

БГМама наръчник

Проект по "Извличане на информация и откриване на знания"

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

| | |
|---------------------------------------------------------------------------|----|
| Декларация за липса на плагиатство..... | 1 |
| Мотивация, Задача на курсовата работа | 3 |
| Мотивация | 3 |
| Идея | 3 |
| Задача за курса "ИИОЗ" | 3 |
| Задача за курса "ПОЕЗ" | 3 |
| Решение | 3 |
| Извличане на хотели ("ИИОЗ") | 3 |
| Семантичен анализ ("ПОЕЗ") | 5 |
| Програмна реализация..... | 5 |
| Предварителна обработка на данните | 5 |
| Парсване на коментар - пример: | 6 |
| Общ модел | 7 |
| Извличане на хотели ("ИИОЗ") | 8 |
| Скриншот: | 9 |
| Семантичен анализ ("ПОЕЗ") | 9 |
| Използван алгоритъм за построяване на "Наивен Бейсов класификатор": | 10 |
| Примерни обучаващи коментари:..... | 10 |
| Примерен тестов коментар:..... | 10 |
| Резултати от експерименти..... | 11 |
| Извличане на хотели ("ИИОЗ") | 11 |
| Семантичен анализ ("ПОЕЗ") | 11 |
| Заключение и бъдещо развитие | 0 |
| Извличане на хотели ("ИИОЗ") | 0 |
| Семантичен анализ ("ПОЕЗ") | 0 |
| Разпределение на задачите | 0 |
| Код на проекта | 1 |
| Литература и използвани източници | 1 |

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.

Ето защо решихме да направим "БгМама наръчник", който по подаден списък от коментари, ще намери тези с положителни отзиви и тези с отрицателни, ще ги маркира и ще извади желаните данни от тях.

Идея

Идеята е взета от предложените проекти от курса по "Извличане на информация", но е обогатена и доразвита.

Тъй като много пъти ни се е случвало да си търсим хотели, в които да пренощуваме по време на почивка, ние решихме да се съсредоточим именно върху извличането на този тип информация.

Идеята ни е да извлечем мнения за хотели от БГМама, да ги класифицираме по положителни и отрицателни, да извлечем хотелите като локации и да ги отбележим като локации в Google maps.

Задача за курса "ИИОЗ"

Задачата за курса по "Извличане на информация и откриване на знания" включва:

- Извличане на коментарите за хотели от json файла, предоставен ни от БгМама
- Разпознаване на хотелите като локации
- Добавяне на линкове към Google Maps за съответните хотели

Задача за курса "ПОЕЗ"

Задачата за курса по "Подходи за обработка на естествен език" включва:

- Семантичен анализ на коментарите за хотели
- Отбелязването на коментарите като положителни или отрицателни

Решение

Извличане на хотели ("ИИОЗ")

Реализацията на задачата постигнахме чрез интеграция с GATE. За целта свалихме Gate Embedded и построихме приложение, използвайки ресурси предоставени от фреймуърка. Приложенията включва:

- GATE Unicode Tokeniser - този, който идва директно с GATE без допълнителна модификация
- LingPipe Sentence Splitter - също по-подразбиране.
- BGLingPipe POS Tagger - ПОС тагер за български език. Включен е във GATE Embedded и използва BulTreeBank wordnet.

- ANNIE Gazetteer, съдържащ списък с градове (всички в България и някои извън България), познати курорти и списък с някои имена на хотели.
- JAPE Transducer - JAPE правила за извод на хотели.

