

Лабораторная работа 1. Язык программирования Python.

Разрешённые библиотеки: вся стандартная библиотека python, numpy, scipy.sparse, requests или подобная для задания 5.

1. Дан массив вида [a,b,c,d,e,a,a,b,c]

а) Преобразуйте его в словарь вида

```
{  
  a: [0,5,6],  
  b: [1,7],  
  ...  
  
  e: [5]  
}
```

б) Сделайте это в одну строку кода (разрешается с меньшей эффективностью)

2. Метрика Жаккара (Jaccard similarity) - способ измерения сходства между множествами. Вычисляется как $\frac{|A \cap B|}{|A \cup B|}$, т.е. размер пересечения поделить на размер объединения множеств.

Реализуйте вычисление метрики не используя методов intersection и union типа set и не создавая промежуточных множеств.

UPD: использовать тип set. Не принимаются решения, делающие $|A| * |B|$ сравнений или похожие переборы.

3*. Minhash - техника быстрого приближенного вычисления метрики Жаккара.

Выбирается k хэш-функций.

Каждое множество S представляется как массив из k значений (Sk), где i-й элемент - это минимальное значение i-й хэш-функции на элементах множества S.

Пример для множества $S=\{a,b,c,d,e\}$ и $k=2$

$Sk[0] = \min(h_0(a), \dots, h_0(e)) = \min(14643, 8586553, 707604, 4431, 555768) = 4431$

$Sk[1] = \min(h_1(a), \dots, h_1(e)) = \min(1664231, 546045, 94382, 452566, 638422) = 94382$

Пусть множествам A и B соответствуют массивы из минимальных хэшей A_k и B_k .

Приблизительное значение метрики Жаккара - это пропорция равных соответствующих элементов $A_k[i]$ и $B_k[i]$.

Реализуйте алгоритм.

Сгенерируйте множество из 20000 уникальных строк.

Сгенерируйте 100 случайных его подмножеств размером 1000-15000.

Вычислите сходство между всеми парами подмножеств точным и приближительным методом. Сравните время выполнения и оцените точность приближения. Проверьте, как влияет на точность значение k.

4. Дан [json-файл](#) вида:

```
[
  {
    "item": "ITEM1",
    "sales_by_country": {
      "Russia": {
        "2010": 23,
        "2011": 40
        ...
        "2019": 300
      },
      "Ukraine": {
        ...
      },
      ...
    }
  },
  {
    "item": "ITEM2",
    "sales_by_country": {
      "Russia": {
        "2010": 43,
        ...
        "2019": 30
      },
      "Ukraine": {
        ...
      },
      ...
    }
  },
  ...
]
```

Преобразуйте данные в CSV-файл вида:

```
item, country, year, sales
ITEM1, Russia, 2010, 23
ITEM1, Russia, 2011, 40
...
ITEM2, Russia, 2010, 40
....
```

5. Используя API центрального банка (<https://www.cbr.ru/development/>) соберите данные по курсам валют (по отношению к рублю) с 1 марта 2020 по 1 июля 2020

Необходимые поля:

Дата

Доллар США

Евро

Японская иена

Украинская гривна

6*. Дан граф друзей из соцсети VK

(https://www.dropbox.com/s/bjoji6y47vwr90d/arcs_refined.txt?dl=0):

Необходимо:

1) Определить число уникальных пользователей

2) Для каждого id пользователя определить число его друзей из данного графа.

Вывести топ 15. Учитывать неориентированность графа.

3) Для всех возможных пар пользователей (декартова произведения):

Для длин L 1..6 определить пропорцию пар пользователей, для которой **кратчайший** путь между пользователями в паре составляет L .

Определить пропорцию, для которой этот путь больше 6 (*или не существует*).

Убедиться, что полученные пропорции суммируются в 1.

7*. Даны параллельные субтитры на русском и болгарском языке

(<https://drive.google.com/file/d/1ZRKBizPGILg6TRS4fgh55UKdeyhBabv2/view?usp=sharing>)

Определите наиболее часто встречающиеся двойки и тройки символов на русском и болгарском языке. Выделите двойки и тройки, наиболее различающиеся по частоте употребления между языками. Подумайте, как использовать эту статистику для определения принадлежности текста к тому или иному языку.