به نام خدا

آشنایی با زبان‌شناسی رایانشی (۱۲۱-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تاریخ تحویل: ۱۴۰۲/۰۳/۰۶

تمرین شماره ۳

دانشکده علوم و فنون نوین

۱. (٪۵) [پژوهش: ابهام گرامری و مدل‌های زبانی] هدف این تمرین آشنایی با ابهام گرامری است. ابهام دستوری/ساختاری پدیده‌ای است که در آن یک عبارت، جمله یا متن بسته به زمینه می‌تواند بیش از یک معنی یا ساختار داشته باشد. برای مثال؛

- They are flying planes. (Are they piloting planes, or are they throwing paper planes?)
- He saw the man with the binoculars. (Did he use the binoculars to see the man, or did he see a man who had binoculars?)

ابهام گرامری می‌تواند به دلیل ابهام لغوی (چند معنایی بودن کلمات) یا ابهام ساختاری (چند روش برای تجزیه و تحلیل جمله) ایجاد شود. این مسئله می‌تواند برای هم انسان‌ها و هم سامانه‌های هوشمند که باید بدون اطلاعات بیشتر معنای قصد شده را تعیین کنند، چالش‌برانگیز باشد. مدل‌های زبانی زمینه‌ای، مدل‌هایی هستند که می‌توانند معنای کلمات را براساس بافت متن تنظیم کنند، مانند BERT.

۱. یک مقاله اخیر (از سال ۲۰۲۰ تا کنون) در مورد این موضوع مطالعه کرده و خلاصه‌ای از ایده اصلی[1]، روش‌شناسی[2]، انگیزه[3] و نتایج آن ارائه کرده و در نهایت **نقاط قوت و ضعف** آن را از **دید خودتان** ارزیابی کنید (نیازی به تسلط به مفاهیم و جزئیات روش مانند جزئیات شبکه‌های عصبی برای این سوال ندارید).

---

۲. (٪۵۰) [پیاده‌سازی: مدل زبانی n-تایی[4]] هدف از این تمرین آشنایی و تمرین با عبارات منظم[5] است.

۲-۱ تولید n-تایی‌ها، در این بخش، شما یک مدل زبانی n-تایی ساده ایجاد خواهید کرد که برای تولید متن تصادفی مشابه متن یک سند واقعی، می‌تواند استفاده شود (در این بخش شما **فقط می‌توانید از ماژول‌های داخلی پایتون استفاده کنید**، برای مثال کتابخانه‌های نامپای[6]، جعبه‌ابزار پردازش زبان طبیعی[7] و ... شامل **نمی‌شود**).

۱. برای تولید n-تایی مدنظر، تابعی بنویسید که مرتبه n و متن را به عنوان آرگومان ورودی می‌گیرد و یک لیست از همه n-تایی‌ها با اندازه مشخص شده از متن ورودی تولید می‌کند. هر n-تایی باید شامل یک دوتایی[8] (زمینه[9]، حرف) باشد که زمینه، یک رشته به طول n بوده و از n کاراکتر قبل از کاراکتر فعلی ساخته شده است (جمله باید در ابتدا با $n \sim$ پر شود برای این منظور تابعی برای اضافه کردن $n \sim$ به تعداد دلخواه بنویسید، همچنین لازم است که $n \geq 0$ در نظر بگیرید). به طور کلی زمینه دنباله ای از کلمات است که قبل از کلمه دیگری در جمله قرار می‌گیرد. برای مثال، در جمله

[1] Objective
[2] Methodology
[3] Motivation
[4] N-Gram Language Model
[5] Regular Expression
[6] Numpy
[7] NLTK
[8] Tuple
[9] Context

h.veisi@ut.ac.ir

«او دوست دارد کتاب بخواند»، زمینه برای کلمه «کتاب» عبارت «او دوست دارد بخواند» است. در یک مدل n-تایی، زمینه برای تخمین احتمال یک کلمه با توجه به کلمه‌های قبلی آن استفاده می‌شود.

۲-۲ در این بخش لازم است که یک کلاس تحت عنوان مدل زبانی n-تایی با توجه به موارد خواسته شده طراحی کنید و برای هر قسمت، تست‌کیس‌های موجود در فایل `CL-HW3-test.py` را اجرا کنید تا از عملکرد و درستی هر قسمت اطمینان حاصل پیدا کنید.

