Разработка рекомендательной системы на основе ИТ-компетенций

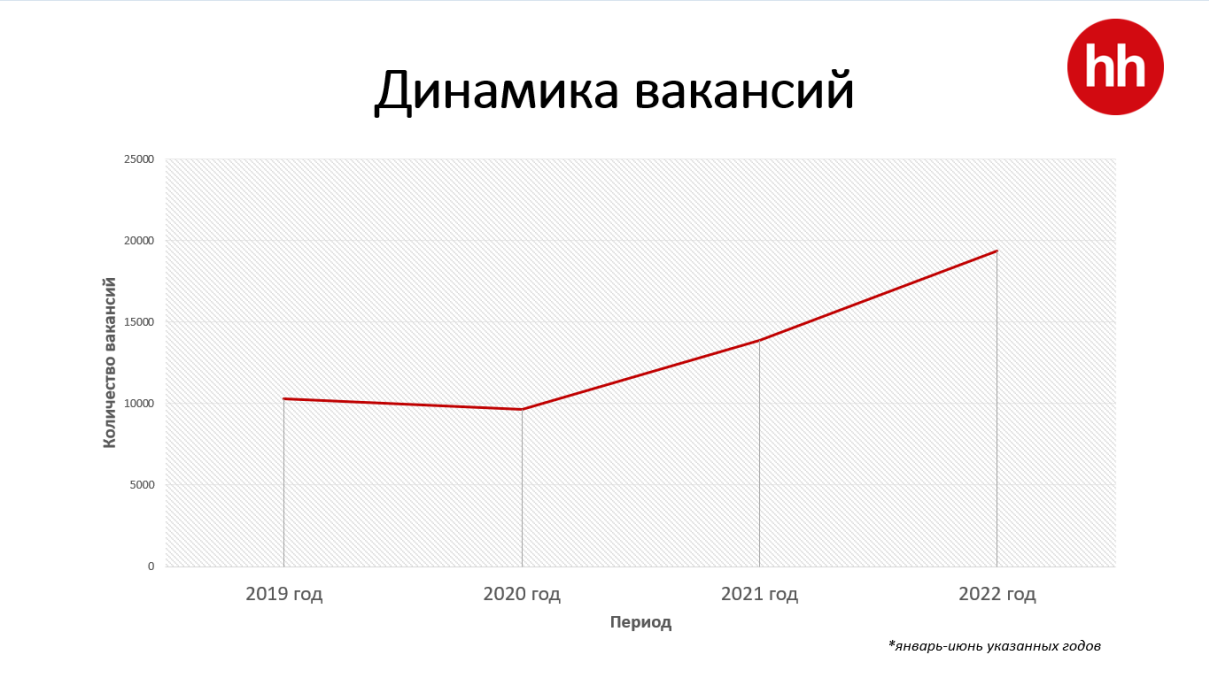
Аширали А.
Шойынбек А.А., PhD ассоц. профессор

NARXOZ UNIVERSITY

Введение

Рост ИТ-сектора в Казахстане требует более эффективных методов подбора персонала, так как спрос на специалистов превышает предложение. Необходимо разработать систему, которая сможет автоматически анализировать ИТ-компетенции кандидатов и предлагать наиболее подходящие вакансии, основываясь на этих данных. Это поможет сократить время и затраты на найм и повысит эффективность подбора персонала.

Динамика роста ИТ-сектора



Источник: https://kapital.kz/tehnology/107740/top-10-vostrebovannykh-it-spetsialistov-v-kazakhstane.html (Дата обращения: 25.05.2023)

Актуальность

Рекомендательная система является важным шагом в области улучшения процесса поиска работы и подбора персонала в ИТ-сфере. Предлагаемая система будет основываться на анализе ИТ-компетенций пользователей, включая их знания, навыки. С помощью современных методов машинного обучения и анализа данных, система будет способна предоставлять рекомендации, соответствующие профессиональным интересам и потребностям каждого пользователя.

