Автоматизация процесса подбора персонала с помощью математического моделирования.

ВКР, Копчев В.

2022

Введение
Два типа резюме

Был на сайте сегодня в 11:49

Мужчина, 32 года, родился 13 января 1991
Москва, м. Новоясеневская, готов к переезду, готов к командировкам

Аналитик **30 000 руб. на руки**

Специализации:
— Аналитик

Занятость: полная занятость
График работы: полный день

Опыт работы 6 лет 4 месяца

Высшее образование

2013 **Российский государственный торгово-экономический университет, Москва**
Международная торговля, ВЭД предприятий и фирм

Знание языков

Русский — Родной

Английский — C2 — В совершенстве

Немецкий — C1 — Продвинутый

Французский — C1 — Продвинутый

Гражданство, время в пути до работы

Гражданство: Россия
Разрешение на работу: Россия
Желательное время в пути до работы: не имеет значения

Введение
Два типа резюме

Обо мне

Личные качества
• Огромное стремление к профессиональному совершенствованию
• Коммуникабельность
• Презентабельная внешность
• Доброжелательность
• Профессиональная наблюдательность
• Способность длительное время сохранять устойчивое внимание

Компетенции
• Анализ потребностей бизнеса и рынка, поиск точек роста, подготовка сбалансированных решений, которые затем реализуют системные интеграторы -
Это позволяет не заниматься автоматизацией хаоса, экономит время и деньги.
• Диагностика, описание, оптимизация, валидация и регламентация бизнес-процессов.
• Управление знаниями.
• Разработка и внедрение систем сбалансированных показателей (BSC) и оценки полезности сотрудников (ITVE).
• Формирование требований для автоматизации бизнес-процессов (Vision, BRD, FRD).
• Подготовка к сертификации и проведение аудитов на соответствие требованиям стандартов серии:
ГОСТ Р ИСО 9000-2015, ГОСТ ИСО/МЭК 17025-2009, ГОСТ Р ИСО 10002-2009.
Использование инструментов и методик: BPM CBOK, PMBOK, TQM, TOC, ТРИЗ, 5С, Бережливого производства …
(software, hardware): Microsoft Office, Visio, Project, ARIS Express, Business Studio, All Fusion Process Modeler, Mindjet MindManager, Flying logic, 1С, Евфрат, Директум, M-Files, Bizagi, Comindware …
Моделирование бизнес-процессов в нотациях: BPMN (предпочтительно), EPC, IDEF.

Отраслевой опыт
• Разработка IT приложений, Риэлтерская деятельность, Общественное питание, Медицина, Проектирование монтаж и обслуживание слаботочных сетей, Производство (сухих строительных смесей, мебели, меховых изд., радио-эл. изд...), Фармацевтика (испытание лекарственных средств), Оптовая и розничная торговля, Транспортная логистика. Имею опыт работы в зарубежной компании в качестве директора регионального подразделения (DAEWOO Электроникс Сибирь)

Общественная деятельность
• Действительный член Ассоциации BPM-профессионалов ABPMP Russian Chapter www.abpmp.org.ru
• Действительный член Международного клуба «Менеджмент»
www.club-management.org

Введение

Как решить задачу автоматически и нет?

- HR-менеджер размечает в результате собеседования или еще на этапе просмотра резюме в виде анкеты
- Или анализирует ключевые слова в тексте
- Автоматизация: классификация и тематическое моделирование, объединения методов

О работе

Тема, объект и предмет исследования

- Исследовательская ВКР
- Тема: Автоматизация процесса подбора персонала с помощью математического моделирования.
- Объект: задача автоматизации отбора персонала с помощью ML
- Предмет: алгоритмы ML,позволяющие анализировать резюме (нейросети, кластеризация, текстовая аналитика)
- Цель: изучить алгоритмы, разработать свою систему отбора персонала
- Ключевые слова: Big Data, clustering, semi-supervised learning, talent analytics, text mining

О работе
Актуальность темы

- Вакансии аналитиков данных становятся все более популярными, мы видим постоянные рекламы курсов
- Необходимо справляться с большим количеством соискателей
- Можно применить методы NLP и ML для автоматизации отбора персонала
- 500 вакансий (https://career.habr.com/vacancies/analitik?ysclid=lbdj4jt8sd440109047), ~20000 резюме (собрано мной)

Обзор литературы
Статьи, на которые опирается работа

- HR-аналитика: сетевой анализ данных, временные ряды, машинное обучение (ранжирование, скоринг, классификация, рекомендательные системы, нейросети, ML Engineering), онтологии.
- There are several works that explore the topic resumes clas- sification in talent analytics. In [11] classification is performed using the SVM, Random Forest, Naive Bayes algorithms on the dataset of 10,000 entries. These entries form an unstruc- tured dataset, in which data passed classical procedure of data cleanup, tokenization (after which each document was trans- lated into a list of words), stemming and lemmatization. After it, tf-idf vectors were formed and classification algorithms were applied. In [9] different models were applied: kNN, MLP (Multi-Layer Perceptron), LR (Logistic Regression) and SVM. Especially interesting is the fact that MLP (Multi-Layer Per- ceptron) was applied — this paper shows that neural networks could be efficiently applied in people analytics classification tasks too.