۱. جهت پیاده‌سازی مدل زبانی n-تایی، یک کلاس ایجاد کرده و مقادیر n، k و دیگر متغیر[ها] (در صورت نیاز) را در تابع سازنده[1] مقداردهی اولیه کنید (n مرتبه مدل n-تایی و k پارامتر هموارسازی به حالت اضافه-k [2] است).

۲. تابعی بنویسید که مجموعه‌ای از تمام کاراکترهای استفاده شده در واژگان توسط این مدل را برگرداند.

۳. تابعی بنویسید که یک رشته به عنوان آرگومان می‌گیرد و n-تایی‌های آن را محاسبه می‌کند، برای هر زیررشته، حرف آخر را به مجموعه حروف ممکن اضافه کنید و تعداد دفعات رخداد زمینه با طول $(n-1)$ و دوتایی (زمینه، حرف) را آپدیت کنید.

۴. تابعی بنویسید که یک زمینه به طول n و یک حرف به عنوان آرگومان بگیرد و احتمال شرطی حرف با توجه به زمینه را با استفاده از یک فرمول هموارسازی بر می‌گرداند:

$$P(\text{char}|\text{context}) = \frac{\text{frequency}[(\text{context},\text{char})] + k}{\text{frequency}[\text{context}] + k \times V}$$

اگر به متن جدیدی برخوردید، آنگاه احتمال هر حرف را، پیش‌فرض $\frac{1}{V}$ در نظر بگیرید.

۵. تابعی بنویسید که یک کاراکتر تصادفی بر اساس توزیع احتمالی زمینه داده شده، برگرداند. به طور مشخص، فرض کنید واژگان $V = \langle v_1, v_2, \cdots, v_n \rangle$ بر اساس ترتیب لغوی مرتب شده‌اند و r یک عدد تصادفی بین ۰ و ۱ است. خروجی تابع کارکتر v_i است، به طوری که

$$\sum_{j=1}^{i-1} P(v_j \mid \text{context}) \leq r < \sum_{j=1}^{i} P(v_j \mid \text{context})$$

باشد (در نظر داشته باشید که برای تولید r، تنها یک‌بار تابع random.random() را فراخوانی کنید).

۶. تابعی بنویسید که از کارکترهای تصادفی ایجاد شده (تابعی که پیش‌تر پیاده‌سازی کردیم) یک رشته به طول $n \sim$ کاراکتر و زمینه شروع همیشه باید آرگومان ورودی برگرداند. زمینه باید با تولید کاراکترها به روز شود. رشته‌ی کامل در خروجی از تولید کارکترهای تصادفی تا رسیدن به تعداد طول مشخص شده از کارکترها ادامه می‌یابد (در صورتی که مقدار n برابر با صفر بود، زمینه شما همیشه رشته خالی است).

[1] __init__ method
[2] Add-k

h.veisi@ut.ac.ir

به نام خدا

آشنایی با زبان‌شناسی رایانشی (۱۲۱-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تاریخ تحویل: ۱۴۰۲/۰۳/۰۶

تمرین شماره ۳

دانشکده علوم و فنون نوین

۷. مدل زبانی n-تایی بدون هموارسازی، و بدون برهم‌نهی بر پیکره اشعار (فایلی با نام `Poems.txt` در اختیار شما قرار گرفته است) را با مرتبه‌های ۲، ۳، ۴، ۷ و ۹ آموزش دهید و متنی به اندازه ۵۰۰ کارکتر تولید کنید، خروجی تولید شده را در گزارش خود ارائه کنید و در مورد تفاوت خروجی‌های حاصل شده بحث کنید؟ آیا مورد قابل توجه‌ای به چشم‌تان می‌خورد، دلیل رخ دادن این موضوع را بیان کنید.

۸. حال بخش ۸ را مجددا با پیکره ژورنال‌ها و فیلم‌ها (فایلی با نام `Journals.txt` و `Movies.txt` در اختیار شما قرار گرفته است) آزمایش کرده و نتایج را در گزارش خود ارائه کنید، همچنین در مورد تفاوت خروجی‌های حاصل شده بحث کنید؟ آیا مورد قابل توجه‌ای به چشم‌تان می‌خورد، دلیل رخ دادن این موضوع را بیان کنید.

