Кластеризация вакансий по их описанию с использованием машинного обучения и методов анализа текста

Д. А. Фомичев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
savior.7@yandex.ru

***Аннотация*. В рамках исследования рассматриваются методы машинного обучения и методы анализа текста для кластеризации вакансий по их названию, ключевым навыкам и описанию. Описано применение таких методов, как doc2vec, tf-idf матрицы, трансформеров для векторизации описания, применение таких алгоритмов кластеризации, как DBSCAN, K-Means.**

Ключевые слова: обработка естественного языка, кластеризация

I. Введение

На различных порталах имеется огромное количество вакансий, сосредоточенных на различных навыках и умениях специалистов. На примере IT-сферы можно выделить WEB-разработчиков, тестировщиков, специалистов машинного обучения и анализа данных и другие профессии, объединенные в одну сферу, но имеющие разные пререквизиты – набор навыков и умений, покрывающих данную вакансию. Это существенно затрудняет как анализ востребованности профессий, так и работу с ними, например, в разработке образовательных программ и построении индивидуальных траекторий студентов. Определяя вакансии в кластеры по определенным признакам, можно в дальнейшем упростить решение связанных задач.

Для решения данной задачи необходимо проанализировать информацию, полученную из названия, ключевых навыков и описания вакансии для правильной кластеризации, а также для дальнейшего извлечения пререквизитов (технологий, методов, алгоритмов, понятий) внутри кластеров.

В основе решения лежит обработка естественного языка для извлечения информации из текстовых данных и использование ее в алгоритмах машинного обучения.

II. Способы решения

Прежде всего, необходима предварительная обработка текстовой информации. На основе данных о вакансиях можно применить различные методы обработки для очистки данных:

- Удаление тегов из описания.
- Удаление знаков препинания.
- Удаление стоп-слов из текста.
- Приведение слов к начальной форме (лемматизация).

После предобработки текстовых данных необходимо провести токенизацию – разбиение текста на отдельные токены (в случае решаемой задачи – слова) для дальнейшей векторизации полученных токенов. В процессе изучения предметной области были рассмотрены следующие способы векторизации текста:

- TF-IDF матрица [1] – метод, основанный на получении математической матрицы, описывающей частоту встречающихся терминов. В матрице столбцы соответствуют запросам, а строки терминам.
- Представление в векторной форме на основе doc2vec – инструмент для представления документов в виде вектора. Каждый абзац и каждое слово представляется в виде уникальных векторов. Вектор абзаца и векторы слова усредняются или объединяются, чтобы предсказать следующее слово в контексте.
- Представление в векторной форме на основе трансформеров [3] – языковые модели, обучаемые на большом количестве текстовых данных. Обычно состоят из двух элементов: элемент кодирования и декодирования. В данной задаче можно использовать элемент кодирования для трансформации текста в вектор. Существуют предобученные модели на разных языках, которые можно найти на Hugging Face Hub [4]. Данный нейросетевой метод способен решать больше задач, чем классические методы, а также чаще показывает лучшие результаты на бенчмарках.

III. Описание методов решения

Набор данных для анализа представляет собой информацию по опубликованным на порталах поиска работы вакансиям, принято решение использовать профессии в области IT-технологий для данного исследования. Для решения данной задачи рассмотрены следующие признаки: имя, ключевые навыки и описание вакансии. Можно выделить несколько основных этапов:

- Предобработка текстовой информации выбранных признаков.
- Векторизация текстовой информации и извлечение ключевых признаков.
- Кластеризация векторов и оценка.

Схема решения поставленной задачи представлена на рис. 1.

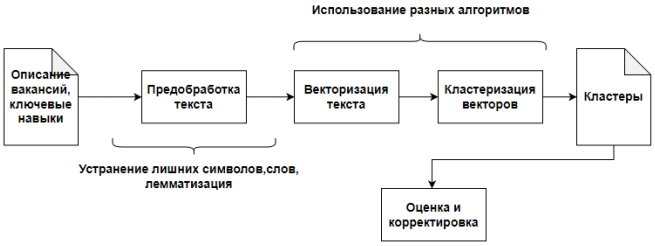

Рис. 1. Схема метода решения

A. *Предобработка текстовой информации*

Для максимального извлечения информации из текста, а также для того, чтобы очистить его от невалидных и некорректных данных, которые впоследствии навредят векторизации и получению ключевых признаков, необходимо провести предпроцессинг. В него входят:

- очистка данных от тегов (например, если это описание на порталах, описанное с помощью html-тегов);
- очистка текста от лишних символов (например, тире, скобки и прочие знаки, ненесущие полезности);
- очистка текста от геоданных (нередко в описаниях вакансий встречаются адреса, локализация места работы и прочее, что тоже не влияет на вид деятельности и ключевыми признаками профессии);
- применение лемматизации для приведения слов к начальной форме.

Далее производилась очистка текста от возможных дефектов (двойные пробелы, новые строки) и приведение к нижнему регистру.

