

# Frequently Based Chunking прототипирование

# Цель и задачи

Цель:

1) Анализ прототипа FBC алгоритма

Задачи:

1) Построение простейших словарей перебором

2) Поиск кандидатов на чанкирование из словаря

3) Новое чанкирование

# Особенности реализации

- 1) Выделение чанков фиксированного размера
- 2) Полный перебор частотных чанков
- 3) Отсутствие оптимизаций

# Структура прототипа

class Analyser fields: file\_in, file\_out

МЕТОДЫ:

+ new

+ reset

+ deduplication

# Реализация метода deduplication

- + fixed\_size\_chunker
- + append\_dict <- in: chunk
- + fbc\_dedup

# Предварительные результаты

На одном конкретном тексте словарь получилось  
уменьшить на 16.5%

# План работы

- 1) Написание тестов, проверка корректности алгоритма, получение определенных результатов
- 2) Рефакторинг кода
- 3) Реализация оптимизаций