

# Curse Words Detector project

Беляев Станислав

СПб АУ РАН

Русская Весна 2016

# Цели и задачи

- 1 Модуль на питоне для блокировки слов
- 2 Тесты
- 3 Профиллирование работы
- 4 Поместить это внутрь Node.js

# Цели и задачи

- 1 Модуль на питоне для блокировки слов
- 2 Тесты
- 3 Профиллирование работы
- 4 Поместить это внутрь Node.js

# Цели и задачи

- 1 Модуль на питоне для блокировки слов
- 2 Тесты
- 3 Профиллирование работы
- 4 Поместить это внутрь Node.js

# Цели и задачи

- 1 Модуль на питоне для блокировки слов
- 2 Тесты
- 3 Профиллирование работы
- 4 Поместить это внутрь Node.js

# Архитектура проекта

Дано слово  $w$ , нужно найти ближайшее к нему "плохое"  $c$

$$\operatorname{argmax}_c P(c|w) = \operatorname{argmax}_c P(w|c) \frac{P(c)}{P(w)}$$

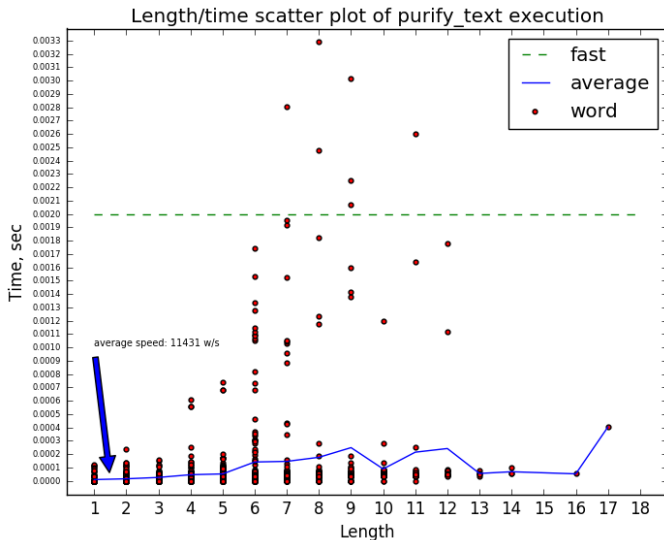
- 1  $P(c)$  - "how likely is  $c$  to appear in text?"
  - частотные списки взял у opencorpora
- 2  $P(w)$  - "same for every  $c$ "
- 3  $P(w|c)$  - вероятность того, что автор написал  $w$ , подразумевая  $c$ 
  - edit distance  $\leq 2$
- 4  $\operatorname{argmax}_c$  - максимум по всем  $c$

# Архитектура проекта

Модули на python3:

- 1 Purifier - main
- 2 Analizier - работа с нормальной формой + предиктор
- 3 Neurister - эвристическая замена цифр и латинских букв на русские
- 4 Statisticer & Tester - классы для сбора статистики и тестов

# Результаты





# Конец

Спасибо за внимание

GitHub: <https://github.com/StasBel/CurseWordsDetector>