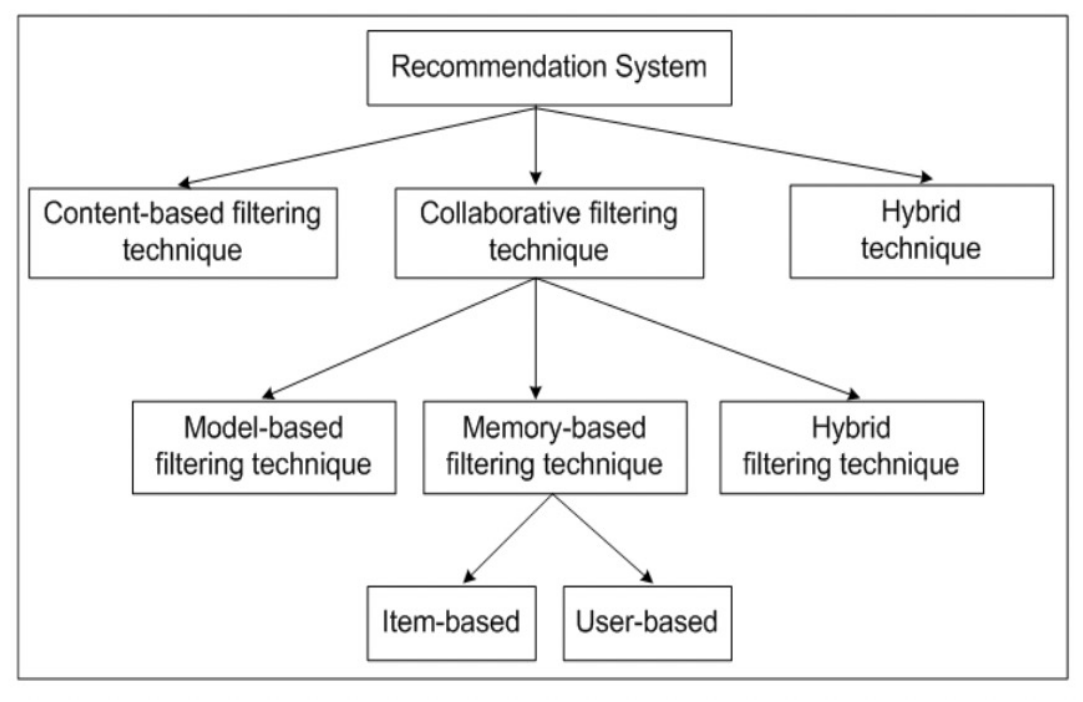
Цель диссертационного исследования

Целью настоящей магистерской диссертации заключается в разработке рекомендательной системы, основанной на ИТ-компетенциях пользователей, с целью предложения наиболее подходящих вакансий

Задачи

1. Анализ современных методов оценки ИТ-навыков и рекомендательных систем.
2. Сбор данных для обучения модели определяющей ИТ-навыки
3. Создание метода автоматизированного извлечения ИТ-навыков из текстовых описаний.
4. Проектирование и тестирование прототипа рекомендательной системы на основе ИТ-навыков.

Типы рекомендательных систем

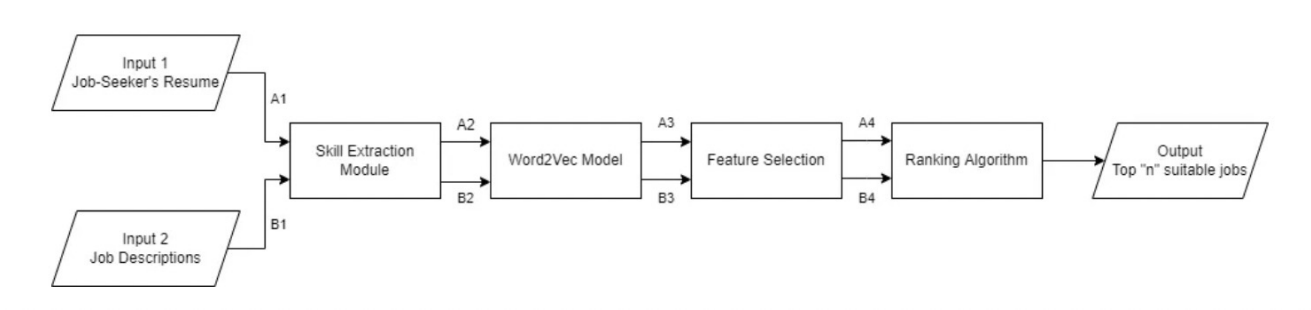


Классификация рекомендательных систем (F.O. Isinkaye et al., 2015)

Методы рекомендательных систем	**Content-based filtering**	**Collaborative filtering technique**	**Hybrid filtering**
Количество пользователей	На основе одного пользователя	Основываясь на многих пользователей, имеющих аналогичный интерес	Комбинация содержания на основе и совместной фильтрация
Преимущества	Пользовательская независимость, Прозрачность	Улучшение рекомендационной производительности	Преодоление проблемы холодного старта, проблема разреженности
Недостатки	Ограниченное содержание анализа, Новый пользователь	Разреженность данных, Масштабируемость	Увеличение сложности, дорогостоящий в реализации

Применение рекомендательных систем в подборе персонала

Методы рекомендательных систем, включая фильтрацию на основе контента, коллаборативную фильтрацию и гибридные методы, имеют различные преимущества и недостатки. Рекомендательные системы на основе текста используют комбинацию техник и служат важным инструментом в подборе персонала.



Общий процесс рекомендательной системы (Kara A. et al., 2023).

