Senty

Балакший Андрей Сухочев Александр
Куратор:Юрий Курочкин

19 мая 2015

О проекте

Поиск упоминаний и сентиментальная разметка коротких и очень коротких текстов(с использованием "чистой" разметки обучающего множества).

Вдохновение

Прочитали статью ребят из Стэнфорда о сентиментальной разметке твитов. Ребята производили "грязную" разметку твитов(разметка по смайликам) и обучались по ней.

Вдохновение

Прочитали статью ребят из Стэнфорда о сентиментальной разметке твитов. Ребята производили "грязную" разметку твитов(разметка по смайликам) и обучались по ней.
Захотелось чего-то аналогичного, но с русским языком и с "чистой" разметкой(разметкой вручную).

Цель

Хотим по короткому тексту понимать настроение человека, написавшего его.

Что планируем сделать

- Найти источник для обучения, содержащий высказывания людей на различные темы.
- Придумать разнообразные фичи для текстов из нашего источника(признаки, которые помогут определять настроения людей, написавших данный текст) и реализовать по ним машинное обучение.

Этапы выполнения

Получение материала для обучения

Необходимо получить материалы для обучения.

Этапы выполнения

Получение материала для обучения

Необходимо получить материалы для обучения.

Решение:

Парсим цитаты (далее называемые башами) с bash.im с помощью Beautiful Soup и сохраняем их в формате json.

bash.im
ЦИТАТНИК РУНЕТА

Этапы выполнения

Надо сделать чистую разметку

Как это лучше делать?

