**Программирование в системах информационной безопасности  
 Лабораторный практикум**

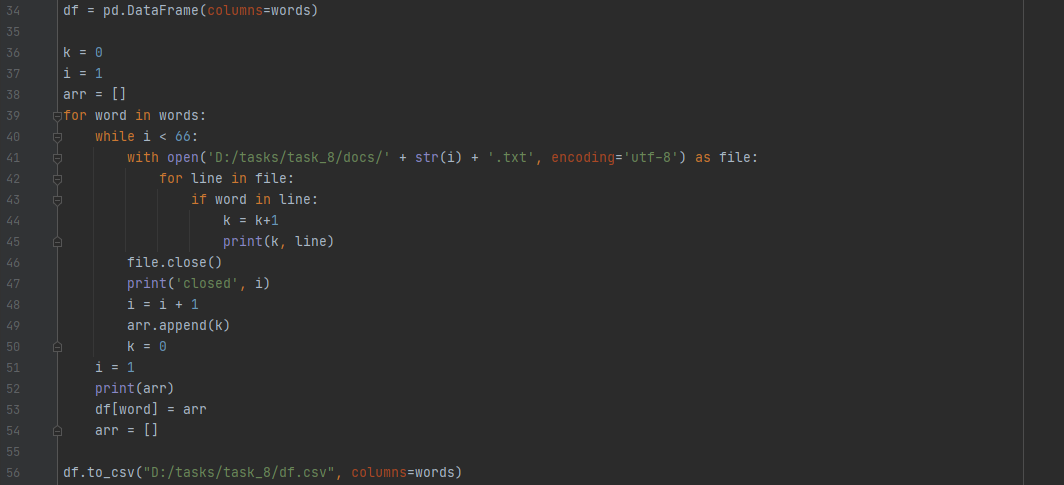
Васюткин Александр Олегович, группа 181-331

**Задание 8.**

1. Выработать словарь терминов (не менее ста), характерных для литературы по информационной безопасности.



1. Составить датасет.

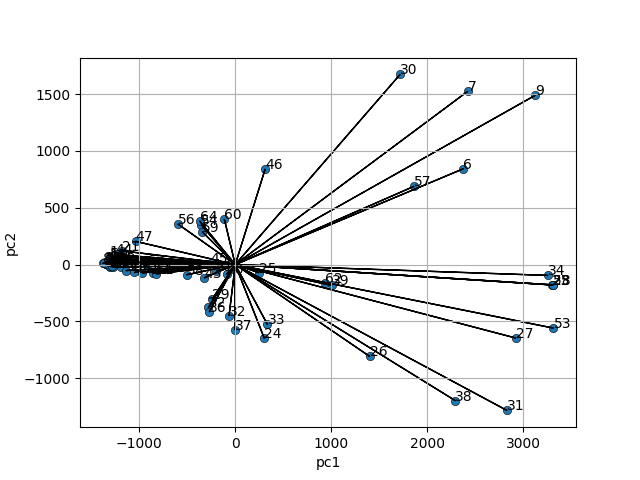


Поочередно перебирая слова, ищем их во всех документах. Если встретилось слово, прибавляем к счетчику 1. Когда программа переходит к рассмотрению следующего документа, значение счетчика записывается в массив, и счетчик обнуляется. Когда мы нашли, сколько раз встречается первое слово в каждой книге, мы заполняем первую строку DataFrame значениями массива (куда записывали значения счетчика). Когда DataFrame будет заполнен, мы сохраняем его в файл формата csv.

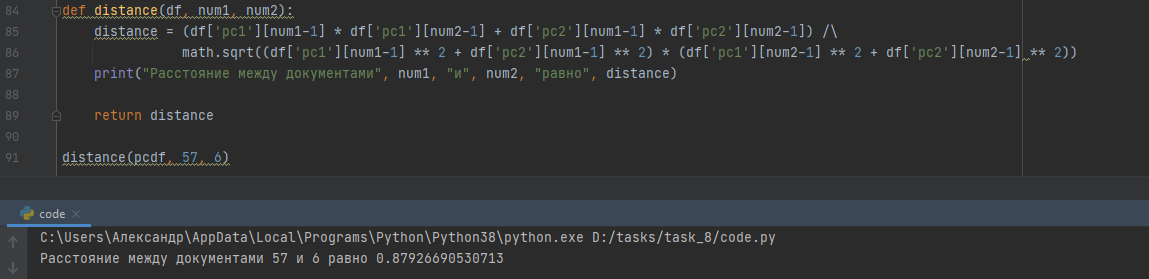
1. Применить к данным метод главных компонент и визуализировать данные в пространстве первых двух главных компонент: рядом с каждой точкой указать номер документа, от начала координат до каждой точки провести отрезок (радиус-вектор).



Применяем метод главных компонент, оставляя только две первые. Визуализируем данные в пространстве двух главных компонент. Строка 73 отвечает за отображение номера документа, а 74 – за отображение радиус-вектора.



1. Реализовать метод, вычисляющий расстояние между двумя произвольными документами (по номерам). За расстояние между документами следует принять косинус угла между соответствующими радиус-векторами в пространстве первых двух главных компонент.



Создаем функцию, которой будем подавать на вход наш DataFrame и два числа – номера документов. На выходе получаем расстояние, равное косинусу угла между этими документами.