Задание за самостоятелна разработка към тема Модели на езика, Вероятностни модели.

Отговорете в писмен вид и, когато е нужно, приложете схеми, на следните въпроси.

1. Запишете верижното правило за условна вероятност.
2. Какво се нарича Модел на езика.
3. Запишете приемането на Марков.
4. Запишете, на базата честоти, статистическа оценка за условните вероятности за биграмите в корпус, съставен от пет изречения, като тези изречиния са взети ( ваш избор кoи точно) от вашия мейл-обмен на тема дистанционно обучение.
5. Запишете какво е ентропия по Шанон, като обясните означенията с думи.
6. Запишет как се извежда каква е ентрапията на равнорапрадалена случайна величина.
7. Запишете в свободен текст какво е перплексност.
8. Запишете в свободен текст защо според вас перплексността измерена за биграмен модел на езика е много по-малка от тази за униграмен.

Предайте просто файл с текст и схеми.

1.  
P(x1, x2, x3, … xn) = P(x1) \* P(x2|x1) \* P(x3|x1,x2) … \* P(xn|x1, … xn-1)

2.   
Целта на вероятностните езикови модели е да се присъди вероятност на дадено изречение. Модел, който пресмята P(W) или P(wn|w1,w2…wn-1) се нарича модел на езика.

3.   
Вместо да изчисляваме вероятността на дадена дума при условие предходните пет думи,  
изчисляваме вероятността само по отношение на предходната дума или пък по отношение на предходните две или N думи.  
P(wi | w1,w2,w3....wi-1) ≈ P(wi | wi-1)  
P(wi | w1,w2,w3....wi-1) ≈ P(wi | wi-1, wi-2)  
P(wi | w1,w2,w3....wi-1) ≈ P(wi | wi-1, wi-2, ... wi-k)

4.  
Използвайте тази тема във форума за да задавате въпроси.  
Както знаете лекциите ще бъдат провеждани дистанционно.  
За въпроси и коментари към лекциите, ще бъдат отворени теми във форума.  
Моля, пишете, ако имате въпроси за процедурата от форума!  
Лекциите ще бъдат качвани в полетата за съответната тема във времето, когато има такива.

P(wi | wi-1) = брой(wi – 1, wi) / брой(wi – 1)  
  
Използвайте тази тема във форума за да задавате въпроси.  
P(използвайте | <s>) = 1 / 5  
P(тази | използвайте) = 1 / 1  
P(тема | тази) = 1 / 2  
P(във | тема) = 2 / 2  
P(форума | във) = 2 / 3  
P(за | форума) = 1 / 3  
P(да | за) = 1 / 1  
P(задавате | да) = 1 / 1  
P(въпроси | задавате) = 1 / 3  
  
По същия начин следващите изречения.  
  
5.  
Ентропия е мярка на неопределеност. Увеличаването на ентропията води до повече (по-стойностна) информация, а намаляването - до по-малко. Ентропията мери средната информация.  
Ентропията Н за дискретна случайна величина Х, която заема стойности х1, х2, х3... e   
H(X) = − ∑ P(X=x) log (P(X=x))  
 x

6.  
H(X) = − log (

7.  
Перплексност или недоумението е мярка за прогнозирането на вероятностен модел. Може да се използва за сравняване на вероятностни модели. Високата перплексност показва, че моделът е лош в прогнозирането на тестови данни, а ниската перплексност – че е добър.

8.  
Защото има повече контекст за предсказване на следващата дума. Ако има две, три или повече думи в дадения контекст, е много по-лесно предвидимо каква ще бъде следващата дума и по този начин намаля перплексността си.