Этапы выполнения

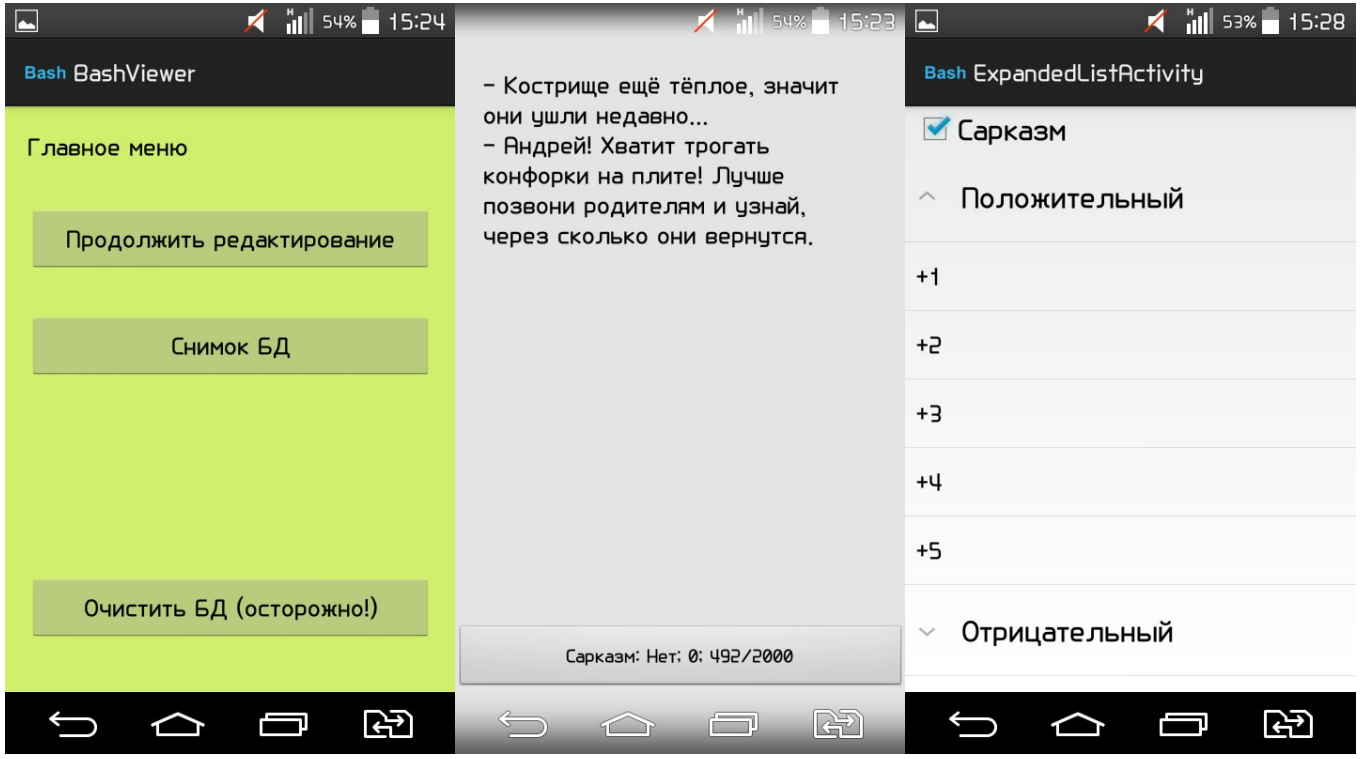
Надо сделать чистую разметку

Как это лучше делать? Было принято решение написать приложение под android для чистой разметки на Java

Этапы выполнения

Надо сделать чистую разметку

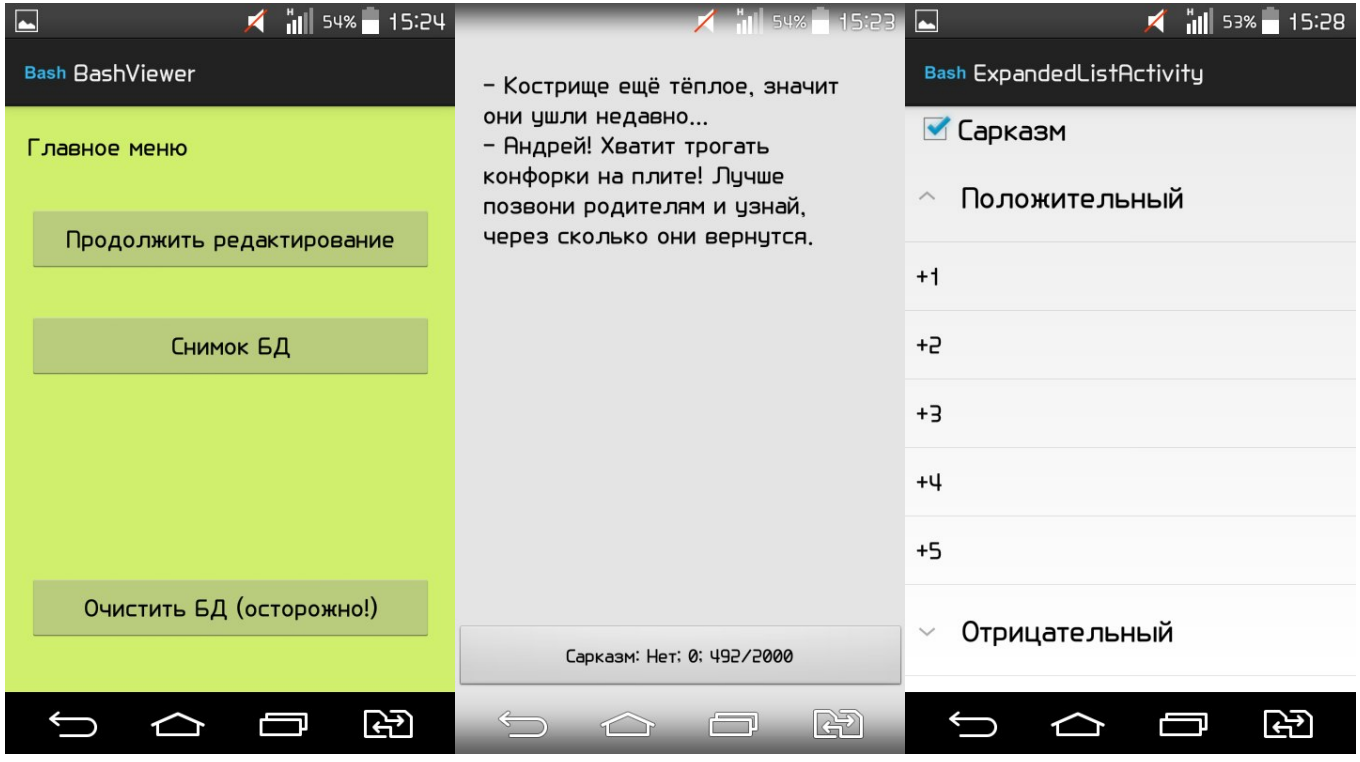
Как это лучше делать? Было принято решение написать приложение под android для чистой разметки на Java