B. *Векторизация текста*

Для дальнейшего анализа текстовой информации необходимо векторизовать полученный текст. В процессе исследования для этого были различные попытки. Применение инструмента doc2vec не дало высоких результатов, как показала практика, его использование для данной задачи помогало справляться только с очевидными различиями векторов (очевидными различиями между признаками профессии). Так, например, с его помощью можно было отличить вакансию «Инженер-программист C++» от вакансии «Тестировщик ПО», однако различить более узкие вакансии, имеющие одну область применения, не удавалось («Инженер-программист C++» – «Программист микроконтроллеров»). На рис. 2 можно увидеть некачественное распределение кластеров с использованием doc2vec при текущей обработке данных.

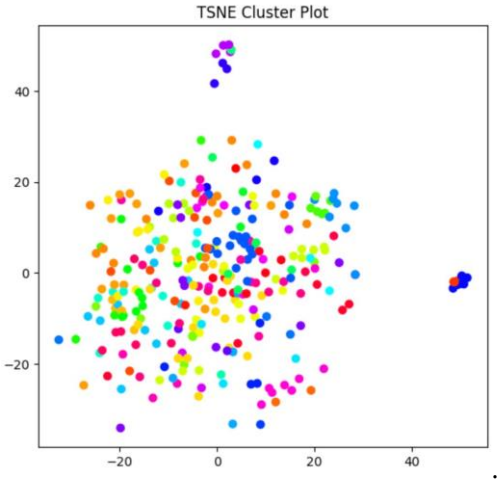

Рис. 2. Распределение кластеров при использовании doc2vec

Сказался также тот факт, что описания вакансий содержали общую информацию о графике работы, общие описания об оформлении на работу, предоставление различных бонусов и прочее. При использовании трансформеров прослеживалась та же проблема: несмотря на то, что нейросети могли быть обучены на более профильной тематике, полученные векторы довольно плохо кластеризовались. Формировались кластеры, содержащие в себе либо весь поднабор данных, либо количество кластеров достигало количества векторов при разных настройках алгоритмов кластеризации.

Наиболее эффективным инструментом для решения данной задачи было использование метода TF-IDF. Определение метода выглядит следующим образом:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \epsilon d} f_{t',d}};\ idf(t,D) = log\frac{|D|}{|\{d\ \epsilon\ D : t\ \epsilon\ d\}|}$$

$$tfidf(t,d,D) = tf\ t,d \cdot idf(t,D),$$

где t – токен, d – документ, D – корпус документов, $f_{t,d}$ – число вхождения t в d, $\sum_{t' \epsilon d} f_{t',d}$ – количество токенов в документе, $|D|$ – число документов, $|\{d\ \epsilon\ D : t\ \epsilon\ d\}|$ – число документов хотя бы с одним вхождением t.

Подбор оптимальных значений (в том числе использование N-gramm) для использования этого метода позволил осуществить более точечную кластеризацию векторов.

C. *Кластеризация векторов*

Для кластеризации векторов были рассмотрены два алгоритма:

- DBSCAN [5]
- K-Means [6]

Применение DBSCAN не принесло успехов, возможно, это связано с тем, что векторы имели довольно большую размерность для применения в данном алгоритме. Подбор параметров не дал результаты, кластеры получались несбалансированные, либо для большинства кластеров имело место шумная метка.

В рамках данного исследования для кластеризации векторов применялся алгоритм K-Means (MiniBatch K-means [7]). Его основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике (обычно Евклидово расстояние). На рис. 2 представлен «метод локтя» для поиска оптимального количества кластеров. И хотя он не дал четкого представления об оптимальном количестве кластеров (нет четко выраженной точки изгиба), в исследовании были рассмотрены разбиения на 44–56 кластеров. По субъективной оценке, сформированные кластеры представляли собой довольно выраженные группы вакансий, определенных уникальным поднабором признаков.

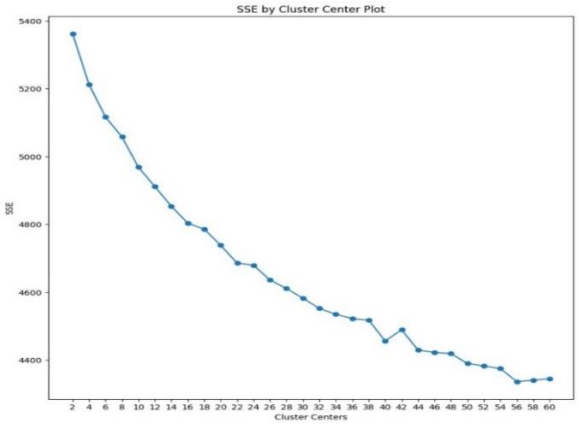

Рис. 3. Применение «метода локтя» для поиска оптимального количества кластеров

Пример кластера (несколько профессий):

- *Data engineer\analyst (развитие источников данных Big Data)*
- *Junior Data-scientist*
- *Data Scientist*
- *Senior Data Scientist (Marketplace Efficiency)*
- *Data Analyst*
- *Senior Data Scientist*
- *ML-разработчик в группу исследований машинного обучения*
- *DataEngineer (Дата-инженер/Специалист по работе с данными)*

Ключевые признаки:

математический,data,алгоритм,обучение,данные, python,машинный обучение,машинный,ml,модель

На рис. 4 изображен график распределения некоторых случайных кластеров с использованием алгоритма TSNE [8] для понижения размерности векторов и сохранения их ключевых признаков.

