

Ранжирование вакансий по схожести с опытом кандидата

Набойщиков Илья

Июнь'22



Проект ПЕРЕЗАПУСК

О себе

- *Набойщиков Илья*
- *Университет ИТМО, Интеллектуальные системы в гуманитарной сфере, бакалавр*
- *Блок «Технологии», Департамент «Финансы», разрабатываю аналитическую отчетность (Qlik Sense), ETL-решения (Pentaho, NiFi)*
- *Санкт-Петербург, в данный момент не готов к переезду*
- *naboishikov@gmail.com*

Описание проекта

- *Опишите кратко суть проекта и функционал. В рамках проекта необходимо было собрать вакансии с портала hh.ru и реализовать их ранжирование на основе схожести описания вакансий с описанием опыта кандидата.*
- https://github.com/naboi/data_analyst_track/blob/main/final_project.ipynb

Бизнес-логика

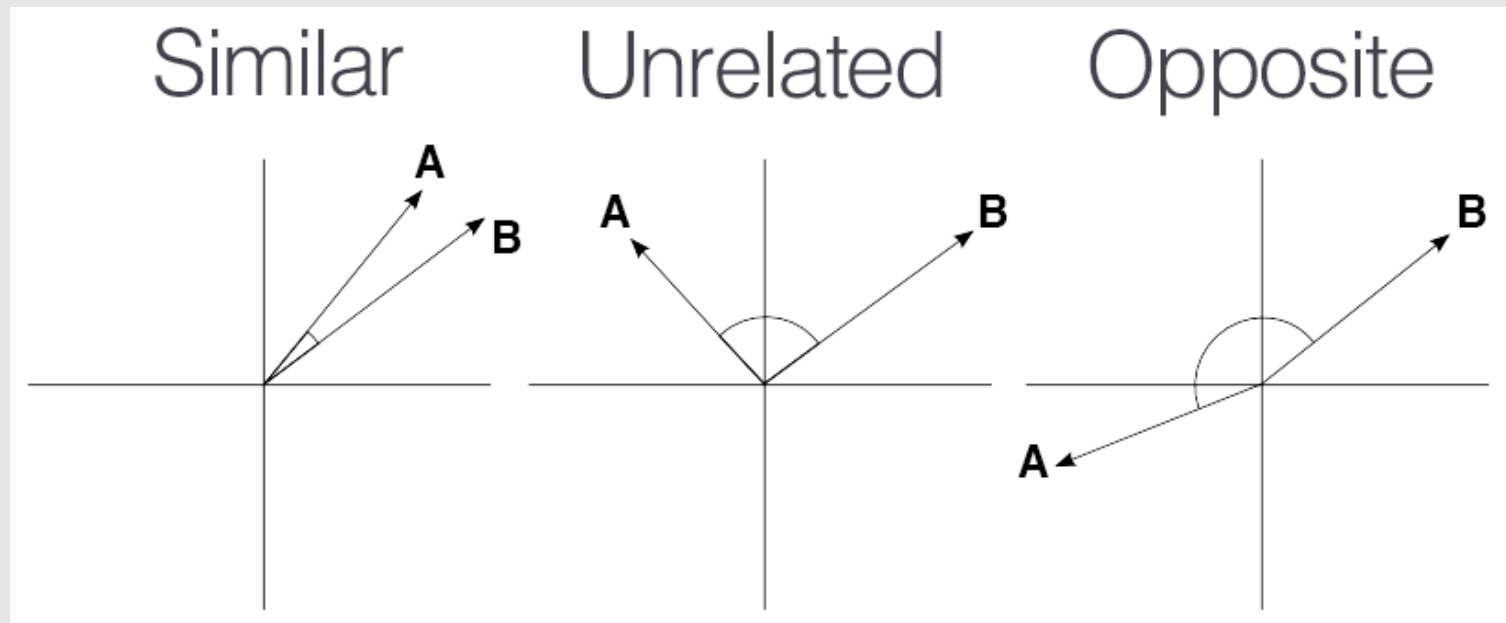
- Сбор вакансий с помощью API портала hh.ru;
- Сохранение массива вакансий для большей отказоустойчивости;
- Добавление описания опыта кандидата к массиву вакансий;
- Расчет косинусного расстояния для полученной матрицы;
- Ранжирование вакансий на основе рассчитанных косинусных расстояний;
- Отображение наиболее подходящих вакансий

Несколько подходов к переводу текста в векторный формат

- *Мешок слов*
 - *результат - разреженная матрица с «0» и «1»*
- *Схема TF-IDF.*
 - *Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.*
- *Word2Vec*
 - *использует методы глубокого обучения и нейронных сетей;*
 - *учитывает контекст предложения;*
 - *применяется для поиска похожих слов*

Косинусное расстояние

- Косинусное подобие - это метрика, которая измеряет косинус угла между двумя векторами, проецируемыми в многомерное пространство. Чем меньше угол между двумя векторами, тем они больше похожи друг на друга.*



Результаты выполнения

- В результате выполнения пользователю выводится список наиболее подходящих вакансий:

	name	employer.name	cos_similarity
81	Аналитик корпоративной риск-отчетности	Сбербанк	0.128459
98	Аналитик хранилищ данных	Сбербанк	0.116713
45	Руководитель направления аналитики (DWH)	Сбербанк	0.108375
87	Java-разработчик	Сбербанк	0.090008
1	Python Developer	Сбербанк	0.089354
10	Ведущий специалист Отдела бюджетирования и прогнозирования Управления по персоналу	Газпром нефть	0.086191
34	Разработчик 1C:ЗУП [id106401]	Газпром нефть	0.085036
2	Ruby middle developer (Platform V)	SberTech	0.083173
14	Product Owner в Platform V Dictionaries	SberTech	0.079612
13	Системный аналитик	Сбербанк	0.078823

Результаты выполнения

- Наиболее встречающиеся ключевые слова в результирующем наборе данных:

	опыт	данных	работы	разработки	знание	greenplum	задач	это	команда	субд	процессов	дмс	разработке	участие	команде
0	14	7	11	8	1	1	2	2	1	1	2	1	3	1	1
1	3	5	5	0	3	0	0	0	1	0	2	1	0	1	0
2	6	4	7	4	1	1	2	2	1	1	1	1	2	1	3
3	2	2	2	1	1	0	0	0	0	0	2	1	0	1	0
4	0	4	0	0	0	0	0	0	1	0	0	1	0	0	0
5	5	4	3	5	1	8	3	2	1	7	2	1	1	2	1
6	2	1	2	0	2	0	1	1	3	0	0	1	0	2	1
7	3	4	4	0	1	1	1	1	0	1	1	1	1	0	2
8	4	6	2	2	0	0	1	1	2	0	0	1	0	0	0
9	1	0	0	2	2	0	1	2	0	0	0	0	2	1	1

Используемые технологии

- *Python (pandas, requests, sklearn и др.)*
- *Для разработки использовался jupyter notebook*
- *Были реализованы функции сбора и ранжирования*

СПАСИБО ЗА ВНИМАНИЕ!