

Senty

Балакший Андрей Сухочев Александр
Куратор: Юрий Курочкин

19 мая 2015

О проекте

Сентиментальная разметка коротких текстов с использованием "чистой" разметки обучающего множества.

Вдохновение

Прочитали статью исследователей из Стэнфордского университета о сентиментальной разметке твитов. Они производили "грязную" разметку твитов(разметка по смайликам) и обучались по ней.

[http://cs.stanford.edu/people/alecmgo/papers/
TwitterDistantSupervision09.pdf](http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf)

Вдохновение

Прочитали статью исследователей из Стэнфордского университета о сентиментальной разметке твитов. Они производили "грязную" разметку твитов(разметка по смайликам) и обучались по ней.

<http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf>

Захотелось чего-то аналогичного, но с русским языком и с "чистой" разметкой (разметкой вручную).

Цель

Хотим по короткому тексту понимать настроение его автора.

Этапы выполнения

- Найти источник материалов для обучения, содержащий высказывания людей на различные темы.
- Разметить эти материалы вручную на положительные, отрицательные и нейтральные.
- Придумать различные признаки, которые помогут определять настроения авторов текстов (далее для упрощения будем называть эти признаки фичами).
- Реализовать машинное обучение с учётом этих признаков.

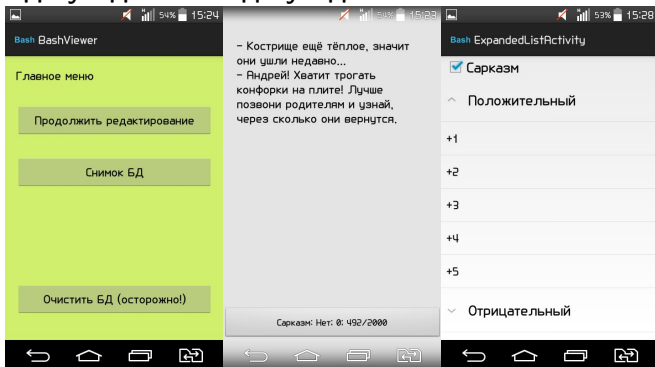
Получение материалов для обучения

Материалы для обучения спарсили с цитатника рунета bash.im. Для парсинга воспользовались питоновской библиотекой Beautiful Soup.



"Чистая" разметка обучающего множества

Для удобства "чистой" разметки текстов (далее башей) написали приложение под Android, чтобы размечать баши где угодно и когда угодно.



Фичи

Было создано множество фич, являющихся надстройками над стандартным экстрактором термов. В стандартном экстракторе мы просто убрали знаки препинания и нормализовывали слова с помощью питоновской библиотеки `rumystem3`.

Примеры работы некоторых фич

- **Исходный текст:**

"УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((... и по вторникам тоже)))"

Примеры работы некоторых фич

- **Исходный текст:**

"УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((... и по вторникам тоже)))"

- **Работа стандартного экстрактора с rymystem3:**

"ужааасно не любить вставать по понедельник и идти на работа ((((и по вторник тоже)))"

Примеры работы некоторых фиц

- **Исходный текст:**

"УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((... и по вторникам тоже)))"

- **Работа стандартного экстрактора с rymystem3:**

"ужааасно не любить вставать по понедельник и идти на работа ((((и по вторник тоже)))"

- **Добавление фици, убирающей предлоги, союзы и местоимения:**

"ужааасно не любить вставать понедельник идти работа ((((вторник тоже)))"

Примеры работы некоторых фиц

- **Исходный текст:**

"УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((... и по вторникам тоже)))"

- **Работа стандартного экстрактора с rymystem3:**

"ужааасно не любить вставать по понедельник и идти на работа ((((и по вторник тоже)))"

- **Добавление фици, убирающей предлоги, союзы и местоимения:**

"ужааасно не любить вставать понедельник идти работа ((((вторник тоже)))"

- **Добавление фици, убирающей повторяющиеся буквы:**

"ужасно не любить вставать понедельник идти работа ((((вторник тоже)))"

Примеры работы некоторых фич

- **Исходный текст:**

"УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((... и по вторникам тоже)))"

- **Добавление фичи, склеивающей "не" со словом, которому эта частица принадлежит:**

