учебного процесса и производственную деятельность.
3. Организация проектной деятельности в условиях промышленных задач, формируемых при непосредственном участии производственных организаций.

Приведенные методы не формируют генеральную совокупность используемых подходов к повышению практикоориентированности учебных программ, однако, демонстрируют общую направленность деятельности учебных заведений.

Отдельным направлением деятельности учебных заведений в парадигме подготовки кадров для производства является формирование и актуализация учебных программ для специальностей и направлений подготовки.
Формирование состава учебных программ, по мнению коллектива авторов настоящей статьи, является одним из наиболее ответственных и значимых направлений деятельности высших учебных заведений в направлении подготовки специалистов.

На практике, формирование учебных программ основано на экспертном подходе профессорско-преподавательского состава кафедры с учетом нормативно-правовых требований регуляторов (в первую очередь, требований ФГОС ВО), а также консультаций с представителями конечных предприятий-заказчиков специалистов.

Экспертный подход продолжительное время являлся безальтернативным источником консенсуса по различным вопросам как в рамках различных производственных организаций, так и в рамках учреждений высшего образования.

Несмотря на продолжительность использования такого подхода во всех направлениях деятельности человека, на текущий момент возможно рассмотрение альтернативного подхода, основанного на анализе большого количества информации. В литературе и обществе устоялось наименование для такого подхода - ‘Data driven decision’.

Суть метода, основанного на данных, заключается в анализе широкого круга информационных источников и агрегирования до уровня статистически значимых закономерностей и тенденций. Такой подход получил существенное распространение с развитием статистики как науки и развитием информационно-телекоммуникационных сетей. Экспоненциальный прирост данных для анализа привнес Интернет и его популяризация в современном мире. Вместе с накоплением информации развиваются методологические направления обработки этих данных, что нашло реальное применение в промышленной эксплуатации.

Одним из примеров стратегии принятия решений на основе данных является задача банковского скоринга. Суть задачи скоринга заключается в формировании дискриминантной карты признаков по потенциальному заемщику с последующим принятием решения в выдаче или отказе в выдачи займа. На текущий момент данная задача решается в автоматическом режиме в большинстве коммерческих банков мира посредством методов машинного обучения. Решение данной задачи методом, основанным на данных, обусловлено следующими факторами:

1. Объем ручного анализа данных при принятии решения и, как следствие, риски реализации человеческого фактора
2. Объем данных по каждой кредитной заявке, значительно превышающий комфортное количество объектов для запоминания человеков на краткосрочном уровне (т.н. “Кошелек Миллера”)
3. Высокая степень детерминированности решений и возможность автоматизации принятия решений вследствие принципа обезличенности процесса в пользу следования действующей политике принятия рисков коммерческого банка.

Вышеприведенные тезисы будут использованы в будущем в качестве репрезентативного примера выгод от использования метода принятия решений, основанного на данных.

Возвращаясь к основной теме данной работы - формирования высококвалифицированных кадров, востребованных на рынке труда, рассмотрим возможности применения метода принятия решений, основанного на данных.

Несмотря на то, что высшие учебные заведения обычно собирают сведения о фактическом трудоустройстве своих выпускников, на текущий момент в России и мире отсутствуют централизованные источники информации и общепринятые методологии ответа на вопрос, насколько комплексно удовлетворяет работодателей подготовка этих выпускников, какие навыки были задействованы, а какие были сформированы непосредственно на производстве.

Для формирования объективной картины о соответствии запросов работодателей и формируемой компетентностной системы выпускников необходимо наладить систематический сбор новой статистики со всех предприятий и государственных учреждений. При гипотетической реализации введения подобных опросных форм, агрегирование и разбор собранной информации имел бы значительные человеческие трудозатраты.

В качестве альтернативного решения данной проблемы коллектив авторов предлагает аппроксимацию реального распределения востребованности формируемых

компетенций путем детального анализа учебных программ (на примере программ Донского государственного технического университета), запросов рынка труда и сопоставлении этих выборок между собой.

