Изготвили:

Иван Капукаранов, ФН: 24958, 1 курс, ИИОЗ

Сияна Славова, ФН: 24963, 1 курс, ИИОЗ

БГМама наръчник

Проект по "Извличане на информация и откриване на знания"

# 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](#_Toc443638774)

[Мотивация, Задача на курсовата работа 3](#_Toc443638775)

[Мотивация 3](#_Toc443638776)

[Идея 3](#_Toc443638777)

[Задача за курса "ИИОЗ" 3](#_Toc443638778)

[Задача за курса "ПОЕЗ" 3](#_Toc443638779)

[Решение 3](#_Toc443638780)

[Извличане на хотели ("ИИОЗ") 3](#_Toc443638781)

[Семантичен анализ ("ПОЕЗ") 5](#_Toc443638782)

[Програмна реализация 5](#_Toc443638783)

[Предварителна обработка на данните 5](#_Toc443638784)

[Парсване на коментар - пример: 6](#_Toc443638785)

[Общ модел 7](#_Toc443638786)

[Извличане на хотели ("ИИОЗ") 8](#_Toc443638787)

[Скрийншот: 9](#_Toc443638788)

[Семантичен анализ ("ПОЕЗ") 9](#_Toc443638789)

[Използван алгоритъм за построяване на "Наивен Бейсов класификатор": 10](#_Toc443638790)

[Примерни обучаващи коментари: 10](#_Toc443638791)

[Примерен тестов коментар: 10](#_Toc443638792)

[Резултати от експерименти 11](#_Toc443638793)

[Извличане на хотели ("ИИОЗ") 11](#_Toc443638794)

[Семантичен анализ ("ПОЕЗ") 11](#_Toc443638795)

[Заключение и бъдещо развитие 0](#_Toc443638796)

[Извличане на хотели ("ИИОЗ") 0](#_Toc443638797)

[Семантичен анализ ("ПОЕЗ") 0](#_Toc443638798)

[Разпределение на задачите 0](#_Toc443638799)

[Код на проекта 1](#_Toc443638800)

[Литература и използвани източници 1](#_Toc443638801)

