Инженерный вестник Дона, №7 (2022)
ivdon.ru/ru/magazine/archive/n7y2022/7783

Метод анализа вакансий строительных специальностей с целью модернизации образовательного курса

В.М. Дронов, Т.С. Рогожина

Санкт-Петербургский государственный архитектурно-строительный университет, Санкт-Петербург

Аннотация: Широкое распространение Educational Data Mining (EDM) позволяет осуществлять сбор данных в образовательной сфере для анализа и корректировки образовательного процесса. На основе выполненного анализа направлений подготовки строительных специальностей был сделан выбор запросов вакансий с сайта hh.ru. Построена цепочка узлов в программе Knime для анализа вакансий строительных специальностей. На основе запросов Knime проведён семантический анализ требований работодателей по строительным специальностям. Была рассмотрена зависимость результата анализа требований работодателей от количества терминов в теме. Была проведена подготовка материала для сравнения данных полученных из рабочих программ с данными, извлечёнными с сайта для поиска вакансий hh.ru.
Ключевые слова: анализ, данные, рабочая программа, образование, запрос, термин, таблица, требование работодателя.

Всё увеличивающийся поток информации электронных ресурсов, «цифровизация» жизни влияет и на систему образования [1], при этом наибольшее внимание уделяется преобразованию образовательного процесса [2]. В то же время, не менее важным является мониторинг трудоустройства выпускников, их востребованность как специалистов и, соответственно, квалификация. Это можно делать, снимая и анализируя цифровой след в различных сферах (социальные сети, экономика, промышленность), объём которого постоянно увеличивается [3,4]. Такая информация создает предпосылки для отслеживания тенденций требований, предъявляемых выпускникам, программам их подготовки.

При разработке учебных программ с учетом требований работодателей с одной стороны, нетрудно учесть требования работодателей в нужном объеме, так как они не регламентируются ФГОС, а с другой в статье 76 ФЗ «Об образовании в Российской Федерации» говорится, что «программы ... разрабатываются на основании установленных квалификационных требований, профессиональных стандартов …». Оптимально, чтобы рабочая

программа одновременно удовлетворяла требованиям рынка и соответствовала ФГОС. Это может быть осуществлено на основании обращений работодателей в образовательное учреждение и анализа потребностей рынка труда.

Курсы и их рабочие программы согласно ФГОС, могут носить научный (академический) характер и слабо касаться интересов работодателей, предъявляемых к выпускникам. Кроме того, знания о нанимателе и его запросах позволят студентам лучше подготовиться, успешно найти работу и эффективно трудиться [3,4]. Следовательно, исследование требований работодателей и их учет при составлении рабочих программ является актуальным.

Для такого анализа были выбраны широко развивающиеся методы Data Mining ("добывание данных"), а далее Educational Data Mining. (EDM) – анализ данных в образовательной сфере. Анализ данных в (EDM) охватывает область исследований, включающую применение интеллектуального анализа данных, методы машинного обучения [5], статистическую обработку информации, полученную из образовательных учреждений.

В качестве инструмента анализа была выбрана программа Knime, которая представляет собою модульную платформу с открытыми исходными кодами, предназначенную для анализа данных, в том числе и текста [6]. Knime позволяет пользователю с помощью цепочки узлов, которые выполняют определённую последовательность действий, строить модули для получения, обработки и анализа данных. К системе можно подключать различные модули, с которыми она совместима: среда машинного обучения WEKA, язык статистических вычислений Project, интерпретатор языка программирования Python, система обработки изображений и их анализа ImageJ, а также базы данных. Можно выполнять запросы с сайтов. Для нашей цели, с помощью Knime, можно легко сделать запросы о вакансиях с сайта

hh.ru и провести их анализ.

Целью настоящего исследования является анализ вакансий строительных специальностей на основе данных портала hh.ru с целью модернизации образовательного курса согласно требованиям работодателя.

Для достижения заявленной цели необходимо решить следующие задачи:

1. Проанализировать направления подготовки и возможное трудоустройство студентов строительных специальностей для выбора параметров запроса.

2. Осуществить экспорт и обработку массива вакансий, полученных с портала hh.ru.

3. Выявить общее и частное в ходе сравнительного анализа.

4. Полученные результаты использовать для рекомендации при составлении и корректировке рабочих программ с целью улучшения ориентации содержания курса требованиям работодателей.

Для анализа были выбраны рабочие программы, читаемые на кафедре «строительной физики и химии» Санкт-Петербуржского государственного строительного университета. Программы составлены, согласно ФГОС.