چگونه می‌دانیم که یک مدل زبانی خوب است؟ دو روش پایه‌ای وجود دارد!

- **ارزیابی خارجی (مبتنی بر وظیفه)**: که در آن مدل زبانی به عنوان بخشی از وظیفه دیگری مانند شناسایی خودکار گفتار، ویراستاری و یا سامانه *OCR* که نوشته‌های نامرتب امتحان یک دانشجو را به متن تبدیل کند، استفاده می‌کنیم.
- **ارزیابی درونی/ذاتی**: در این مدل ارزیابی خوب بودن مدل زبانی به طور مستقیم توسط کیفیت تفسیر مجموعه آزمون از پیش دیده‌نشده به وسیله توزیع‌های احتمالاتی که مدل تخمین می‌زند، محاسبه می‌شود.

در کتاب مرجع اینگونه آمده است [1, Ch. 4]:

برای ارزیابی ذاتی یک مدل زبانی، ما به یک مجموعه آزمون نیاز داریم. همانطور که در بسیاری از مدل‌های آماری در حوزه زبان‌شناسی رایانشی دیده می‌شود، احتمال‌های یک مدل n-تایی از پیکره‌ای که بر آن آموزش داده می‌شود، بدست می‌آید. با این حساب می‌توانیم کیفیت یک مدل n-تایی را با عملکرد آن روی برخی از داده‌های دیده‌نشده به نام مجموعه/پیکره آزمون اندازه‌گیری کنیم. پس اگر یک پیکره متنی به ما داده شود و بخواهیم دو مدل n-تایی متفاوت را با هم مقایسه کنیم، داده‌ها را به دو قسمت آموزش و آزمون تقسیم می‌کنیم، با یادگیری هر دو مدل بر دادگان آموزش و مقایسه چگونگی مناسب بودن دو مدل آموزش دیده شده بر دادگان آزمون، انتخاب خود برای مدل بهتر را انجام می‌دهیم. اما منظور از «مناسب بودن بر دادگان آزمون» چیست؟ جواب ساده است: مدلی که بالاترین احتمال را به مجموعه آزمون اختصاص می‌دهد، مدل بهتری است.

h.veisi@ut.ac.ir

به نام خدا

۳-۲ هدف از این بخش آشنایی و پیاده‌سازی برخی از روش‌های بهبود مدل‌های زبانی (سرگشتگی، هموارسازی و برهم‌نهی[1]) است.

۱. برای پیاده‌سازی مدل زبانی n-تایی خود، تابعی بنویسید که متن ورودی را گرفته و سرگشتگی را حساب کند. سرگشتگی یک مدل زبانی بر روی یک مجموعه آزمون، معکوس احتمال مجموعه آزمون است و با تعداد کاراکترها نرمال‌سازی می‌شود. برای یک مجموعه آزمون $W = w_1 w_2 \dots w_N$ داریم:

$$\text{Perplexity}(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i \mid w_1 \dots w_{i-1})}}$$

چند نکته را به خاطر داشته باشید:

- کمبود حسابی[2] یک چالش خواهد بود، بنابراین استفاده از لگاریتم را در نظر بگیرید.
- سرگشتگی برای متنی که مدل زبانی به آن احتمال صفر اختصاص می‌دهد، تعریف نشده است. در این صورت کد شما باید مثبت بی‌نهایت $\text{float}(\infty)$ را برگرداند.
- در مدل‌هایی که هموارسازی نکرده‌اید، قطعاً برای مجموعه آزمون احتمال‌های صفر خواهید داشت. در پیاده‌سازی لازم است سرگشتگی را بر مجموعه آموزش و همچنین برای مدل‌های زبانی خود که از هموارسازی و برهم‌نهی استفاده می‌کنند، محاسبه کنید.

