Сравнение методов векторизации и кластеризации вакансий

Д. А. Фомичев

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

savior.7@yandex.ru

А. А. Кочешков

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

kocheshkovgia@yandex.ru

***Аннотация*. В данной работе рассматриваются методы векторизации вакансий на основе их описаний (или ключевых навыков), а также методы кластеризации на основе полученных векторных представлений текстов. Используются такие методы векторного представления текста, как tf-idf матрица, doc2vec и построение эмбеддингов текста на основе трансформеров. Представленная работа являет продолжением предыдущих исследований.**

Ключевые слова: векторизация текста, кластеризация

I. Введение

Регулярный прогресс в современном мире несет за собой появление новых, уникальный профессий на рынке труда. Любая профессия предполагает наличие определенных навыков и умений – пререквизитов. Для того, чтобы иметь представление о текущей ситуации на рынке труда, необходимо понимать, какие именно пререквизиты стоят за той или иной профессией и какой круг вакансий доступен соискателю с его бэкграундом. Таким образом, целью данной работы является формирование кластеров схожих между собой вакансий. Это, в свою очередь, подразумевает наличие тех или иных навыков, объединяющих конкретные вакансии. Кроме того, такой анализ в совокупности с сопоставлением учебных программ и сформированных кластеров способен сформировать учебную траекторию в ВУЗах для каждого студента и помочь ему определиться с дальнейшей профессией. В данном исследование проводится сравнение таких методов векторизации как TF-IDF, doc2vec, использование трансформеров, а также таких алгоритмов кластеризации как K-Means, DBSCAN и его оптимизированную версию HDBSCAN. Сравнение проводится на подготовленном наборе данных, в его основе лежит получение индекса Рэнда [3] и других метрик для валидации кластеризации.

В отличии от предыдущей работы [1] в текущем исследование также использованы методы выделения ключевых навыков [2] из описания вакансий для дальнейшей векторизации конкретных ключевых понятий, полученных из описания.

II. Набор данных

Для исследования и сравнения методов был подготовлен набор данных на основе того, что использовался в предыдущей работе [1]. Важным отличие является то, что в текущей работе использовался проаннотированный вручную набор данных с выделением принадлежности вакансии к определенной категории, среди которых: *Мобильный разработчик, Аналитик, Инженер-данных, Backend-разработчик, Fullstack-разработчик, Системный администратор, GameDev, Разработчик САУ, Дизайнер, Frontend-разработчик, Специалист технической поддержки, Специалист информационной безопасности, ML-разработчик, Тестировщик, Менеджер продукта.*

Проаннотированный набор поможет осуществить субъективную оценку качества кластеризации векторов на основе метрик кластеризации, одна из которых индекс Рэнда [3].

Для оценки работы метода с реальными данными также используется набор, содержащий 108.000 непроаннотированных вакансий с портала «Работа России» [4].

III. Ход решения

A. *Предобработка данных*

В первую очередь необходимо осуществить предобработку данных. Как и в случае с предыдущим исследованием были проделаны следующие операции: удаление тегов, удаление стоп-слов, лемматизация.

Кроме того, как упоминалось ранее, в данной работе применялись модели для извлечения ключевых навыков из описаний вакансий. После применения данный нейросетевых методов производилась постобработка полученных навыков с целью очищения от лишних токенов и нормализации полученной строки.

Пример обработанной строки ключевых навыков: *kubernetes дистрибуция десктопный софт массовый онлайн сервис python django flask mysql оптимизация запрос django rest framework unittest pytest mock mongodb redis influxdb jenkins kubernetes технически грамотный github bitbucket*

B. *Векторизация и кластеризация*

Из полученных строк ключевых навыков формировались векторы на основе исследуемых методов векторизации. При кластеризации использовались два алгоритма: DBSCAN и K-Means.

Для алгоритма K-Means оптимальное значение кластеров находилось с помощью «метода локтя» и силуэта кластера [5–6]. Для алгоритма DBCSAN

необходимо было подобрать параметры epsilon и min_samples, чтобы добиться наилучшего распределения векторов на кластеры. Для поиска оптимальных параметров проводились эксперименты для получения максимального значения индекса Рэнда от сочетания вышеупомянутых параметров.

Далее представлены примеры результатов кластеризации алгоритмами K-Means и DBSCAN для исследуемых методов векторизации:

- TF-IDF (K-means)

На рис. 1 представлено разбиение кластеров с помощью K-Means для tf-idf (оптимальное количество кластеров по методу локтя – 15) Средний коэффициент силуэта составляет 0.336.

