Класификатор на коментари

Проект по "Откриване на знания в текст"

ИЗГОТВИЛИ:
ИВАН КАПУКАРАНОВ, ФН: 24958, 1 КУРС, ИИОЗ
СИЯНА СЛАВОВА, ФН: 24963, 1 КУРС, ИИОЗ

Декларация за липса на плагиатство

1. Тази курсова работа е моя работа, като всички изречения, илюстрации и програми от други хора са изрично цитирани.

2. Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.

3. Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.

Иван Капукаранов, ФН: 24958, 1 курс, ИИОЗ

Сияна Славова, ФН: 24963, 1 курс, ИИОЗ

Съдържание

Мотивация, Задача на курсовата работа

Мотивация

БгМама е един от най- разпространените сайтове в момента в България за търсене на информация. Там може да намериш всичко от как се гледа определен сорт цветя до кой хотел е най - подходящ за лятната ти почивка. Цялата тази информация обаче е във вид на форум и за да намериш, каквото търсиш, трябва да изчетеш всички коментари.

От защо решихме да направим "БгМама наръчник", който по подаден списък от коментари, ще намери тези с положителни отзиви и тези с отрицателни.

Идея

Идеята ни е да извлечем мнения за хотели от БГМама, да ги класифицираме по положителни и отрицателни.

Съществуващо решение

Реализацията на задачата постигнахме чрез използването на "Наивен Бейсов Класификатор", който реализирахме сами. За целта използвахме сет с обучаващи данни (120 коментара) и сет с тестови данни (44 коментара). Тестовите и обучаващите данни са реални коментари за хотели, взети от следните сайтове:

http://www.booking.com

http://www.zahotelite.com/

Избрахме именно тези сайтове, защото в тях, когато човек пише мнение за хотел, трябва задължително да отбележи дали това мнение е положително или отрицателно. Също така използвахме няколко различни източника за коментарите, тъй като би ни дало по -разнообразно множество от обучаващи данни.

Версия 2

Във версия 2 решихме да си обогатим обучаващото множество, за да постигнем по –добро класифициране на данните. Основната ни идеята е да вземем данни от английски и чрез превеждане да получим достатъчно добро обучаващо множество на български език.

Тъй като на английски език има много подходящи множества, ние избрахме едно такова, а именно:

http://sifaka.cs.uiuc.edu/~wang296/Data/index.html.

За превеждане на данните използвахме Bing translator:

https://www.microsoft.com/en-us/translator/getstarted.aspx

https://github.com/boatmeme/microsoft-translator-java-api

Идеята ни за версия 2 е да направим различни експерименти, с които да постигнем възможно най-добро подобрение на алгоритъма.

Решение

Програмна реализация

Предварителна обработка на данните

Първоначалният json файл с данни от БгМама беше около 3 ГБ. Тъй като нашата реализация е свързана само с намирането на имена на хотели, решихме да извадим от този файл само мнения от теми, свързани с хотели. За целта търсихме в името на темата дали се среща думата "хотел" под някаква форма и взехме коментарите само за тази форма.

Другата предварителна обработка, която направихме, бе да махнем html таговете от текста на коментара, тъй като те не ни носят никаква информация нито за класификацията, нито за откриването на хотел. По този начин се подобри и успеваемостта ни, тъй като махнахме излишния "шум" от данните.

Също така получените дани бяха в json формат и съдържаха доста атрибути (като час на коментар, име на потребител и други), които не са релевантни към нашата задача, така че решихме да не ги парсваме при работа с данните. От релевантните атрибути формирахме нов файл в xml формат, който да съдържа елементи "коментар" и "име на тема" и атрибути "id" на коментар и "категория", която в момента на създаване на файла е "unknown". Същите елементи имат и файловите с тестовите и обучаващите данни, необходими ни за класификацията на коментари.

От първоначалния json файл чрез приложените подобрения на данните накрая достигнахме до файл само с коментари за хотели, които е около 4 МГ и съдържа над 4800 различни коментара.