Цялостната апликация изглежда по следния начин:

| Selected Processing resources | | |
|-------------------------------|-------------------------------------|-------------------------------|
| ! | Name | Type |
| | Document Reset PR | Document Reset PR |
| | GATE Unicode Tokeniser_00047 | GATE Unicode Tokeniser |
| | LingPipe Sentence Splitter PR_0000B | LingPipe Sentence Splitter PR |
| | BG LingPipe POS Tagger | LingPipe POS Tagger PR |
| | Dictionary | ANNIE Gazetteer |
| | MatchHotelAndPlace | JAPE Transducer |
| | MatchHotelName | JAPE Transducer |
| | MatchHotelNamesFromList | JAPE Transducer |
| | MatchHotelTypeAndName | JAPE Transducer |
| | MatchHotelTypeAndGeoLocation | JAPE Transducer |
| | MatchHotelParks | JAPE Transducer |
| | MatchParkAndPalace | JAPE Transducer |
| | MatchHotelAndPlace2 | JAPE Transducer |

А крайния резултат от изпълнението на апликацията върху извадка от няколко коментара изглежда по този начин:

The screenshot displays the GATE Developer 8.1 build 5169 interface. The left sidebar shows the project structure, including the 'Processing Resources' section with various JAPE transducers like 'MatchHotelAndPlace2', 'MatchParkAndPalace', 'ANNIE OrthMatcher', 'ANNIE NE Transducer', 'ANNIE POS Tagger', and 'ANNIE Sentence Splitter'. The central text editor shows a document with a paragraph of text. The right sidebar displays a list of annotations, including 'Hotel_Name', 'Lookup', 'Sentence', 'SpaceToken', 'Token', and 'Original markings'. The document view shows a paragraph of text with several words highlighted in red, indicating matches found by the JAPE transducers. The annotations list on the right shows the details of these matches, including the type of match (e.g., Hotel_Name, Lookup, Sentence, SpaceToken, Token) and the original text.

След като бяха постигнати задоволителни резултати, приложението се запазва и се извиква програмно от java кода, като се подават различни параметри за документа, с който ще работи предварително запазената апликация. Резултатите, които GATE връща се ползват за по-нататъшна обработка от приложението.

Семантичен анализ ("ПОЕЗ")

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.

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 данни 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 файловете 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:

- 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#msg29806558 date=1420132114]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 УЖАСНА ОРГАНИЗАЦИЯ за Нова Година! P.S. Честита Нова Година на всички! Желая ви повече приятни моменти в велинградските хотели![/quote]ловеч Здравейте, момичета! &lt;br /&gt;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; :mrgreen: ОРГАНИЗАЦИЯ за Нова Година! &lt;br /&gt;P.S. Честита Нова Година на всички! <b>ХТМЛ</b>Желая ви повече приятни моменти 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:

