Практическое задание для студентов 4 курса кафедры СП. Осень 2014

Постановка задачи

Задание 1

Целью задания является создание системы, позволяющей выделять из отзывов мнения людей о заданном объекте и его характеристиках. В данной работе под мнением понимается пара **[характеристика объекта, тональность (эмоциональная окраска)].** Определять объект не требуется.

Пример:

> *Отзыв: "Сделал заказ в магазине. Оперативно перезвонил менеджер, договорились о доставке, способе оплаты и выдачи накопительной карты. Курьер привез все в оговоренные сроки"*

Ответ системы:

> *[колл-центр, позитивная],*
> *[доставка, позитивная].*

Таким образом, система должна выявлять эмоциональные оценки авторов из частей отзывов и определять по отношению к какой характеристике объекта была выражена эмоция.

Вводятся следующие ограничения на задачу:

1. Объект фиксирован - это некоторый магазин.
2. Список характеристик заранее известен и фиксирован (см. приложение 1)
3. Отзывы взяты с Яндекс.Маркет

Задание 2

Второе задание состоит в разработке системы для автоматического определения частей речи. Разрабатываемая система должна определять следующие части речи[1]:

S — существительное (*яблоня, лошадь, корпус, вечность*)
A — прилагательное (*коричневый, таинственный, морской*)
NUM — числительное (*четыре, десять, много*)
A-NUM — числительное-прилагательное (*один, седьмой, восьмидесятый*)
V — глагол (*пользоваться, обрабатывать*)
ADV — наречие (*сгоряча, очень*)
PRAEDIC — предикатив (*жаль, хорошо, пора*)
PARENTH — вводное слово (*кстати, по-моему*)
S-PRO — местоимение-существительное (*она, что*)
A-PRO — местоимение-прилагательное (*который, твой*)

[1] http://ruscorpora.ru/corpora-morph.html

ADV-PRO — местоименное наречие (*где, вот*)
PRAEDIC-PRO — местоимение-предикатив (*некого, нечего*)
PR — предлог (*под, напротив*)
CONJ — союз (*и, чтобы*)
PART — частица (*бы, же, пусть*)
INTJ — междометие (*увы, батюшки*)

Решение задачи

Практические аспекты

Решения должны быть написаны на языке Python (версия 2.7). Можно использовать все стандартные библиотеки, а также

- NLTK - инструменты для обработки текстов
- scikit-learn - алгоритмы машинного обучения
- numpy - работа с многомерными массивами

Доступ в Интернет на проверяющей машине закрыт.

