# АОТ, Отчет, ДЗ-2. Карпов Алексей, 402

Была написана программа, которая выводит статистические характеристики текстов в консоль. Программа написана на JavaScript и протестирована под Windows и Linux. Использовался морфопроцессор mystem.

1. В исходном тексте все символы переноса строки \n заменяются на пробелы.
2. Mystem обрабатывает текст, возвращая отчет в формате json.
3. Отчет обрабатывается программой, статистика выводится в консоль.

## Инструкция по запуску программы

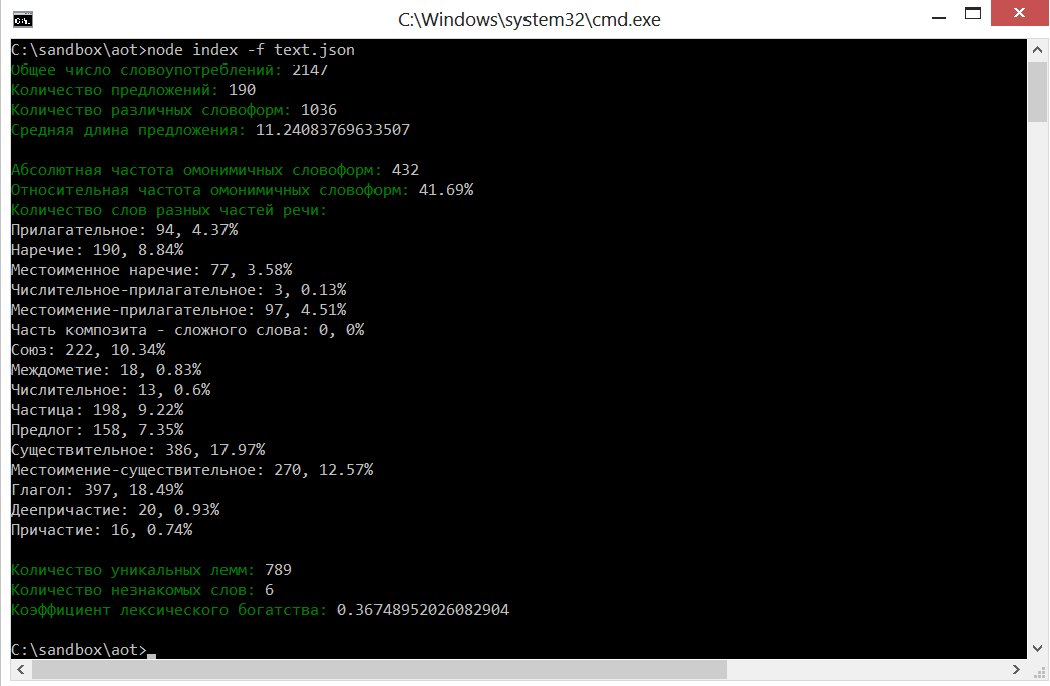
Программа доступна для загрузки: <https://github.com/cherurg/aot>

О том, как генерировать json, можно понять из флагов mystem: https://tech.yandex.ru/mystem/doc/index-docpage/

После того, как был сгенерирован json, можно запустить программу. Понадобится nodejs.

1. Установка
   * Windows: скачать и установить <http://nodejs.org/>
   * Linux: sudo apt-get install nodejs, sudo apt-get install npm
2. Перейти в папку проекта. Выполнить команду npm install
3. Запуск
   * Windows: node index -f <filename>
   * Linux: nodejs index -f <filename>

Будет выведена статистика как на скриншоте



## Особенности формата json, генерируемого mystem

Строго говоря, файл, который генерирует mystem при указании опции –format json не является файлом в формате json. Он состоит из литерала [], внутрь которого заключены все словоупотребления исходного текста.

В случае если в исходном тексте встречаются символы переноса строки, то в json появляются новые пустые литералы [] для каждого переноса строки. По этой причине символы ‘\n’ заменялись на пробелы.

Литерал [] был заключен в литерал {} с указанием имени “o”. Таким образом сгенерированный файл стал файлом в формате json.