Парсване на коментар – пример (БГ мама):

- Първоначален вид:


```
[
   {
      "msgcontent":{
         "msg":{
            "idmsg":29804053,
            "idtopic":798438,
            "topicname":"ДУБАЙ 7-ма тема",
            "idboard":199,
            "boardname":"На път в чужбина",
```

```
            "msgtime":1420070516,
            "msgsubject":"Re: ДУБАЙ 7-ма тема Хотел",
            "msgbody":"[quote author=beny_nn link=topic=793144.msg29806558#msg298
06558 date=1420132114]Здравейте, момичета! Аз ще споделя само с едно изречение мо
ите впечатления от Роял Спа: Никога преди не съм била в толкова невероятно хубав
 хотел, с толкова УЖАСНА ОРГАНИЗАЦИЯ за Нова Година! P.S. Честита Нова Година на в
сички! Желая ви повече приятни моменти в велинградските хотели![/quote]ловеч Здра
вейте, момичета! &lt;br /&gt;Аз ще споделя само с едно изречение моите впечатлени
я от Роял Спа: Никога преди не съм била в толкова невероятно хубав хотел, с толко
ва УЖАСНО НЕПРОФЕСИОНАЛНА&amp;nbsp;  :mrgreen: ОРГАНИЗАЦИЯ за Нова Година! &lt;br
 /&gt;P.S. Честита Нова Година на всички! <b>ХТМЛ</b>Желая ви повече приятни моме
нти в велинградските хотели!",
            "topicreplies":759,
            "topicviews":43929,
            "topiclikes":6,
            "msgcount":697,
            "msglikes":0
         },
         "member":{
            "mid":425569,
            "mreg":1366753597,
            "mposts":2060,
            "mlastlogin":1450131803,
            "mname":"MayyaI",
            "mbdate":"0001-01-01",
            "mgender":"female"
         }
      }
   }
]
```


- След премахване на html таговете:


```
"msgbody Здравейте, момичета! Аз ще споделя само с едно изречение моите
впечатления от Роял Спа  Никога преди не съм била в толкова невероятно
хубав хотел, с толкова УЖАСНО НЕПРОФЕСИОНАЛНА nbsp;    ОРГАНИЗАЦИЯ за
Нова Година!  br / P.S. Честита Нова Година на всички!  ХТМЛ Желая ви
повече приятни моменти в велинградските хотели!”
```


- В xml вид:


```
<comment category="unknown" id="29804053">
Здравейте, момичета! Аз ще споделя само с едно изречение моите
впечатления от Роял Спа  Никога преди не съм била в толкова невероятно
хубав хотел, с толкова УЖАСНО НЕПРОФЕСИОНАЛНА nbsp;    ОРГАНИЗАЦИЯ за
Нова Година!  br / P.S. Честита Нова Година на всички!  ХТМЛ Желая ви
повече приятни моменти в велинградските хотели!
<topic>Re: ДУБАЙ 7-ма тема Хотел</topic>
</comment>
```

Данните от trip advisor за версия 2 също трябваше да бъдат предварително преработени, защото не бяха в подходящия за нас вид. Освен това те не бяха разделени по сентимент – положителен, отрицателен, а по рейтинг. Ние ги разпределихме, като тези с цялостен рейтинг <= 3 взехме за отрицателни, а тези с оценка над 3 – за положителен.

Парсване на коментар – пример (Trip advisor):

- Първоначален вид:

```
{
         "Content":"Great We stayed for a week golfing this year and
even though it was in an industrial area it was easy to get to the
freeways from there. The staff was very nice and helpful. The hot
breakfast every morning was served made to order with a nice selection
of items. The rooms were very clean and quiet. Restuarants were in
short driving distances from the hotel.",
         "Date":"Mar 28, 2005",
         "ReviewID":"UR522022148",
         "Ratings":{
            "Service":"4",
            "Business service":"-1",
            "Cleanliness":"5",
            "Check in / front desk":"-1",
            "Overall":"3",
            "Value":"3",
            "Rooms":"3",
            "Location":"-1"
         },
         "Author":"A TripAdvisor Member"
      }
```

За построяване на Наивен Бейсов класификатор, използваме алгоритъмът, посочен тук: http://www.cs.cmu.edu/~tom/book.html. При разделянето на коментарите на отделни думи не сме взели под внимание препинателните знаци и цифрите.