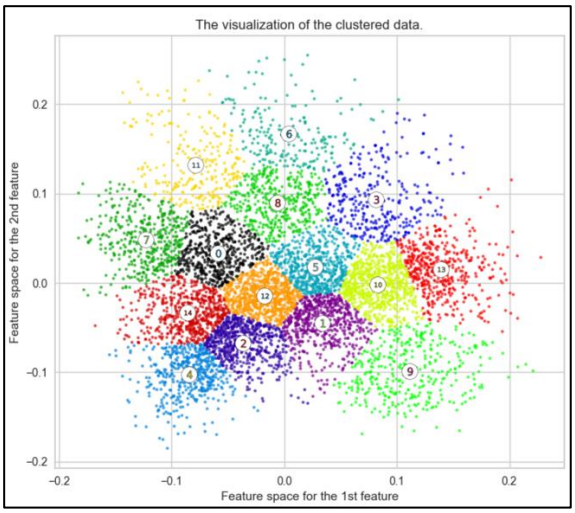

Рис. 1. Разбиение кластеров для TF-IDF метода K-Means алгоритмом

- Doc2vec (DBSCAN)

На рис. 2 представлено разбиение кластеров с помощью DBSSCAN для doc2vec.Средний коэффициент силуэта составляет -0.1.

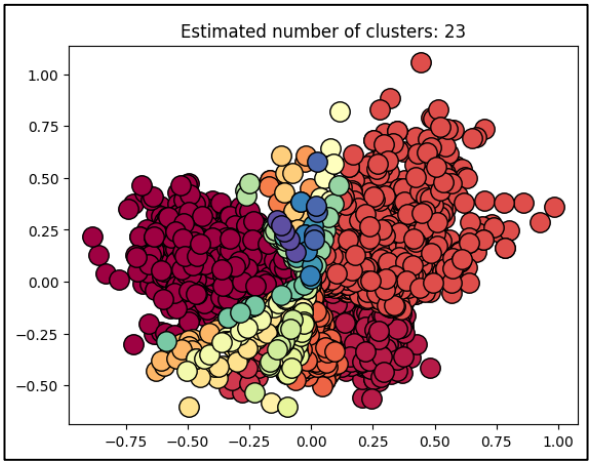

Рис. 2. Разбиение кластеров для doc2vec метода DBSCAN алгоритмом

При кластеризации также использовался метод главных компонент [7] с целью уменьшения размерности полученных векторов и визуализации в двумерном пространстве.

C. Оценка и результаты

Для оценки полученной кластеризации использованы следующие метрики: Скорректированный индекс Рэнда [3], скорректированная взаимная информация, метрика однородности, метрика полноты маркировки.

Результаты оценки приведены в табл. 1.

ТАБЛИЦА I.

Алгоритм	K-Means			DBSCAN		
Метод векторизации	tfidf	doc2vec	transformers	tfidf	doc2vec	transformers
Adjusted rand score	0.13	0.13	0.04	0.16	0.14	0.04
Adjusted mutual info score	0.29	0.24	0.08	0.20	0.19	0.05
Homogeneity score	0.31	0.26	0.09	0.23	0.20	0.06
Completeness score	0.30	0.25	0.09	0.24	0.24	0.08

Оптимальное значение *epsilon* для DBSCAN алгоритма колеблется в пределах $[10^{-5};10^{-4}]$ при *min_samples* от 2 до 10 в зависимости от метода векторизации.

Исходя из полученных результатов можно сделать вывод о том, что самым оптимальным способом разбить вакансии на кластеры является сочетание tf-idf и алгоритма K-Means. Однако, важно понимать, что оценка производилась на размеченном вручную наборе данных и сильно субъективна, в том числе по причине того, что заранее создаются рамки (определенное число предложенных вакансий), из-за которых результаты валидации сильно зависят от количество выходных уникальных меток. Таким образом, сочетание методов, которые разбивают вакансии на узконаправленные кластеры, становятся обреченными на низкую оценку. Кластеризация эмбеддингов, полученных при помощи трансформеров, на обработанном наборе данных не дает высоких результатов, потому далее рассмотрим другой подход с применением эмбеддингов.

IV. Ход решение без использования предобработки текстов

В данном разделе пойдет речь о методе векторизации, основанном на извлечении эмбеддингов с использованием модели RuBERT-tiny2. В рассматриваемом подходе текстовые данные не обрабатываются перед использованием, в отличие от ранее рассмотренного подхода.

A. Предобработка данных

Использование трансформеров в задачах обработки текста существенно снижает необходимость в предварительной обработке данных. В необработанных текстах трансформеры способны извлекать смысловые элементы и абстрагироваться от шумовых компонентов.