Поскольку выпускники имеют широкие возможности трудоустройства по данной специальностям, полный перебор по всем вакансиям будет слишком трудозатратен, поэтому для исследования были сделаны следующие запросы:

специальность 08.05.01 Строительство уникальных зданий и сооружений – строка запроса вакансии «инженер-строитель»

21.03.02 Землеустройство и кадастры – строка запроса вакансии «кадастровый инженер»

27.03.01 Стандартизация и метрология – строка запроса вакансии «инженер-метролог»

20.04.01 Техносферная безопасность – строка запроса вакансии «техносферная безопасность»

В первой части работы проведено исследование по наличию вакансий по этим направлениям и уровень зарплат в работе по данным должностям. Для решения этой задачи была построена следующая цепочка Knime. Для апробации метода было решено анализировать вакансии в четырех городах и их областях двух мегаполисах Москва и Московская область, Санкт-Петербург и Ленинградская область, а также двух крайних точках России Калининград и его область и Владивосток и его область.

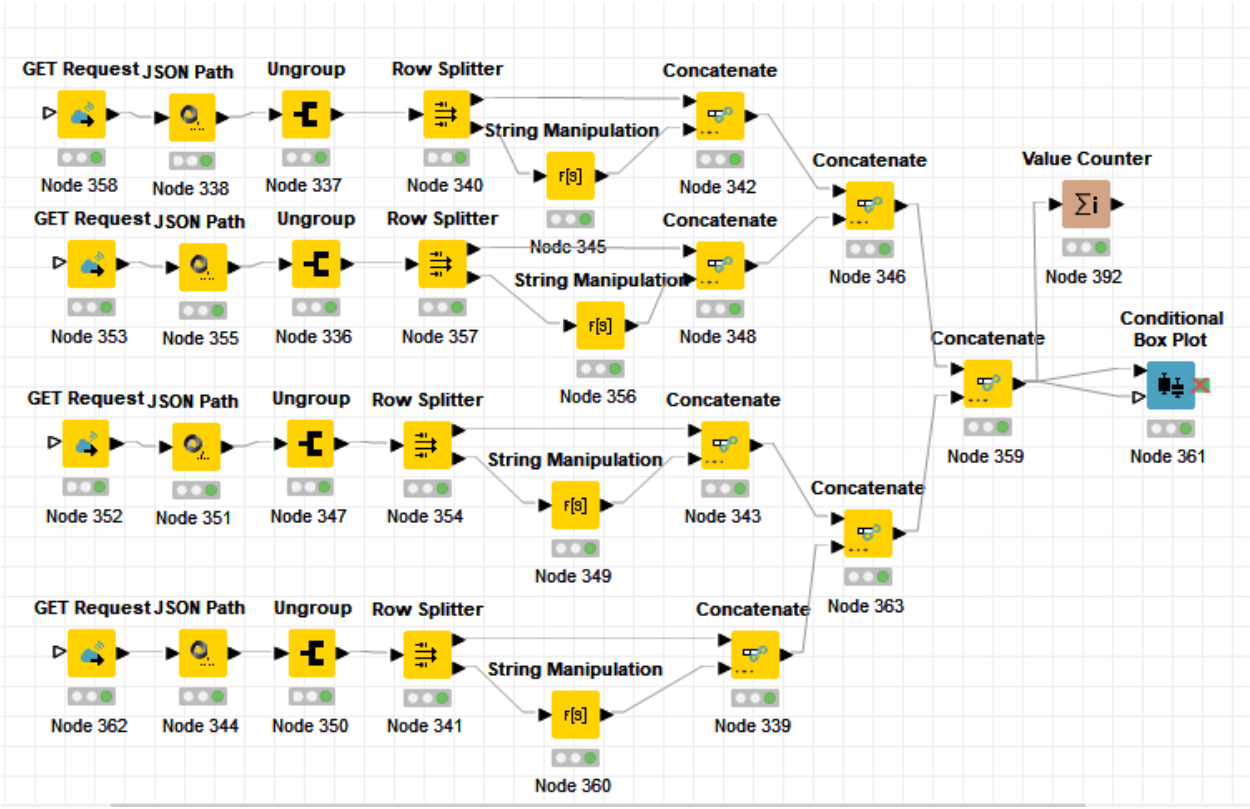


Рис. 1. – Схема Knime для построения ящиковой диаграммы запроса вакансий по четырем городам

1.GET Request, с помощью этого узла делались запросы о вакансиях на сайт hh.ru.

2. Результаты запросов JSONPath преобразуются в выбранный тип KNIME.

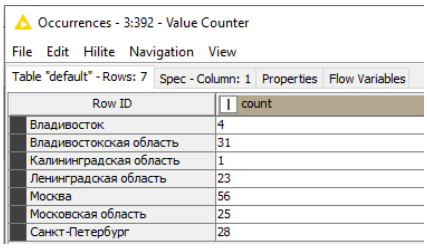
3. Было выяснено, что результаты запросов содержали вакансии в городе и области. Было принято решение с помощью узла Ungroup разделить позиции о вакансиях в городах и их областях.

4. Следующий узел Row Splitter – узел, который фильтрует строки по условию и выдает две таблицы, одна содержит строки, содержащие условие, вторая все остальные. В работе узел разделяет вакансии Владивостока и все остальные.

5. String Manipulation – узел, с помощью которого все пустые строки колонки «sitys» и строки, содержащие разные населённые пункты области переименовали соответствующую область.