# 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

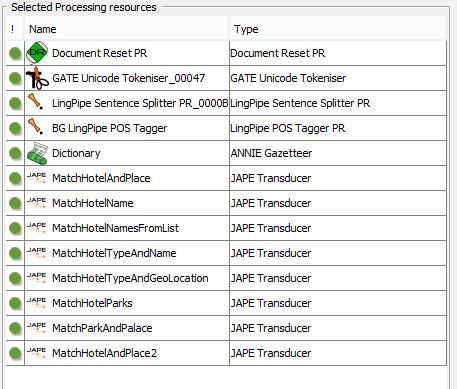
# Решение

## Извличане на хотели ("ИИОЗ")

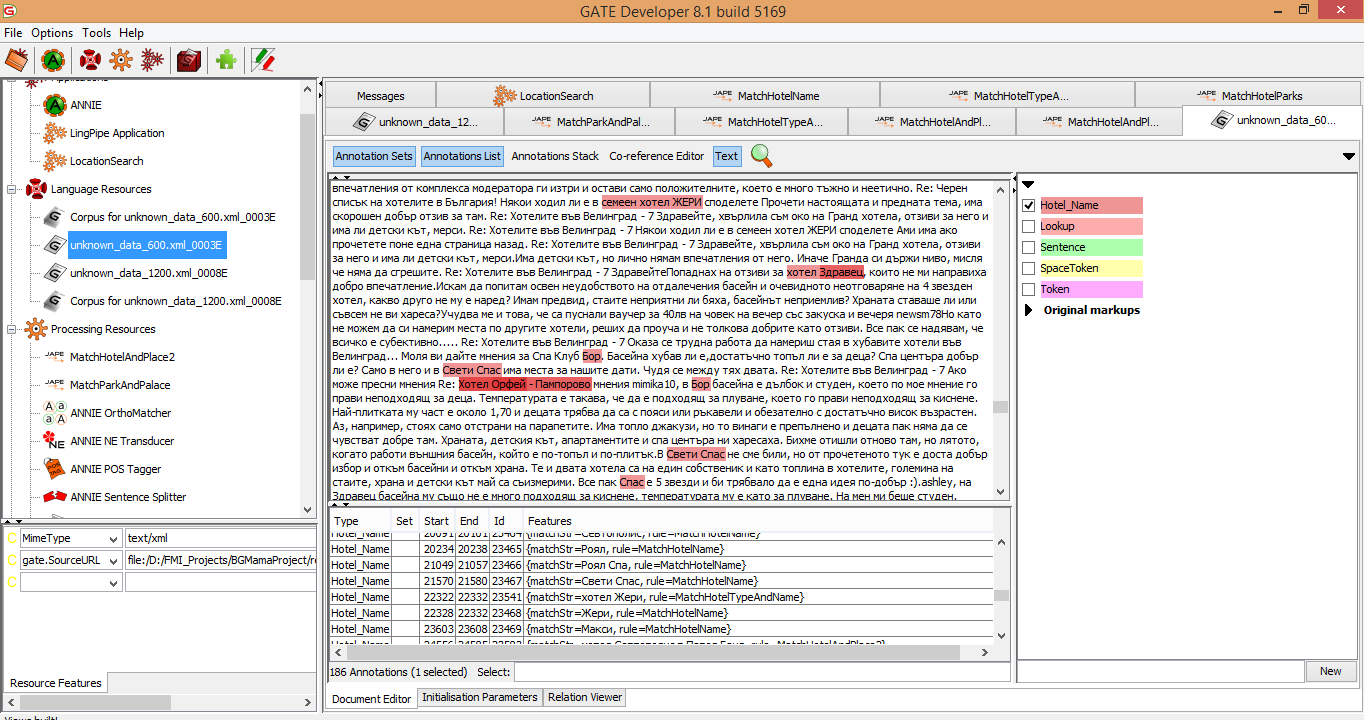
Реализацията на задачата постигнахме чрез интеграция с GATE. За целта свалихме Gate Embedded и построихме приложение, използвайки ресурси предоставени от фреймуърка. Приложените включва:

* 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), по-известни курорти и списък с някои имена на хотели.
* JAPE Transducer - JAPE правила за извод на хотели.

Цялостната апликация изглежда по следния начин:



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



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 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">

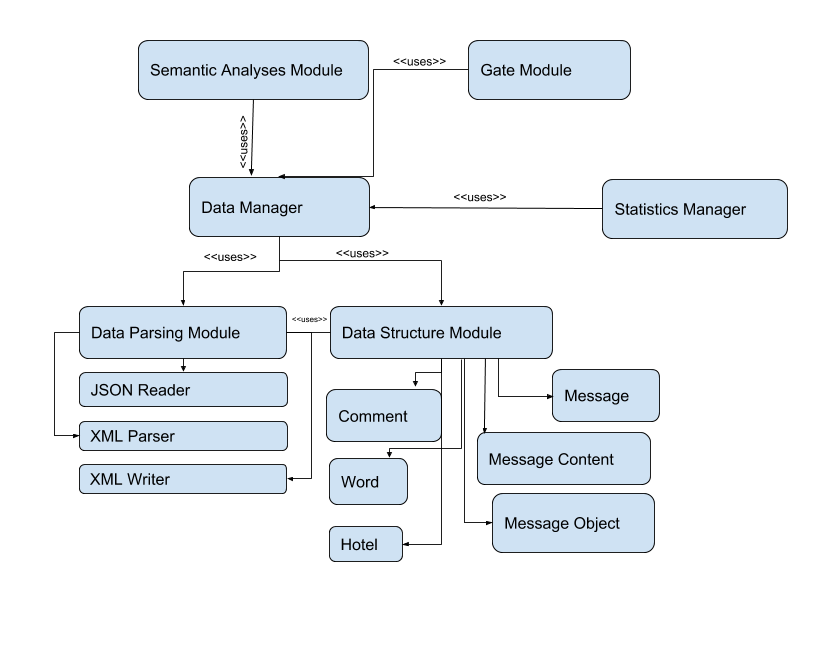
Здравейте, момичета! Аз ще споделя само с едно изречение моите впечатления от Роял Спа Никога преди не съм била в толкова невероятно хубав хотел, с толкова УЖАСНО НЕПРОФЕСИОНАЛНА nbsp; ОРГАНИЗАЦИЯ за Нова Година! br / P.S. Честита Нова Година на всички! ХТМЛ Желая ви повече приятни моменти в велинградските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

