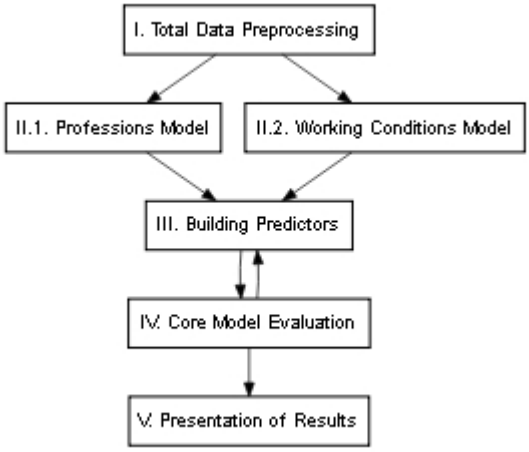
Алгоритм

Голощапова Ирина

26 августа 2015 г.

Общая картинка



I. Total Data Preprocessing

1. Вытащить данные и понять структуру

2. Определить, нужна ли вся выборка для анализа. Каким образом будет осуществляться проверка результатов?

3. Разбиение на train/test/.validation sets.

4. Разбивка профилей заявок работодателей на части (подразделы):

- должность
- обязанности
- требования (aka навыки)
- условия работы
- описание компании

5. Векторизация текста каждой части

6. Обработка и чистка:

- нижний регистр
- удаление союзов, предлогов, знаков препинания
- контроль “ч”, “ё”
- удаление стоп-слов и всякого мусора

II.1. Professions Model

Построение категориальной переменной профессий с помощью классификации раздела “Требования”. Unsupervised ML methods.

- строим
- смотрим есть ли какой-то смысл в кластерах при сопоставлении с реальными должностями
- выбираем оптимальное количество кластеров профессий

II.2. Working Conditions Model

Построение категориальной переменной условий работы с помощью классификации раздела “Условия работы”. Unsupervised ML methods.

Building Predictors.

Конструируем предикторы от головы. Для каждого подраздела профилей заявок работодателей.

В зависимости от качества подгонки моделй этапа II строим много или не очень предикторов, относящихся к разделу “Требования” и “Условия работы”.

1. **Само объявление**.

- на английском или нет
- tag профобласть (если есть в данных)
- ...

2. **Должность**.

- уровни: ведущий(lead), senior, junior
- manager, менеджер
- ...

3. **Обязанности**.

- управлять, управление...
- ...

4. **Требования (aka навыки)**.

- уровень английского: intermediate/upper-intermediate/fluent
- опыт работы
- топ-вузы
- высшее образование или нет
- ...

5. **Условия работы**.

- страховка/нет
- график работы: гибкий график/удаленная работа/полный день
- местоположение: Москва/МО/Санкт-Петербург/ЦФО/зарубеж/остальное
- оформление ТК РФ, белая зарплата/остальное
- соцгарантии, соцпакет/остальное
- тип занятости: стажировка/проект/полная/частичная/волонтер
- ...

6. **Описание компании**.

- на английском или нет
- стартап,"молодая, успешная компания"/остальное
- прямой работодатель/нет
- ...

IV. Core model evaluation

1. Чистка переменной Wage. Все ли данные в рублях. В зависимости от ответа: оценка 2 моделей, удаление выбросов, перевод в рубли.

2. Разбиение wage на диапазоны.

3. Сбор предикторов этапа II (models output variables) и III в таблицу данных.

4. Оценка модели supervised ML.
   4.1. Прогон разных алгоритмов МО с кросс-валидацией - предсказание на train.
   4.2. Оценка предсказания на test.
   4.3. Tuning top-3.
   4.4. Выбор лучшей модели.

V. Presentation of Results

- Если будет позволять время, могу сделать в Rmd. Если нет - что-то на коленке.