1. Разработать на языке python (версии 3.6 и выше) в среде разработки PyCharm программу, выполняющую:
   1. Загрузку текстовых выборок
   2. Предварительную обработку текстовой информации (удаление стоп-слов, чисел, стемминг и т.п.)
   3. Разбиение выборок на обучающие и тестовые части (80% и 20% от общего числа текстов в выборке, соответственно)
   4. Классификацию текстов из тестовой части выборок двумя методами в соответствии с заданным вариантом из табл.1
   5. Провести сравнение полученных результатов по точности и быстродействию классификации.
2. Оформить отчет (титульный лист, формализованную постановка задачи, теоретические разделы описывающие выбранные методы и подходы для решения задачи, разделы описывающие результаты обработки текстовой информации и классификации с выводами, список использованной литературы, приложение с кодом программы). Объем отчета без приложений: 10-20 стр.

**Таблица 1. Исходные данные**

|  |  |  |
| --- | --- | --- |
| **Вариант** | **Выборки** | **Методы** |
| 21 | rus1, rus2, rus3 | Метод к-ближайших соседей, метод опорных векторов |

**Пояснения**

Имеется 15 выборок с названиями и аннотациями статей на английском и русском языке. В выборках статьи распределены по семи классам (англ. статьи) и 5 классам (русск. статьи). Названия классов перечислены в строке, начинающейся со слова «ClNames:», после которого через точку запятую перечислены названия классов. В строке, начинающейся со слова «CLASS:» указан порядковый номер класса статьи из списка названий «ClNames». В строке, начинающейся со слова «Title:» указано название статьи. В строке, начинающейся со слова «Abstract:» начинается аннотация статьи.

Студент имеет право настраивать параметры методов классификации для достижения наилучшей точности и использовать готовые библиотеки и функции.