Построихме два класификатора – един с обучаващо множество от български данни и друг с обучаващо множество от английски данни. Направихме различни експерименти върху тях, описани в секцията „Резултати от експерименти“.
За класификатора, трениран върху английските обучаващи данни, премахнахме стоп думите.

Използван алгоритъм за построяване на "Наивен Бейсов класификатор":

LEARN_NAÏVE_BAYES_TEXT(*Примери, V*)

Примери е множество от текстови документи заедно с техните класификации. V - е множеството от възможни класификации (стойности на целевия атрибут). Процедурата научава вероятностите $P(w_k|v_j)$, описващи вероятност, че случайно избрана дума от документа с клас v_j ще бъде дума w_k. Тя също така научава и априорните вероятности на класове $P(v_j)$.

1. **Събери всички думи и знаци за пунктуация, намиращи се в *Примери***
   - *Речник* ← множеството от всички различни думи и знаци за пунктуация, срещани в текстови документи на *Примери*
2. **Изчисли необходимите $P(v_j)$ и $P(w_k|v_j)$**
   - За всяка стойност на целевия атрибут v_j направи:
     - $docs_j$ ← подмножество на документи от *Примери*, за които стойността на целевия атрибут е v_j.
     - $P(v_j) \leftarrow \frac{|docs_j|}{|Примери|}$
     - $Text_j$ ← един общ документ, получен чрез обединение на всички членове на $docs_j$
     - n ← общия брой на различни позиции на думи в $Text_j$
     - за всяка дума w_k от *Речник* направи:
       - n_k ← броя на срещане на думата w_k в $Text_j$
       - $P(w_u \mid v_j) = \frac{n_k + 1}{n + |Речник|}$

CLASSIFY_NAÏVE_BAYES_TEXT(*Документ*)

Предсказва значение на целевия атрибут на некласифициран *Документ*. a_i означава думата, намерена в i-та позиция на *Документа*.

- *позиции* ← всички позиции на думи от *Документа*, които се срещат в *Речник*.
- Върни v_{NB}, където

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i \in позиции} P(a_i \mid v_j)$$

Примерни обучаващи коментари:

- Позитивен

```
<comment id = '1' category = 'positive'>
    Много отзивчив и внимателен персонал,чудесно разположение до метро станцията и НДК.
    Приятна гледка от просторния апартамент към нощна София!
</comment>
```

- Негативен

```
<comment id = '2' category = 'negative'>
    Дори и в новото крило матраците са от стария тип с пружини и са доста неудобни!
    Малко е шумно през нощта ,когато започнат да си форсират колите нагоре по "Черни връх".
</comment>
```

Примерен тестов коментар:

```
<comment id = '8' category = 'positive'>
    Невероятен хотел!Дизайнерско обзавеждане,перфектно обслужване!
</comment>
```

Резултати от експерименти

Версия 1

Получените статистики са резултат от изпълнение на приложението върху тестови документи. Тестовите данни са 44.

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0.77273	Precision negative: 0.95238	Precision overall : 0.86047
Recall positive: 0.94444	Recall negative: 0.80000	Recall overall : 0.86047
F1 positive: 0.85000	F1 negative: 0.86957	F1 overall : 0.86047

Версия 2

Във версия 2 използвания тестов сет съдържа 87 ръчно анотирани коментара.

Тестване с преведени коментари върху класификатора, трениран с обучаващо множество на английски език с премахнати стоп думи:

- Трениращо множество с разпределение положителни – негативни коментари приблизително 2:1 (~6000).

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,85106	Precision negative: 0,75000	Precision overall : 0,80460
Recall positive: 0,80000	Recall negative: 0,81081	Recall overall : 0,80460
F1 positive: 0,82474	F1 negative: 0,77922	F1 overall : 0,80460

- Трениращо множество с относително равномерно разпределение положителни – негативни коментари (~4000).

