

**Word Lister** [X]

Source: Dickens, Charles ! A Tale of Two Cities.txt ...

Output: output.txt ...

Stop list: stop\_words.txt ...

Word size	Word break	Stop words	Output	Method
Min: 1	<input checked="" type="checkbox"/> Apostrophe	<input type="radio"/> None	<input checked="" type="checkbox"/> Word size	<input type="radio"/> Unconcern
Max: 255	<input checked="" type="checkbox"/> Hyphen	<input type="radio"/> Internal list	<input checked="" type="checkbox"/> Word count	<input type="radio"/> Strong
	<input checked="" type="checkbox"/> Underscore	<input checked="" type="radio"/> External file	<input type="checkbox"/> Case insensitive	<input checked="" type="radio"/> Rigorously

(c) gazlan@gmx.de 2011

Go!

# **Word Lister (c) gazlan@gmx.de 2011**

## *English version*

Программа предназначена для составления списка уникальных слов, встречающихся в файле. Обрабатываются файлы любых форматов (включая бинарные).

Под словом при этом понимается непрерывная последовательность латинских букв (в обоих регистрах) в диапазоне длин от MIN до MAX (зависит от настроек).

Три специальных символа: Апостроф, Подчеркивание и Тире (и их "типографские" варианты а la M\$word), в зависимости от настроек, трактуются либо как часть слова (считаются за латинскую букву) либо как разрыв между словами.

Дополнительно, может быть использован фильтр т.н. "стоп-слов" (часто встречающихся слов, таких как предлоги, союзы, междометия и т.д.), нерелевантных содержанию документа и бесполезных для поискового запроса, встроенный в программу, либо использующий слова из специально сформатированного файла (Word List).

В программе реализовано три метода обработки:

### **1. Unconcern**

Наиболее быстрый. Для проверки слов на уникальность вычисляется хэш-функция высокой диффузностью. Для хранения вычисленных хэшей используется вращающийся файл.

### **2. Strong**

Медленный. Метод аналогичен предыдущему, но использована Strong Crypt function, гарантирующая отсутствие коллизий (теоретическая оценка вероятности коллизии -  $1 / 2^{80}$  - пренебрежимо малая величина).

### **3. Rigorously**

Самый медленный. Хэши вычисляются так же, как в предыдущем методе, но дополнительно, создается временный файл статистики для подсчета общего количества вхождений для каждого слова.

Файл отчета создается в стандартном формате Word List (word per line). По необходимости, дополнительно может быть выведена длина слова и общее количество вхождений (только для метода Rigorously).