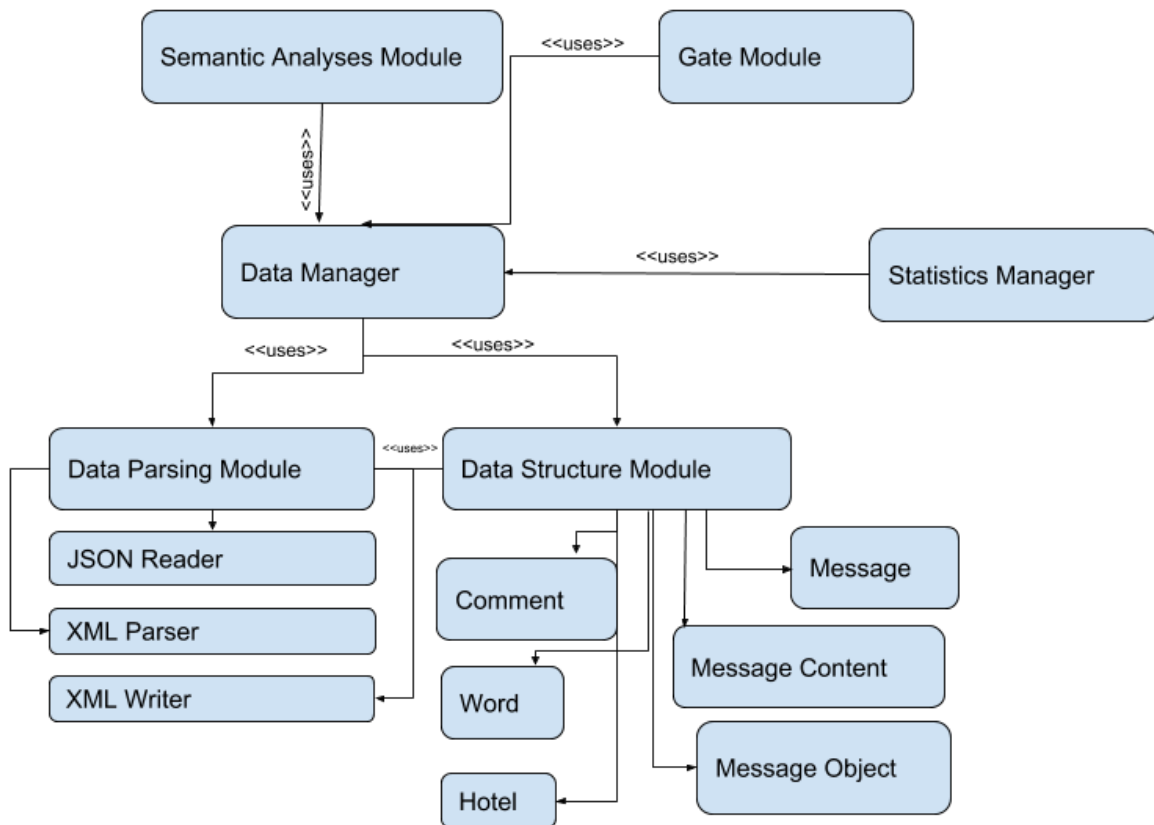
"msgbody Здравейте, момичета! Аз ще споделя само с едно изречение моите впечатления от Роял Спа Никога преди не съм била в толкова невероятно хубав хотел, с толкова УЖАСНО НЕПРОФЕСИОНАЛНА nbsp; ОРГАНИЗАЦИЯ за Нова Година! br / P.S. Честита Нова Година на всички! ХТМЛ Желая ви повече приятни моменти в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

Общ модел

Общият модел на архитектурата на приложението е показан на следната диаграма:



Както се вижда от схемата, и двата модула - този за семантичния анализ и този за извличането на хотели използват общ модел.

Извличане на хотели ("ИИОЗ")

Модулът за извличане на хотели работи с апликацията, създадена през Gate. Пайплайна на самата апликация е отбелязан по - горе в документацията.

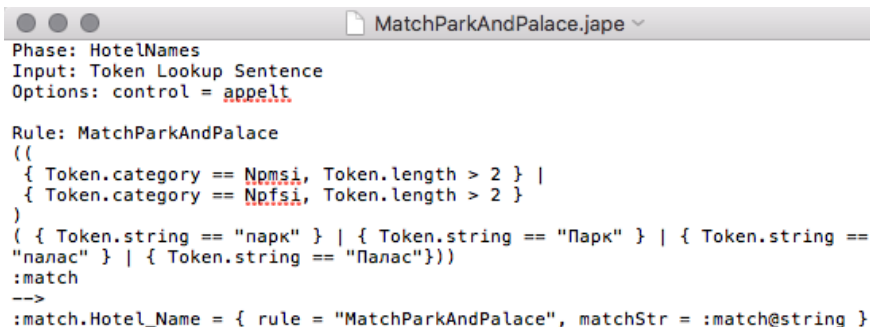
За Tokenizer и Sentence splitter сме използвали стандартни такива, който Gate предлага, без никакви модификации.

За POS targer използвахме LingPipe POS - tagger за български текст. Той е трениран върху BulTreeBank-DP [[Osenova & Simov 04](#), [Simov & Osenova 03](#), [Simov et al. 02](#), [Simov et al. 04a](#)].