6. С помощью Concatenate – последовательно объединили таблицы. Сначала выполняется объединение таблиц вакансий города и области, затем двух таких таблиц и, наконец, получаем объединенную таблицу по вакансиям.

7.Узел Values Counter – позволяет рассчитать число вакансий по городам. См. рис.2. Из таблицы рисунка видно, что такие вакансии в Калининграде и области практически отсутствуют.


Occurrences - 3:392 - Value Counter

File Edit Hilite Navigation View

Table "default" - Rows: 7 | Spec - Column: 1 | Properties | Flow Variables

Row ID	count
Владивосток	4
Владивостокская область	31
Калининградская область	1
Ленинградская область	23
Москва	56
Московская область	25
Санкт-Петербург	28

Рис. 2. – Пример действия узла Values Counter. Для вакансий «инженер-строитель»

Conditional Box Plot – позволяет визуализировать результаты запроса в виде ящиковой диаграммы и провести оценку востребованности вакансий по регионам и уровень зарплат. В качестве категориальной колонки для построения диаграмм выбрана «sitys» (см.рис.3-6.).

Результаты запросов по количеству вакансий сведены в таблицу 1.

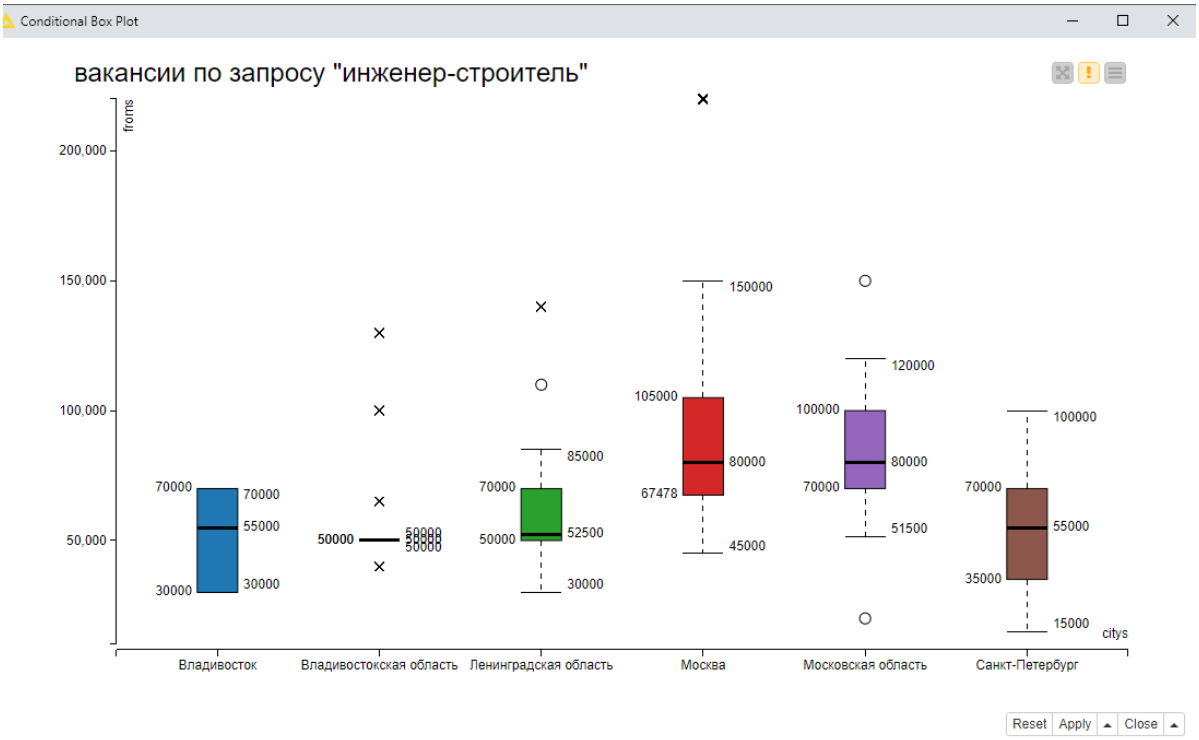


Рис. 3. – Диаграмма Conditional Box Plot (ящиковая диаграмма) по запросу «инженер строитель»