۲. هموارسازی لاپلاس یک واحد به هر شمارش اضافه می‌کند (بنابراین نام دیگر آن هموارسازی اضافه-یک[3] است). از آنجا که V کاراکتر در واژگان وجود دارد و هر یک افزایش یافته است، ما نیز باید مخرج را تنظیم کنیم تا V مشاهده اضافی را در نظر بگیریم.

$$P_{\text{Laplace}}(w_i) = \frac{\text{count}_i + 1}{N + |V|}$$

یک نوع از هموارسازی لاپلاس به نام هموارسازی اضافه-k یا هموارسازی اضافه-ایپسیلون نامیده می‌شود. کلاس مدل زبانی n-تایی خود را به‌روزرسانی کرده و هموارسازی اضافه-k را پیاده‌سازی کنید.

[1] Interpolation
[2] Underflow
[3] Add-One

h.veisi@ut.ac.ir

به نام خدا

آشنایی با زبان‌شناسی رایانشی (۱۲۱-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تاریخ تحویل: ۱۴۰۲/۰۳/۰۶

تمرین شماره ۳

دانشکده علوم و فنون نوین

۳. برهم‌نهی یک روش برای تخمین احتمال یک n-تایی از طریق ترکیب احتمال‌های n-تایی‌های مرتبه پایین‌تر است. در این روش برای مثال اگر ۴-تایی مدنظر را در دادگان آموزش خود ندیده باشید، باز هم می‌توانید از مشکل صفر شدن احتمال‌ها جلوگیری کنید، بدین صورت که از n-تایی‌های مرتبه پایین‌تر به عنوان یک پشتیبان استفاده می‌کند. فرم ریاضی این روش به صورت زیر خواهد بود:

$$P_{interpolation}(w_i|w_{i-2}w_{i-1}) = \lambda_1 P(w_i|w_{i-2}w_{i-1}) + \lambda_2 P(w_i|w_{i-1}) + \lambda_3 P(w_i)$$

که در آن مجموع لامبداها برابر یک خواهد بود $\lambda_1 + \lambda_2 + \lambda_3 = 1$

برای پیاده‌سازی مدل زبانی n-تایی با برهم‌نهی، یک کلاس دیگر تعریف کرده که از کلاس مدل زبانی n-تایی ارث‌بری می‌کند. بسته به تمیز بودن کد/ پیاده‌سازی‌ها [1]، لازم است که تعدادی کمی از تابع‌ها بازنویسی [2] شوند (**پیشنهاد شما چیست؟ چرا؟**). پارامتر n که به تابع سازنده پاس داده می‌شود، بالاترین مرتبه n-تایی است که باید توسط مدل در نظر گرفته شود (برای مثال $n = 2$ سه n-تایی با طول‌های مختلف را در نظر می‌گیرد).

هنگامی که هموارسازی اضافه-k را اعمال می‌کنید، باید فقط برای هر مرتبه از n-تایی انجام دهید و نه برای احتمال ترکیب کل، برای مثال، اگر بخواهید احتمال یک ۴-تایی را با استفاده از برهم‌نهی و هموارسازی اضافه-k تخمین بزنید، باید اول احتمال‌های ۴-تایی، ۳-تایی، ۲-تایی و ۱-تایی را به طور جداگانه با استفاده از هموارسازی اضافه- k هموار کنید، سپس آن‌ها را با استفاده از برهم‌نهی ترکیب کنید (**توجه: شما نباید احتمال نهایی ترکیب شده را دوباره با استفاده از هموارسازی اضافه-k هموار کنید**).

۴. به طور پیش‌فرض لامبداها (λ) را با وزن‌های برابر تنظیم کنید (لازم است تابعی بنویسید که این پیش‌فرض را بازنویسی کند).