Задание 1

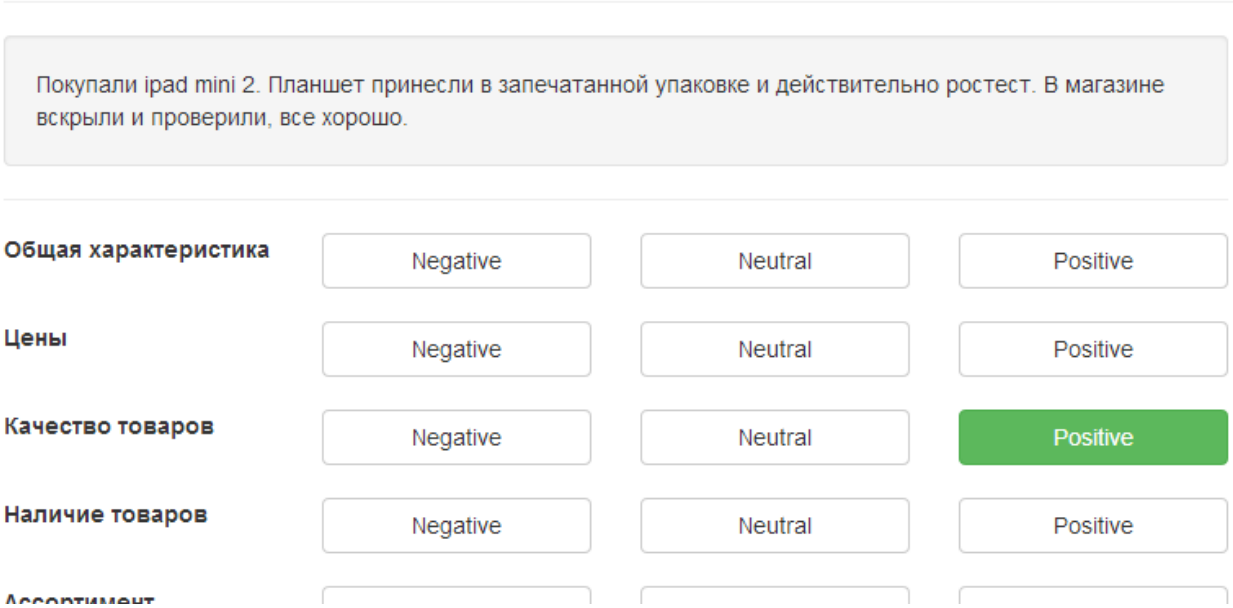
Теоретические аспекты

Для решения первой задачи рекомендуется использовать методы машинного обучения с учителем. Для обучения алгоритмов требуется придумать признаки и дать ему на вход правильные примеры - обучающий корпус. Считается, что чем больше обучающий корпус, тем лучше работают алгоритмы. Однако создание большого обучающего корпуса - довольно трудоемкая задача, непосильная одному человеку. Поэтому предлагается создать его с помощью коллективной работы. Чтобы облегчить эту работу, был сделан сайт: http://reviews.at.ispras.ru.

Разметка обучающего корпуса

Для разметки корпуса необходимо зарегистрироваться на сайте http://reviews.at.ispras.ru. Пожалуйста, вводите правильные данные, так как они будут использоваться при выставлении зачетов. Вне рамок практикума эти данные использоваться не будут.

Далее система будет показывать отзывы и список характеристик магазинов, которые могут быть описаны в отзыве. Для тех характеристик, которые действительно описываются в отзыве, необходимо выбрать тональность (позитивная, негативная или нейтральная). Если характеристика не обсуждается в отзыве, все кнопки тональности, соответствующие этой характеристике, должны быть отжаты. Это касается и случаев, когда в отзыве почему-то описывается товар, а не магазин. Также возможны случаи, когда в одном отзыве в разных местах описывается одна характеристика, но с разной тональностью. В этом случае необходимо нажать несколько кнопок.



В систему загружено 2500 отзывов. Каждому человеку предлагается разметить 100 отзывов, случайно выбранных из всех загруженных. Система перестает показывать отзыв после того как отзыв был размечен 3-мя разными людьми. Информацию о различиях в разметке можно использовать при обучении алгоритмов (например, обучать только на характеристиках выделенных всеми людьми).

После того, как будет размечено не менее 100 отзывов, появится кнопка, позволяющая скачать размеченные отзывы (см. раздел "тренировочный корпус").

Рекомендуется размечать максимально честно, так как от этого будет зависеть качество всех классификаторов. Если есть сомнения, как разметить отзыв, то его стоит пропустить, обновив страницу.

Тренировочный корпус

Тренировочный корпус будет доступен для скачивания в формате json. Для извлечения информации из этого файла рекомендуется использовать стандартную библиотеку Python с одноименным названием.

Для синхронизации обучения и тестирования в течение недели, корпус будет состоять из отзывов, размеченных автором классификатора, плюс все отзывы, размеченные в течение предшествующей недели.

Тестирование

Вместе с кнопкой скачивания тренировочного корпуса появится ссылка на форму для загрузки файла и личную страницу со статистикой. На личной странице находится статистика со всеми результатами в т.ч. результатами последнего тестирования (дата, описание, точность, полнота, F_1-мера).

