Изготвили:

Иван Капукаранов, ФН: 24958, 1 курс, ИИОЗ

Сияна Славова, ФН: 24963, 1 курс, ИИОЗ

БГМама наръчник

Проект по "Извличане на информация"

# 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](#_Toc443683479)

[Мотивация, Задача на курсовата работа 3](#_Toc443683480)

[Мотивация 3](#_Toc443683481)

[Идея 3](#_Toc443683482)

[Задача за курса "ИИ" 3](#_Toc443683483)

[Задача за курса "ПОЕЗ" 3](#_Toc443683484)

[Решение 3](#_Toc443683485)

[Извличане на хотели ("ИИ") 3](#_Toc443683486)

[Семантичен анализ ("ПОЕЗ") 5](#_Toc443683487)

[Програмна реализация 5](#_Toc443683488)

[Предварителна обработка на данните 5](#_Toc443683489)

[Парсване на коментар - пример: 6](#_Toc443683490)

[Общ модел 7](#_Toc443683491)

[Извличане на хотели ("ИИ") 8](#_Toc443683492)

[Скрийншот: 9](#_Toc443683493)

[Семантичен анализ ("ПОЕЗ") 9](#_Toc443683494)

[Използван алгоритъм за построяване на "Наивен Бейсов класификатор": 10](#_Toc443683495)

[Примерни обучаващи коментари: 10](#_Toc443683496)

[Примерен тестов коментар: 10](#_Toc443683497)

[Резултати от експерименти 11](#_Toc443683498)

[Извличане на хотели ("ИИ") 11](#_Toc443683499)

[Семантичен анализ ("ПОЕЗ") 11](#_Toc443683500)

[Заключение и бъдещо развитие 0](#_Toc443683501)

[Извличане на хотели ("ИИ") 0](#_Toc443683502)

[Семантичен анализ ("ПОЕЗ") 0](#_Toc443683503)

[Разпределение на задачите 0](#_Toc443683504)

[Код на проекта 0](#_Toc443683505)

[Литература и използвани източници 1](#_Toc443683506)

# 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 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

. За целта използвахме сет с обучаващи данни (120+ коментара) и сет с тестови данни (87 коментара)

-използване на данни на ае от Trip advisor

- различни експерименти

- refactoring

- богатяване на тестовото множество

# 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:

* 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&amp;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!

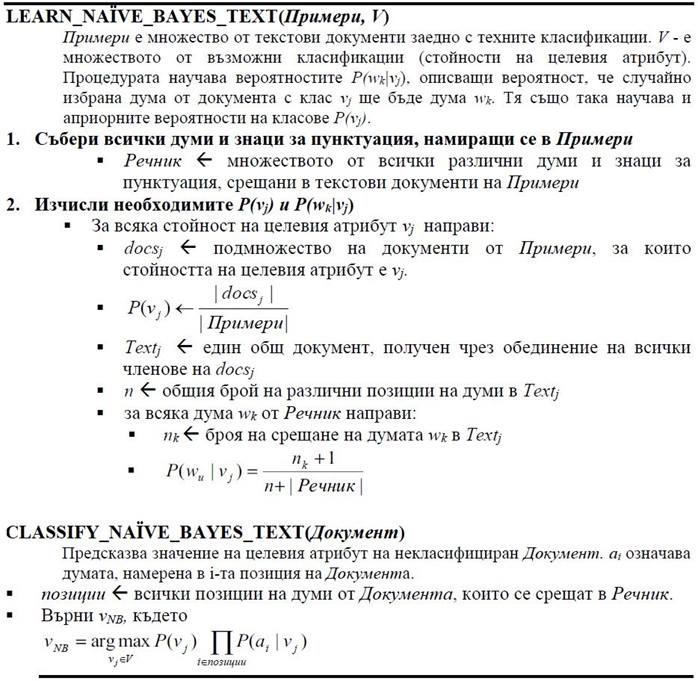
<topic>Re: ДУБАЙ 7-ма тема Хотел</topic>

</comment>

Същото и за trip advisor данните

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. Образуваме речника само от български думи.

### Използван алгоритъм за построяване на "Наивен Бейсов класификатор":



### Примерни обучаващи коментари:

* Позитивен



* Негативен



### Примерен тестов коментар:



# Резултати от експерименти

## Версия 1

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

Статистика за положителните коментари

--------------------------------------

Precision positive: 0.77273

Recall positive: 0.94444

F1 positive: 0.85000

--------------------------------------

Статистика за негативните коментари

--------------------------------------

Precision negative: 0.95238

Recall negative: 0.80000

F1 negative: 0.86957

--------------------------------------

Статистика за всички коментари

--------------------------------------

Precision overall : 0.86047

Recall overall : 0.86047

F1 overall : 0.86047

--------------------------------------

## Версия 2

Добавяне на още данни в тестовия сет

* Преведени коментари в/у ЕН класификатор без премахване на стоп думите
* Преведени коментари в/у ЕН класификатор с премахване на стоп думите
* Миксиране на двата класификатора, като взимаме по –голямата вероятност от двата (погрешно)
* Миксиране като избираме макс от пол1+пол2 и нег1+нег2
* Класифициране от класификатор с обучаващо множество от класифицираните данни от АЕ класификатор

# 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

# 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/>
* Алгоритъм за „Наивен Бейсов класификатор“ – лекции по „Машинно самообучение“, Г. Агре
* Библиотека за парсване на Json - https://github.com/google/gson