به نام خدا

دانشگاه علم و صنعت ایران

دانشکده‌ی مهندسی کامپیوتر

کلاس حل تمرین درس پردازش زبان‌های طبیعی – شنبه، ۱۳۹۶/۸/۱۳ استاد درس: دکتر بهروز مینایی

۱- با استفاده از احتمالات موجود در اسلایدهای درس برای دادگان رستوران برکلی، تعیین کنید کدام یک از دو جمله‌ی زیر صحت بیشتری دارد؟ احتمال هر جمله را نیز به دست آورید و سرگشتگی (Perplexity) را برای جمله‌ی محتمل‌تر محاسبه کنید. محاسبات را یک بار دیگر در فضای لگاریتمی انجام دهید.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

P(I | <s>) = 0.25

P(</s> | food) = 0.11

- I want eat Chinese food
- I want to eat Chinese food

حل:

ابتدا به آغاز و پایان جمله، تگ‌های <s> و </s> را اضافه می‌کنیم و احتمال توأم جمله را به صورت حاصل ضرب احتمالات شرطی آن می‌نویسیم:

<s> I want eat Chinese food </s>

$$P(< s >. I. want. eat. Chinese. food. </s >) =$$

$$P(I| < s >) \times P(want|I) \times P(eat|want) \times P(Chinese|eat) \times P(food|Chinese) \times P(</s > |food)$$
$$= 0.25 \times 0.33 \times 0.0011 \times 0.021 \times 0.52 \times 0.11 = 1.1 \times 10^{-7}$$

<s> I want to eat Chinese food </s>

$$P(< s >. I. want. to. eat. Chinese. food. </s >) =$$

$$P(I| < s >) \times P(want|I) \times P(to|want) \times P(eat|to) \times P(Chinese|eat) \times P(food|Chinese) \times P(</s > |food) = 0.25 \times 0.33 \times 0.66 \times 0.28 \times 0.021 \times 0.52 \times 0.11 = 1.8 \times 10^{-5}$$

بنابراین جمله‌ی پایینی که احتمال وقوع بیشتری دارد. حال سرگشتگی را برای آن محاسبه می‌کنیم:

$$PPL = \sqrt[n]{\frac{1}{P(W)}} = \sqrt[7]{\frac{1}{1.8 \times 10^{-5}}} = 4.76$$

توجه می‌کنیم که برای مقدار n در رابطه‌ی بالا، </s> را نیز یک توکن در نظر گرفته‌ایم (ولی <s> را در نظر نمی‌گیریم).

در فضای لگاریتمی:

جمله‌ی اول:

$$\log P(W) = (-0.6) + (-0.48) + (-2.96) + (-1.68) + (-0.28) + (-0.96) = -6.96$$

جمله‌ی دوم:

$$\log P(W) = (-0.6) + (-0.48) + (-0.18) + (-0.55) + (-1.68) + (-0.28) + (-0.96) = -4.73$$

۲- فرض کنید در یک پیکره‌ی متنی (Corpus) تعداد تکرارهای کلمات را به صورت زیر داریم:

کلمه	تکرار	کلمه	تکرار	کلمه	تکرار	کلمه	تکرار
Brown	29	Tree	1	Napkin	9	Syrup	9
Fox	34	Skim	4	Cheap	22	Short	28
Lazy	18	Neat	49	Fork	10	Options	13
Dog	1	Syzygy	33	Nickel	1	Car	14
Plenty	41	Missing	12	Chocolate	5	Concinnity	0

Sum: 333

مقدار احتمالات زیر را مشخص کنید:

- P_{MLE}(short)
- P_{MLE}(concinnity)

حل:

$$P_{MLE}(short) = \frac{count(short)}{sum} = \frac{28}{333}$$

$$P_{MLE}(concinnity) = \frac{count(concinnity)}{sum} = \frac{0}{333}$$

اکنون فرض کنید که از روش هموارسازی لاپلاس (Laplace smoothing) استفاده کرده‌ایم. مقدار احتمالات زیر را مشخص کنید:

- $P_{Laplace}$(lazy)
- $P_{Laplace}$(concinnity)

حل:

$$P_{Laplace}(lazy) = \frac{count(lazy) + 1}{sum + |V|} = \frac{18 + 1}{333 + 20} = \frac{19}{353}$$

$$P_{Laplace}(concinnity) = \frac{count(concinnity) + 1}{sum + |V|} = \frac{0 + 1}{333 + 20} = \frac{1}{353}$$

حال فرض کنید از روش هموارسازی Good Turing استفاده می‌کنیم. مقدار احتمالی را که به تکرارهای صفر کلمات اختصاص می‌دهیم چقدر است؟

حل:

$$P^*_{GT}(zeros) = \frac{N_1}{N} = \frac{3}{333}$$

اگر بدانیم مقادیر داده‌شده بخشی از یک مجموعه‌ی داده‌ی بزرگ‌تر با فراوانی تکرارهای (Frequency of frequencies) زیر هستند،

N_1 = 112849

N_2 = 41018

N_3 = 15608

N_4 = 5704

N_5 = 2111

N_6 = 754

N_7 = 283

N_8 = 104

N_9 = 37

N_{10} = 14

اکنون تکرار c^* را برای کلمات زیر محاسبه کنید:

- C^*(skim)
- C^*(syrup)

حل:

$$c^*(\boldsymbol{skim}) = \frac{[c(\boldsymbol{skim}) + 1]N_{c+1}}{N_c} = \frac{5 \times N_5}{N_4} = \frac{10555}{5704}$$

$$c^*(\boldsymbol{syrup}) = \frac{[c(\boldsymbol{syrup}) + 1]N_{c+1}}{N_c} = \frac{10 \times N_{10}}{N_9} = \frac{140}{37}$$

موفق باشید