21. Регулярні вирази

Т21.1 У текстовому файлі є дати, задані у форматі dd.mm.yyyy або у форматі yyyy/mm/dd. Також день та/або місяць може містити одну цифру, а не 2. Привести всі дати до єдиного формату dd.mm.yyyy.
Вказівка: використати регулярні вирази та функцію (метод) sub.

Т21.2 За допомогою регулярних виразів розбити текст у текстовому файлі на речення.

Т21.3 У текстовому файлі, що містить текст українською мовою, є дати (або сторіччя), записані римськими цифрами. Виділити всі такі числа та перевірити їх правильність. Число, записане римськими цифрами, використовує символи M (1000) D (500) C (100) L (50) X (10) V (5) I (1). При цьому, 900 записують як CM, 400 – CD, 90 – XC, 40 – XL, 9 – IX, 4 – IV. Наприклад, 1984 записується MCMLXXXIV.
Підказка: використати два шаблони регулярних виразів: один – для виділення числа, інший – для перевірки правильності.

Т21.4 У текстовому файлі містяться дати у форматі dd.mm.yyyy або підкреслення для запису дат вручну __.__.____ Знайти всі дати у тексті. Замість підкреслень вставити поточну дату. Зберегти оновлений текст.

Т21.5 У текстовому файлі містяться дати у форматі dd.mm.yyyy або yyyy-mm-dd або yyyy/mm/dd
Привести дати до єдиного формату за допомогою шаблону(ів), що містить(ять) іменовані підгрупи.

Т21.6 У текстовому файлі міститься переписка декількох осіб електронною поштою. Скласти список контактів (адрес електронної пошти) осіб, що фігурують у даній переписці. Використати регулярні вирази.

Т21.7 У текстовому файлі міститься текст українською мовою. «Стиснути» цей текст, видаливши у словах усі голосні літери. Якщо у слові тільки голосні літери або слово має довжину не більше 2 символів, - голосні не видаляти. Використати регулярні вирази.

Т21.8 У текстовому файлі містяться, крім іншої інформації, дійсні числа у форматі з фіксованою крапкою. При цьому, частина чисел не містить 0 перед крапкою, якщо число менше 1 за модулем (.253) або після крапки, якщо число ціле (5891.). Виділити всі дійсні числа, записані у файлі, та вставити 0 у тих числах, у яких його немає. Зберегти оновлений текст.

Т21.9 У текстовому файлі записано переписку у чаті, яка окрім слів містить «смайли». Виділити всі слова та всі смайли та порахувати «індекс емоційності» переписки, як відношення кількості смайлів до кількості слів.

Т21.10 В умовах задачі Т21.9 розділити смайл за емоційністю на «просто емоційні» та «винятково емоційні». Просто емоційним смайлам присвоїти коефіцієнт 1, а винятково емоційним, - 2. Порахувати «індекс емоційності» переписки, враховуючи ці коефіцієнти.

Т21.11 Так звана «олбанська мова» - це інтернет-діалект російської мови з навмисним перекручуванням звичайних слів. У деяких словах окремі дзвінкі приголосні замінюються глухими (аффтар) або ‘в’ на ‘фф’, глухі, - дзвінкими (йазыг), ‘а’ на ‘о’, ‘о’ на ‘а’, ‘я’ на ‘йа’, ‘и’ на ‘ы’, ‘и’ на ‘е’. У текстовому файлі записано текст російською мовою. На основі цього тексту за допомогою регулярних виразів побудувати текст «олбанською мовою». Для побудови проводити від 1 до 2 замін у частині слів. Випробувати заміну 50%, 40%, 30%, 20% слів. Запропонувати додатково власні правила заміни.

Т21.12 За допомогою регулярних виразів проаналізувати синтаксичну правильність простих арифметичних виразів, що містять числа, знаки операцій, дужки. Наприклад, 2 – 57*(33 + 25/4).

