ИНСТИТУТ КРИПТОГРАФИИ, СВЯЗИ И ИНФОРМАТИКИ

АКАДЕМИИ ФЕДЕРАЛЬНОЙ СЛУЖБЫ БЕЗОПАСНОСТИ

РОССИЙСКОЙ ФЕДЕРАЦИИ

КАФЕДРА 714

МЕТОДЫ АНАЛИЗА ДАННЫХ.

Лабораторная работа №3.1 по теме:

«Изучение свойств массивов текстов и методов задания

весов информационных признаков»

**Выполнил:**

слушатель гр. №7153

ряд. Братухин И.М.

МОСКВА

2017г.

# Постановка задачи

Цель работы – изучение основных характеристик массивов текстов и методов вычисления весов информационных признаков.

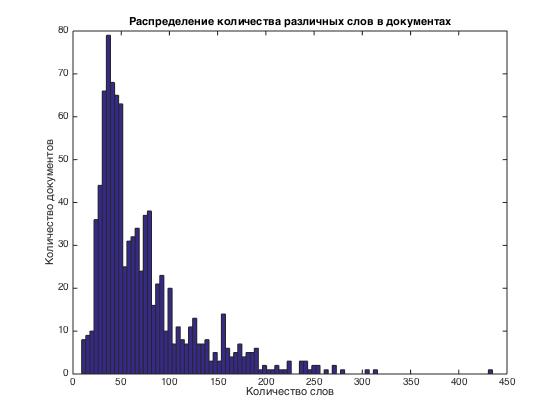
# Исходные данные

Исходные данные – индекс массива документов, содержащий информацию о частотах встречаемости слов в текстах и распределении текстов по рубрикам.

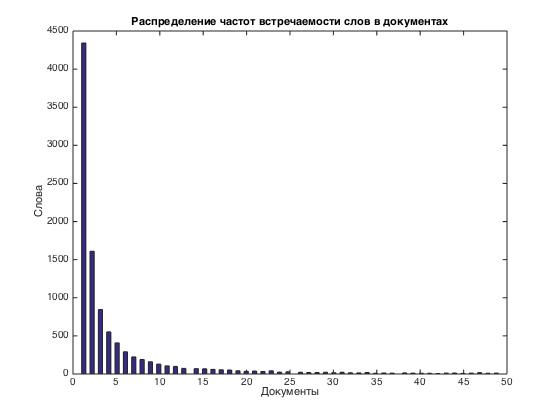
|  |  |  |
| --- | --- | --- |
| **№** | **Характеристика** | **Значение** |
| 1 | Название массива текстов | Reuters |
| 2 | Общее число различных слов во всех текстах | 10167 |
| 3 | Количество текстов в массиве | 935 |
| 4 | Количество рубрик (количество текстов в отдельных рубриках) | 6 |

# Гистограммы

Гистограмма распределения различных слов в документах:



Гистограмма распределения частот встречаемости слов в документах:

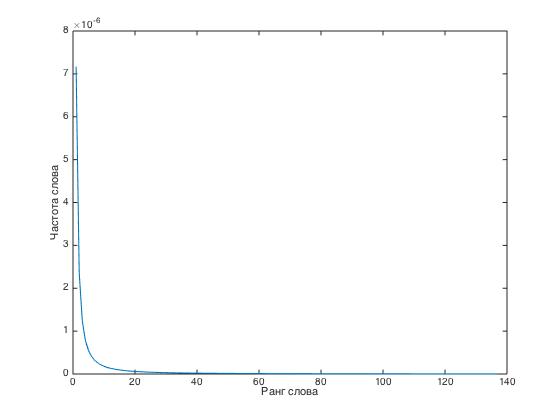


# Закон Ципфа

Закона Ципфа для частоты встречаемости слов в документе утверждает, что частота встречаемости слова в тексте обратно пропорциональна его рангу (порядковый номер в возрастающей последовательности частот слов в тексте)

Выполнение этого закона наглядно видно на гистограмме распределения частот встречаемости слов в документах. Это можно проверить, построив график закона Ципфа.

В нашем случае:



# Описание наборов слов с наибольшим весом для каждой рубрики

Приведем по 10 слов с наибольшим весом для каждого класса.

Вес признаков будем рассчитывать по следующей формуле:

пусть j, где – класс j, – частота встречаемости признака i в тексте j, n – количество слов в словаре, – количество классов.