Этапы выполнения

Надо сделать чистую разметку

Как это лучше делать? Было принято решение написать приложение под android для чистой разметки на Java



Отмечу, что хоть приложение и называется BashViewer, однако туда можно запихнуть тексты из любого источника.

Этапы выполнения

Выбор классификаторов

- Naive Bayes
- SVM
- MaxEntropy

Этапы выполнения

Выбор классификаторов

- Naive Bayes
- SVM
- MaxEntropy

scikit-learn.org - спасибо, что ты есть!

Этапы выполнения

Экстрактор и фичи

Для выделения термов из башей был написан экстрактор, который выделял слова, отметая знаки препинания и прочие разделители, и приводил их к нормальной форме с помощью Mystem. Далее было придумано множество различных фич, которые представляли собой дополнительные опции экстрактора и отвечали за создание новых термов и удаление из рассмотрения уже существующих. Их можно использовать в различных комбинациях для повышения точности машинного обучения.

Этапы выполнения

Примеры работы различных фич

- **Исходный текст:**
  "УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((,... и по вторникам тоже))"

Этапы выполнения

Примеры работы различных фич

- **Исходный текст:**
  "УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((,... и по вторникам тоже))"
- **Работа стандартного экстрактора с Mystem:**
  "ужаааасно не любить вставать по понедельник и идти на работа ((((и по вторник тоже))"

Этапы выполнения

Примеры работы различных фич

- **Исходный текст:**
  "УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((,... и по вторникам тоже))"
- **Работа стандартного экстрактора с Mystem:**
  "ужаааасно не любить вставать по понедельник и идти на работа ((((и по вторник тоже))"
- **Добавление фичи, убирающей предлоги, союзы и местоимения:**
  "ужаааасно не любить вставать понедельник идти работа ((((вторник тоже))"

Этапы выполнения

Примеры работы различных фич

- **Исходный текст:**
  "УЖАААСНО не люблю вставать по понедельникам и идти на работу ((((,... и по вторникам тоже))"
- **Работа стандартного экстрактора с Mystem:**
  "ужаааасно не любить вставать по понедельник и идти на работа ((((и по вторник тоже))"
- **Добавление фичи, убирающей предлоги, союзы и местоимения:**
  "ужаааасно не любить вставать понедельник идти работа ((((вторник тоже))"
- **Добавление фичи, убирающей повторные буквы:**
  "ужасно не любить вставать понедельник идти работа ((((вторник тоже))"

Этапы выполнения

Примеры работы различных 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"

Этапы выполнения

Примеры работы различных 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"