Обзор литературы
Статьи, на которые опирается работа

- Besides classification algorithms, ranking algorithms could be applied in talent analytics problems. As we see in [13], we can apply more complex neural networks for ranking problems based on semantic representations. In [6] we see that both ranking and classification-based approaches give us good results, combined with semantic representations, obtained with help of neural networks. Paper [15] shows how NLP (Natural language processing) and classification in machine learning could be blended together to build a pipeline that helps score every resume based on skills, education, work experience using scores of how useful these applicant’s characteristics for the company. For example, work experience could be not “4 years”, but “48/100”.

Обзор литературы
Статьи, на которые опирается работа

- Problems in talent analytics, that could be solved my machine learning, are not limited to classification or ranking. One of the greatest problems in this field is building recom- mendation systems of resumes. In paper [14] candidates' CVs are ranked using kNN and cosine similarity distance to build a recommendation system of CVs. In [1] recommendation systems in HR analytics are investigated even further. In this paper, concepts from IR (information retrieval), NLP (Re- current, Convolutional, Graph Neural Networks, Transformer architecture) are observed in relation to the task of building resume and job description matching system.
- Besides recommendation systems building, classification and ranking problems, machine learning and neural networks could be used in people analytics in more specific way. In [12] Ability-aware Person-Job Fit Neural Network (APJFNN) model is presented to improve job-applicant fitting. In [5], a general framework ResumeNet is proposed for automatization of resumes' quality assessment task.
- Moreover, as [10] shows, network data analysis could be applied in talent analytics too for studying career progression of different individuals. In this work dataset of online profiles from anonymous OPN is presented. In this dataset authors analyzed distribution of job level in years by different coun- tries, job age range, centralities of top jobs and companies. Besides machine learning and network science, as we see in [7], ontologies could be applied in people analytics.

Обзор литературы
Достоинства и недостатки методов

- Supervised: не все данные, переобучение; решается semi-supervised
- Topic modelling: слова, а не n-граммы; решается кластеризацией n-грамм
- While HR managers usually look not at single words like “supervised”, “University”, “machine”, “Prince- ton”, “learning” in resumes, but at n-grams like “Studied at Princeton University”, “supervised machine learning”, topic modeling algorithms do not work well if we break text not into words, but into n-grams.

Обзор литературы
Достоинства и недостатки методов

- Пример проблемы: Допустим, в кластере есть слово “microsoft”.
- Microsoft Word и уверенный пользователь ПК? Или Microsoft Excel и Microsoft Power BI?

Цели и задачи
Цели

- Исследовать алгоритмы ML и возможность их использования для автоматизации задач HR-отдела (HR-аналитики), различные подходы к работе с текстом
- Проверить гипотезу: кластеризация n-грам и semi-supervised 3-label классификация лучше скоринга

Цели и задачи
Задачи

- Собрать данные с hh.ru с помощью библиотек Python
- Предобработка текста: n-граммы, стоп-слова, извлечение сущностей, приведение к единому виду, обработка шумов с помощью mBART, эмбецдинги (BERT, doc2vec, средний word2vec)
- Кластеризация (KMeans, HDBSCAN, Gaussian Mixture Model)
- Обучение semi-supervised learning модели на частично размеченных данных (~400 строк)
- Валидация моделей: метрики, анализ факторов

Результаты

Полученные и планируемые

- Собраны факторы:
- Исследовательская ВКР
- Фичи Link
- City. Is applicant willing to relocate? Is applicant prepared for business trips?
- Work experience in years and month
- Text of “About me” section of resume
- Education
- Interests
- Skills
- Bachelor degree information
- Master’s degree information
- Level of education
- Count of degrees
- Last job
- Second last job
- Is university applicant graduated from Russian top university? (Yes, No)
- Is resume written in English?
- Do analyst words (Python, SQL, Tableau, etc.) present in applicant’s interests?
- Skills by list
- Relocation (Yes, No)
- Occasional business trips (Yes, No)
- City