Загрузка решения. Загружаемый файл должен представлять собой zip архив с любым именем. Архив должен обязательно содержать:

- Решение в файле solution.py. В файле должен содержаться класс Solution. В классе должны присутствовать методы
    - train(self, training_corpus), где training_corpus - это пара из двух списков [text, opinions]. Параметр opinions – это список списков кортежей (tuple) вида ([(характеристика 1, тональность2), (Характеристика2, тональность2)], [...], ...]). Внимание: метод train будет вызываться отдельно, так что не стоит вызывать его в конструкторе класса.
    - getClasses (self,texts), который получает на вход список текстовых сообщений и возвращает список списков кортежей (как параметр options и предыдущем пункте). Каждому документу соответствует один элемент верхнего списка. Во внутренних списков содержатся 0 или более ответов классификатора вида (характеристика, тональность). (Пример: Пусть есть два отзыва, тогда ответ может быть примерно такой: [[(общая, positive), (цены, negative)], [(общая, positive)]])
- описание применяемых методов в файле description.txt
- все используемые внешние библиотеки, кроме библиотек пакетов NLTK, scikit-learn и numpy (они доступны автоматически).

Результаты тестирования появятся на личной странице, как только закончится обучение и тестирование. При загрузке нового классификатора обучение будет производиться на корпусе из отзывов, размеченных автором решения, плюс все отзывы, размеченные в течение предшествующей загрузке недели.

В течение недели студенты не видят прогресс своих коллег и могут посмотреть только свой результат. В конце каждой недели (каждый вторник в 23.59.59) будет производиться переобучение последнего присланного решения от каждого студента на новом корпусе, а результаты тестирования будут показаны в сводной таблице.

Оценка качества

Для оценки качества используются F1-мера, вычисляемая как среднее геометрическое точности и полноты.

$F_1 = \frac{2pr}{p+r}$

Точность равна количеству правильных ответов к общему числу ответов системы

$\text{precision} = \frac{correct\ answers}{total\ answers\ by\ system}$

Полнота равна отношению количества правильных ответов к общему числу правильных ответов

$\text{recall} = \frac{correct\ answers}{total\ answers\ in\ test\ set}$

Описание в документации к библиотеке scikit-learn: http://scikit-learn.org/stable/modules/model_evaluation.html

Baseline

Baseline. В качестве нижней границы используется один из стандартных алгоритмов классификации с N-граммами в качестве признаков. Этот классификатор будет тренироваться на том же корпусе, что и присланные алгоритмы, и его достоверность будет меняться соответственно.

Задание 2

Во втором задании каждому слову во входном тексте необходимо поставить в соответствие тэг, обозначающий часть речи.

Загрузка решения

В качестве решения принимается zip архив с произвольным именем, содержащий

- Решение в файле solution.py. В этом файле должен содержаться класс Solution, который должен быть наследником класса `nltk.tag.api.TaggerI`. В классе должны присутствовать:
  - Конструктор, получающий на вход тренировочный набор в виде списка пар (токен, тэг)
  - Метод tag, получающий на вход список токенов и возвращающий список тэгов для этих токенов
- описание применяемых методов в файле description.txt
- все используемые внешние библиотеки, кроме библиотек пакетов NLTK, scikit-learn и numpy (они доступны автоматически).

Пример

```
#coding=CP1251
import nltk
class Solution(nltk.tag.api.TaggerI):
   def __init__(self, train):
      None
   def tag(self, tokens):
      return nltk.DefaultTagger("S").tag(tokens)
```

Тестирование

Для тестирования вызывается метод evaluate у присланного класса (наследуется от `nltk.tag.api.TaggerI`):

```
Solution(train).evaluate(test)
```

Тренировочный и тестовый корпуса

Тренировочный и тестовый корпуса получены из Национального корпуса русского языка. Тренировочный корпус доступен для скачивания на сайте практического задания.

Внимание! В соответствии с лицензией НКРЯ, тренировочный корпус можно использовать только в рамках данного практического задания. По вопросам использования корпуса в других целях необходимо связаться с правообладателями корпуса (http://ruscorpora.ru/corpora-usage.html).

Тренировочный корпус предоставляется в сериализованном с помощью библиотеки cPickle виде. Для чтения используйте функцию cPickle.load(). Полученный таким образом список можно подавать на вход конструктору класса Solution для обучения.

Baseline

В качестве нижней границы используется система, основанная на модели n-грамм с откатами (см. лекцию 3. Реализация основана на библиотеке nltk и на тестовых данных показывает точность 0.911143690166

Теоретические аспекты решения

При реализации можно пользоваться любыми средствами и библиотеками. Ваша цель побить baseline и предложить лучшее решение. Все используемые подходы должны быть подробно описаны в description.txt. При этом приветствуются оригинальные, отличные от других решения, как показатель самостоятельной работы.

