Класификатор на коментари

Проект за курса „Откриване на знания в текст“

СИЯНА СЛАВОВА, 24963

ИВАН КАПУКАРАНОВ, 24958

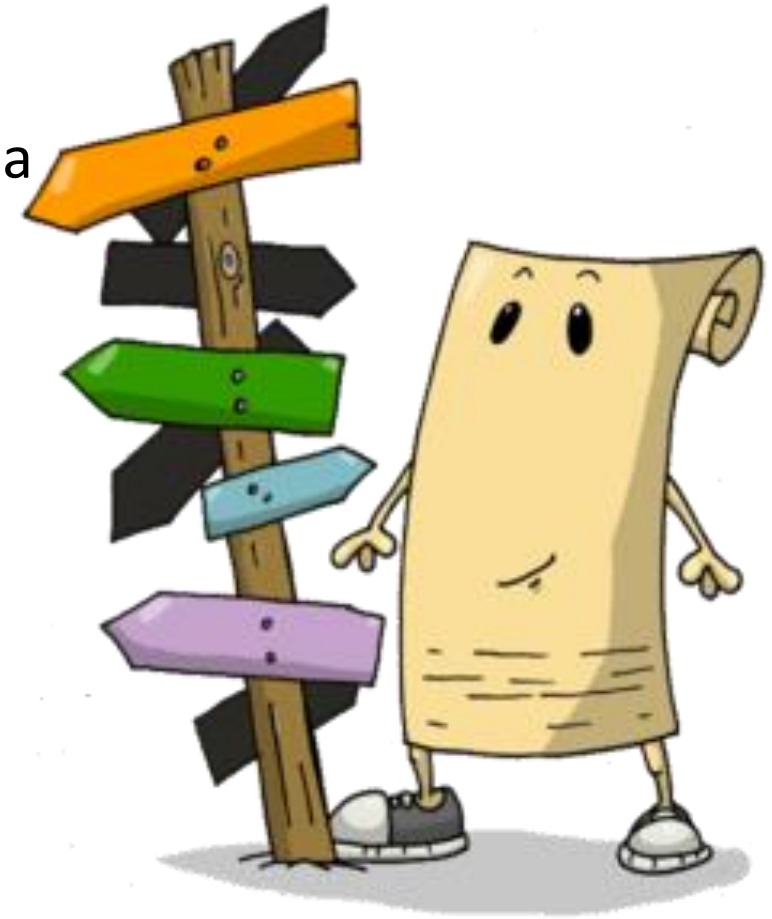
Наследството от миналия семестър

- Покрай наръчника за BG-Mamma и откриването на локации в него направихме и прототип на класификатор.

- "Наивен Бейсов Класификатор", който реализирахме сами
  - Обучаващи данни (120 коментара)
  - Тестови данни (44 коментара).

Продължението

- Обогатихме обучаващото множество
  - Взехме данни от английски и ги преведохме
    - Превеждахме чрез Bing translator api.
  - На английски език има много подходящи множества
    - Използвахме данни то Trip advisor.
- Нов класификатор за английски език



Защо?

- Прекарахме 1/3 от времето за предишния проект в ръчно търсене и класифициране на коментари.
- Не беше много приятно.
- Нека си направим наши класификатори.

Програмна реализация

- Данните от БГ-Мама вече бяха парснати.
- Данните на английски се нуждаеха от нов парсер.
  - Категоризиране
    - Преобразувахме оценките 1,2,3,4,5 до позитивен и негативен клас.

Класификаторите

- Обучаващ се на английски
  - Ползва обучаващия сет на английски
    - 6000 - 2:1 положителни към отрицателни
    - 4000 - равномерно разпределени
  - Премахнахме стоп думите
- Обучаващ се на български
  - Преведохме българските некласифицирани коментари от БГ мама
    - 4000 коментара
  - Класифицирахме преведените коментари с английския класификатор
  - Пуснахме ги за обучаващо множество заедно с оригиналните 120 български коментара