Название	Коммандировка	Опыт	О себе	Образование	
Analyst	Moscow, willing to relocate, prepared for occa...	Work experience 7 years 4 months	Responsible, communicable, quick study and det...	Higher education\n2015\nMOSCOW STATE UNIVERSIT...	Specializations:\nSales n
BI аналитик	Москва, не готова к переезду, готова к редким ...	Опыт работы 1 год 5 месяцев	В последние годы проходила обучение без возмож...	Высшее образование (Бакалавр)\n2022\nНациональ...	Специализации анал
BI аналитик	Москва, не готова к переезду, готова к команди...	Опыт работы 3 года 3 месяца	---	Высшее образование (Бакалавр)\n2019\nСамарский...	Специализации: анал
Аналитик	Санкт-Петербург, м. Гражданский проспект, гото...	Опыт работы 7 лет 2 месяца	Имею экономическое образование. Продвинутый по...	Высшее образование (Бакалавр)\n2015\nСанкт-Пет...	Специализации:\nАнали
Программист-разработчик	Москва, м. Петровско-Разумовская, не готов к п...	Опыт работы 1 год 1 месяц	Имеется опыт в создании Android-приложений на ...	Высшее образование (Бакалавр)\n2022\nЧелябинск...	Специализации: разр
...	...	...	...	...	
Analyst	Moscow, metro station Kantemirovskaya, willing...	Work experience 14 years 3 months	Hardworking and well-organized specialist with...	Higher education\n2022\nЯндекс Практикум\nИнже...	Specializations:\nTester\
Аналитик-	Москва, м. Раменки, не	Опыт работы 3	Ответственна к выполнению работы, имею навыки	Высшее образование	Специализации

Результаты
Полученные и планируемые

- Исследованы самые популярные 3-граммы, построены эмбеддинги (средний word2vec), на них проведена кластеризация методом KMeans
- Most popular 1-grams were analyzed: ”work”, ” experience”, ”knowledge”, ”business”, ”project”. Most popular 2-grams: ”work experience”, ”work skills”, ” business process”, ”MS Office”, ”command work”, ” microsoft office”. Our hypothesis is true: 2-grams are more meaningful than 1-grams. Now we need to check whether it means that it could give us more precise results.
- Now we need to label more data, fit semi-supervised model, perform clusterisation and vuild embeddings using different algorithms, evaluate and interpret model.

Рис. 1. Результаты кластеризации

Список литературы

[1] Barrak, A., Adams, B., & Zouaq, A. (2022). Toward a traceable, explainable, and fairJD/Resume recommendation system. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2202.08960

[2] Gusev, I. (2020). Dataset for Automatic Summarization of Russian News. Communications in Computer and Infor- mation Science, 122–134. https://doi.org/10.1007/978-3-030- 59082-6 9

[3] Lewis, M. (2019). BART: Denoising Sequence- to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ArXiv.Org. https://doi.org/10.48550/arXiv.1910.13461

[4] Liu, Y., Gu, J., Goyal, N., Li, X. C., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2001.08210

[5] Luo, Y., Zhang, H., Wang, Y., We, Y., & Zhang, X. (2018). ResumeNet: A Learning-based Framework for Auto- matic Resume Quality Assessment. ArXiv (Cornell Univer- sity). https://doi.org/10.48550/arxiv.1810.02832

Список литературы

[6] Menacer, M. A. (2021). An interpretable person-job fitting approach based on classification and ranking. ACL Anthology. https://aclanthology.org/2021.icnlsp-1.15/

[7] Mishra, R. (2020). An AI based talent acquisition and benchmarking for job. ArXiv.Org. https://doi.org/10.48550/arXiv.2009.09088

[8] Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

[9] Naidu, P. V., Bommu, V. M. R., Pallapothu, V. D., Janapamula, S. R. N., & Kommuri, N. L. (2021). Resume Screening Using Machine Learning. Lecture Notes in Net- works and Systems, 745–751. https://doi.org/10.1007/978-3- 030-84760-9 63

[10] Oentaryo, R. J., Lim, E., Ashok, X. J. S., Prasetyo, P. K., Ong, K., & Lau, Z. T. (2018). Talent Flow Analytics in Online Professional Network. Data Science and Engineering, 3(3), 199–220. https://doi.org/10.1007/s41019-018-0070-8

Список литературы

[11] Pal, R., Shaikh, S., Satpute, S., & Bhagwat, S. (2022). Resume Classification using various Machine Learning Algorithms. ResearchGate. https://doi.org/10.1051/itmconf/20224403011

[12] Qin, C., Zhu, H., Xu, T., Zhu, C., Jiang, L., Chen, E., & Xiong, H. (2018). Enhancing Person-Job Fit for Talent Recruitment. The 41st International ACM SIGIR Confer- ence on Research & Development in Information Retrieval. https://doi.org/10.1145/3209978.3210025

[13] Ramanath, R., Inan, H., Polatkan, G., Hu, B., Guo, Q., Ozcaglar, C., Wu, X., Kenthapadi, K., & Geyik, S. C. (2018). Towards Deep and Representation Learning for Talent Search at LinkedIn. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1809.06473

[14] Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A Machine Learning approach for automation of Resume Recommendation system. Procedia Computer Science, 167, 2318–2327. https://doi.org/10.1016/j.procs.2020.03.284

[15] Zimmermann, T. (2016). Data-driven HR - Resume Analysis Based on Natural Language Processing and Machine Learning. ArXiv.Org. https://doi.org/10.48550/arXiv.1606.05611