Подходы для определения ИТ-компетенций

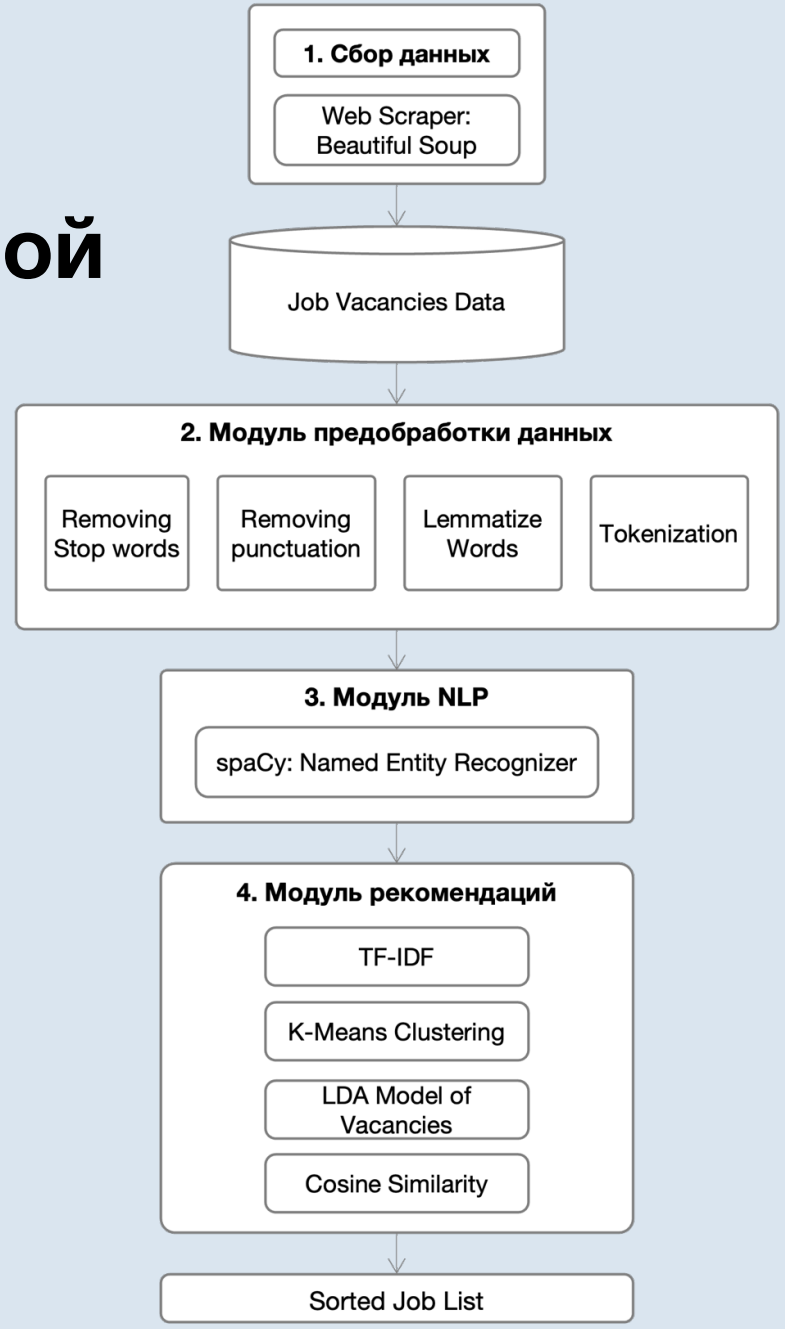
Метод/Подход	Описание
Анализ вакансий и требований к навыкам	Анализ текстов вакансий и требований к навыкам работодателей с помощью алгоритмов обработки естественного языка и методов машинного обучения.
Анализ профилей и резюме	Обработка профилей и резюме соискателей с использованием алгоритмов обработки естественного языка. Анализ ключевых слов и навыков для определения наличия и уровня ИТ-компетенций у соискателей.
Опросы и анкетирование	Проведение опросов и анкетирования среди ИТ-специалистов и работодателей для определения необходимых ИТ-компетенций и их уровней.
Экспертные оценки	Экспертные оценки со стороны ИТ-экспертов, таких как руководители проектов, менеджеры или преподаватели.

Сравнение моделей для выявления сущностей

Модель	Основа	Преимущества	Недостатки
spaCy NER	Rule-based, CNN	Легкая и быстрая; хорошо подходит для английского языка; возможность дообучения и настройки	Может быть менее точной для некоторых языков и доменов
BERT	Transformers	Высокая точность и обобщающая способность; множество предобученных моделей; многоязычность	Требует больших вычислительных мощностей и памяти; может быть медленным при инференсе
GPT-2, GPT-3	Transformers	Высокая точность; хорошо работает с контекстом; многоязычность	Требует больших вычислительных мощностей и памяти; может быть медленным при инференсе; ограниченный доступ к GPT-3
RoBERTa	Transformers	Очень высокая точность; основана на BERT; больше данных для обучения; многоязычность	Требует больших вычислительных мощностей и памяти; может быть медленным при инференсе
Stanza (Stanford NLP)	RNN, LSTM	Хорошая точность; поддержка множества языков; модульность	Может быть медленным при инференсе; требует больших вычислительных мощностей и памяти

Выбранные инструменты для анализа

Процесс рекомендательной системы



Сбор данных

Набор данных состоит из 3 310 ИТ вакансии от компании с сайта https://career.habr.com/ и содержит следующие столбцы:

title: Заголовок / название вакансии
company: Компания – Работодатель
requirements: Требования вакансии
description: Описание вакансии
time: Дата вакансии
salary_min: Минимальная вилка ЗП
salary_max: Максимальная вилка ЗП
salary_currency: Валюта ЗП