Для оценки качества кластеризации рассчитаем косинусное расстояние [9] (наиболее популярное решение в задачах анализа текста) между каждым вектором и центром кластера, таким образом, определим схожесть между ними. Если найдется центр кластера, для которого схожесть с вектором выше, чем у центра, определенного алгоритмом, зафиксируем это и рассчитаем количество ошибочных с точки зрения косинусной схожести векторов. Из 6349 векторов 265 оказались схожи к другим центрам, таким образом, качество кластеризации в рамках выбранных путей решения можно оценить на 96 %. Однако, оценка очень субъективная, но она позволяет отловить выбросы и перераспределить некоторые векторы, если есть необходимость.

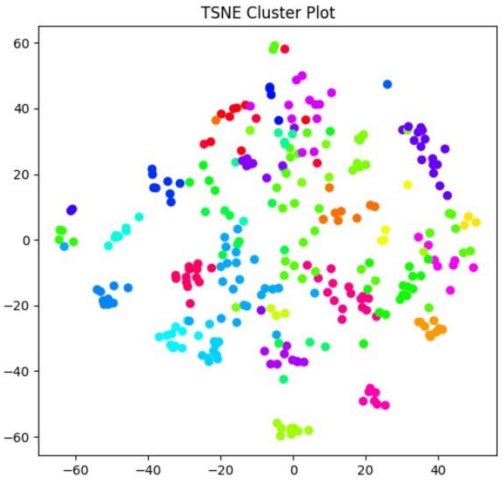

Рис. 4. Распределение кластеров с использованием TSNE

IV. Основные проблемы

Одной из главных проблем, возникающих при решении данной задачи, является наличие универсальных фраз в описании текста вакансии. Например: «предоставляется отпуск», «оформление согласно ТК РФ», «график работы с до » и прочие фразы, присутствующие в описании ко многим вакансиям. Пока прямых решений найти не удавалось, для попытки получить необходимые признаки применялось извлечение ключевых навыков [10], однако и с применением этого метода попадали универсальные фразы и сейчас ведется исследование для возможных решений данной проблемы или уменьшения зависимости от нее. Помимо этого, очень часто подобные друг другу вакансии описываются с помощью синонимов и похожих между собой терминов, также на иностранном языке. Возможное решение данной проблемы (помимо модернизации обработки текста) – подбор подходящего трансформера, обученного на данных, схожих с предметной областью решаемой задачи.

В текущей реализации обработки текста применялось использование стоп-слов при векторизации документов, но, таким образом, можно обработать лишь явно выраженные токены, не относящиеся к предметной области.

V. Заключение

В процессе исследования данной предметной области были рассмотрены методы обработки естественного языка для кластеризации вакансий по их описанию. Предварительно, в процессе изучения самой предметной области, выбор остановился на применении метода TF-IDF и K-means кластеризации. Однако, это не означает, что метод является наиболее эффективным для решения данной задачи в целом. В дальнейшем, постепенно углубляясь в обработку естественного языка, планируется исследовать применение трансформеров,

так как нейросетевые методы являются многообещающими. Но для этого необходимо улучшить предобработку текстовой информации описания вакансий. Как упоминалось ранее, основная проблема извлечения ключевых слов – это наличие в описании вакансий текста, не относящегося и не раскрывающего предметную область профессии, которая появляется из раза в раз практически в каждом описании. Доработка данной проблемы позволит открыть использование нейросетевых методов в полном объеме и, вероятно, улучшит результаты кластеризации.

Список литературы

[1] Scikit-learn, TfidfVectorizer URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (дата обращение 01.03.2023)

[2] Doc2Vec Model URL: https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html (дата обращения 05.03.2023)

[3] Tunstall L., Von Werra L., Wolf T. Natural language processing with transformers. O'Reilly Media, Inc., 2022, 571 с.

[4] Wolf T. и др. HuggingFace’s Transformers: State-of-the-art Natural Language Processing // arXiv preprint arXiv:1910.03771. 2020.

[5] Scikit-learn, DBSCAN URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html (дата обращения 16.03.2023)

[6] Swati P. K-means Clustering Algorithm: Implementation and Critical Analysis, Scholars' Press, 2019, 68с.

[7] Scikit-learn, MiniBatch K-means URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html (дата обращения 17.03.2023)

[8] Laurens van der M., Geoffrey H. Visualizing Data using t-SNE // Journal of Machine Learning Research 9, 2008

[9] Jiawei Han, M. Kamber, J. Pei, Getting to Know Your Data. Data Mining (Third Edition), 2012

[10] Korytov P. V., Kholod I. I. Application of Text Analysis Methods to Recommend Student Choices // 2022 XXV International Conference on Soft Computing and Measurements (SCM). СПб.: IEEE, 2022. С. 107–110.