Здравствуйте, я Дмитрий Пасько, капитан команды АК. Мы разработали алгоритм решения первого кейса через ориентированный граф понятий.

Алгоритм базируется на идее, что одно понятие (навык) может быть выражено несколькими фразами и при этом быть связанным с другими понятиями. Например, фразы «люблю общаться» или «грамотная речь» относятся к навыку «коммуникабельность», а навык ведения переговоров связан как с коммуникабельностью, так и с умением убеждать.

Таким образом, сами навыки и фразы, которыми они выражаются, связаны друг с другом как узлы графа; и производя поиск на графе, можно реальные выражения переводить навыки.

Сначала производится обработка текста (вакансии или резюме): удаляются лишние символы, пустые строки, стоп-слова, ссылки и т. п. В результате от текста остаётся набор фраз, которые могут нести какой-то смысл.

После этого из каждой фразы выделяются униграммы и биграммы, производится их стемминг, то есть выделение основы.

Все найденные основы объединяются в некоторое облако информации, характеризующее текст. Из этого облака через поиск на графе мы можем выделить интересующие нас понятия, если они там присутствуют.

Для каждого понятия, представляющего интерес, происходит его поиск в графе понятий. В случае успеха от найденного понятия происходит движение по графу к его родителям и дальше, пока это возможно. Понятия, через которые прошло это движение, объединяются и выводятся как результат. Например, если в тексте встречается слово «эксель», движение по графу пройдет через понятия “MS Excel” и “MS Office”.

Вдобавок, поскольку используется стемминг, нет необходимость фиксировать все вариации слова с учётом падажей и частей речи. Слова, связанные с английским языком, из фраз "хорошо знаю английскИЙ", "читаю статьи на англискОМ", "владею англискИМ языком" будут сведены к одной и той же основе "англиск" и отнесены к навыку владения английским.

Также есть дополнительный алгоритм, использующий АПИ Википедии. Он будет применяться к фразам, с которыми не справился основной алгоритм (предположительно, это будут IT-термины). Поскольку этот функционал хорошо кешируется, можно время от времени фиксировать появление каких-то популярных навыков, с которыми не справилась основная модель, и добавлять их в основную модель. Таким образом, в процессе эксплуатации модели будет происходить её улучшение.

Для визуализации был разработан АПИ, который включает этот алгоритм как сервис + фронтенд-часть для визуализации.

Для улучшения этого проекта можно проделать следующие шаги:

* разработать систему, которая упрощает процесс создания графа понятий; в перспективе это — умный компилятор, указывающий на ошибки и неоднозначности, подсветка синтаксиса, компоновка файлов. Вдобавок, есть несколько идей, как упростить сам подбор понятий и т. д. Здесь нужны 1-2 человека, имеющих опыт с редакторами кода и создании интерпретаторов/компиляторов (но это если нужна будет действительно большая система).
* определять опыт с помощью нейронных сетей, но для этого нужен достаточно крупный (не меньше 10 000 наблюдений) набор размеченных резюме вида "резюме/вакансия" --> категория опыта (неизвестно, нет опыта, до 3 лет, от 3 до 6 и т. п.)

Необходимые ресурсы:

* Система уже сейчас выдает хорошие результаты, для расширения проекта требуются HR-специалисты (3-5 человек) из разных представляющих интерес областей (например, общепит, документооборот, строительство). Их задачей будет помощь в расширении графа понятий, которое приведет к более масштабной структуре проекта.