	title	company	requirements	description	time	salary_min	salary_max	salary_currency
0	Главный эксперт кибербезопасности	Магнит	['Инженер по безопасности', 'Ведущий (Lead)', ...	Стек технологий: cloud: mail,yandex,sber,azure...	03.05.2023	NaN	NaN	NaN
1	PHP программист (Bitrix фреймворк, удаленно)	Ньютон Технологии	['Фулстек разработчик', 'Средний (Middle)', 'H...	ООО «Ньютон Технологии» — продуктовая IT-комп...	28.04.2023	NaN	NaN	NaN
2	Администратор приложений (Atlassian JIRA и Con...	СберКорус	['Системный администратор', 'Младший (Junior)'...	СБЕР КОРУС - IT-компания, разработчик и провай...	06.05.2023	100000.0	140000.0	₽
3	Администратор БД PostgreSQL	ИНГОССТРАХ	['Администратор баз данных', 'PostgreSQL', 'DB...	Обязанности:\nПолный спектр задач DBA:\nУстано...	21.04.2023	NaN	NaN	NaN
4	Системный инженер	Rambler&Co	['Системный администратор', 'Средний (Middle)'...	омандеRambler&Co занимает первое место среди м...	11.05.2023	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...
3305	Системный аналитик	TINKOFF	['Системный аналитик', 'SQL', 'REST', 'Apache ...	Тинькофф Банк — это один из самых успешных и г...	10.05.2023	NaN	NaN	NaN
3306	Бизнес-архитектор [Governance]	МТС	['Архитектор программного обеспечения', 'Ведущ...	МТС – это мультисервисная цифровая экосистема....	11.05.2023	NaN	NaN	NaN
3307	PHP Backend разработчик	Andagar	['Бэкенд разработчик', 'Средний (Middle)', 'PH...	омандеВ компанию разработчик электронной торго...	03.05.2023	NaN	NaN	NaN

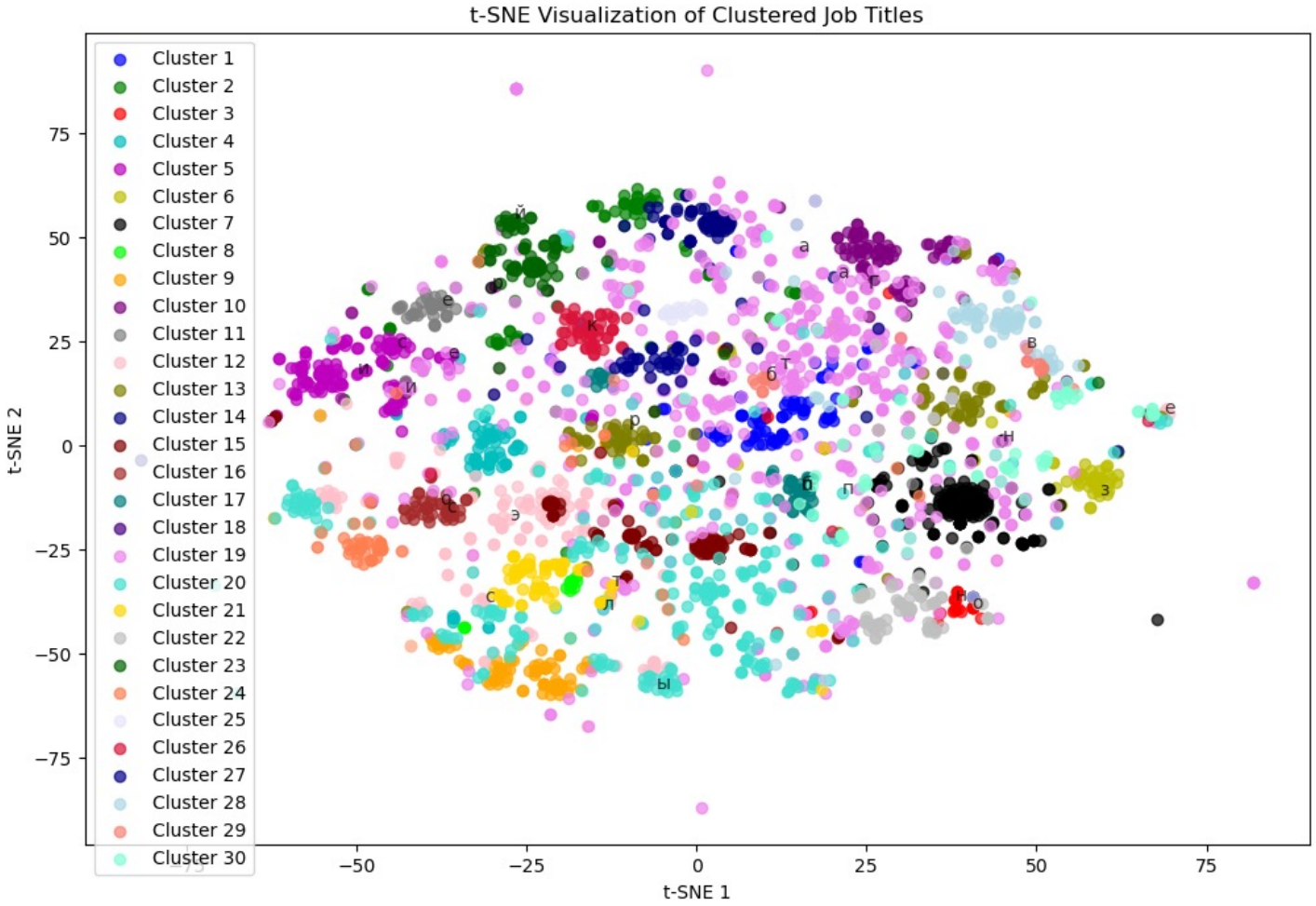
Обучение NER модели

Обучение модели включало подачу обучающих примеров на вход модели, которая пыталась выявить общие закономерности и характеристики, связанные с ИТ-навыками, чтобы эффективно классифицировать их в тексте.