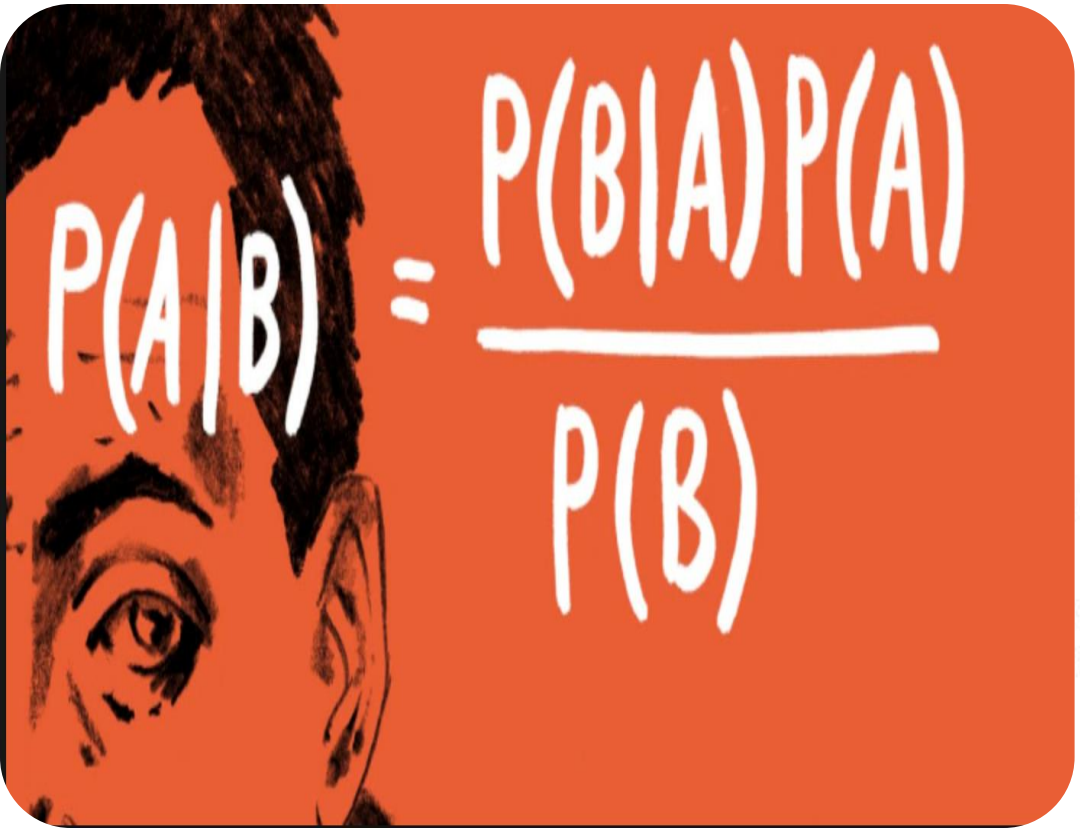
Класификаторите

- Миксиран
  - Взимаме сумата от предсказаните положителни и отрицателни вероятности на двата класификатора.

$$max\ (positive_1 + positive_2\ , negative_1 + negative_2)$$

Класификаторите

- 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

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

Резултати

- 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

Резултати

- Обучаващо множество на българския класификатор, предварително класифицирано от английския:
  - разпределение положителни – негативни коментари приблизително 2:1 (~6000).

Статистика за положителните коментари

Precision positive: 0,79630

Recall positive: 0,86000

F1 positive: 0,82692

Статистика за негативните коментари

Precision negative: 0,78788

Recall negative: 0,70270

F1 negative: 0,74286

Статистика за всички коментари

Precision overall : 0,79310

Recall overall : 0,79310

F1 overall : 0,79310

Резултати

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

Статистика за положителните коментари

Precision positive: 0,80000

Recall positive: 0,88000

F1 positive: 0,83810

Статистика за негативните коментари

Precision negative: 0,81250

Recall negative: 0,70270

F1 negative: 0,75362

Статистика за всички коментари

Precision overall : 0,80460

Recall overall : 0,80460

F1 overall : 0,80460

Бъдещо развитие

- Могат да бъдат включени различни категории
  - могат да се включат и неутрални коментари.
- Коментарите могат да се разделят по степен на негативност/ позитивност.

Благодарим ви за вниманието!