В целях дополнительного детерминирования задач исследования была дополнительно определена сфера подготовки специалистов, а именно сфера информационных технологий. Выбор данного сегмента реального сектора обусловлен следующими факторами:

1. Глобализация рынка IT, что позволяет повысить объем выборки, включая информацию из различных стран, при этом сохраняя применимость запросов между ними.
2. Высокая детерминированность запросов к специалистам. В связи с дороговизной стоимости работ специалистов области IT, компании вынуждены уделять дополнительное внимание анализу компетенций действующих и запрашиваемых специалистов, что систематизирует имеющиеся сведения в данной области.
3. Высокая волатильность компетенций внутри сферы, что актуализирует проблему исследования и прогнозирования рынка труда IT-специалистов для содействия высшим учебным заведениям в актуализации динамично устаревающих навыков и компетенций.

Решение обозначенных задач является основной целью комплексного исследования коллектива авторов, включающее в себя несколько созависимых процессов:

1. Анализ существующих решений по анализу компетентностных моделей рынка труда и сопоставления ее с моделью образовательных организации
2. Выбор источника и подготовка данных для дальнейшего проведения анализа
3. Моделирование компетентностной картины рынка труда
4. Сопоставление компетентностных моделей рынка труда и образовательных организаций.
5. Прогнозирование востребованности компетенций и навыков

Настоящая работа призвана обозначить используемый подход к решению поставленной задачи и раскрыть промежуточные результаты в их соотносимости с предметом исследования.

Стоит отметить, что анализ существующих решений приведен в отдельной работе, посвященной обзору подходов и методов различных российских и зарубежных исследователей.

2. Выбор источника и подготовка данных для дальнейшего проведения анализа

Одной из наиболее приоритетных целей исследования является получение наиболее репрезентативной картины рынка труда, что является фундаментальной основой для дальнейших работ.

Благодаря существованию и популярности особых агрегаторов по поиску вакансий и предложений о работе в сети Интернет, получение репрезентативного набора данных для дальнейшего обучения (далее - датасета) является выполнимой задачей.

В связи с приоритетом формирования дальнейшего исследования на примере российского рынка труда, первым ресурсом для дальнейшего изучения стал крупнейший в России веб-портал по поиску работы HeadHunter. В процессе парсинга данных с ресурса были обработаны более 70 миллионов вакансий, из которых отобраны более 4 миллионов вакансий в сфере информационных технологий за период публикаций с октября 2003 года по март 2023 года.
Объем выборки в более 4 миллиона вакансий является наибольшим из аналогичных исследований, основанных на анализе вакансий.

В данном датасете было обнаружено при помощи метода веб-скрапинга и парсинга более 1,5 миллионов первично уникальных навыков. Под первично уникальными навыками подразумеваются навыки, уникальные в рамках исходного выбора из исходного источника.

Поскольку дальнейшее изучение данных чувствительно к многообразию выборки, необходимо устранить несущественные различия в данных путем первичной предобработки (препроцессинга). Понижение размерности исходного словаря приводит к улучшению конечных качеств моделей и меньшему реагированию на дискриминацию незначительных с фактической точки зрения различий. Для вышеописанных целей была проведена очистка данных, включающая в себя:

1. Приведение текстовой информации к нижнему регистру;
2. Замена всех чисел на единообразные символы;
3. Исключение символов пунктуации и специальных символов;
4. Исключение излишних (двойных, тройных и т.д.) пробелов;
5. Удаление семантически незначимых слов русского и английского языков (так называемых в обработке естественного языка “стоп-слов”)
6. Удаление одиночно стоящих символов в связи с их семантической бесполезностью

После приведения строковых данных в унифицированное состояние в датасете все еще остаются однокоренные слова, различные по падежам, полу и числу, например, “работа”, “работе”, “работу”. Для унификации состояния таких подмножеств слов внутри лексических групп необходимо провести одно из двух преобразований стемминг или лемматизацию. В первом случае слова “обрезаются” по основе слова, а во втором приводятся в нормальную форму. В связи с особенностью данных и качеством имеющихся программных инструментов, было принято решения для слов на русском языке проводить операцию стемминга, а на прочих языках - лемматизацию.
Гибкий подход к анализу различных языков обеспечивает более направленное сохранение семантического единообразия обработанных данных.