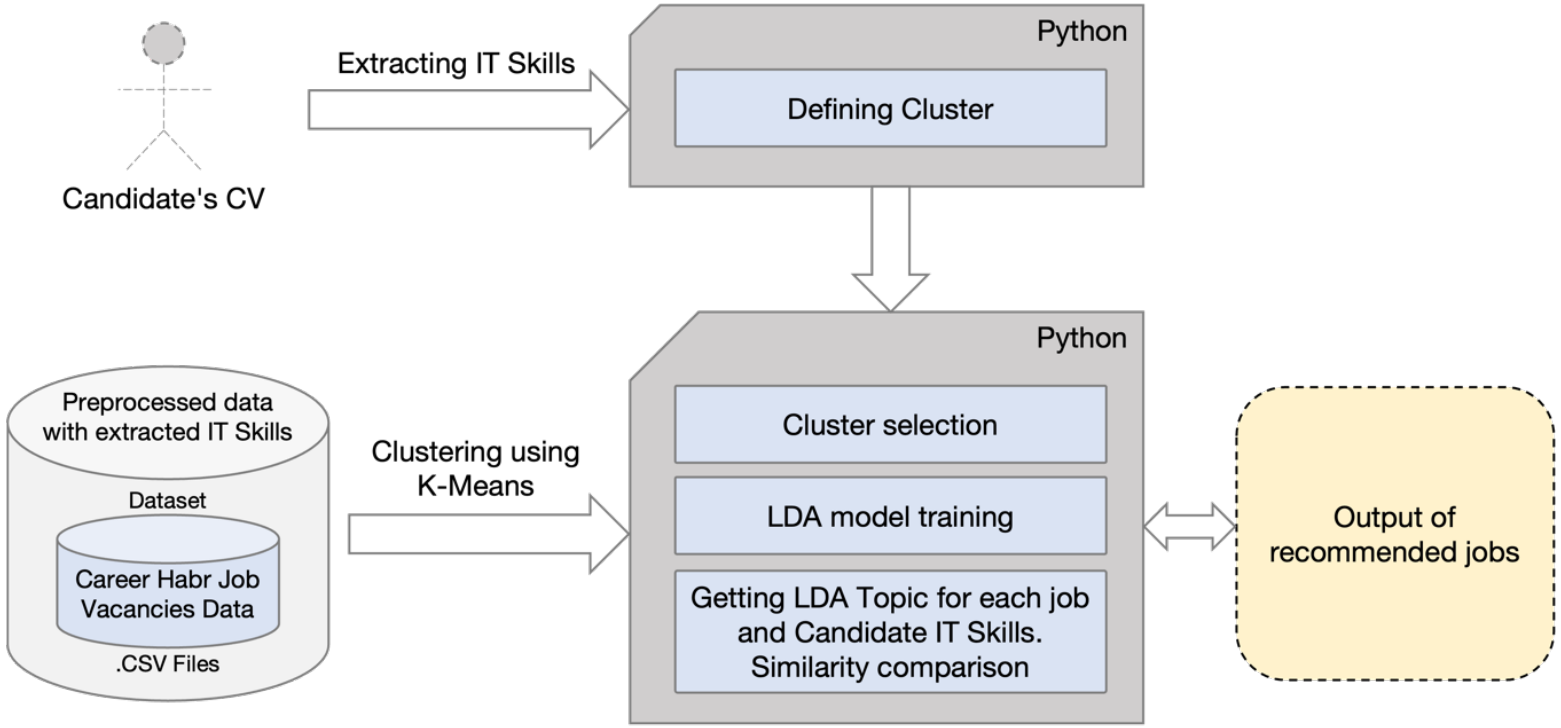
Для тренировочного набора данных был создан набор кортежей из 107 тренировочных данных следующего вида:

("Хорошие знания PHP, Laravel и Node.js",
{"entities": [(15, 18, "IT-SKILL"), (20, 27, "IT-SKILL"), (30, 37, "IT-SKILL")]})

