

Тестовое задание (Data Scientist)

Горизонтальная линия

# Описание задачи

## Jooble - сервис для поиска работы. Для решения многих задач для нас важно решить проблему репрезентации вакансии в виде набора разнотипных (числовых, строковых и т.д) факторов(признаков).

## В данном тестовом задании Вам требуется реализовать пакет по предобработке признаков вакансий.

# Входные данные

<https://drive.google.com/drive/folders/1lCVt_UtxwfLyu16rgmQm-ej68XztFEtN>

В директории python-dev-test/data находится два файла: test.tsv, train.tsv одного формата.Каждая строка файла является характеристикой для одной вакансии.

В файле две колонки:

* id\_job - (integer) идентификатор вакансии
* features - (string) конкатенация из признаков вакансии определенного типа (Объединенные через символ “,”).
  + Первый элемент списка - код набора признаков. (В данном примере “2”)
  + Остальные элементы - числовые характеристики данного типа. В рамках этой задачи это 256 integer чисел (итого каждый признак можно пронумеровать и название колонок может иметь вид : “feature\_2\_{i}”, где i - индекс элемента в массиве)(\*).

# Требования к выходным данным

В качестве тестирования к модулю должен быть прикреплен скрипт, который генерирует файл test\_proc.tsv который содержит следующий набор колонок (признаков) для каждой вакансии из test.tsv:

1. id\_job - (integer) идентификатор вакансии (размерность : 1);
2. feature\_2\_stand\_{i} - (double) результат стандартизации (z-score нормализация) входного признака feature\_2\_{i} (См. Входные данные(\*)) (размерность : 256);
   1. Ссылка на определение стандартизации :
      1. <https://ru.wikipedia.org/wiki/Z-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0>
      2. <https://en.wikipedia.org/wiki/Feature_scaling>
   2. Для проведения этой операции требуется оценить две статистики для каждой колонки feature\_2\_{i} на данных из файла train.tsv:
      1. mean(feature\_2\_{i}) - среднее значение по всем вакансиям для признака feature\_2\_{i};
      2. std(feature\_2\_{i}) - среднеквадратическое отклонение по всем вакансиям для признака feature\_2\_{i};
3. max\_feature\_2\_index - (integer) индекс i максимального значения признака feature\_2\_{i} для данной вакансии (размерность : 1);
4. max\_feature\_2\_abs\_mean\_diff - (double) - абсолютное отклонение признака с индексом max\_feature\_2\_index от его среднего значения mean(feature\_2\_{max\_feature\_2\_index}) (размерность : 1);

# Требования к реализации

1. Реализацию требуется произвести на Python 3.7
2. При проектировании модуля требуется учитывать следующие факторы:
   1. Размеры данных файла test.tsv может достигать нескольких миллионов строк
   2. При использовании модуля в будущем планируется добавление факторов новых типов (в нашем тесте только признаки типа “2”).
   3. При использовании модуля в будущем планируется добавление новых способов нормализации признаков (в нашем тесте только z-score нормализация).
3. Пакет требуется загрузить на github репозиторий.