После очистки данных и максимально возможной их унификации необходимо перевести текстовые представления слов в векторных вид, поскольку алгоритмы машинного обучения, используемые на следующих стадиях, в абсолютном большинстве работают непосредственно с числовыми представлениями.

Выбор модели для векторизации является эмпирическим принципом, поскольку различные современные модели с таким назначением обладают относительно схожими метриками качества. Для решения задачи векторизации была выбрана

имплементация модели FastText под лицензией свободно распространяемого ПО. Как следует из названия имплементации, данная модель ставит своим преимуществом скорость обработки данных, что стало решающим фактором при обработке имеющегося значительного корпуса слов. В конечном итоге, после векторизации была получена модель, векторизующая более 42 миллионов слов.

Векторизованное представление требуемых навыков вакансиях позволяет проводить дополнительные исследования данных и определять группы схожих наборов навыков (что является задачей кластеризации в терминологии машинного обучения). Получение кластеров требуемых навыков из более 4 миллионов вакансий позволяет вручную анализировать значимость и тенденции данных кластеров, определяя логическую связь элементов внутри групп и, что необходимо для дальнейших исследований, динамику востребованности и изменчивости этих групп. В случае, если кластер стабильно востребован на рынке труда IT-специалистов и его состав не меняется, это может стать целесообразным рассмотрением кластера для включения в учебную программу “как есть”. В случае же, если кластер является популярным в вакансиях, однако, отдельные его элементы стремительно меняются, целесообразным будет формирование фундаментальных навыков, позволяющих овладевать запрашиваемыми инструментами.

Кластеризация навыков может быть проведена при помощи различных инструментов, в данном исследовании выбраны алгоритмы ‘k-means’ и ‘DBSCAN’ в связи с высоким качеством кластеризации и различным подходом к определению опорных элементов для формирования кластеров, а также алгоритм ‘T-SNE’ для формирования визуально наглядных кластеров, демонстрирующих возможности использования данного метода для генерализации суждения о применимости подхода в анализе рынка труда и его связи с образовательными программами.

По итогам проведения кластеризациями методами ‘k-means’ были получены 60 кластеров, содержащие более более 14 миллионов навыков для 4 миллионов исходных вакансий. Целевой метрикой для сравнения кластеров служит универсальная метрика Силуэта, описывающая усредненное качество кластеризации каждого объекта.

По результатам анализа, метод кластеризации ‘k-means’ показал удовлетворительные метрики схожести, что позволяет использовать результаты обработки для дальнейшего исследования кластеров.

Для получения наглядной картины расстояния между кластерами был использован метод T-SNE, демонстрирующий расхождения между кластерами.