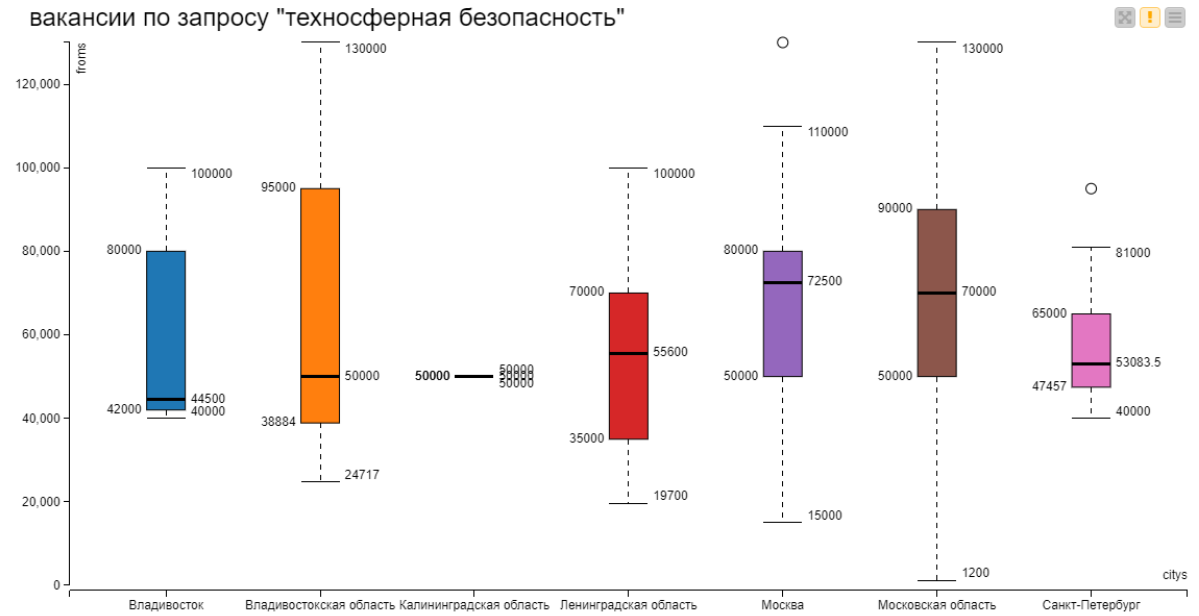


Рис. 4. – Диаграмма Conditional Box Plot (ящиковая диаграмма) по запросу «техносферная безопасность»

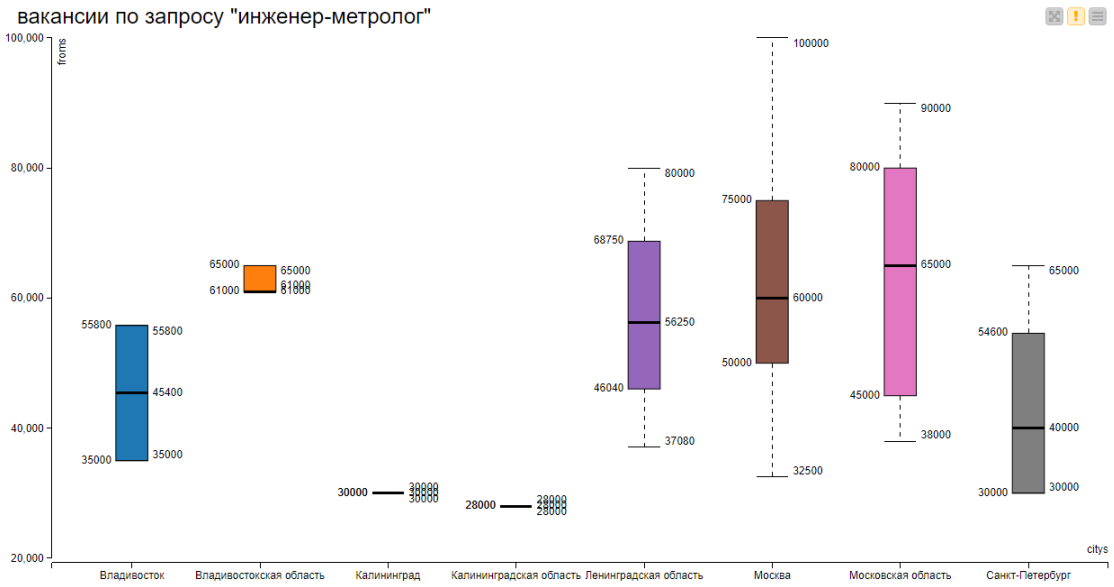


Рис. 5. – Диаграмма Conditional Box Plot (ящиковая диаграмма) по запросу «инженер-метролог»

