

Определение языка сообщений социальной сети Twitter

Выполнил студент 327 группы
Михаил Кольцов

Научный руководитель:
Владимир Майоров

20 мая 2014 г.

Введение

Twitter — социальная сеть, в которой более 200 миллионов пользователей обмениваются сообщениями, которые называются «твитами»:

Пример твита

RT @TheWritingDiva: #BlogPaws #blogpawty fiona waits for treats @susandaffron pic.twitter.com/p41wMgEdq8

Ретвит, хэштег, ссылка, упоминание.

Введение. Почему Twitter?

- Отслеживание общественного мнения (новый фильм, реформа, ...)
- Обнаружение событий (землетрясения, пробки, ...)
- Выявление социальных слоёв (либералов, вегетарианцев, ...)
- И многое другое

Введение. Трудности

- Длина сообщения — до 140 символов
- Пользователи допускают ошибки
- Произвольное проявление эмоций

Необходимая мера — *нормализация*:

Пример нормализации

"The Comeback" is coming back to HBO and DREAMS
DO COME TRUE AFTER ALL!!!!!!! → the comeback is
coming back to hbo and dreams do come true after all

Постановка задачи

Цель работы — исследование и сравнение методов автоматического определения языка сообщений Twitter

- 1 Исследовать существующие методы определения языка текста
- 2 Провести совместное сравнительное тестирование некоторых методов
- 3 Исследовать зависимость качества классификации от метода нормализации и кол-ва твитов в обучающей выборке
- 4 Исследовать возможность улучшения какого-либо метода

Сравниваемые системы

- 1 **TextCat** — создан в 1997 г., основан на подсчёте частоты 1..4-грамм
- 2 **Google CLD2** — создан в 2011 г., встроен в Google Chrome и Google Translate; основан на Naive Bayes
- 3 **Langid** — создан в 2011 г.; основан на Naive Bayes
- 4 **LIGA** — реализованный в рамках курсовой работы подход, основанный на графах, предложен в 2011 г.
- 5 **LIGAv2** — предлагаемое улучшение метода LIGA
- 6 **LogR** — реализованный в рамках курсовой работы метод, использующий логистическую регрессию, предложен в 2012 г.

Корпус. Источники

- 1 Научные статьи
- 2 Twitter API
- 3 Indigenous Tweets

18 языков, 4 различных стиля написания:

- Основанные на кириллице: болгарский, чувашский, русский, татарский, украинский (всего 205 175 твитов)
- Арабские: арабский, персидский (фарси), урду (всего 4605 твитов)
- Латинские: нидерландский, французский, английский, немецкий, итальянский, испанский, турецкий (всего 13 382 твита)
- Деванагари: хинди, маратхи, непальский (всего 4041 твит)

Система тестирования

Написан набор скриптов на Bash и Python, обеспечивающих лёгкое добавление/удаление тестируемой системы или изменение параметров нормализации, предназначенный для сбора характеристик сравниваемых систем.

Два результирующих представления для твитов:

- plain text
- plain text + извлечённая метаинформация (имя пользователя, местоположение, ...)

Схема работы с твитами

plain_text_getters → merge_files → normalize_text →
gen_stat → main → результат

Тестирование. Результаты

Мера	CLD2	langid.py	LIGAv2	LIGA	TextCat	LogR
<i>Точность</i>	85.2%	78.4%	94.3%	90.8%	93.8%	29.5%
<i>Полнота</i>	79.4%	78.7%	94.0%	89.3%	90.5%	29.9%
<i>F1-мера</i>	82.2%	78.6%	94.2%	90.1%	92.1%	29.7%
<i>Accurasy</i>	79.5%	78.5%	94.1%	89.6%	90.5%	29.4%

Тестирование. Примеры ошибок

- ❶ **Похожесть языков:** *lang leve ikea (nl → en), state department condemns concerted campaign intimidate international journalists cairo (en → fr).*
- ❷ **Наличие имён собственных на другом языке:** *moyes ponders toffees selection everton boss david moyes will have decisions make both ends the pitch (en → fr).*
- ❸ **Фразы на другом языке:** *харесах видеоклип watch mario balotelli failed backheel (bg → en).*
- ❹ **Неверная разметка:** *встиг підстригтися уже наступний раз записався (ru → uk).*
- ❺ **Малая длина:** *кис, прекратите, lesles, haa dus, lool grave.*

Результаты работы

- 1 Исследованы методы автоматического определения языка сообщений Twitter и произведено сравнительное тестирование некоторых из них
- 2 Реализовано два метода в рамках курсовой работы
- 3 Предложено улучшение для одного из методов, которое показывает качество классификации выше всех остальных
- 4 Реализована система для тестирования методов классификации твитов
- 5 Чтобы показать трудность решения задачи определения языка применительно к Twitter, проведён анализ ошибок классификации

Спасибо за внимание!