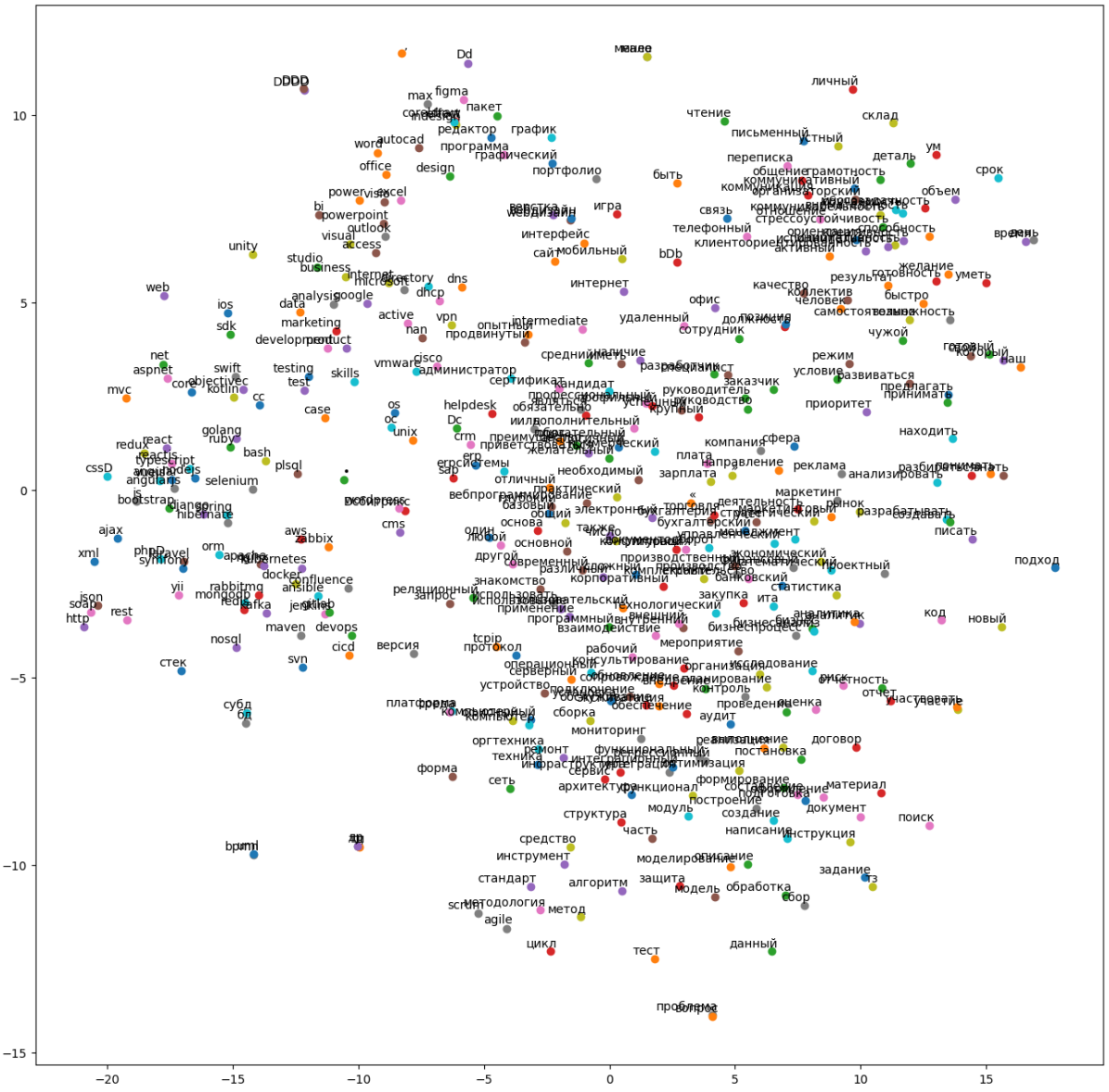

Рис. 1. Кластеризация ключевых слов в навыках при помощи метода T-SNE

Получение и первичная обработка требуемых навыков на рынке труда дает возможность проведения дальнейшего подробного изучения полученных кластеров и решения следующих проблем:

1. Прогнозирование востребованности навыков IT-специалистов;
2. Изучение связи образовательных программ университетов и запросов рынка труда
3. Подготовка автоматизированной рекомендательной платформы по отслеживанию изменений спроса на определенные группы навыков на рынке труда IT-специалистов

Решение данных проблем будет приведено в последующих публикациях цикла.

Сведения об авторах:
Кадомцев Максим Игоревич - к.т.н, зав. каф. Медиатехнологии, Донской Государственный Технический Университет, 2023
Борисов Дмитрий Витальевич - студент магистр, Донской Государственный Технический Университет, 2023
Цеменко Олег Игоревич - студент магистр, Донской Государственный Технический Университет, 2023
Омельченко Андрей Олегович- студент магистр, Донской Государственный Технический Университет, 2023
Омельченко Сергей Евгеньевич - студент магистр, Донской Государственный Технический Университет, 2023
Данилов Даниил Вадимович - студент магистр, Донской Государственный Технический Университет, 2023