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

</commentt>

## 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.

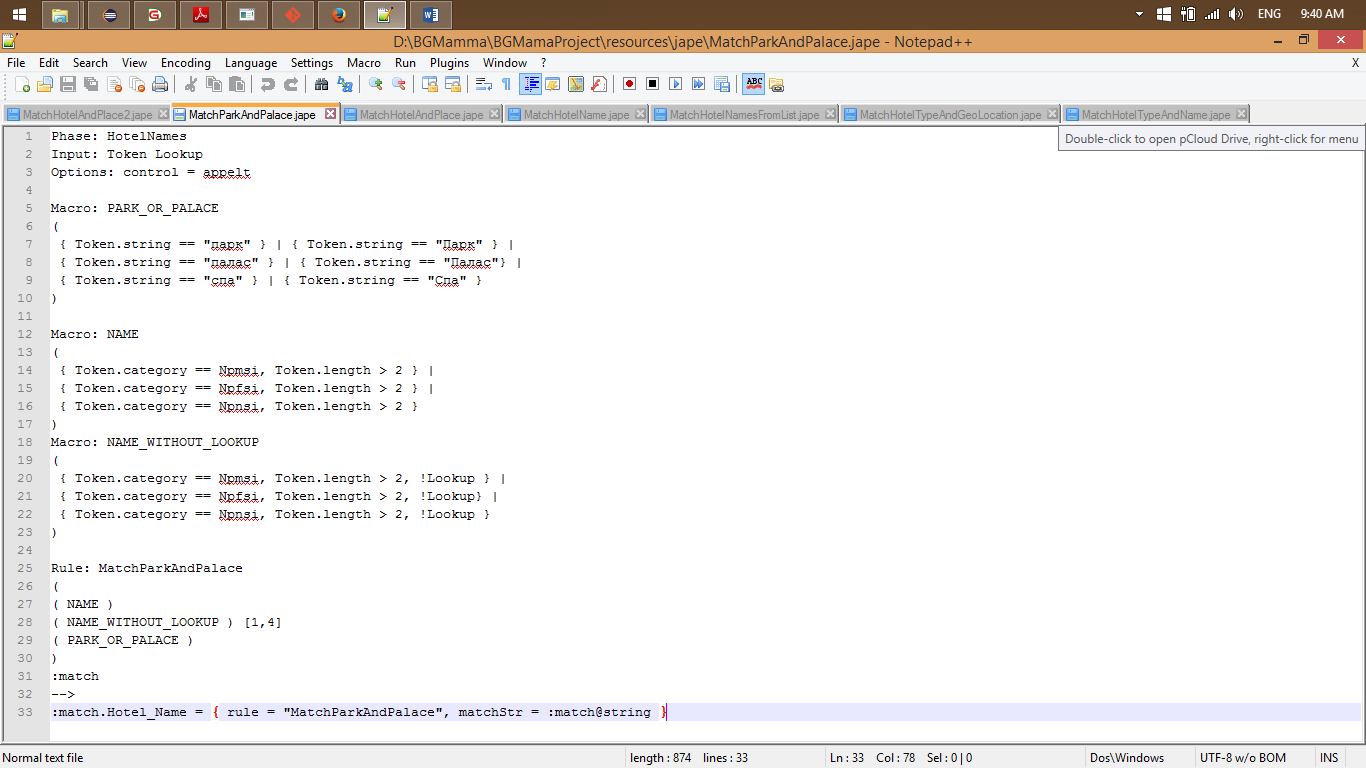
За Tokenizer и Sentence splitter сме използвали стандратни такива, който Gate предлага, без никакви модификации.

За POS тагер използвахме LingPipe POS - tagger за български текст. Той е трениран върху BulTreeBank-DP [[Osenova & Simov 04](https://gate.ac.uk/sale/tao/splitli2.html#Xbtb-stylebook), [Simov & Osenova 03](https://gate.ac.uk/sale/tao/splitli2.html#Xbtb-2003), [Simov *et al.* 02](https://gate.ac.uk/sale/tao/splitli2.html#Xbtb-2002), [Simov *et al.* 04a](https://gate.ac.uk/sale/tao/splitli2.html#Xbtb-2004)].