В рамках данного метода также применялись модели для извлечения ключевых навыков из описаний вакансий. Извлеченные ключевые навыки добавлялись к текстам вакансий с целью более точной кластеризации.

B. Векторизация и кластеризация

Преобразуем исходные строки в векторные представления, используя модель RuBERT-tiny2 [8]. После преобразования, размерность полученных векторов слишком высокая для проведения кластеризации. Для снижения размерности данных был использован метод UMAP. Этот метод позволяет сохранить важные структурные характеристики данных при снижении их размерности.

Для стимулирования кластеризации к определенным темам применяется полу-контролируемое тематическое моделирование (Semi-supervised Topic Modeling [9]). Этот метод позволяет указать метки только для части документов. Эти метки затем используются для направления кластеризации в соответствии с темами, к которым относятся эти документы. Для данной работы были выбраны 15 категорий вакансий, из каждой категории взято по десять вакансий, суммарно 5 % от всех вакансий. Метки, соответствующие этим категориям, используются в полу-контролируемом тематическом моделировании. Результаты экспериментов показывают рост индекса Рэнда, при использовании данного метода.

При использовании HDBSCAN возможно появление документов, которые не попадают в какую-либо из созданных тем. Для уменьшения количества выбросов применяется метод, который заключается в вычислении представлений c-TF-IDF для этих выбросов и присвоении им наиболее подходящих тематических представлений c-TF-IDF [10].

В процессе кластеризации были применены два алгоритма: HDBSCAN и K-Means.

- RuBERT-tiny2 (K- Means)

На рис. 3 представлены кластеризованные эмбеддинги методом K-Means.

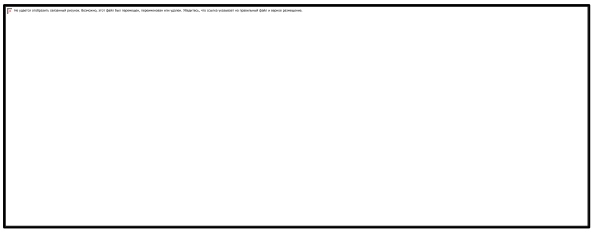

Рис. 3. Кластеризованные эмбеддинги методом K-Means

- RuBERT-tiny2 (HDBSCAN)

На рис. 4 представлены кластеризованные эмбеддинги методом HDBSCAN до уменьшения числа выбросов.

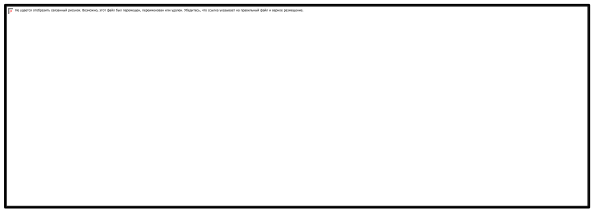

Рис. 4. Кластеризованные эмбеддинги методом HDBSCAN до уменьшения числа выбросов

На рис. 5 представлены кластеризованные эмбеддинги методом HDBSCAN после уменьшения числа выбросов.

Рис. 5. Кластеризованные эмбеддинги методом HDBSCAN после уменьшения числа выбросов

Для визуализации вектора понижали размерность с помощью метода UMAP.

C. Оценка и результаты

Для оценки полученной кластеризации, также как и в прошлом методе, использованы следующие метрики: Скорректированный индекс Рэнда, скорректированная взаимная информация, метрика однородности, метрика полноты маркировки. Результаты оценки приведены в табл. 2.

ТАБЛИЦА II.

Алгоритм	K-Means	HDBSCAN	
Semi-supervised Topic Modeling	Есть	Нет	Есть
Adjusted rand score	0.13	0.13	0.18
Adjusted mutual info score	0.17	0.20	0.22
Homogeneity score	0.17	0.19	0.21
Completeness score	0.19	0.23	0.25

По результатам исследования можно сделать вывод, что наиболее эффективным методом кластеризации векторных представлений вакансий (полученных с использованием RuBERT-tiny2) является HDBSCAN.

Однако следует отметить, что HDBSCAN имеет проблему выбросов, составляющих 8 % от всего набора данных. Эти выбросы требуют удаления или обработки, так как их присутствие может быть связано с субъективностью разметки набора данных или наличием вакансий, которые трудно отнести к существующим кластерам.

Без предварительной обработки текста при использовании эмбеддингов возникает проблема, когда вакансии, написанные на разных языках, могут быть разбиты на отдельные кластеры в зависимости от языка, и при этом не учитывать требования к вакансии. На рисунках 4-6 виден кластер, который наиболее удален от остальных; он содержит вакансии на английском языке, в отличие от остальных вакансий на русском.

