**Лабораторная работа №1.  
Обработка текстовых данных**

## Регулярные выражения

**Задание 1 (0,5 балла).**

Найти в тексте запись об указании времени в формате чч:мм:сс. В заданном формате чч - целое число из диапазона от 00 до 23, мм и сс - целые числа из диапазона от 00 до 59 (если какая-то часть формата нарушена, то данная подстрока в качестве верного значения не рассматривается).

**Задание 2 (0,5 балла).**

Найти в тексте знаки препинания (точка, запятая, точка с запятой, двоеточие, тире («минус», длинное и короткое), вопросительный и восклицательный знаки, многоточие (три идущие подряд точки и отдельный символ «…» ) .

**Задание 3 (0,5 балла).**

В тексте могут содержаться IP-адреса компьютеров в формате d.d.d.d, где d - целое число из диапазона от 0 до 255. Найти все IP-адреса, содержащиеся в тексте.

**Задание 4 (0,5 балла).**

Проверить надежно ли составлен пароль. Пароль считается надежным, если он состоит из 8 или более символов. Пароль должен содержать хотя бы одну заглавную букву, одну маленькую букву одну цифру, хотя бы один специальный символ («!», «@», «#», «$», «%», «^», «&», «\*») и не является последовательностью («123456Qwerty», «Ps123456789», «Qwerty1»).

**Задание 5 (1 балл).**

Выделить в тексте прямую речь с опорой на пунктуацию (см. <http://www.gramota.tv/biblio/textbooks/gramota/punct/343-45-192?ysclid=lne281u9yq246756449>).

**Задание 6 (1 балл).**

Выделить в тексте предложения с обобщающими словами. Рассмотреть все варианты пунктуации. (Например: «Всякий день стал приносить старый Мосеич **разную крупную рыбу**: щук, язей, голавлей, линей, окуней.») (см. <http://gramota.ru/class/coach/punct/45_190?ysclid=lne2kwryum489092268>).

**Задание 7 (1 балл).**

Дан текст с географическими названиями населенных пунктов (г. Пермь, п. Скобелевка, …). Найти вхождения населенных пунктов:

* г - город;
* пгт - поселок городского типа;
* рп - рабочий поселок;
* кп - курортный поселок;
* к - кишлак;
* дп - дачный поселок;
* п - поселок;
* п. ст - поселок при станции (поселок станции);
* с - село;
* д - деревня;
* сл - слобода;
* ст - станция;
* ст-ца - станица;
* х - хутор;
* у - улус;
* рзд - разъезд;
* зим - зимовье.

**Задание 8 (1 балл)**

Решить задачи (20 заданий) на ресурсе <https://www.spoj.com/SHORTEN/problems/regex/>. (Ресурс доступен через VPN. В качестве результата в отчете представить снимок экрана и пояснить логику решения.)

# Отработка текстов

**Задание 9 (1 балл)**

Написать программу (Python, С# или любой другой язык программирования) для построения частотного словаря текста на русском языке без учета стоп-слов. Входной текст хранится в текстовом файле, построенный словарь также должен быть сохранен в текстовый файл в формате (словоформа, количество\_вхождений\_словоформы) с сортировкой по количеству вхождений. Для деления текста на токены и удаления стоп-слов использовать библиотеку nltk, а для нормализации использовать библиотеку Pymorphy2.

**Задание 10 (1 балл)**

Реализовать вторую версию программы (см. задание 1) построения частотного словаря, в которой для нормализации используется стеммер (найти необходимую библиотеку самостоятельно). Сравнить построенные частотные словари.

**Задание 11 (1 балл)**

Собрать корпус документов по выбранной теме и по основы программ, разработанных в заданиях 1 и 2, подсчитать метрику TF-IDF для ключевых слов документа.

**Задание 12 (1 балл)**

Реализовать алгоритм автоматического реферирования (квазиреферирование) на основе статистического подхода, алгоритм приведен ниже. Входные данные: исходный текст и коэффициент сжатия. Единицей реферирования (фрагментом) должно являться предложение. Выходные данные: список ключевых слов с весами, список предложений с весами, текст реферата.

**Алгоритм**:

1. Разбить текст на предложения.
2. Разбить текст на слова произвести их нормализацию.
3. Удалить стоп-слова.
4. Подсчитать веса слов (tf или tf-idf).
5. Определить веса предложений, рассчитанный как сумма весов, входящих в предложение слов.
6. Отсортировать предложения по убыванию веса.
7. В отсортированном списке оставить те предложений, которые входят в задаваемый процент сжатия. Например, если процент сжатия 10%, а в исходном тексте 50 предложений, то в реферате будет 5 предложений.
8. Сформировать текст реферата из отобранных на предыдущем шаге предложений в порядке, котором они встретились в исходном тексте.