За създаване на gazetteer използвахме ANNIE Gazetter, на който сме подали различни предварително подготвени списъци с градове, курорти, типове хотели и някои имена на хотели, както и Blacklist списък с често грешени думи. .def файла изглежда по следния начин:

```
BGcities.txt:geo:city
BGResorts.txt:geo:resort
HotelTypes.txt:hotel_type
HotelNames.txt:name
Blacklist.txt:blacklist
```

За намиране на имена на хотели използвахме JAPE Transducer с правила. Правилата изглеждат по следния начин:



```
Phase: HotelNames
Input: Token Lookup Sentence
Options: control = appelt

Rule: MatchParkAndPalace
((
  { Token.category == Npmsi, Token.length > 2 } |
  { Token.category == Npfsi, Token.length > 2 }
)
{ { Token.string == "парк" } | { Token.string == "Парк" } | { Token.string ==
"палас" } | { Token.string == "Палас"}}}
:match
-->
:match.Hotel_Name = { rule = "MatchParkAndPalace", matchStr = :match@string }
```

В това правило следим за категории Npmsi и Npfsi, които означават съответно:

- Npmsi - Noun, proper noun, masculine, singular, indefinite
- Npfsi - Noun, proper noun, feminine, singular, indefinite

След това, ако следващата дума е "парк" или "палас" махваме целия стринг като един хотел. Това правило ще магне име на хотел, изписано по следния начин: "Елина Палас".

Скриншот:

| Хотел | Линк |
|--------------------------|--------------------------------------------------|
| ХОТЕЛ ЕВРЕДИКА ПАМПОРОВО | www.google.bg/maps/search/ХОТЕЛЕВРЕДИКАПАМПОРОВО |
| хотел Сарай | www.google.bg/maps/search/хотелСарай |
| хотел Севтополис | www.google.bg/maps/search/хотелСевтополис |
| хотел Севтополис | www.google.bg/maps/search/хотелСевтополис |
| хотел Викони | www.google.bg/maps/search/хотелВикони |
| хотел Старосел | www.google.bg/maps/search/хотелСтаросел |
| Гранд хотел Поморие | www.google.bg/maps/search/ГрандхотелПоморие |
| Гранд хотел София | www.google.bg/maps/search/ГрандхотелСофия |
| хотел Сенс | www.google.bg/maps/search/хотелСенс |
| Гранд хотел София | www.google.bg/maps/search/ГрандхотелСофия |
| хотел Сенс | www.google.bg/maps/search/хотелСенс |
| хотел Севтополис | www.google.bg/maps/search/хотелСевтополис |
| хотел Севтополис | www.google.bg/maps/search/хотелСевтополис |
| хотел Севтополис | www.google.bg/maps/search/хотелСевтополис |
| хотел Севтополис | www.google.bg/maps/search/хотелСевтополис |
| Хотел Велинград | www.google.bg/maps/search/ХотелВелинград |
| хотел Марая | www.google.bg/maps/search/хотелМарая |
| хотел Сарай | www.google.bg/maps/search/хотелСарай |

| Хотел | Линк |
|------------------------------------------|-----------------------------------------------------------------|
| Кемпински, Банско | www.google.bg/maps/search/Кемпински,Банско |
| Лъкшъри Банско | www.google.bg/maps/search/ЛъкшъриБанско |
| България Лъки | www.google.bg/maps/search/БългарияЛъки |
| Кемпински, Банско | www.google.bg/maps/search/Кемпински,Банско |
| Лъкшъри Банско | www.google.bg/maps/search/ЛъкшъриБанско |
| хотел Севтополис в Павел Баня | www.google.bg/maps/search/хотелСевтополисвПавелБаня |
| хотел Севтополис в Павел Баня | www.google.bg/maps/search/хотелСевтополисвПавелБаня |
| хотел Севтополис в Павел Баня | www.google.bg/maps/search/хотелСевтополисвПавелБаня |
| хотел Севтополис в Павел Баня | www.google.bg/maps/search/хотелСевтополисвПавелБаня |
| Свети СпасАкватоникГранд Хотел Велинград | www.google.bg/maps/search/СветиСпасАкватоникГрандХотелВелинград |
| хотел Севтополис в Павел Баня | www.google.bg/maps/search/хотелСевтополисвПавелБаня |
| Кемпински Гранд Арена Банско | www.google.bg/maps/search/КемпинскиГрандАренаБанско |
| Пирин Сандански | www.google.bg/maps/search/ПиринСандански |
| Кемпински Гранд Арена Банско | www.google.bg/maps/search/КемпинскиГрандАренаБанско |
| Пирин Сандански | www.google.bg/maps/search/ПиринСандански |
| Кемпински Гранд Арена Банско | www.google.bg/maps/search/КемпинскиГрандАренаБанско |
| Пирин Сандански | www.google.bg/maps/search/ПиринСандански |
| Спа Централ в Хисаря | www.google.bg/maps/search/СпаЦентралвХисаря |