۵. تنظیم لامبداها می‌تواند به صورت رهیافت آنی [3] یا با استفاده از دادگان اعتبارسنجی [4] انجام شود، برای این منظور تغییرات لازم را اعمال کنید. رهیافت‌های آنی به طور کلی بر اساس بینش یا تجربه هستند. برای مثال، n-تایی مرتبه بالاتر ممکن است وزن بیشتری داشته باشند زیرا دارای زمینه و ویژگی‌های [5] بیشتری هستند. مجموعه اعتبارسنجی بخشی از داده‌ها است که برای تنظیم پارامترهای مدل استفاده می‌شود. برای مثال، می‌تواند با حداکثر کردن درستنمایی یا حداقل کردن سرگشتگی، به انتخاب بهترین وزن‌ها کمک کند. درستنمایی [6] نشان‌دهنده این است که چقدر یک دنباله از کلمات طبق مدل، محتمل است. سرگشتگی، معکوس میانگین درستنمایی هر کلمه است، و نشان می‌دهد که مدل وقتی می‌خواهد کلمه بعدی را پیش‌بینی کند، چقدر انتخاب دارد و در نهایت سرگشتگی پایین‌تر به معنای اطمینان بیشتر و گزینه‌های کمتر است.

---

[1] Clean Code
[2] Override
[3] Heuristic
[4] Validation Set/Development Set
[5] Specificity
[6] Likelihood

h.veisi@ut.ac.ir

به نام خدا

۶. مدل‌های n-تایی را با و بدون برهم‌نهی، هموارسازی و سرگشتگی مقایسه کنید. این روش‌های چگونه بر دقت و تعمیم‌پذیری مدل‌ها برای متون مختلف و مقادیر n، k و λ تأثیر می‌گذارند؟ توضیات خود را به همراه جدول‌های مرتبط در گزارش خود ارائه کنید.

توجه: بخش‌های مشخص شده با **رنگ قرمز** مواردی هستند که باید در پیاده‌سازی‌ها ارائه شوند.

---

۳. (۵٪) **[نظری: سرگشتگی مدل]** در یک مجموعه داده آزمون که هر کلمه میانگین ۶ حرف دارد، سرگشتگی مدل برای هر کلمه باید دقیقا ۶ برابر سرگشتگی مدل برای هر حرف باشد. درستی یا نادرستی گزاره بیان شده را بررسی کنید (برای ساده‌سازی فرض می‌کنیم که فضاهای خالی شمارده نمی‌شوند، در صورت نیاز برای اثبات درستی یا نادرستی، محاسبه‌های لازم را ارائه کنید).

---

۴. (۱۰٪) **[نظری: پرتاب‌های مشکوک و احتمال‌ها]** قصد داریم احتمال شیر آمدن در یک سکه ناسالم که تعداد رخداد شیرها (Heads) اغلب بیشتر از خطها (Tails) را تخمین بزنیم.

۱. ستون‌های زیر نشان دهنده دنباله‌های واقعی از پرتاب‌هایی است که بدست آورده‌ایم. برای هر دنباله، دو تخمین خواسته شده را پر کنید.

دنباله مشاهده شده	HHHH	HHHH	THHH	HTHH	THHT	HTHT
تخمین $p(H)$ بدون هموارسازی						
تخمین $p(H)$ با هموارسازی لاپلاس						

۲. در واقعیت ممکن است فقط یکی از این دنباله‌های پرتاب را مشاهده کنید.

۱. با میانگین‌گیری این ۶ سناریو، تخمین متوسط بدون هموارسازی شما از $p(H)$ چه خواهد بود؟

۲. تخمین متوسط با هموارسازی لاپلاس چطور؟

۳. کدام تکنیک:

۱. با هموارسازی یا بدون هموارسازی کمتر دارای بایاس است؟

۲. با هموارسازی یا بدون هموارسازی کمتر دارای واریانس است؟

۳. کدام میانگین از قسمت (۲) احتمالاً نزدیک‌تر به واقعیت است؟

۴. فرض کنید شما به شکلی برخی از دادگان آموزش را تخمین می‌زنید که $p(H) = 0.8, p(T) = 0.2$ است. شما این مدل ۱-تایی را با یک دنباله آزمون (۶ پرتاب HHHHTT) ارزیابی می‌کنید.

۱. آنتروپی متقاطع هر کاراکتر از مدل تخمین زده شده شما برای این مجموعه آزمون چقدر است؟ (برحسب بیت پاسخ دهید)

h.veisi@ut.ac.ir

به نام خدا

آشنایی با زبان‌شناسی رایانشی (۱۲۱-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تاریخ تحویل: ۱۴۰۲/۰۳/۰۶

تمرین شماره ۳

دانشکده علوم و فنون نوین

۵. دنباله آزمون شما (۶ پرتاب HHHHTT) ممکن است به طور عادی رخ ندهد. در اینصورت، اگر شما دنباله آزمون دیگری داشتید، آنتروپی متقاطع متفاوتی (یعنی واریانسی در روش تخمین‌زدن آنتروپی متقاطع[1] وجود دارد) خواهید داشت.