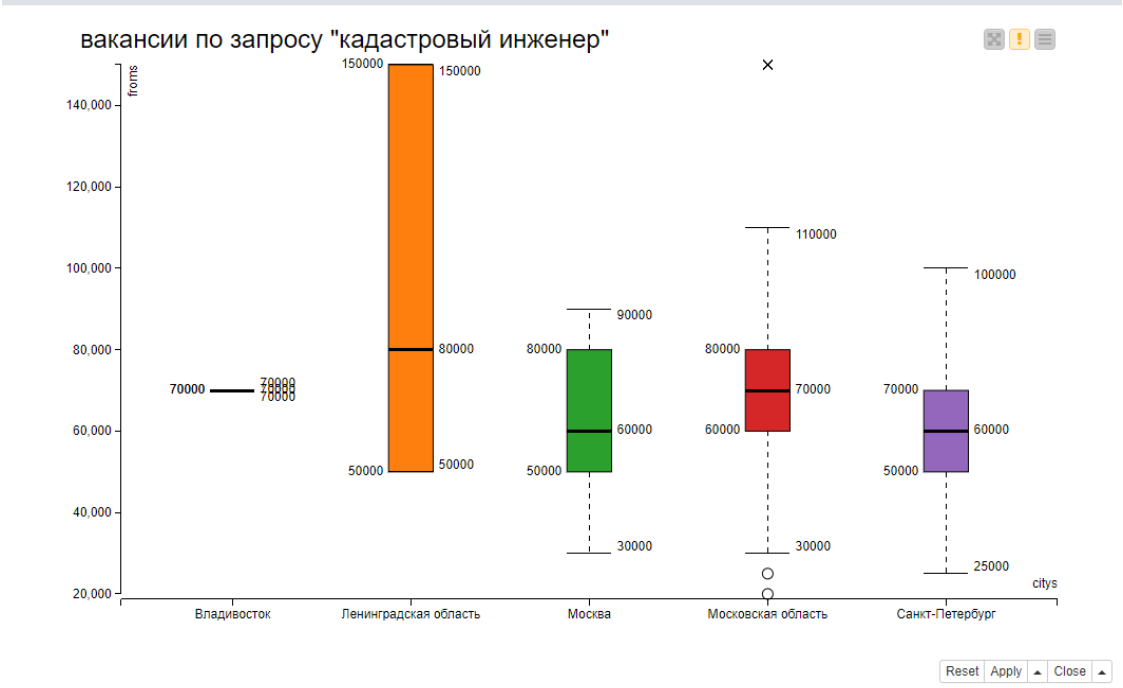


Рис. 6. – Диаграмма Conditional Box Plot (ящиковая диаграмма) по запросу «кадастровый инженер»

Следует отметить, что данные сводной таблицы постоянно меняются, вакансии могут добавляться и исчезать. Приведенные тут данные относятся к 10-20 июля 2021 года.

Таблица 1.

Сводная таблица запроса количества вакансий по 4 направлениям подготовки.

	направление			
регион	08.05.01 Строительство уникальных зданий и сооружений	20.04.01 Техносферная безопасность	27.03.01 Стандартизация и метрология	21.03.02 Землеустройство и кадастры
Владивосток	4	8	2	1
Владивостокская область	31	15	3	0
Калининград	0	0	1	0
Калининградская область	1	1	1	1
Москва	56	62	26	32
Московская область	25	38	22	22
Санкт-Петербург	28	37	23	18
Ленинградская область	23	21	4	3
Всего	168	182	82	77

Анализ вакансий с сайта hh.ru проходил по следующей схеме (см. Рис.7), которая является продолжением схемы, изображённой на рисунке 1 (Рис.1), входной таблицей является итоговая таблица по всем исследуемым регионам от которой строилась «ящиковая диаграмма».

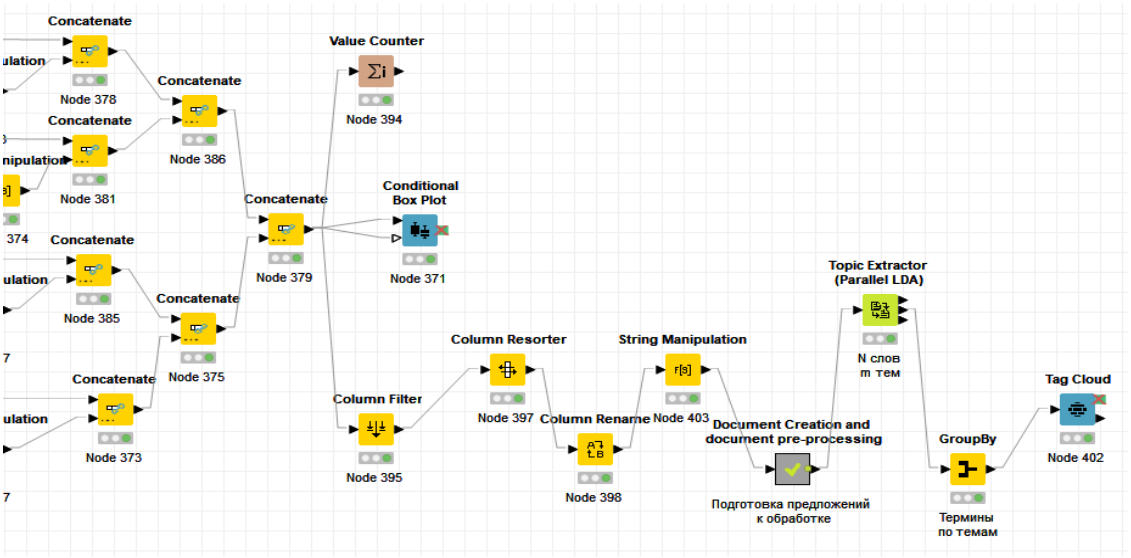


Рис. 7. – Схема анализа вакансий Knime.