Семантичен анализ ("ПОЕЗ")

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. Образоваме речника само от български думи.

Използван алгоритъм за построяване на "Наивен Бейсов класификатор":

LEARN_NAIVE_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* \leftarrow 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:
 - $docs_j \leftarrow$ подмножество на документи от *Примери*, за които стойността на целевия атрибут е v_j .
 - $P(v_j) \leftarrow \frac{|docs_j|}{|Примери|}$
 - $Text_j \leftarrow$ един общ документ, получен чрез обединение на всички членове на $docs_j$
 - $n \leftarrow$ общия брой на различни позиции на думи в $Text_j$
 - за всяка дума w_k от *Речник* направи:
 - $n_k \leftarrow$ броя на срещане на думата w_k в $Text_j$
 - $P(w_k | v_j) = \frac{n_k + 1}{n + |Речник|}$

CLASSIFY_NAIVE_BAYES_TEXT(Документ)

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

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

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

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 форсирват колите нагоре по "Черни връх".
</comment>
```

Примерен тестов коментар:

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

Скрийншот:

| Коментар | Категория |
|---------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| - Прави сте за тази болест. Аз научих, че случаите на болни кучета зачестяват и преди да отидем на море, кучето беше заведено на ветери... | ClassTypeNegative |
| - За Роял Спа или хубаво или нищо в тази тема, иначе ви изкарват капризни, необективни и т.н. ... | ClassTypeNegative |
| - За Роял Спа или хубаво или нищо в тази тема, иначе ви изкарват капризни, необективни и т.н. ... | ClassTypeNegative |
| - Вероятно по стар български обичай уикенда гледат нещата да са на шест, а през седмицата не се стараят толкова. Кое е разочароващ... | ClassTypeNegative |
| - Наистина имаше дни в които водата е по хладка във външния басейн , предполагам че я подгряват Джакузита във бяха студени !Най ... | ClassTypeNegative |
| - Ето го джакузито - горе в дясноТова се пада на гърба на Руската баня и се вижда от една зала със затоплени каменни легла, която винаг... | ClassTypeNegative |
| - Ето го джакузито - горе в дясноhttpxarmsD.Не знам кога е направено Ахааа, ясно, това нещо лятото беше в строеж, беше оградено с дъ... | ClassTypeNegative |
| - fatalina, благодаря много за хубавия и изчерпателен отговор!това ми е най важното хотела да е чист и топъл,храната да става,басейна и... | ClassTypeNegative |
| + От прочетеното из нета - България, да, но за Рим, аз даже съм го хвалила тук, месеци напред не можеш да намериш места в него, прозвъ... | ClassTypePositive |
| - evhen,мерси много за информацията!Направо съм в чуденка вече а нормално по евтините оферти да привличат хората! | ClassTypeNegative |
| - Нямам предвид конкретен хотел. Така на англо и пренавито казани думи за Казабланка, може да сподели само човек, който завижда на ... | ClassTypeNegative |
| - fatalina, благодаря много за хубавия и изчерпателен отговор!това ми е най важното хотела да е чист и топъл,храната да става,басейна и... | ClassTypeNegative |
| + Благодаря ти много, вече на 100% се навих да избира този хотел и още сега ще направя резервация!Все пак като пари е изключително д... | ClassTypePositive |
| + Кики не е късно,ще чакам с нетърпение отзив за хотел Жери тъй като точно на него съм се спряла и утре мислех да правя резервация м... | ClassTypePositive |
