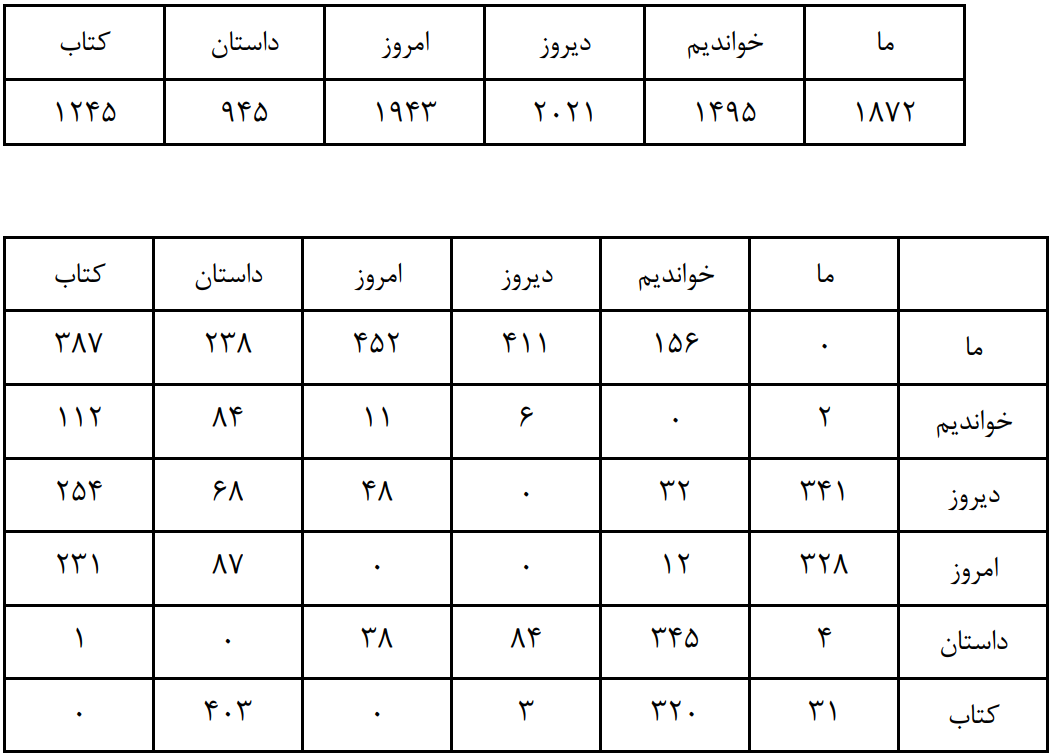
به نام خدا

تكليف سري چهارم NLP بهاره كاوسي نژاد – 99431217

سوالات تئوری

1. دو جدول زیر تعداد Unigram ها و Bigram ها در یک پیکره فرضی را نشان می دهد.



احتمال رخداد جملات تست زیر را محاسبه کنید. فرض بر این است که جملات تست در وسط یک رشته هستند. یعنی در نظر گرفتن احتمال بندهای شروع و پایان جمله الزام نیست.

جمله تست ۱ : ... ما امروز کتاب خواندیم ...

1. ابتدا احتمال رخداد هر کلمه را با توجه به unigram محاسبه می کنیم:

* P(ما) = Count(ما) / Total Count of Unigrams = 1872 / (1872 + 1495 + 2021 + 1943 + 945 + 1245) = 1872 / 9521 = 0.196
* P(امروز) = Count(امروز) / Total Count of Unigrams = 1943 / 9521 = 0.204
* P(کتاب) = Count(کتاب) / Total Count of Unigrams = 1245 / 9521 = 0.130
* P(خواندیم) = Count(خواندیم) / Total Count of Unigrams = 1495 / 9521 = 0.157

1. احتمال رخداد هر bigram را با توجه به جدول bigramها محاسبه می کنیم:

* P(امروز|ما) = Count(ما امروز) / Count(ما) = 452 / 1872 = 0.241
* P(کتاب|امروز) = Count(امروز کتاب) / Count(امروز) = 231 / 1943 = 0.118
* P(خواندیم|کتاب) = Count(کتاب خواندیم) / Count(کتاب) = 320 / 1245 = 0.257

