# Отчет

Задача, которую решает наш токенизатор сформулирована следующим образом:

*"В рамках построения лингвистической системы анализа новостных статей требуется токенизатор. Известно, что новостные порталы пользуются правилом одна строка - один абзац."*

### 1. Проблемы дефисов и другие

Все дефисы внутри слов не трогаются, дефисы выделенные пробелами считаются токенами. Кроме точки все остальные знаки препинания считаются токенами.

Для каждой точки решается задача классификации: является ли она самостоятельным токеном или же ее нужно присоединить к соседу слева.

### 2. правила золотого стандарта

Для подготовки корпуса таких статей был обработан корпус OpenCorpora согласно правилам золотого стандарта:

* Точки после сокращений, длина которых более 5 букв - один токен
* Точки в конце абзацев и параграфов - один токен
* Остальные точки не являются токенами
* Остальные знаки препинания - один токен

### 3. общая организация ваших программ

* CorporaExtractor.py - извлечение корпуса по правилам золотого стандарта
* FeatureExtractor.py - алгоритм извлечения признаков из контекста вокруг точки
* MaxEntClassifier.py - обучение, тренировка, сохранение, метрики для MaxEnt
* Tokenizer.py - токенизатор.

### 4. функциональное тестирование - (1) проблема, (2) тест, (3) результат, (4)коммент

В процессе разработки токенизатора было решено много проблем, но некоторые остались:

|  |  |  |  |
| --- | --- | --- | --- |
| Проблема | Тест | результат | комментарий |
| Даты | 09/12/94 | '09', '/', '12', '/', '94' |  |
| даты | 09.12.94 | '09.', '12.', '94' |  |
| дефисы | Сказал-иди-ка сюда | 'Сказал-иди-ка', 'сюда' | Ошибки в прямой речи и вообще - редкое явление для новостей. |
|  |  |  |  |
|  |  |  |  |

### 5. результат тестирования - точность, полнота, f-мера:

Точность 0.986

Полнота 0.987

f-мера 0.986