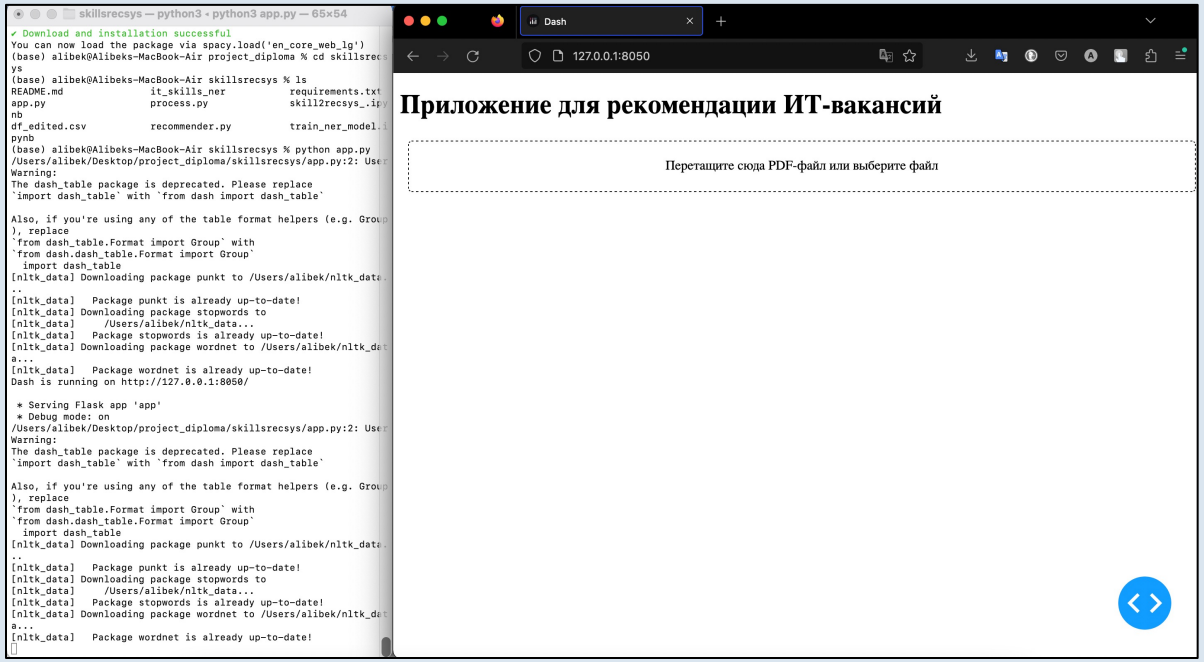
Кластеризация ИТ-вакансии

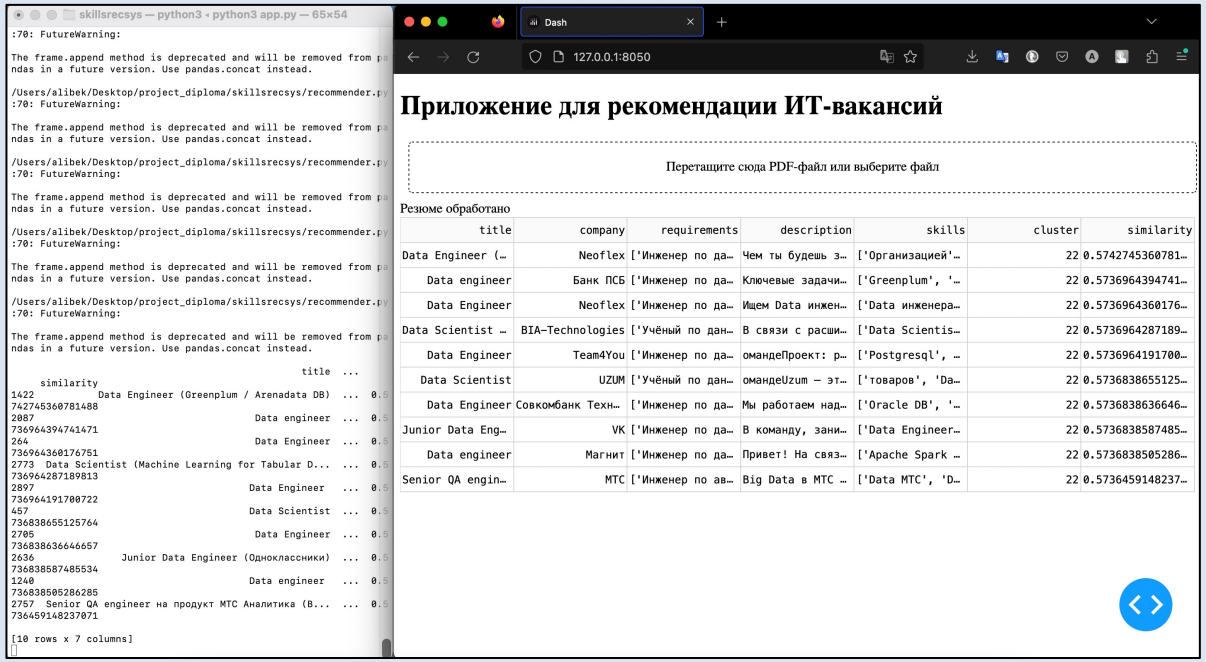


Дизайн рекомендательной системы



Развертывание веб-приложения





Результаты рекомендательной системы

Кандидат	Precision	Recall	F1	Coverage
1	0.4	0.57	0.47	1.42
2	0.6	0.66	0.63	1.11
3	0.5	0.71	0.58	1.42
4	0.4	0.57	0.47	1.42
5	0.2	0.25	0.22	1.25
Средний результат	0.42	0.552	0.474	1.324

$$Coverage = \frac{n}{N} \cdot 100$$

n = количество рекомендуемых элементов

N = количество элементов в списке

Дальнейшие исследования

Для дальнейшего улучшения системы рекомендаций возможно проведение следующих дополнительных исследований:

1. Расширение модуля NER для учета уровня владения знаниями и технологиями ИТ-компетенций.
2. Анализ взаимодействий между соискателями и вакансиями для улучшения персонализации рекомендаций.
3. Использование дополнительных мер оценки качества, таких как перплексия или когерентность тем, для настройки LDA модели.
4. Улучшение интерфейса системы и увеличение удобства использования для пользователей.

Спасибо за внимание!