Ограничения

1. каждую неделю можно послать только 10 версий ***каждой*** программы (Внимание! Итоговое тестирование будет проводиться на последнем загруженном решении)
2. размер архива не может превышать 15Мб
3. Ограничение времени на проверку задания 10 минут

В связи с первым ограничением, для тестирования на локальной машине рекомендуется использовать метод перекрестной проверки (http://en.wikipedia.org/wiki/Cross-validation_(statistics)). В библиотеке scikit-learn есть функции, которые могут помочь в использовании этого метода (например, KFold()).

Подсчет очков

Как было сказано выше, в конце каждой недели вы сможете посмотреть, насколько хороший классификатор вы сделали по сравнению с другими предложенными решениями. Эти результаты нужны только для понимания текущей ситуации.

В течение семестра будет три дедлайна, когда текущие результаты преобразуются в очки, которые повлияют на итоговую оценку за курс.

Расписание дедлайнов:

1. 14 октября (учитываются все решения, присланные до 23:59:59 14 октября)
2. 18 ноября
3. 16 декабря

При наступлении дедлайнов, так же как и в конце обычной недели производится обучение и тестирование всех присланных решений. Далее производится ранжирование результатов (по мере достоверности) и начисляются очки: за 1 место – 10 очков, 2-9 и т.д. Все программы выше baseline получают по 2 очка, ниже baseline, но с работающим решением - по одному очку. После этого результаты становятся доступны всем на главной странице.

Задание 1 необходимо сдать до 18 ноября. При выставлении оценок будут учитываться результаты 1 задания, полученные до первого и второго дедлайна, и результаты второго задания полученные до второго и третьего дедлайна.

Выставление оценок

После 17 декабря будут выставляться итоговые оценки за практикум, а также проводится экзамен по курсу. Первое и второе задания оцениваются раздельно, дал получения положительных оценок необходимо решить оба задания. Оценка за практикум будет выставляться по следующим критериям:

- Для получения отметки "**Отлично**" - необходимо набрать минимум 4 балла за каждое задание и не менее 9 баллов в сумме(быть всегда лучше baseline и хотя бы раз и попасть в top-8).
- "**Хорошо**" ставится за 6-8 баллов: минимум 3 балла за каждое задание (надо побить baseline по каждому заданию).
- Для получения отметки "**Удовлетворительно**" необходимо набрать минимум 2 балла за каждое задание (вовремя прислать рабочие решения).

Все, кто получил оценку «неудовлетворительно» за практическую работу, также не допускаются к экзамену. Кроме того, при получении оценки «удовлетворительно» за практическую работу максимальная оценка за экзамен по курсу может быть только «хорошо».

Дополнительные вопросы

- Все вопросы, кроме технических, задавайте на сайте http://modis.ispras.ru/tpc, либо пишите одновременно на turdakov@ispras.ru и vmayorov@ispras.ru
- Все технические вопросы относительно проверки заданий просьба присылать на laguta@ispras.ru либо спрашивать в разделе сайта, посвященном практикуму.
- Для установки внешних модулей (NLTK, scikit-learn, BeautifulSoup) рекомендуется использовать easy_install из пакета setuptools.

Вспомогательная литература

- Тоби Сегаран, “Программируем коллективный разум” (Книга про прикладное применение некоторых технологий искусственного интеллекта, включая машинное обучение, в Web 2.0 с огромным количеством примеров на Python).
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python(Книга про обработку текста с помощью библиотеки NLTK для языка Python. Доступна на сайте NLTK)
- Daniel Jurafsky, James H. Martin. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (Одна из лучших книг про обработку текстов)
- Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing (Книга содержит хорошие примеры применения машинного обучения для обработки текстов)

Приложение 1. Характеристики магазинов для задачи 1

- Общая (характеристика магазина в целом)
- Цены
- Качество товаров
- Наличие товаров
- Выбор
- Возврат/обмен товаров
- Скидки/бонусы
- Интернет сайт (актуальность информации на сайте, удобство)
- Операторы (колл-центр)
- Доставка (скорость, аккуратность)
- Магазин (offline)/пункт выдачи
- Персонал/продавец (вежливость, подготовка)
- Время работы

Примеры отзывов

Общая

- Всё удобно, покупал там не раз.
- Качеством услуг доволен
- Магазин хорош до того момента, пока покупатель не отдал деньги за товар. После этого момента про вежливость и лояльное отношение можно забыть.