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,87179	Precision negative: 0,66667	Precision overall : 0,75862
Recall positive: 0,68000	Recall negative: 0,86486	Recall overall : 0,75862
F1 positive: 0,76404	F1 negative: 0,75294	F1 overall : 0,75862

Тестване с коментари върху класификатора, трениран с обучаващо множество на български език от версия 1 (170 коментара) :

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,83333	Precision negative: 0,56140	Precision overall : 0,65517
Recall positive: 0,50000	Recall negative: 0,86486	Recall overall : 0,65517
F1 positive: 0,62500	F1 negative: 0,68085	F1 overall : 0,65517

Миксиране на двата класификатора, като взимаме по –голямата вероятност от двата (погрешно допускане):

- Трениращото множество на английски език е с разпределение положителни – негативни коментари приблизително 2:1 (~6000).

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,77500	Precision negative: 0,59574	Precision overall : 0,67816
Recall positive: 0,62000	Recall negative: 0,75676	Recall overall : 0,67816
F1 positive: 0,68889	F1 negative: 0,66667	F1 overall : 0,67816

- Трениращо множество с относително равномерно разпределение положителни – негативни коментари (~4000).

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,81250	Precision negative: 0,56364	Precision overall : 0,65517
Recall positive: 0,52000	Recall negative: 0,83784	Recall overall : 0,65517
F1 positive: 0,63415	F1 negative: 0,67391	F1 overall : 0,65517

Използване на обучаващо множество на българския класификатор, предварително класифицирано от английския:

- Предварително класифицирано обучаващо множество от английски класификатор с разпределение положителни – негативни коментари приблизително 2:1 (~6000).

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,79630	Precision negative: 0,78788	Precision overall : 0,79310
Recall positive: 0,86000	Recall negative: 0,70270	Recall overall : 0,79310
F1 positive: 0,82692	F1 negative: 0,74286	F1 overall : 0,79310

- Трениращо множество с относително равномерно разпределение положителни – негативни коментари. (~4000)

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,79630	Precision negative: 0,78788	Precision overall : 0,79310
Recall positive: 0,86000	Recall negative: 0,70270	Recall overall : 0,79310
F1 positive: 0,82692	F1 negative: 0,74286	F1 overall : 0,79310

Миксиране на двата класификатора, като взимаме сумата от предсказаните положителни и отрицателни вероятности на двата класификатора, т.е max(пол1+пол2, нег1+нег2):

- Трениращото множество на английски език е с разпределение положителни – негативни коментари приблизително 2:1 (~6000).

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,75806	Precision negative: 0,88000	Precision overall : 0,79310
Recall positive: 0,94000	Recall negative: 0,59459	Recall overall : 0,79310
F1 positive: 0,83929	F1 negative: 0,70968	F1 overall : 0,79310

- Трениращо множество с относително равномерно разпределение положителни – негативни коментари (~4000).

Статистика за положителните коментари	Статистика за негативните коментари	Статистика за всички коментари
Precision positive: 0,80000	Precision negative: 0,81250	Precision overall : 0,80460
Recall positive: 0,88000	Recall negative: 0,70270	Recall overall : 0,80460
F1 positive: 0,83810	F1 negative: 0,75362	F1 overall : 0,80460

Заключение и бъдещо развитие

Относно семантичния анализ могат да бъдат включени различни категории (например, могат да се включат и неутрални коментари). Също така, коментарите могат да се разделят по степен на негативност/ позитивност.

Разпределение на задачите

Сияна Славова се зае със рефактуриране на стария проект и нагаждане на архитектурата му към новата версия. Също така обработи новите данни за английския класификатор.
Иван капукаранов се зае с интегрирането на АПИ-то на бинг за превод на коментарите. Също така част от експериментите се проведоха от него.

Код на проекта

Кодът на проекта е качен в GitHub на следния адрес:

https://github.com/skeleta/BGMamaProject.git

Литература и използвани източници

- Мнения, използвани за тестово и обучаващо множество:
  - http://www.booking.com
  - http://www.zahotelite.com/
  - http://sifaka.cs.uiuc.edu/~wang296/Data/index.html
- Мнения, използвани за категоризация и извличане на хотели:
  - http://www.bg-mamma.com/
- Алгоритъм за „Наивен Бейсов класификатор“ – лекции по „Машинно самообучение“, Г. Агре
- Библиотека за парсване на Json - https://github.com/google/gson
- Превод:
  - https://www.microsoft.com/en-us/translator/getstarted.aspx
  - https://github.com/boatmeme/microsoft-translator-java-api