۱. با این حساب ممکن است نتوانید به عدد آنتروپی متقاطعی که حاصل شده است، اعتماد کنید، علت این نگرانی را بیان کنید؟

۲. راه آسان برای به دست آوردن تخمین دقیق‌تر و قابل تکرار از آنتروپی متقاطع مدل شما نسبت به رفتار واقعی (اما ناشناخته) این سکه چیست؟

---

۵. (۱۵٪) [نظری: مارکوف و هماهنگی کلمات] هدف از این تمرین یادگیری نحوه استفاده و بهبود یک مدل مارکوف مرتبه دوم برای وظایف زبانی است.

۵-۱ قصد داریم یک مدل زبانی بسازیم که از فرض مارکوف مرتبه دوم استفاده می‌کند؛

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid X_{i-2} = x_{i-2}, X_{i-1} = x_{i-1})$$

۱. به عبارت دیگر ما فرض می‌کنیم که کلمه i-ام X_i مستقل از $X_1 \dots X_{i-3}$ و مشروط به X_{i-2} و X_{i-1} است، دو نمونه در زبان انگلیسی و فارسی (مجموعا چهار مثال) برای این بخش ارائه کنید.

۲. دو نمونه در زبان انگلیسی و فارسی (مجموعا چهار مثال) ارائه کنید که نشان دهد **فرض استقلال نقض می‌شود**، برای هر جمله، توضیح دهید چرا این فرض نقض شده است و چگونه بر احتمال جمله تأثیر می گذارد؟

۵-۲ حال می‌خواهیم از یک مدل مارکوف مرتبه دوم برای تولید جملات به زبان انگلیسی یا فارسی استفاده کنیم که از نظر دستور زبان و هماهنگی صحیح باشند.

۱. چگونه می‌توان چالش وابستگی های راه دور[2]، که در آن انتخاب یک کلمه به کلمه دیگر وابسته است و با فاصله از هم در جمله واقع شده‌اند، برطرف کرد؟

۲. چگونه می توانیم مدل مارکوف مرتبه دوم را برای درک این وابستگی ها تغییر دهیم؟

۳. چه مصالحه‌ای[3] بین استفاده از مقادیر بزرگتر یا کوچکتر N وجود دارد؟

---

۶. (۱۵٪) [نظری: بیشینه و کمینه سرگشتگی] هدف از این تمرین آشنایی بیشتر با مفهوم سرگشتگی است. در این تمرین شما باید با استفاده از مجموعه‌های آموزش و آزمون مناسب، سرگشتگی مدل‌های ۲-تایی و ۳-تایی را بررسی کنید.

[1] Cross Entropy
[2] Long-Distance Dependencies
[3] Trade-off

h.veisi@ut.ac.ir

به نام خدا

سرگشتگی، به طور کلی، یک معیار از چگونگی پیش‌بینی یک مدل احتمالاتی برای یک نمونه یا یک مجموعه آزمون است. در پردازش زبان طبیعی، سرگشتگی یک راه برای ارزیابی مدل‌های زبانی است.

سرگشتگی را می‌توان به دو روش معادل تعریف کرد:

- سرگشتگی به عنوان احتمال وارون نرمال شده مجموعه آزمون: احتمالی که مدل به مجموعه آزمون اختصاص می‌دهد، به توان منفی ۱ بر تعداد کلمه‌های مجموعه آزمون می‌رسانیم. این روش به ما احتمال تولید مجموعه آزمون توسط مدل را می‌دهد (سرگشتگی پایین‌تر به معنای احتمال بالاتر است).

