Секция «Психология труда и инженерная психология»

Кластеризация профессии аналитика в ИТ на основе вакансий, представленных на платформе HeadHunter

Научный руководитель – Савельев Владимир Вадимович

Лалетина Ю.В.[1], *Иванов И.С.*[2]

1 - Уральский федеральный университет имени первого Президента России Б.Н.Ельцина, Уральский гуманитарный институт, Екатеринбург, Россия, *E-mail: laletinajulia@list.ru*; 2 - Уральский федеральный университет имени первого Президента России Б.Н.Ельцина, Уральский гуманитарный институт, Екатеринбург, Россия, *E-mail: vrangelorient@yandex.ru*

В последнее время профессия аналитика стала очень популярна. При этом названия, которые содержат слово «аналитик», могут звучать совершенно по-разному. Более того, в каждой компании под словом «аналитик» могут иметься в виду разные наборы компетенций. В настоящее время нет четко сформированных границ между профессиями системного аналитика, бизнес-аналитика, продуктового аналитика и т.д. В своей работе мы предприняли попытку кластеризовать реальные актуальные данные по вакансиям аналитика в ИТ, чтобы выявить основные группы компетенций на основе статистического метода LDA.

Существуют исследования [2, 3], которые применяют алгоритм LDA для кластеризации профессий, но они основывались на данных должностных инструкций. Но проблема данного подхода заключается в несоответствии между традиционными описаниями обязанностей работника и реальными требованиями работодателей. Чтобы преодолеть этот разрыв, мы провели тематическое моделирование основываясь не на текстах нормативных документов, а на текстах вакансий с сайта HeadHunter, актуальных на конец февраля 2021 года по России, в названии которых есть слово «аналитик». Мы выбрали для анализа профессию аналитика в сфере ИТ, так как она требует разнообразных компетенций: как хороших коммуникативных навыков, так и знаний основ в сферах разработки и маркетинга, понимания бизнес-процессов. Для разных направлений аналитики преобладающими будут разные компетенции.

В исследовании мы воспользовались языком Python и его библиотеками: requests, scikit-learn, nltk и др. Мы собрали данные, обработали их и применили на подготовленных данных латентное размещение Дирихле, или LDA, чтобы автоматически выделить основные тематические блоки. Тематическое моделирование выполняет задачи кластеризации текстов таким образом, что каждый кластер содержит в себе тексты со схожими темами.

Наш алгоритм анализа текстов вакансий:

1. Получили данные с сайта Headhunter через API посредством библиотек requests и pandas.

2. Предобработали тексты:

- удалили слова на основе библиотеки nltk, не указывающих на конкретные элементы деятельности. К ним относятся наиболее частотные «стоп-слова»: предлоги, союзы, частицы, числительные, местоимения и другие;

- провели лемматизацию на основе библиотеки pymorphy2, то есть привели слова к начальной (словарной) форме.

3. Векторизовали слова на основе библиотеки scikit-learn, используя схему взвешивания TF-IDF, которая позволяет определить наиболее важные слова в коллекции документов.

4. Применили метод LDA на основе библиотеки scikit-learn на предобработанном тексте, после которого мы получили 8 кластеров.

5. Интерпретировали полученные данные.

В результате мы получили 8 кластеров, которые достаточно четко отличаются друг от друга. Мы охарактеризовали их таким образом:

1. Кластер «Продуктовый анализ». Пример содержания кластера: рынок, гипотеза, метрика, анализировать и др.

Нахождение зависимостей, работа с гипотезами, исследование рынка, продукта. Такой тип вакансий мы можем охарактеризовать как «продуктовый аналитик».

2. Кластер «Визуализация данных, отчеты». Пример содержания: power, bi, визуализация, расчёт, таблица. Компетенции данного кластера можно выделить как в отдельную профессию визуализатора данных, так и в важную компетенцию аналитика, работающего с данными.

3. Кластер «Базовые качества аналитика». Пример содержания: мышление, речь, готовность, склад, ум. Данный кластер можно охарактеризовать как «тест на адекватность» кандидата.

4. Кластер «Маркетинг, клиентский сервис». Пример содержания: клиентский, коммерческий, рекламный, маркетинговый, клиент, продажа.

5. Кластер «Системный анализ». Пример содержания: коммуникация, модель, проектирование, постановка, разработчик, взаимодействие.

6. Кластер «Взаимодействие со стейкхолдерами». Пример содержания: обследование, интеграция, уточнение, функциональный, реализация, согласование, ТЗ, доработка, внедрение.

7. Кластер «Soft skills/фасилитация». Пример содержания: сотрудник, согласование, организация, роль, презентация, английский, проектный, управление, обучение, проведение. Внутри данного кластера нам видится смешение soft skills как навыков, необходимых для любой вакансии, так и компетенций роли человека в команде, который занимается работой с командой.

8. Кластер «Менеджер продукта/проекта». Пример содержания: постановка, внутренний, внешний, архитектура, цикл, систематизация, запрос, выявление, задание.

Мы выполнили поставленную задачу. У нас получилось выявить четкие кластеры и тем самым приблизиться к более объективному разделению профессии аналитика на подразделы, приближенному к реальным требованиям рынка труда. Интересным оказалось то, что в нашей подборке слабо представлены конкретные навыки, а основу составляют размытые понятия. Из этого делаем вывод, что на рынке нет строгого единства даже в рамках кластеров, которые были выявлены в ходе нашего исследования.

Далее мы планируем классифицировать вакансии по их близости к выявленным кластерам. Практическая значимость исследования заключается в том, что с помощью данного алгоритма легко можно воспроизвести наше исследование на любых других профессиях либо продолжить его в динамике.

Источники и литература

1) Koltcov S., Koltsova O., Nikolenko S. Latent dirichlet allocation: stability and applications to studies of user-generated content //Proceedings of the 2014 ACM conference on Web science. – 2014. – С. 161-165.

2) Савельев В.В. Профессия как вектор в n-мерном пространстве элементов деятельности // Психологический журнал. 2018. Т. 39. № 5. С. 37-45.

3) Савельев В.В. Применение методов тематического моделирования к профессиональным стандартам для реализации модульного подхода к классификации профессий // Педагогическое образование в России. 2017. № 3. С. 22-28.