Первые четыре узла обеспечивали предварительную подготовку данных для удобства дальнейшей обработки.

1. С помощью узла Column Filter оставляем id вакансий их название и требования для них из общей таблицы.

2. Для удобства работы с таблицей используются узы пересортировка колонок - Column Resorter, а затем их переименование Column Rename. В результате мы получили таблицу из трех столбцов, однако в тексте колонки «предложения» оказалось служебное слово для подсветки текста вакансий «<highlighttext>», чтобы его удалить, пришлось использовать узел String Manipulation – с помощью которого извлекли лишний термин по команде replaceChars($предложения$,"<highlighttext>",""). Таким образом, данные были подготовлены к дальнейшей обработке с помощью составного узла «Document Creation and document pre-processing».

3. «Document Creation and document pre-processing» - узел подготавливает извлечённые предложения для обработки и поиска часто встречающихся терминов. Эта процедура состоит из ряда узлов, последовательно преобразующих документ (см. рисунок 8).

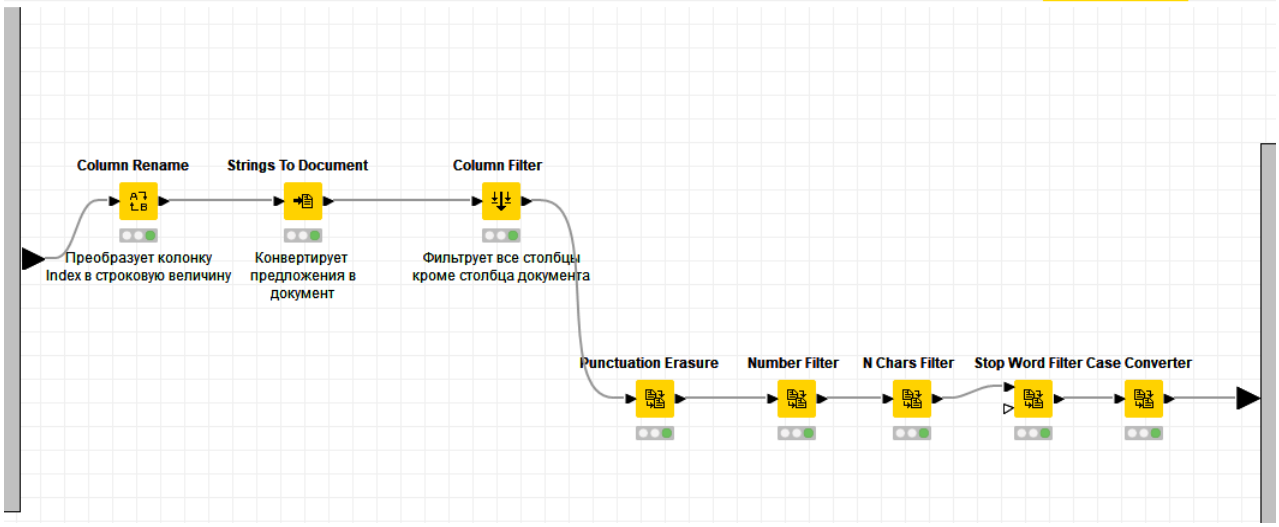


Рис. 8. – «Document Creation and document pre-processing»

3.1. Узел Column Rename превращает целые значения колонки индекс в строковые переменные.

3.2. String to Document – конвертирует строковые предложения, извлеченные из рабочей программы в документ.

3.3. Column Filter – фильтрует все столбцы, кроме столбца документа, полученного из предыдущего действия.

Следующие узлы позволяют избавиться от служебных символов пунктуации и т.д., чтобы не обрабатывать их в дальнейшем, фильтруют все термины в документе, которые указаны в списке стоп-слов. В Knime предусмотрены списки стоп-слов для разных языков, в нашем случае нашёлся русский, заменяют заглавные буквы.

4. Подготовленная таблица, каждая строка которой является документом, поступает на обработку узлом «Topic Extractor (Parallel LDA)», для поиска терминов и тем.

Модель машинного обучения (Латентное распределение Дирихле) LDA [7] служит для обнаружения тем в группе документов, и пытается их классифицировать по обнаруженным темам. Темы выдаются в виде списка слов или терминов, взятых вместе (то есть отдельных слов или фраз) которые, предполагают общую тему. LDA является развитием более старого алгоритма вероятностного латентного семантического анализа (pLSA). При равномерном априорном распределении Дирихле модель pLSA работает, как LDA. Чтобы воспользоваться этими методами пользователю необходимо указать количество тем, которые будут обнаружены до начала обучения (как и в случае кластеризации К-средних) [6].

преимущества LDA перед pLSA:

- более точное распределение документов по темам и лучшее устранение неоднозначности слов для LDA.

- позволяет осуществлять "генеративный" процесс при вычислении вероятностей, с помощью которого может быть сгенерирована коллекция новых “синтетических документов”, которые будут точно отражать

статистические характеристики исходной коллекции [8,9].