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:



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

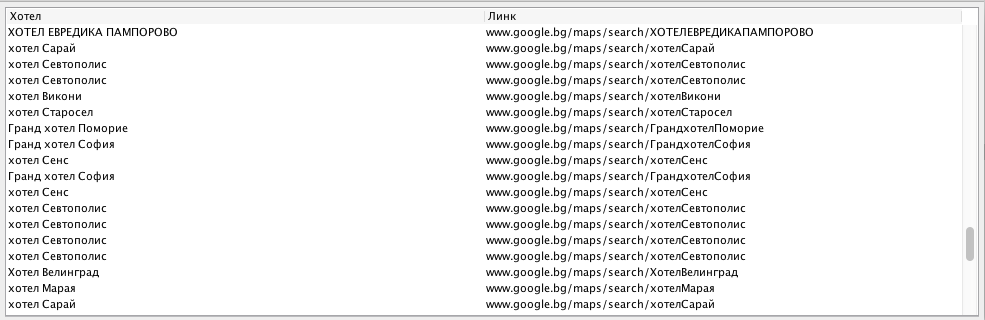


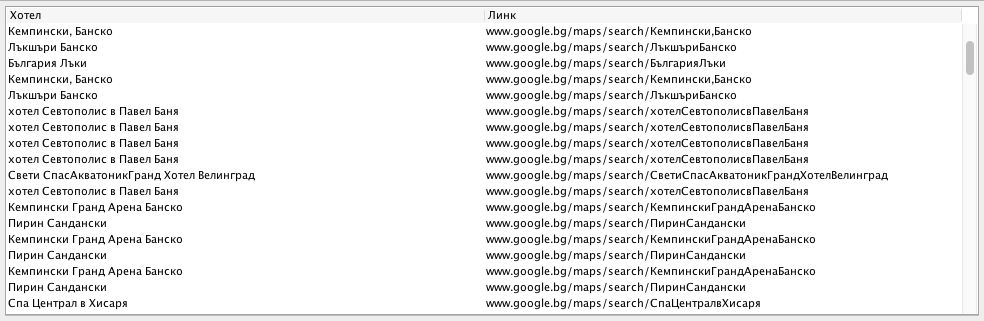
В това правило следим за категории Npmsi и Npfsi, които означават съответно:

* Npmsi - Noun, proper noun, masculine, singular, indefinite
* Npfsi - Noun, proper noun, feminine, singular, indefinite
* Npnsi - Noun, proper noun, neuter, singular, indefinite

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

### Скрийншот:

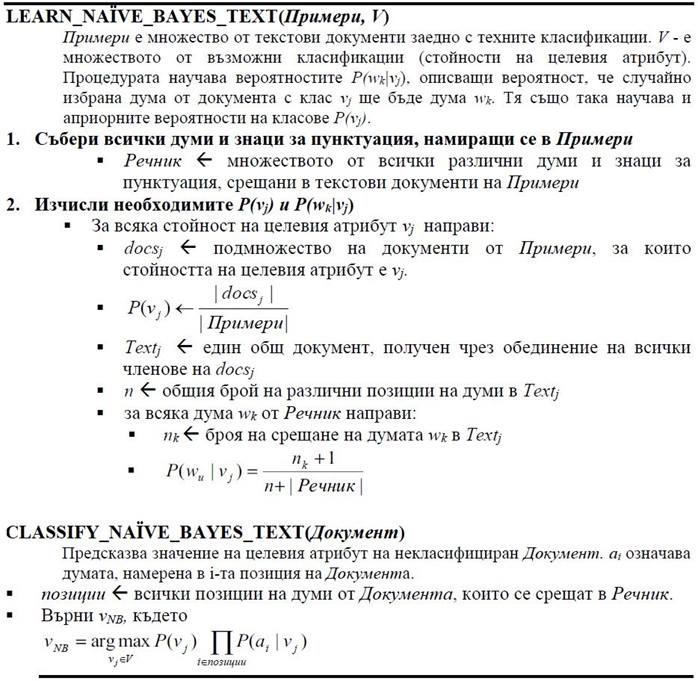




## Семантичен анализ ("ПОЕЗ")

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. Образуваме речника само от български думи.