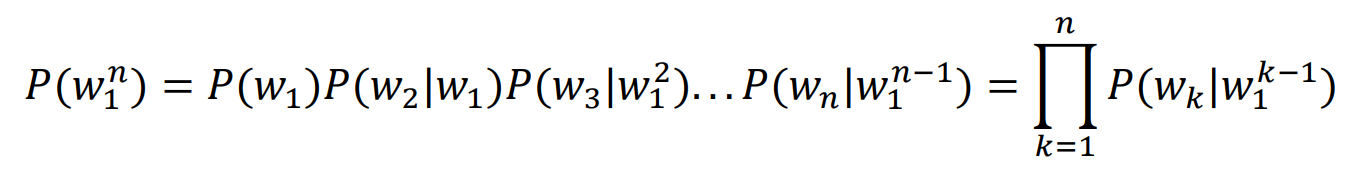
1. احتمال نهایی را با ضرب احتمالات کلمات یا bigramها بدست می آوریم:

* P(ما امروز کتاب خواندیم) = P(ما) \* P(امروز|ما) \* P(کتاب|امروز) \* P(خواندیم|کتاب) = 0.196 \* 0.241 \* 0.118 \* 0.257 = 0.001

جمله تست ۲ : ... ما دیروز داستان خواندیم ...

* ابتدا احتمال رخداد هر کلمه را با توجه به unigram محاسبه می کنیم:
* P(ما) = Count(ما) / Total Count of Unigrams = 1872 / (1872 + 1495 + 2021 + 1943 + 945 + 1245) = 1872 / 9521 = 0.196
* P(دیروز) = Count(دیروز) / Total Count of Unigrams = 2021 / 9521 = 0.212
* P(داستان) = Count(داستان) / Total Count of Unigrams = 945 / 9521 = 0.099
* P(خواندیم) = Count(خواندیم) / Total Count of Unigrams = 1495 / 9521 = 0.157
* احتمال رخداد هر bigram را با توجه به جدول bigramها محاسبه می کنیم:
* P(دیروز|ما) = Count(ما دیروز) / Count(ما) = 411 / 1872 = 0.219
* P(داستان|دیروز) = Count(دیروز داستان) / Count(دیروز) = 68 / 2021 = 0.033
* P(خواندیم|داستان) = Count(داستان خواندیم) / Count(داستان) = 345 / 945 = 0.365
* احتمال نهایی را با ضرب احتمالات کلمات یا bigramها بدست می آوریم:
* P(ما دیروز داستان خواندیم) = P(ما) \* P(دیروز|ما) \* P(داستان|دیروز) \* P(خواندیم|داستان) = 0.196 \* 0.219 \* 0.033 \* 0.365 = 0.0005

1. رابطه زیر را اثبات کنید.

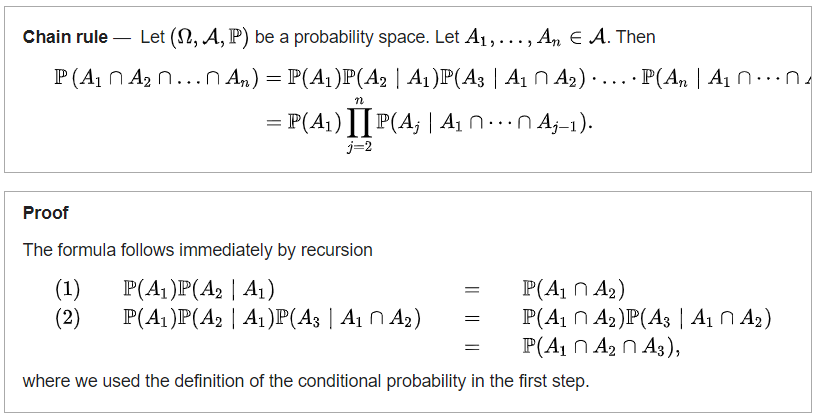


دنباله n کلمه را یا به صورت w1…wn و یا به صورت w1n نمایش می دهیم. بنابراین عبارت w1n-1 به معنای رشته w1, w2, …, wn-1 است. برای joint probability هر کلمه در یک دنباله که به این صورت

P(X1 = w1, X2 = w2, X3 = w3, …, Xn=wn)

نوشته می شود از P(w1,w2,...,wn) استفاده می کنیم.

برای محاسبه احتمال کل دنباله مانند P(w1,w2,...,wn) می توانیم از chain rule of probability استفاده کنیم.



با اعمال chain rule به کلمات خواهیم داشت:

Chain rule ارتباط بین محاسبه joint probability یک دنباله را نشان می دهد و محاسبه احتمال شرطی یک کلمه با کلمات قبلی. معادله بالا نشان می دهد که ما می توانیم joint probability یک دنباله کامل کلمات را با ضرب تعدادی احتمال شرطی در یکدیگر تخمین بزنیم.

**منابع:**

* <https://en.wikipedia.org/wiki/Chain_rule_(probability)>