Т21.13 У текстовому файлі зберігається текст лінійної програми у Python. У програмі використовується тільки команда присвоєння. У правій частині присвоєнь розташовані арифметичні вирази, що використовують арифметичні операції та містять константи та змінні. За допомогою регулярних виразів побудувати словник змінних програми (у формі <ім’я>:<значення>). Перевірити, чи є зміні, що використовуються раніше, ніж визначаються.

Т21.14 У текстовому файлі зберігається стаття. Треба скласти програму, яка перевіряє, чи має ця стаття позитивну чи негативну спрямованість. Про спрямованість статті свідчить використання слів: при позитивній спрямованості кількість позитивних слів суттєво перевищує кількість негативних. Використати словник, що складається з позитивних та негативних слів. Побудувати регулярні вирази для знаходження різних словесних форм позитивних та негативних слів.

Т21.15 Ви працюєте у компанії, що надає послуги. Компанія має багато філій. Робітники на місцях зібрали дані про боржників компанії. Усі дані містять прізвище, ім’я та по-батькові, адресу боржника, його телефон та борг. Але порядок слідування інформації у різних філій – різний. Так само, розрізняється формат окремих частин інформації. Наприклад, адреса може починатися словом «адреса» або «адр.» у верхньому або нижньому регістрі.

Так само, для багатоквартирних будинків може бути вказано «кв.» та номер квартири, а для приватних будинків номер квартири не вказаний. Замість прізвища, ім’я та по-батькові може бути надане прізвище та ініціали. Усю інформацію зібрано у текстовий файл. Інформація про кожного боржника займає 1 рядок файлу.
Всього боржників порядку 10 000. Ваше керівництво доручило Вам до наступного ранку підготувати до відправки листи всім боржникам такого змісту:

«<Адреса>
Шановна(ий) ____<П.І.Б>_____
Сума Вашого боргу за послуги складає <Борг>.
Просимо сплатити борг протягом місяця. У іншому випадку, надання послуг буде припинено.»

Відомо , що адреса починається з одного зі слів вище, телефон починається з слів «телефон» або «тел.» (літери можуть бути у верхньому або нижньому регістрі. Частини інформації розділяються одним або декількома пропусками, сума боргу йде останньою у рядку.
Використайте регулярні вирази для розв’язання задачі та складіть програму для формування листів та запису їх у текстовий файл.

Т21.16 В умовах завдання Т21.15 Вам доручили також терміново надати список телефонів 100 найбільших боржників для їх подальшого інформування телефоном.
Список повинен мати формат:
<П.І.Б> <Телефон> <Борг>
Використайте регулярні вирази для розв’язання задачі та складіть програму для формування списку.

Т21.17 Ви працюєте у департаменті інформаційної безпеки компанії. Відомо, що працівникам компанії заборонено на роботі користуватись рядом сайтів (є список адрес заборонених сайтів). На запит вашої компанії інтернет-провайдер надав інформацію про відвідування сайтів співробітниками за декілька місяців. Інформація – у вигляді великого текстового файлу, у якому, зокрема, є трійки:
<адреса комп’ютера> <відвіданий сайт> <дата та час>
<адреса комп’ютера> - рядок у форматі XXX.XXX.XXX.XXX, де X – цифри (між крапками може бути й менше 3 цифр.
<відвіданий сайт> - рядок у форматі http:// <адреса сайту>
<дата та час> - рядок у форматі dd.mm.yyyy hh:mm:ss
У Вас є список співробітників разом з адресами їх комп’ютерів.
Вам потрібно підготувати звіт щодо відвідування співробітниками заборонених сайтів, впорядкувавши його за кількістю відвідувань.

Використайте регулярні вирази для розв’язання задачі та складіть програму для формування списку.

Т21.18 У текстовому файлі зберігається авторська стаття. Треба скласти програму, яка аналізує цю статтю та виділяє найбільш часто використовувані слова, а також словосполучення з 2 та 3 послідовних слів.
Порівняти результат цієї статті з результатами для інших статей цього ж автора, а також для статей інших авторфів.
Використати регулярні вирази.