Узел использует библиотеку моделирования тем "MALLET: инструментарий машинного обучения для языка".

В узле можно варьировать количество тем и количество слов в теме. На выходе мы получали три таблицы: 1) Таблица документов с названиями тем и вероятностью принадлежности каждого документа к определенной теме, 2) Термины тем и их весов для каждой темы, 3) Таблица со статистикой для каждой итерации.

6. Следующий узел GroupBy – группирует строки таблицы по темам в выбранных столбцах группы. Строки создаются из уникального набора слов темы. Остальные столбцы агрегируются на основе заданных параметров агрегирования [10]. В качестве метода агрегирования выбираем «Concatenate» колонку терминов. Выходная таблица содержит одну строку для каждой уникальной комбинации значений столбцов темы. В результате получаем таблицу терминов по темам (Рис.9). Визуализацию результата можно провести с помощью узла «облако тегов» Tag Cloud.

Group table - 0:406 - GroupBy (Термины)

File Edit Hilite Navigation View

Table "default" - Rows: 8 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	S Topic id	S Concatenate(Term)
Row0	topic_0	объекты, работ, рентабельность, рассчитывать, уход
Row1	topic_1	объектов, документации, объекты, руководство, разработка
Row2	topic_2	документации, работ, составление, проверка, объемов
Row3	topic_3	работ, контроль, организация, качества, объектах
Row4	topic_4	контроль, работы, проекта, задач, поиск
Row5	topic_5	расчет, работ, подготовка, подрядчика, стоимости
Row6	topic_6	архитекторами, систем, взаимодействие, строителями, проектировщиками
Row7	topic_7	выполнение, систем, конструкций, составление, проведение

Рис. 9. – таблица терминов по темам (для заданных 8 тем по 5 слов)

Исследование проводились путем изменения количества тем и количества слов в узле Topic Extractor (Parallel LDA). Анализ зависимости количества тем и терминов показал, что увеличение числа тем приводит к возникновению повторов в терминах, и в окончательной таблице достаточно оставить по 4 или 5 терминов на 8 тем.

Таблица № 2

Сводная таблица анализа программ. (количество тем 8, терминов 4)

Направление подготовки/специальность			
08.05.01	21.03.02	27.03.01	20.04.01
объекты, работ, рентабельность, рассчитывать	сооружений, составление, помещений, проведение	учета, контроль, приборов, компании	проведению, осмотров, принимать, регионах
объектов, документации, объекты, руководство	органами, проектов, межевания, взаимодействие	документации, услуг, продаж, оборудования	организация, производственных, суот, инструктажи
документации, работ, составление, проверка	документации, контроль, проекта, готовить	работы, электронные, электрические, техника	условий, труда, охраны, мероприятий
работ, контроль, организация, качества	кадастровых, егрн, работа, изменений	ema-рассылок, плановых, цепочками, триггерными	безопасности, контроль, охраны, работ
контроль, работы, проекта, задач	планов, подготовка, межевых, участков	контроль, деталей, проектов, ведение	труда, проведение, охране, инструктажей
расчет, работ, подготовка, подрядчика	кадастровых, работ, участие, проектных	метрологической, проведение, документации, измерений	безопасности, труда, пожарной, промышленной
архитекторами, систем, взаимодействие, строителями	составление, результатам, отчетов, работы	измерений, средств, поверки, организация	профессиональных, строительства, случаев, несчастных
выполнение, систем, конструкций, составление	кадастровый, учет, документов, технических	средств, измерений, калибровка, поверка	труда, обеспечение, управления, системы

Таблица № 3

Сводная таблица анализа программ. (количество тем 8, терминов 5)

Направление подготовки/специальность			
1	2	3	4
08.05.01	21.03.02	27.03.01	20.04.01
объекты, работ, рентабельность, рассчитывать, уход	сооружений, составление, помещений, проведение, технических	учета, контроль, приборов, компании, работать	проведению, осмотров, принимать, регионах, медицинских
объектов, документации, объекты, руководство, разработка	органами, проектов, межевания, взаимодействие, территории	документации, услуг, продаж, оборудования, клиентами	организация, производственных, суот, инструктажи, направлению
документации, работ, составление, проверка, объемов	документации, контроль, проекта, готовить, рабочей	работы, электронные, - электрические, техника, кабельная	условий, труда, охраны, мероприятий, безопасных
работ, контроль, организация, качества, объектах	кадастровых, егрн, работа, изменений, инженеров	ema-рассылок, плановых, цепочками, триггерными, прогревающими	безопасности, контроль, охраны, работ, сфере
контроль, работы, проекта, задач, поиск	планов, подготовка, межевых, участков, земельных	контроль, деталей, проектов, ведение, покрытий	труда, проведение, охране, инструктажей, участие
расчет, работ, подготовка, подрядчика, стоимости	кадастровых, работ, участие, проектных, документооборота	метрологической, проведение, документации, измерений, экспертизы	безопасности, труда, пожарной, промышленной, охраны
архитекторами, систем, взаимодействие, строителями, проектировщиками	составление, результатам, отчетов, работы, кадастровые	измерений, средств, поверки, организация, проведение	профессиональных, строительства, случаев, несчастных, выполнением