"ужасно **нелюбить** вставать понедельник идти работа ((((вторник тоже"

Примеры работы некоторых фич

- **Исходный текст:**

"УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((... и по вторникам тоже)))"

- **Добавление фичи, склеивающей "не" со словом, которому эта частица принадлежит:**

"ужасно **нелюбить** вставать понедельник идти работа (((
вторник тоже"

- **Добавление фичи, выделяющей смайлики:**

"ужасно нелюбить вставать понедельник идти работа
плохосмайл вторник тоже **хоросмайл**"

Примеры работы некоторых фич

- **Исходный текст:**

"УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((... и по вторникам тоже)))"

- **Добавление фичи, склеивающей "не" со словом, которому эта частица принадлежит:**

"ужасно **нелюбить** вставать понедельник идти работа (((
вторник тоже"

- **Добавление фичи, выделяющей смайлики:**

"ужасно нелюбить вставать понедельник идти работа
плохосмайл вторник тоже **хоросмайл**"

- **Также были созданы фичи, выделяющие n-граммы, убирающие иностранные слова и многие другие.**

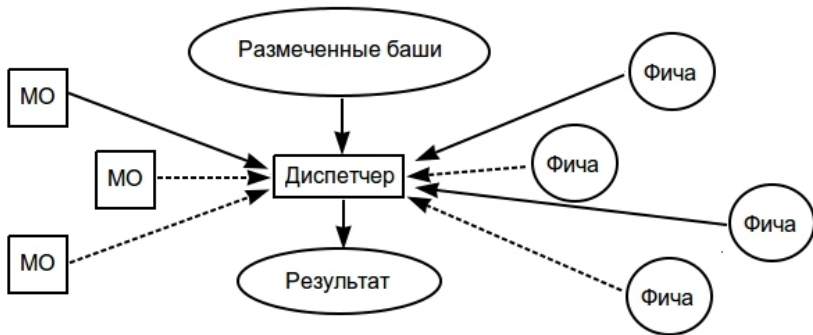
Выбор классификаторов

- Naive Bayes
- SVM
- MaxEntropy

Для их реализации использовали питоновскую библиотеку `scikit-learn`.

Структура проекта

Для использования различных комбинаций фич с различными классификаторами была использована следующая структура.



Результаты(Max Entropy)

Фичи и точность:

- Стандартный экстрактор без фич - 70,58%
- + убираем союзы - 70,85%
- + биграммы с "не" - 70,94%

Итого удалось повысить точность на 0,36%

Результаты(Naive Bayes)

Фичи и точность:

- Стандартный экстрактор без фич - 65,83%
- + убираем повторяющиеся буквы - 66,37%
- + биграммы с "не" - 66,83%
- + убираем иностранные слова - 67,17%
- + выделяем смайлики - 67,26%

Итого удалось повысить точность на 1,43%

Результаты(SVM)

Фичи и точность:

- Стандартный экстрактор без фич - 70,31%
- + убираем повторяющиеся буквы - 70,67%
- + убираем союзы - 70,76%
- + биграммы с "не" - 71,39%

Итого удалось повысить точность на 1,08%

Проблемы

- Малый объём выборки(примерно каждый четвёртый баш является эмоционально окрашенным, много сарказма, "чистая" разметка - дело не быстрое). Всего 9811 различных термов было выделено стандартным экстрактором.
- Сильный "перекос" в сторону отрицательных башей в обучающей выборке: из 1115 башей, имеющих эмоциональную окраску, 732 являются отрицательными (65.65%).

Дальнейшие перспективы

- Двойная классификация: сначала классифицировать на нейтральное/с эмоциональной окраской, а потом уже эмоциональные на положительные/отрицательные.
- Прикрутить поиск по упоминаниям по базе башей для определения (с возможностью получения детальной информации с помощью какого-нибудь фильтра), сколько раз о предмете поиска отзывались положительно/отрицательно/нейтрально.
- Увеличение объёма материала для обучения.

Итоги практики

- Работа в команде
- Изучение Python(pymystem3, Scikit-learn, Beautiful Soup)
- Работа с GitHub
- Работа с \LaTeX (собственно создание самой презентации)



Спасибо за внимание

<https://github.com/cscenter/senty>