Программа-обработчик принимает в качестве аргумента под флагом -f имя файла, который нужно разобрать. Флаг -f с указанием имени файла обязателен.

## Краткое описание работы программы

Основная обработка происходит в файле mystemAnalyser.js. Сначала в него передаются данные (метод setData), а затем поочередно вызываются функции, рассчитывающие статистики. Ниже примеры двух методов.

Метод для подсчета количества словоупотреблений в тексте:

var sentenceEndSymbol = '\\s';

var wordUsages = us.memoize(function () {

my.usagesCounter = 0;

if (!my.sentencesCounter) {

my.sentencesCounter = 0;

}

us.each(my.data, function (el) {

if (el.text !== sentenceEndSymbol) {

my.usagesCounter++;

} else {

my.sentencesCounter++;

}

});

return my.usagesCounter;

});

Метод для подсчета словоформ:

function uniqueElements(arr) {

var set = {};

us.each(arr, function (el) {

set[el] = true;

});

return us.size(set);

}

var wordforms = us.memoize(function () {

var arr = us.map(my.data, function (el) {

return el.text;

});

return uniqueElements(arr);

});

## Отчет по второй главе «Алисы»

**Общее число словоупотреблений:** 2147

**Количество предложений:** 190

**Количество различных словоформ:** 1036

**Средняя длина предложения:** 11.24083769633507

**Абсолютная частота омонимичных словоформ:** 432

**Относительная частота омонимичных словоформ:** 41.69%

**Количество слов разных частей речи:**

Прилагательное: 94, 4.37%

Наречие: 190, 8.84%

Местоименное наречие: 77, 3.58%

Числительное-прилагательное: 3, 0.13%

Местоимение-прилагательное: 97, 4.51%

Часть композита - сложного слова: 0, 0%

Союз: 222, 10.34%

Междометие: 18, 0.83%

Числительное: 13, 0.6%

Частица: 198, 9.22%

Предлог: 158, 7.35%

Существительное: 386, 17.97%

Местоимение-существительное: 270, 12.57%

Глагол: 397, 18.49%

Деепричастие: 20, 0.93%

Причастие: 16, 0.74%

**Количество уникальных лемм:** 789

**Количество незнакомых слов:** 6

**Коэффициент лексического богатства:** 0.36748952026082904

## Второй отчет

Вторым текстом была глава из романа Вернора Винджа «Пламя над бездной»

**Общее число словоупотреблений**: 1798

**Количество предложений**: 192

**Количество различных словоформ:** 1003

**Средняя длина предложения**: 9.316062176165824

**Абсолютная частота омонимичных словоформ**: 484

**Относительная частота омонимичных словоформ**: 48.25%

**Количество слов разных частей речи:**

Прилагательное: 116, 6.45%

Наречие: 142, 7.89%

Местоименное наречие: 42, 2.33%

Числительное-прилагательное: 5, 0.27%

Местоимение-прилагательное: 91, 5.06%

Часть композита - сложного слова: 0, 0%

Союз: 142, 7.89%

Междометие: 4, 0.22%

Числительное: 22, 1.22%

Частица: 89, 4.94%

Предлог: 185, 10.28%

Существительное: 459, 25.52%

Местоимение-существительное: 153, 8.5%

Глагол: 304, 16.9%

Деепричастие: 42, 2.33%

Причастие: 46, 2.55%

**Количество уникальных лемм**: 733

**Количество незнакомых слов**: 0

**Коэффициент лексического богатства**: 0.40767519466073415

## Вывод

Mystem выводит информацию в виде, который не очень удобен для машинного чтения. Во-первых, имеется описанная проблема с файлами json. Во-вторых, описание каждой леммы является строкой, в которой через запятую перечислены все дескрипторы слова. В примере обратите внимание на содержимое литерала “” после “gr”:

{"lex":"чудеситься","qual":"bastard","gr":"V,несов,нп=непрош,ед,изъяв,3-л"}

Такое представление информации потребовало дополнительной обработки результатов, сгенерированных mystem.