Цены

- Цена на нужный товар была низкая
- Дешевле чем в других магазинах
- Цена ниже магазинов в которых привык покупать.
- Удобный сайт, вкусные цены
- Постоянно скачут цены

Качество товаров

- Заказывал ноутбук,доставка была поздняя,спал ребёнок,поэтому не стали проверять, на следующий день включили и увидели битый пиксель,названивать не стали,написали лично директору,
  ответа никакого не последовало,вывод,заказывать больше ничего у них не будем,стоимость бука 33.000р.
- Заказал комплектующие для сборки ПК с доставкой. Меня дома не было , товар приняла жена. Стал собирать комп, а в сокете замяты ножки.
- Качественный товар
- удовлетворительное качество товаров

Наличие товаров

- Понравился быстрый отклик сотрудников на мой заказ и в наличие все оказалось

Выбор

- Реально очень большой выбор аксессуаров для GoPro - смотрел, есть почти все.
- Ооооооочень большой выбор экшн камер, мало кто может такой предоставить
- Огромный ассортимент. Грамотные диспетчера, рассказали как настроить приставку.

Возврат/обмен товаров

- Вернуть товар нельзя, полное нарушение прав потребителя!!! В гарантийном отделе сидят люди, которые при тебе же отрывают плёнку и обвиняют в этом же покупателя. Говорят только то, что им выгодно.

Скидки/бонусы

- Заказывал через интернет. На следующий день приехал и забрал. Только вот немного не понял как потратить бонусы начисленные на карту. К счастью при оплате в кассе мне предложили списать эти бонусы и в результате сэкономил целые 200 рублей. Приятно всё же.

Интернет сайт (актуальность информации на сайте, удобство)

- представлена актуальная информация о ценах и наличие
- Удобный интерфейс на сайте
- Не достоверная информация на сайте и в момент оформления заказа о сроках получения заказанного товара !!!
- Несколько мудреный сайт

Операторы (колл-центр)

- Обратилась за уточнением комплектации МФУ, на что человек на телефоне нагрубил и сказал, что если вы не знаете зачем у меня спрашивайте.

Доставка (скорость, аккуратность)

- Товар доставили в назначенное время, ночью я сделала заказ и вечером его уже привезли
- Покупал эпилятор BRAUN. Курьер привез в тот же день.
- Плохая логистика
- Отсутствие точной информации о дате доставки товара
- Сроки доставки, указанные на сайте, НЕ СООТВЕТСТВУЮТ действительности
- Задержались с доставкой, но заранее вежливо предупредили об этом. В целом, претензий нет.

Магазин (offline)/пункт выдачи

- Выдача без проволочек, вскрыли, проверили, отметили гарантию, упаковали.
- Оформление всех гарантийных талонов, проверка товара, его выдача заняли максимум мин.10.
- Большое число пунктов выдачи, расположенных в шаговой доступности/бонусы/недорогой магазин/большой ассортимент товаров.

Персонал/продавец (вежливость, подготовка)

Имеется ввиду любые сотрудники за исключением операторов колл-центра и курьеров службы доставки.

- Очень вежливый персонал, показали первичное использование, проверили наличие комплектации и работоспособность фотика.
- Сделал заказ, сразу же перезвонила девушка пригласила в выходные за ним приехать. Потратил два часа своего времени и в итоге - пообщался только с охранником, который в грубой форме сказал что в выходные магазин не работает.

Время работы

- Заказывал поздно вечером,на следующий же день утром позвонил сотрудник магазина и вежливо еще раз уведомил о том что заказ можно забрать в удобное для меня время
- время работы пунктов выдачи большое и плюс работают по выходным
- Круглосуточный телефон