

Задача 4

Да се напише програма, която позволява въвеждане име на файл от клавиатура заедно с разширението, след което намира всяка уникално срещаща се дума /токен/ в съответния файл и ги подрежда в нов отделен файл, озаглавен "xxxx_tokenized.txt" (където xxxx е името на файла за обработка без разширението). На всеки ред се пази съответната извлечена дума /токен/ и честотата ѝ на срещане. Файлът за обработка може да съдържа максимум 1000 символа.

Допълнителни указания:

Приемаме за специални символи всички тези, които не са част от английската азбука. Всеки ред от файла е разделен с "Enter".

В изходния файл форматът за запазване е:

токен_1 5

токен_2 9

...

токен_ *N* 17

Важно уточнение!

go \neq **goes** (различен смисъл носят двете думи – третира се като отделни)

fly (муха) = **fly (летя)** (имат различна семантика спрямо контекст, но тук се интересуваме от състава на думата)

Mountain = **mountain** (токените се пазят с малки букви в изходния файл)

Изисквания:

- Да е позволено въвеждането на име на файл за обработка
- Програмата да извежда в изходния файл форматът, посочен в допълнителните указания
- Програмата да извежда общият брой извлечени токени от подадения файл
- Програмата да позволява въвеждане на число *N* (не по-голямо от 300) след което да изведе първите *N* най-често срещащи се токени.

Бонус точки се дават, ако програмата може да обработва файлове без ограничение на размера.

Отнемат се точки за:

- Неоригинален написан код
- Неправилно обработване на файл
- Неправилна работа с файл
- Несъобразяване с допълнителните указания
- Липса на някое горепосочено изискване