Ссылка на репозитории с проектом:
https://github.com/alibekashirali/skillsrecsys

Обучение NER модели

Programming	*Database*	Web Development	Network	Data analyze	Security	Management	Cloud	DevOps
Programming Paradigms	**Types of Databases**	**Front-end Development**	**Network topology**	**Data Analysis Methods**	**Access control**	**Management Concep**	**Cloud Computing Services**	**Continuous Integ**
Imperative Programming	Relational Databases	HTML	Bus topology	Descriptive Statistics	Authentication	Strategic Planning	Infrastructure as a Service (IaaS)	Jenkins
Object-Oriented Programming	Non-Relational Databases	CSS	Star topology	Inferential Statistics	Authorization	Project Management	Platform as a Service (PaaS)	CircleCI
Functional Programming	Document-Oriented Databases	JavaScript	Ring topology	Exploratory Data Analysi	Password policies	Change Management	Software as a Service (SaaS)	Travis CI
Declarative Programming	Key-Value Stores	React.js	Mesh topology	Time Series Analysis	Single sign-on (SSO)	Risk Management	Function as a Service (FaaS)	GitLab CI
Event-Driven Programming	Graph Databases	Angular.js	Tree topology	Machine Learning	Two-factor authentication (2FA	Quality Management	Backend as a Service (BaaS)	AWS CodePipelin
Procedural Programming	Column-Family Stores	Vue.js	Hybrid topology	Deep Learning		Process Improvement	Disaster Recovery as a Service (I	Azure DevOps
				Natural Language Proces	**Cryptography**	Leadership	Security as a Service (SECaaS)	
Programming Languages	**Database Management Systems**	**Back-end Development**	**Network devices**	Predictive Modeling	Encryption	Decision Making	Database as a Service (DBaaS)	**Configuration Ma**
Java	MySQL	Node.js	Router	Data Mining	Decryption	Communication		Puppet
Python	PostgreSQL	Python	Switch		Hashing		**Cloud Deployment Models**	Chef
C++	Microsoft SQL Server	Ruby on Rails	Hub	**Data Types**	Digital signatures	**Project Management**	Public Cloud	Ansible
JavaScript	Oracle Database	PHP	Repeater	Structured Data	Public key infrastructure (PKI)	Work Breakdown Stru	Private Cloud	SaltStack
Ruby	MongoDB	Django	Gateway	Unstructured Data		Gantt Chart	Hybrid Cloud	
PHP	Cassandra	Flask	Bridge	Semi-Structured Data	**Network security**	Critical Path Method ((	Community Cloud	**Containerization**
Swift	Redis	Laravel	Access point	Time Series Data	Firewalls	Agile Methodology		Docker
Kotlin	Neo4j	ASP.NET		Spatial Data	Intrusion detection and preventi	Waterfall Methodology	**Cloud Providers**	Kubernetes
TypeScript	Apache HBase		**Network protocols**		Virtual private networks (VPNs	Scrum	Amazon Web Services (AWS)	OpenShift
Go		**Web Technologies**	TCP/IP	**Data Storage**	Denial of service (DoS) protecti	Kanban	Microsoft Azure	
Rust	**Database Operations**	AJAX	HTTP	Relational Databases	Penetration testing	Lean Six Sigma	Google Cloud Platform (GCP)	**Infrastructure as**
Dart	Create	JSON	FTP	Non-Relational Databases		Earned Value Manage	IBM Cloud	Terraform
	Read	XML	DNS	Data Warehouses	**Application security**		Oracle Cloud	CloudFormation
Coding Concepts	Update	REST	DHCP	Data Lakes	Secure coding practices	**IT Service Managem**	Alibaba Cloud	Ansible
Algorithms	Delete	SOAP	SMTP	Big Data Storage	Web application firewalls (WAF	Incident Management	DigitalOcean	
Data Structures	Indexing	GraphQL	POP3		Code review	Problem Management		**Monitoring and L**
Control Flow	Transactions		IMAP	**Data Manipulation**	Dynamic application security te	Change Management	Cloud Security	Nagios
Error Handling	Backup and Recovery	**Web Design**	SNMP	Data Cleaning	Static application security testin	Service Level Manage	Identity and Access Management	Prometheus
Memory Management	Data Migration	UI/UX design	SSH	Data Preprocessing		Availability Managem	Network Security	Grafana
Input/Output Operations		Responsive design	Telnet	Feature Engineering	**Cloud security**	Capacity Management	Data Security	ELK Stack (Elastic
Regular Expressions	**Data Modeling**	Adobe Photoshop		Dimensionality Reductio	Identity and access managemen	IT Service Continuity	Compliance	
Recursion	Entity-Relationship Model	Adobe Illustrator	**Network security**	Data Aggregation	Virtual private clouds (VPCs)		Encryption	**Version Control**
Unit Testing	Relational Model	Sketch	Firewall	Data Visualization	Data encryption at rest and in tr	**IT Financial Manage**	Vulnerability Management	Git
Debugging	Object-Oriented Model		VPN		Incident response planning	Budgeting	Incident Response	GitHub
	Document Model	**Web Security**	SSL/TLS	**Tools**		Cost Accounting		Bitbucket
Integrated Development Environments (IDEs)	Graph Model	HTTPS	IDS/IPS	Programming Languages	**Physical security**	Asset Management	**Cloud Computing Architecture**	GitLab
Visual Studio Code		SSL/TLS	Access control	Data Analytics Tools (Tal	Building access control	Financial Reporting	Cloud Native Architecture	
IntelliJ IDEA	**Database Design**	OWASP	Encryption	Data Science Platforms (I	Video surveillance	ROI Analysis	Microservices Architecture	**Collaboration**
Eclipse	Normalization	Authentication	Authentication	Cloud-Based Analytics P	Security guards	Cost-Benefit Analysis	Serverless Architecture	Slack