| - Кики не е късно,ще чакам с нетърпение отзив за хотел Жери тъй като точно на него съм се спряла и утре мислех да правя резервация ма... | ClassTypeNegative |
| - х-л Сарай Твърдо НЕ! Била съм там преди три години- мръсен басейн, ама много мръсен, храната горе-долу и в това село няма какво д... | ClassTypeNegative |
| - Не е лъжа, на мен не ми хареса и няма да повторя. Чудя се между Helios Bay и Sol Luna има 25% за ранно записване.Казабланка е за непр... | ClassTypeNegative |
| - Кики много се радвам че се включи със съвсем прясна информация!изчетох написаното от теб и в крайна сметка и аз резервирах за кра... | ClassTypeNegative |
| + Здравейте!Тук за хотел Рим питаше една мама. Той е почти в идеалния център, близо до Рич и до градския басейн долепен почти Има г... | ClassTypePositive |
| + Някакво актуално мнение за хотел Аура? Искам да отседна във Велинград за няколко дни, а не ми се дават луди суми като в хотелите от ... | ClassTypePositive |
| - Кики много се радвам че се включи със съвсем прясна информация!изчетох написаното от теб и в крайна сметка и аз резервирах за кра... | ClassTypeNegative |
| - Кики, това пояснение което даваш за кой хотел се отнася? Благодаря! | ClassTypeNegative |
| - militana, за Аура писахме малко по-назад - 1-2 страници.Може ли да ми дадете и пресни впечатления от Акватоник - храна, как са цен... | ClassTypeNegative |
| - С теб имаме явно еднакъв вкус, така че ще си позволя да ти предложа Сол Луна бей. Аз съм много доволна. Много по-напред е от Казаб... | ClassTypeNegative |
| - С теб имаме явно еднакъв вкус, така че ще си позволя да ти предложа Сол Луна бей. Аз съм много доволна. Много по-напред е от Казаб... | ClassTypeNegative |
| - Кики много много благодаря за страхотното описание и вече съм сигурна че не съм сбъркала при избора на хотел и то за без пари!а от... | ClassTypeNegative |
| - Отзиви за Аспен апартохотел? | ClassTypeNegative |
| - Може ли актуален сайт на Веника Палас!Понеже намирам само през туристически агенции!Благодаря | ClassTypeNegative |
| - Отзиви за Аспен апартохотел? Nadeto ,първо да те попитам за х-л Аспен ли питаш, защото има и х- л Аспен ризорт ,гольф ,ски и спа ... | ClassTypeNegative |
| - Аз съм написала апартохотел Аспен Банско.Става въпрос за тази оферта http://www.vipoferta.bg/offer.php?offer_id=5398 | ClassTypeNegative |
| - Сибирин,мисля че Grand Hotel PaskoУбаво СПА, доста хубаво. Много изгодни цени,отлични услуги,хотелът е в прекрасна обстановка.Мисля че... | ClassTypePositive |

Резултати от експерименти

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

Извличане на хотели ("ИИОЗ")

Precision : 0.83333

Recall : 0.76923

F1 : 0.80000

Семантичен анализ ("ПОЕЗ")

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

Заклучение и бъдещо развитие

Приложението може да се развие по много различни начини.

Извличане на хотели ("ИИОЗ")

Относно извличането на хотели, могат да се направят следните подобрения:

- Да се разпознават хотели и в английски текст
- Да се разпознават хотели в коментари, написани на латиница
- Да се разпознават не само хотели, но и някакви забележителности

Семантичен анализ ("ПОЕЗ")

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



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/>
- Мнения, използвани за категоризация и извличане на хотели:
 - <http://www.bg-mamma.com/>
- BulTreeBank Morphosyntactic Tagset
- Ling Pipe - <https://gate.ac.uk/sale/tao/splitch23.html#x28-57800023.24.3>
- <https://gate.ac.uk/>
- Алгоритъм за „Наивен Бейсов класификатор“ – лекции по „Машинно самообучение“, Г. Агре