1	2	3	4
выполнение, систем, конструкций, составление, проведение	кадастровый, учет, документов, технических, объектов	средств, измерений, калибровка, поверка, величин	труда, обеспечение, управления, системы, охраной

В итоговых таблицах присутствуют термин, который не определён – «суот». Однако, в целом, можно сделать вывод, что использование метода анализа на основе Knime – успешно. Полученные термины можно использовать для создания устойчивых словосочетаний, отражающих требования работодателей. Следует отметить, что данная схема позволяет отслеживать и динамику их изменения. Результаты исследования предполагается в дальнейшем использовать для корректировки рабочих программ строительных специальностей после их соответствующего анализа.

Литература

1. Егорова Е.М. Теоретические основы цифровизации в профессиональном образовании // Вопросы педагогики. 2020. № 6-1. С.100-109. DOI 10.37882/2223-2982.2020.07.10.

2. Попова О.И. Трансформация высшего образования в условиях цифровой экономики // Вопросы управления. 2018. №5(35). С. 158-160.

3. Гриненко С.В. Маркетинговые исследования как инструмент мониторинга миграции выпускников системы профессионального образования // Инженерный вестник Дона, 2012, №2. URL: ivdon.ru/ru/magazine/archive/n2y2012/761.

4. Диков М.Е., Широбокова С.Н. О варианте формализации задачи определения востребованности направлений подготовки и возможных сфер трудоустройства выпускников на основе семантического анализа описаний вакансий // Инженерный вестник Дона, 2022, №5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7631.

5. Какутин Д.Ю., Дмитриев А.С. Формирование и анализ эффективности выборки для обучения языковых моделей распознаванию и анализу исходного кода программ // Инженерный вестник Дона, 2022, №5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7682.

6. Ломакина Л.С., Суркова А.С. Прикладные аспекты концептуального анализа и моделирования текстовых структур // Фундаментальные исследования, 2015. №7-3. С. 540-544. URL: fundamental-research.ru/ru/article/view?id=38775.

7. Blei D., Ng A., and Jordan M.. Latent Dirichlet allocation. // Journal of Machine Learning Research, 2003, vol. 3, pp. 993-1022.

8. Griffiths T. and Steyvers M.. Finding scientific topics. // In Proceedings of the National Academy of Sciences, 2004, vol. 101, pp. 5228-5235, DOI:10.1073 /PNAS.0307752101.

9. Teh Y. W., Jordan M. I., Beal M. J., and Blei D. M.. Hierarchical Dirichlet processes. // Journal of the American Statistical Association, 2006. №101(476), pp. 1566-1581. URL: jstor.org/stable/27639773.

10. Maier D., Niekler A., Wiedemann G., Stoltenberg D.. How Document Sampling and Vocabulary Pruning Affect the Results of Topic Models // Computational Communication Research, Volume 2, Issue 2, Oct 2020, p. 139-152, DOI: 10.5117/CCR2020.2.001.MAIE.

References

1. Egorova E. M. Voprosy pedagogiki [Questions of pedagogy]. 2020. №6-1. pp. 100-109. [in Russian], DOI 10.37882/2223-2982.2020.07.10.

2. Popova O.I. Voprosy upravleniya [Management issues]. 2018. №5(35). pp. 158-160.

3. Grinenko S.V.. Inzhenernyj vestnik Dona, 2012, №2. URL: ivdon.ru/ru/magazine/archive/n2y2012/761.

4. Dikov M.E., Shirobokova S.N.. Inzhenernyj vestnik Dona, 2022, №5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7631.

5. Kakutin D.YU., Dmitriyev A.S.. Inzhenernyj vestnik Dona, 2022, №5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7682.

6. Lomakina L.S., Surkova A.S.. Fundamental′nyye issledovaniya [fundamental-research], 2015. №7-3. pp. 540-544; URL: fundamental-research.ru/ru/article/view?id=38775.

7. Blei D., Ng A., and Jordan M.. Journal of Machine Learning Research, 2003, vol. 3, pp. 993–1022.

8. Griffiths T. and Steyvers M.. Finding scientific topics. In Proceedings of the National Academy of Sciences, 2004, vol. 101, pp. 5228–5235, DOI:10.1073 /PNAS.0307752101.

9. Teh Y. W., Jordan M. I., Beal M. J., and Blei D. M.. Journal of the American Statistical Association, 2006. №101(476), pp. 1566–1581, URL: jstor.org/stable/27639773.

10. Maier D., Niekler A., Wiedemann G., Stoltenberg D.. Computational Communication Research, Volume 2, Issue 2, Oct 2020, pp. 139-152. DOI: 10.5117/CCR2020.2.001.MAIE.