Обучение NER модели

```
#
TRAIN_DATA = [
    ("Хорошие знания PHP, Laravel и Node.js", {"entities": [(15, 18, "IT-SKILL"), (20, 27, "IT-SKILL"), (30, 37, "IT-SKILL")]}),
    ("Знания SQL (PostgreSQL), CI/CD, Docker, kubernetes;", {"entities": [(7, 10, "IT-SKILL"), (12, 22, "IT-SKILL"), (25, 30, "IT-SKILL"), (32, 38, "IT-SKILL
    ("Понимание restapi, OpenAPI (например swagger), kafka;", {"entities": [(10, 17, "IT-SKILL"),(19, 26, "IT-SKILL"),(37, 44, "IT-SKILL"),(47, 52, "IT-SKILL
    ("Опыт администрирования Linux;", {"entities": [(23, 28, "IT-SKILL")]}),
    ("Опыт работы с Git", {"entities": [(14, 17, "IT-SKILL")]}),
    ("Плюсом будет знание Go и/или C#", {"entities": [(20, 22, "IT-SKILL"),(29, 31, "IT-SKILL")]}),
    ("Умение писать понятный код с комментированием и применением ключевых стандартов, рефакторинг при необходимости;", {"entities": [(81, 92, "IT-SKILL")]})
    ("Ответственность, самоорганизованность, коммуникабельность.", {"entities": [(17, 37, "IT-SKILL"),(39, 57, "IT-SKILL")]}),
    ("Знакомство с различными видами тестирования;", {"entities": [(31, 43, "IT-SKILL")]}),
    ("Трекинговые системы Jira, ClickUp, etc;", {"entities": [(20, 24, "IT-SKILL"),(26, 33, "IT-SKILL")]}),
    ("Базовые навыки работы с DevTool, Postman;", {"entities": [(24, 31, "IT-SKILL"),(33, 40, "IT-SKILL")]}),
    ("Понимание HTML, CSS;", {"entities": [(10, 14, "IT-SKILL"),(16, 19, "IT-SKILL")]}),
    ("Знание техник тест-дизайна;", {"entities": [(14, 25, "IT-SKILL")]}),
    ("Уверенное знание Postman, Insomnia;", {"entities": [(17, 24, "IT-SKILL"), (26, 34, "IT-SKILL")]}),
    ("Базовое знание SQL;", {"entities": [(15, 18, "IT-SKILL")]}),
    ("Понимание SOAP, REST;", {"entities": [(10, 14, "IT-SKILL"), (16, 20, "IT-SKILL")]}),
    ("Понимание клиент-серверной архитектуры.", {"entities": [(10, 38, "IT-SKILL")]}),
    ("I'm proficient in HTML, CSS, and JavaScript for web development.", {"entities": [(18, 22, "IT-SKILL"), (24, 27, "IT-SKILL"), (33, 43, "IT-SKILL"), (48,
    ("Developed applications using React, Angular, and Vue.js.", {"entities": [(29, 34, "IT-SKILL"), (36, 43, "IT-SKILL"), (49, 55, "IT-SKILL")]}),
    ("Built RESTful APIs with Node.js and Express.js.", {"entities": [(6, 18, "IT-SKILL"), (24, 31, "IT-SKILL"), (36, 46, "IT-SKILL")]}),
    ("Extensive experience with Ruby on Rails and Django.", {"entities": [(26, 39, "IT-SKILL"), (44, 50, "IT-SKILL")]}),
    ("Implemented machine learning models using TensorFlow and Keras.", {"entities": [(12, 28, "IT-SKILL"), (42, 52, "IT-SKILL"), (57, 62, "IT-SKILL")]}),
    ("Used R for data analysis and Python for data visualization.", {"entities": [(5, 6, "IT-SKILL"), (11, 24, "IT-SKILL"), (29, 35, "IT-SKILL"), (40, 58, "I
    ("Worked with database management systems like MySQL and PostgreSQL.", {"entities": [(45, 50, "IT-SKILL"), (55, 65, "IT-SKILL")]}),
    ("Experienced in cloud technologies such as AWS, Azure, and Google Cloud Platform.", {"entities": [(15, 20, "IT-SKILL"), (42, 45, "IT-SKILL"), (47, 52, "
    ("Implemented cybersecurity measures, including penetration testing and security audits.", {"entities": [(12, 34, "IT-SKILL"), (46, 65, "IT-SKILL"), (70,
    ("Managed Linux servers and wrote shell scripts for automation.", {"entities": [(8, 13, "IT-SKILL"), (14, 21, "IT-SKILL"), (32, 37, "IT-SKILL"), (38, 45,
    ("Applied Agile methodologies like Scrum and Kanban for project management.", {"entities": [(8, 13, "IT-SKILL"), (33, 38, "IT-SKILL"), (43, 49, "IT-SKILL
```

Процесс LDA обучения и подсчета близости