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



### Примерни обучаващи коментари:

* Позитивен



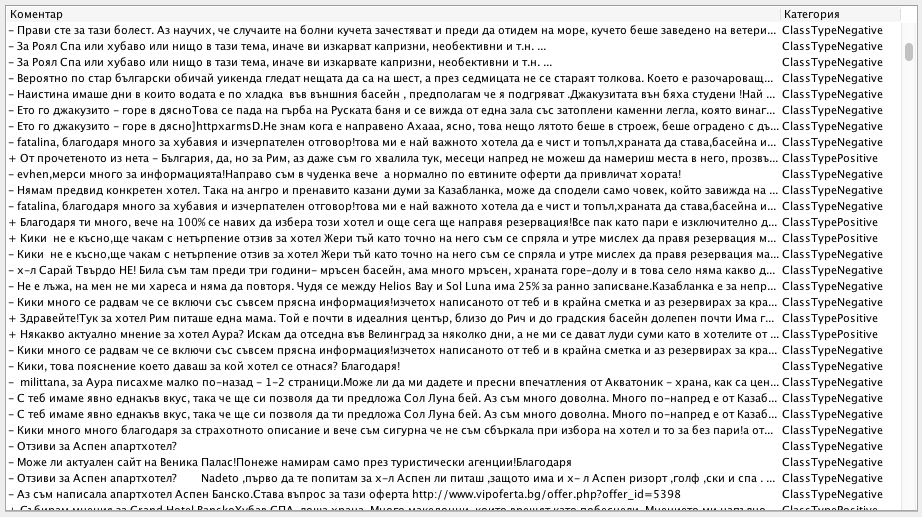
* Негативен



### Примерен тестов коментар:



Скрийншот:



# Резултати от експерименти

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

## Извличане на хотели ("ИИОЗ")

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

Precision : 0.94286

Recall : 0.82500

F1 : 0.88000

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

## Семантичен анализ ("ПОЕЗ")

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

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

# Заключение и бъдещо развитие

Приложението може да се развие по много различни начини.

## Извличане на хотели ("ИИОЗ")

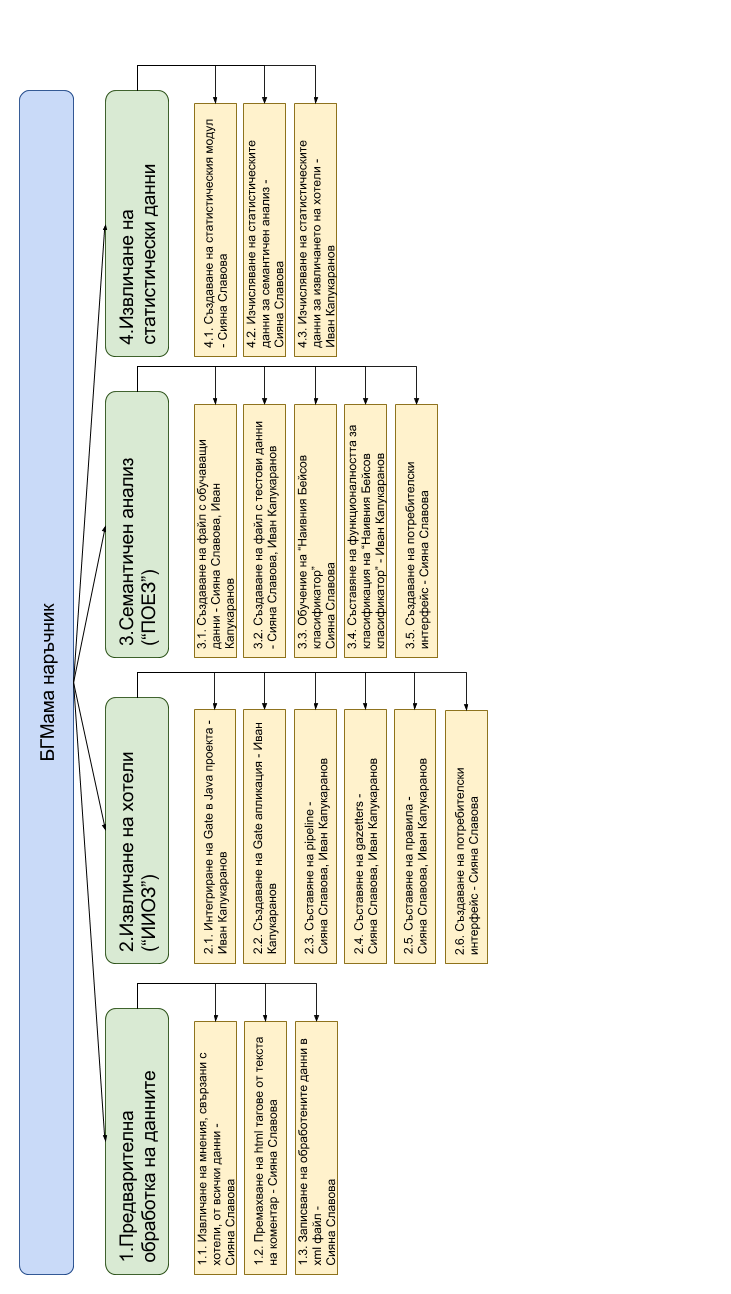
Относно извличането на хотели, могат да се направят следните подобрения:

* 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
* Библиотека за парсване на Json - https://github.com/google/gson