тогда вес признака i в классе j равен

Таблица 1

|  |  |  |
| --- | --- | --- |
| **№ кластера** | **Характерные слова** | **Вес** |
| 1 | 'GASOLINE'  'FUEL'  'OIL'  'OCTANE'  'PETROLEUM'  'CRUDE'  'DISTILLATE'  'REFINERY'  'UNLEADED'  'ENDED' | 0.0106  0.0052  0.0043  0.0041  0.0041  0.0035  0.0033  0.0032  0.0032  0.0032 |
| 2 | 'GROWTH'  'ECONOMY'  'ECONOMIC'  'GROSS'  'PRODUCT'  'GDP'  'INFLATION'  'GNP'  'REAL'  'FORECAST' | 0.0065  0.0048  0.0045  0.0044  0.0036  0.0034  0.0031  0.0029  0.0028  0.0026 |
| 3 | 'GOLD'  'OUNCES'  'OUNCE'  'TON'  'MINING'  'ORE'  'SILVER'  'MINES'  'DEPOSIT'  'PROPERTY' | 0.0174  0.0085  0.0066  0.0052  0.0050  0.0048  0.0045  0.0034  0.0029  0.0029 |
| 4 | 'GAS'  'NATURAL'  'CUBIC'  'FEET'  'OIL'  'ENERGY'  'OFFSHORE'  'BARRELS'  'PIPELINE'  'HOUSTON' | 0.0093  0.0069  0.0054  0.0035  0.0031  0.0025  0.0023  0.0022  0.0020  0.0020 |
| 5 | 'SHIPPING'  'SHIPS'  'GULF'  'SHIP'  'PORT'  'IRAN'  'VESSELS'  'IRANIAN'  'ATTACK'  'KUWAITI' | 0.0051  0.0040  0.0032  0.0032  0.0030  0.0030  0.0029  0.0026  0.0022  0.0022 |
| 6 | 'SUGAR'  'TENDER'  'WHITE'  'LONDON'  'TRADERS'  'EUROPEAN'  'CROP'  'CANE'  'INTERVENTION  'COMMUNITY' | 0.0134  0.0031  0.0030  0.0029  0.0028  0.0026  0.0022  0.0022  0.0022  0.0021 |

# Проведение эксперимента

Для оценки влияния задания весов признаков на ошибку классификации проведем классификацию исходного массива текстов с использованием различных методов задания весов признаков.

Для вычисления весов признаков были выбраны три метода:

* Smart “ltc”
* Простая вероятность
* Log-Entropy

|  |  |  |
| --- | --- | --- |
| **Параметры эксперимента** | | |
| **1** | Соотношение размеров обучающего и тестового множества (без кросс-валидации) | 655/280 (70% / 30%) |
| **2** | Тип создаваемого классификатора | k-ближайших соседей (k=7, Евклидово расстояние) |
| **3** | Размерность входного пространства признаков | 10167 |
| **4** | Количество классов | 6 |

# Результаты эксперимента

Ошибка классификации:

|  |  |
| --- | --- |
| **Способ вычисления весов** | **Ошибка классификации** |
| SMART «ltc» | 0.0679 |
| Простая вероятность | 0.2857 |
| Log-Entropy | 0.0821 |

Макро точность и макро полнота классификации:

|  |  |  |
| --- | --- | --- |
| **Способ вычисления весов** | **Макро точность** | **Макро полнота** |
| SMART «ltc» | 0.9241 | 0.9149 |
| Простая вероятность | 0.7792 | 0.7012 |
| Log-Entropy | 0.9070 | 0.9021 |

Микро точность и микро полнота классификации:

|  |  |  |
| --- | --- | --- |
| **Способ вычисления весов** | **Микро точность** | **Микро полнота** |
| SMART «ltc» | 0.9321 | 0.9321 |
| Простая вероятность | 0.7143 | 0.7143 |
| Log-Entropy | 0.9179 | 0.9179 |

F-мера

|  |  |
| --- | --- |
| **Способ вычисления весов** | **F-мера** |
| SMART «ltc» | 0.9175 |
| Простая вероятность | 0.6873 |
| Log-Entropy | 0.9043 |

# Вывод

В результате рассмотрения и анализа массива Reuters было выделены наиболее характерные представители 6 классов.

Была проведена классификация исходного массива с использованием трех различных вариантов задания весов. Оценки результатов классификации показывают, что наилучшие показатели полноты и точности и F-меры имеет вариант задания весов SMART “ltc”.