Метод ACSI-Matic основывается на том, что представительными являются слова, частота встречаемости которых превосходит среднюю частоту встречаемости слов в документе.

# Общий принцип работы

Для определения объёма реферата общее количество предложений начального текста делится на 10.

Предложения с большими весами подлежат включению в реферат, а со средним − помечаются как «резервные».

При изучении избыточности в предложениях вводился следующий критерий: если число встретившихся в двух предложениях синонимов превышало 25 % от общего количества слов в предложении, то такие предложения считались избыточными и вычеркивались. В этом случае для реферата выбирались предложения из резерва.

Этот процесс длится до тех пор, пока не устранятся избыточные или не закончатся «резервные» предложения.

Общая схема реализации алгоритма

1. Разбить текст на предложения.
2. Определить выходной размер автореферата.

Задается путем подсчета 10 процентов от общего количества предложений в исходном тексте.

1. Назначить каждому слову коэффициент значимости.

Для этого строится матрица частоты вхождения каждого слова в общий текст, то есть TF-матрица.

1. Определить значимость каждого предложения.

Для этого строится IDF-матрица – обратная частота вхождения каждого слова в общем тексте. Необходима для уменьшения веса широкоупотребительных слов. Затем строится матрица TF-IDF, демонстрирующая вес каждого предложения нашего текста.

1. Разбить предложения на две группы: «реферат» и «резерв» по весовому признаку.

Для этого необходимо найти границы разбиения на группы. А именно: между максимальным весом и средним – верхняя граница. Между минимальным и средним – нижняя граница.

Затем в ходе распределения, в группу «реферат» отправляются предложения, вес которых входит в диапазон от верхней границы до максимума, а в «резерв» предложения, вес которых находится в диапазоне от минимума до нижней границы.

1. Если полученный набор предложений из группы «реферат» удовлетворяет заданному объему, то алгоритм завершает свою работу и выводит результат.
2. Если полученный набор предложений из группы «реферат» меньше заданного объема, то мы находим количество недостающих предложений (k) и добавляем из «резерва» k-предложений с максимальными весами в группу «реферат».
3. Если полученный набор предложений из группы «реферат» больше заданного объема, то генерируется матрица количества синонимов для каждого слова в предложениях.

Затем по этой матрице из «реферата» исключаются предложения, содержащие синонимы.

1. Если этого было недостаточно, и объем «реферата» остался больше необходимого, то он отправляется в начало алгоритма и проделывает все шаги выше до полного сжатия.
2. Если объем «реферата» стал меньше необходимого, то по аналогии с пунктом 7, «реферат» отправляется на добавление предложений из «резерва».

# Реализация программы

## Библиотеки:

* import math

Этот модуль обеспечивает доступ к математическим функциям, в частности к вычислению логарифма в матрице IDF

* from nltk import sent\_tokenize, word\_tokenize, SnowballStemmer

NLTK — это ведущая платформа для создания программ для работы с данными человеческого языка.  С этой платформы будет использовано большое количество библиотек и модулей, а именно:

* sent\_tokenize - токенизация предложения: мы используем метод, чтобы разделить документ или абзац на предложения.
* word\_tokenize - Токенизация слов: мы используем метод для разделения предложения на токены или слова.
* SnowballStemmer - удаляет аффиксы из токена.
* nltk.corpus - модули в этом пакете предоставляют функции, которые можно использовать для чтения текстов в различных форматах.
* оттуда же stopwords - Стоп-слова — это общеупотребительные слова, такие как «то», «и», «я» и т. д., которые очень часто встречаются в тексте и поэтому не передают понимание конкретной темы документа.
* from string import punctuation - это предварительно инициализированная строка, используемая в качестве строковой константы. Дает все наборы пунктуации.
* import heapq – в данном модуле необходима Функция [nlargest()](https://docs-python.ru/standart-library/modul-heapq-python/funktsija-nlargest-modulja-heapq/" \o "Функция nlargest() модуля heapq в Python.), которая возвращает список с n самыми большими элементами из набора данных.
* from wiki\_ru\_wordnet import WikiWordnet - это семантическая сеть типа WordNet для русского языка, составленная из данных русского Викисловаря. Выдает синонимы заданных слов.