Этапы выполнения

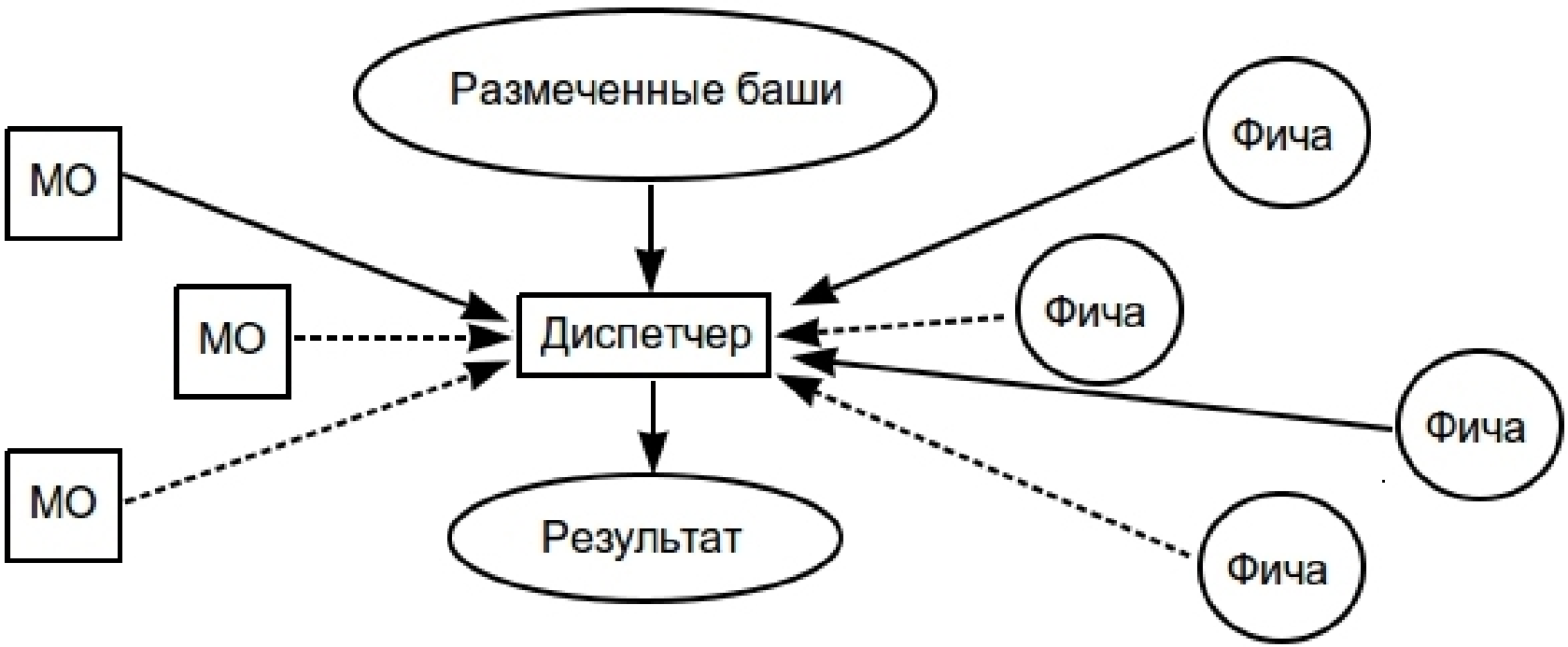
Структура проекта

Теперь нужно как-то объединить всё в одно целое.

Этапы выполнения

Структура проекта

Теперь нужно как-то объединить всё в одно целое.



Проблемы

- Малый объём выборки(примерно каждый четвёртый баш является эмоционально окрашенным, много сарказма, "чистая" разметка - дело не быстрое)

Проблемы

- Малый объём выборки(примерно каждый четвёртый баш является эмоционально окрашенным, много сарказма, "чистая" разметка - дело не быстрое)
- Сильный "перекос" в сторону отрицательных башей в обучающей выборке: из 1115 башей, имеющих эмоциональную окраску, 732 отрицательных (65.65 %)

Дальнейшие перспективы

- Двойная классификация: сначала классифицировать на нейтральное/с эмоциональной окраской, потом уже эмоциональные на положительные/отрицательные.

Дальнейшие перспективы

- Двойная классификация: сначала классифицировать на нейтральное/с эмоциональной окраской, потом уже эмоциональные на положительные/отрицательные.
- Прикрутить поиск по упоминаниям по базе башей для определения в скольких упоминаниях о предмете поиска отзывались положительно/отрицательно/нейтрально, с возможностью получения детальной информации по башам (по какому-то фильтру например)

Дальнейшие перспективы

- Двойная классификация: сначала классифицировать на нейтральное/с эмоциональной окраской, потом уже эмоциональные на положительные/отрицательные.
- Прикрутить поиск по упоминаниям по базе башей для определения в скольких упоминаниях о предмете поиска отзывались положительно/отрицательно/нейтрально, с возможностью получения детальной информации по башам (по какому-то фильтру например)
- Увеличение материала для обучения

Итоги практики

- Работа в команде
- Изучение Python(Mystem, Scikit learn, Beautiful soup)
- Работа с GitHub
- Работа с LaTeX

LaTeX

Спасибо за внимание

https://github.com/cscenter/senty