V. Обработка набора необработанных данных

В данном разделе рассмотрим процесс векторизации и кластеризации набора данных, состоящего из 108 тысяч вакансий.

A. Предобработка данных

Учитывая достигнутую эффективность использования модели RuBERT-tiny2 без удаления тегов, стоп-слов и лемматизации, данный подход рассматривается как предпочтительный. В данном решении не используется модель, извлекающая ключевые навыки, и вместо этого используется только текстовое описание вакансии.

B. Векторизация и кластеризация

Исходные строки преобразованы в векторные представления с использованием модели RuBERT-tiny2. Для снижения размерности данных был применен метод UMAP.

В данном наборе данных нецелесообразно применять полу-контролируемое тематическое моделирование, поскольку размер датасета слишком велик. Для получения значимых результатов необходимо было бы проаннотировать большую часть вакансий, предварительно выделив категории.

Поскольку алгоритм HDBSCAN продемонстрировал более высокую эффективность по сравнению с K-Means и не требует задания числа кластеров перед началом работы, что является важным в контексте такого объемного набора данных, выбор делается в пользу HDBSCAN.

Для сокращения числа выбросов используется метод, который заключается в вычислении c-TF-IDF представлений для этих выбросов и их соотнесении с наиболее подходящими кластерами. Данный метод позволил сократить число выбросов от 58 тысяч (53 %) до 17 тысяч вакансий (16 %).

C. Оценка и результаты

В данном наборе данных отсутствует возможность расчета метрик для оценки, однако возможна субъективная оценка. Кластеры содержат вакансии, собранные на основе требований. Например, один из кластеров содержит 447 вакансий для Frontend-разработчика, а другой – 57 вакансий для музыкального педагога. Однако продолжаются проблемы, связанные с иностранными языками, выделяющимися в отдельные кластеры. Также в данном наборе данных, где заранее не определены категории вакансий, возникают кластеры, связанные с конкретными работодателями или условиями работы, а не с требованиями к кандидатам.

VI. Заключение

В ходе работы рассмотрены сочетание методов векторизации текстовых данных и алгоритмов кластеризации для извлеченный из описания вакансий информации. Исследование проводилось как на подготовленном наборе данных, так и на «сыром». Можно отметить, что в экспериментах с предобработанным набором данных мы получаем невысокий коэффициент Рэнда, особенно при применении DBSCAN алгоритма, где в предустановке нет количества итоговых кластеров. Одна из причин этому – предначертание модели желаемых кластеров. Использование необработанных данных позволяет модели действовать более широко и формировать более узкие кластеры вакансий, однако оценка становится ещё более субъективной.

Дальнейшая работа по улучшению качества кластеризации может быть связана оптимизацией метода с использованием готовых моделей и HDBSCAN алгоритма с целью уменшения числа выбросов, решению вопроса с иностранным описанием и проблемы формирования кластеров, относящихся к работодателю, а не к самой профессии. Кроме того, возможно исследованием с применением LLM моделей посредством формирования промт-запросов и последующей обработки полученной информации.

Итоговое решение будет внедрено в ИС «Индивидуальные образовательные траектории» для регулярного обновление текущего состояния рынка труда.

Список литературы

[1] Фомичев Д.А. Кластеризация вакансий по их описанию с использованием машинного обучения и методов анализа текста. СПбГЭТУ «ЛЭТИ», Санкт-Петербург

[2] Корытов П.В., Грибецкий Я.Ю., Андреева Е.А., Холод И.И. Анализ подходов к извлечению ключевых навыков из вакансий. СПбГЭТУ «ЛЭТИ», Санкт-Петербург

[3] Adjusted Rand Score [Электронный ресурс] URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rand_score.html (Дата обращения: 01.03.2024)

[4] Работа России. Открытые данные [Электронный ресурс]. URL: https://trudvsem.ru/opendata/ (дата обращения: 04.03.2024)

[5] KMeans Silhouette analysis, URL: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html (Дата обращения 05.03.2024)

[6] Elbow Method [Электронный ресурс] URL: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html (Дата обращения 05.03.2024)

[7] Principal component analysis [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (Дата обращения 07.03.2024)

[8] RuBERT-tiny2 [Электронный ресур]. URL: https://huggingface.co/cointegrated/rubert-tiny2 (Дата обращения 02.04.2024)

[9] Semi-supervised Topic Modeling [Электронный ресурс]. URL: https://maartengr.github.io/BERTopic/getting_started/semisupervised/semisupervised.html (Дата обращения: 01.04.2024)

[10] Outlier reduction [Электронный ресурс]. URL: https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html#exploration (Дата обращения: 01.04.2024)