



Сентиментальная разметка ТВИТОВ

Тищенко Дмитрий, 504SE

Руководитель: Курочкин Юрий, Yandex

Суть проблемы

← ↻ 5 ★ 1 ⋮

Expand



Alexey Venediktov @aavst · 22m

Рубль заметно вырос к доллару и евро в начале торгов на московской бирже echo.msk.ru/news/1460278-e...

← ↻ 6 ★ 2 ⋮

Нейтральный твит

Village

The Village @villagemsk · 23m

Рубль сегодня выпался и подрос #эфир the-village.ru/village/situat...

The Village

Положительный твит



Суть проблемы

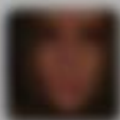


Генерал Ветров @romanvorkuta · 13m

Забыл айфон на ночь в тачке, он замёрз видимо. Но при попытке включить всё равно упорно выдавал "Требуется ОХЛАДИТЬ". Ок, Эпл.



Олег Шварц @olegshvartz · 20m
У меня айфон замёрз на ночь в тачке, он выдавал "Требуется ОХЛАДИТЬ". Ок, Эпл.



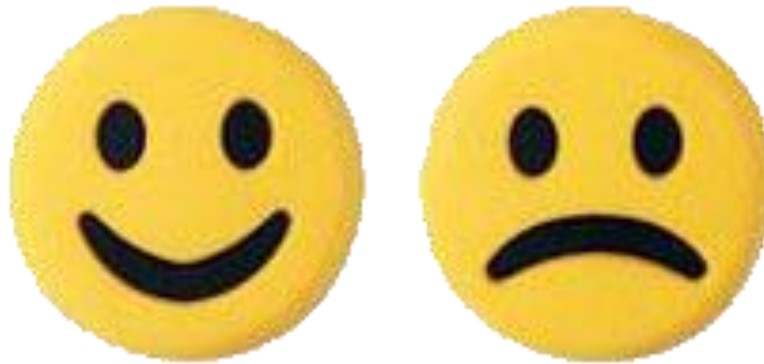
Е. Сидорова @sidorova_e · 20m
Привет айфон замёрз, не включается.



Александр Сидоров @sidorov_alex · 20m
Со мной тоже случилась такая же проблема.

Отзывы о продуктах

Описание задачи



**Машинное обучение для разметки твитов
со смайликами**

Twitter Sentiment Classification using Distant Supervision Go et al., Stanford University, 2009:

- Разметка с точностью 80%
- Использование SVM, Naïve Bayes, Maximum Entropy
- Размер обучающей выборки – 800.000 твитов каждой тональности

Что уже было сделано

A magnifying glass is positioned over an open book. The book's pages contain text in Hindi script. The magnifying glass is centered over a portion of the text, highlighting it. The background is a soft, out-of-focus white.

Повторить аналогично для русского языка:

- **Использовать лемматизацию**
- **Сделать предобработку**
- **Использовать Naïve Bayes с онлайн-обучением**

Способ решения задачи

Результаты

Размер обучающей выборки - 6 миллионов ТВИТОВ

Униграммы:

807/1000 распознанных ПОЗИТИВНЫХ ТВИТОВ

460/1000 распознанных НЕГАТИВНЫХ ТВИТОВ

63.35% - общий процент совпадения

Биграммы:

747 / 1000 распознанных ПОЗИТИВНЫХ ТВИТОВ

629 / 1000 распознанных НЕГАТИВНЫХ ТВИТОВ

68.8% - общий процент совпадения

Униграммы+биграммы:

822 / 1000 распознанных ПОЗИТИВНЫХ ТВИТОВ

556 / 1000 распознанных НЕГАТИВНЫХ ТВИТОВ

68.9% - общий процент совпадения



Код: <https://github.com/flire/Tweets-sentiment-analysis>

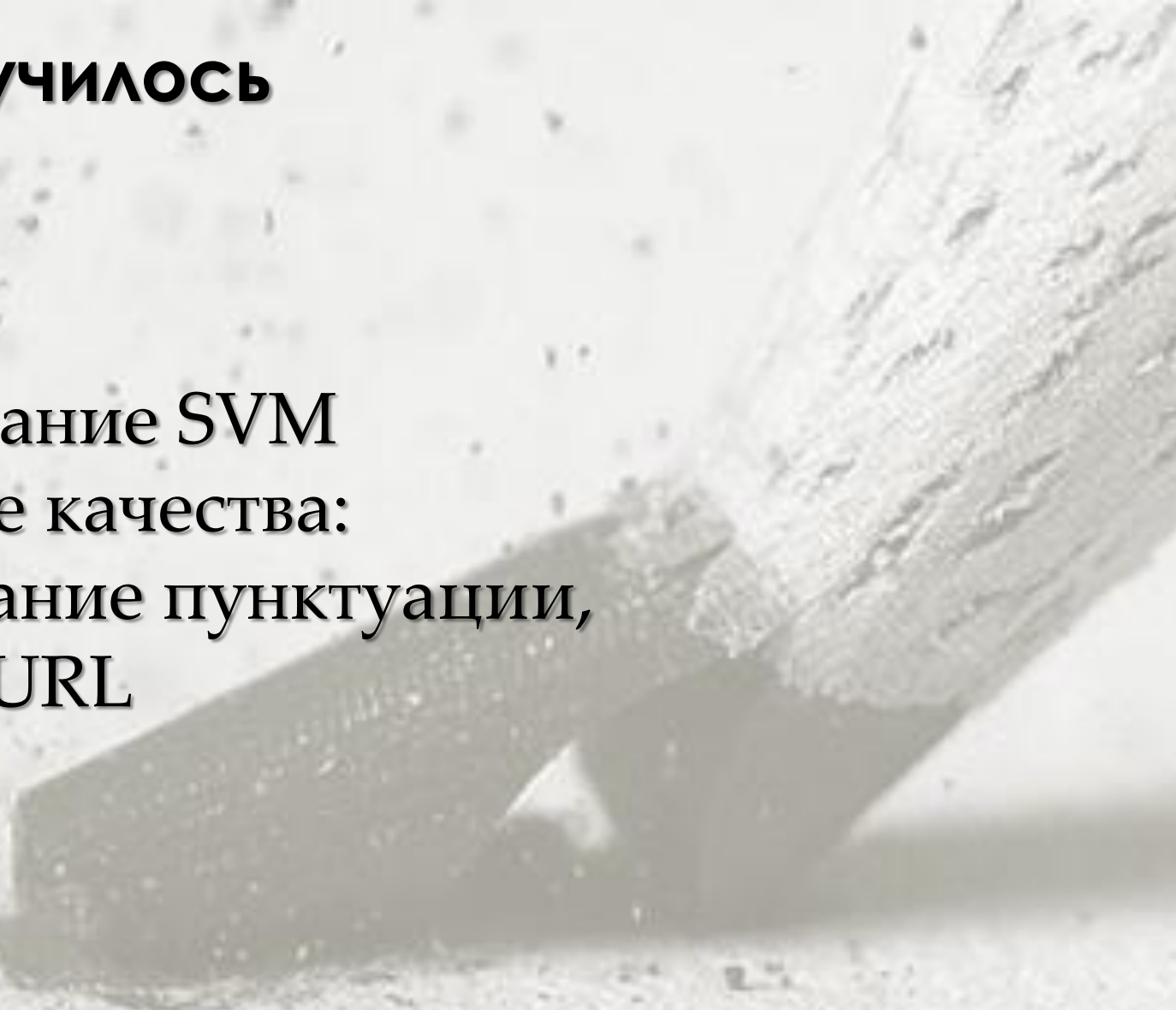
Инструменты

Python:

- Scikit-learn – библиотека машинного обучения
- Python-twitter – библиотека работы с Twitter
- Yandex Mystem - лемматизатор

Что не получилось

- Использование SVM
- Улучшение качества:
использование пунктуации,
хэштегов, URL



Что нового изучили

Опыт в машинном обучении

Python





Сентиментальная разметка ТВИТОВ

Тищенко Дмитрий, 504SE

Руководитель: Курочкин Юрий, Yandex