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i \mid w_1 \dots w_{i-1})}}$$

- سرگشتگی به عنوان نمایی آنتروپی متقاطع: این بدان معناست که ما متوسط منفی لگاریتم احتمال هر کلمه در مجموعه آزمون را به عنوان توان نمایی در نظر می‌گیریم $PP(W) = 2^{-L}$. این تعریف، تعداد بیت‌های لازم برای بازنمایی هر کلمه به طور متوسط با استفاده از مدل را می‌دهد (سرگشتگی پایین‌تر به معنای تعداد بیت‌های کمتر است). بیت‌ها واحدهای اطلاعاتی هستند که میزان کاهش عدم قطعیت را با مشاهده یک رویداد اندازه‌گیری می‌کنند. در نظریه اطلاعات، بیت‌ها برای کمی‌سازی آنتروپی یا تصادفی بودن یک سامانه استفاده می‌شوند. بیت را می‌توان به عنوان مقدار اطلاعات به دست آمده با مشاهده یک متغیر دودویی که با احتمال برابر صفر یا یک است، در نظر گرفت. آنتروپی متقاطع یک مدل زبانی، میانگین تعداد بیت های مورد نیاز برای بازنمایی هر کلمه در مجموعه آزمون با استفاده از مدل است. آنتروپی متقاطع را می‌توان با گرفتن لگاریتم مبنای ۲ از احتمالی که مدل به هر کلمه نسبت می‌دهد، محاسبه کرد. پایه لگاریتم ۲ احتمال را به بیت تبدیل می‌کند و علامت منفی اطمینان می‌دهد که احتمال‌های بالاتر با آنتروپی‌های متقاطع کمتر مطابقت دارد. حال با نگاه دوباره می‌توان گفت، سرگشتگی یک مدل زبانی، نمایی آنتروپی متقاطع است، به عبارت دیگر؛ ما ۲ را به توان آنتروپی متقاطع می‌رسانیم تا سرگشتگی را بدست بیاوریم. سرگشتگی را می‌توان به عنوان ضریب انشعاب وزنی یا تعداد موثر انتخاب‌هایی که مدل در هر مرحله دارد، تفسیر کرد (سرگشتگی کمتر به معنای کمتر بودن انتخاب‌های مدل و اطمینان بیشتر مدل از پیش‌بینی‌های خود است).

با در نظر گرفتن تعریف دوم، سرگشتگی یک مدل زبانی برای یک مجموعه آزمون[1] به این صورت تعریف می‌شود؛

$$PP = 2^{-L}$$

که در آن

[1] Test Set

h.veisi@ut.ac.ir

به نام خدا

آشنایی با زبان‌شناسی رایانشی (۱۲۱-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تاریخ تحویل: ۱۴۰۲/۰۳/۰۶

تمرین شماره ۳

دانشکده علوم و فنون نوین

$$L = \frac{1}{M}\sum_{i=1}^{m} \log_2 p\left(x^{(i)}\right)$$

m تعداد جملات در مجموعه، M تعداد کل کلمات در مجموعه، $\log_2$ لگاریتم با پایه ۲، $x^{(i)}$ جمله i ام در مجموعه و $p\left(x^{(i)}\right)$ احتمال جمله i ام در مدل زبانی است.

۱. بیشترین مقداری که سرگشتگی می‌تواند بگیرد چقدر است؟

۲. کمترین مقداری که سرگشتگی می‌تواند بگیرد چقدر است؟

۳. فرض کنید که ما یک مدل زبان ۲-تایی داریم، که در آن

$$p(w_1 \dots w_n) = \prod_{i=1}^{n} q(w_i|w_{i-1})$$

و $w_0 = *$ و $w_n = <del>$ است. همچنین پارامترها را به این صورت تخمین می‌زنیم

$$q(w|v) = \frac{\text{Count}(v,w)}{\text{Count}(v)}$$

یک مجموعه آموزش و آزمون بنویسید که سرگشتگی مدل زبانی آموزش دیده شده روی مجموعه آزمون بزرگترین مقدار ممکن را بگیرد.

۴. یک مجموعه آموزش و آزمون بنویسید به طوری که سرگشتگی مدل زبانی آموزش دیده شده روی مجموعه آزمون کوچکترین مقدار ممکن را بگیرد (فرض کنید که ما از یک مدل زبانی ۲-تایی استفاده می‌کنیم).

۵. فرض کنید که ما یک مدل زبان ۳-تایی داریم، که در آن

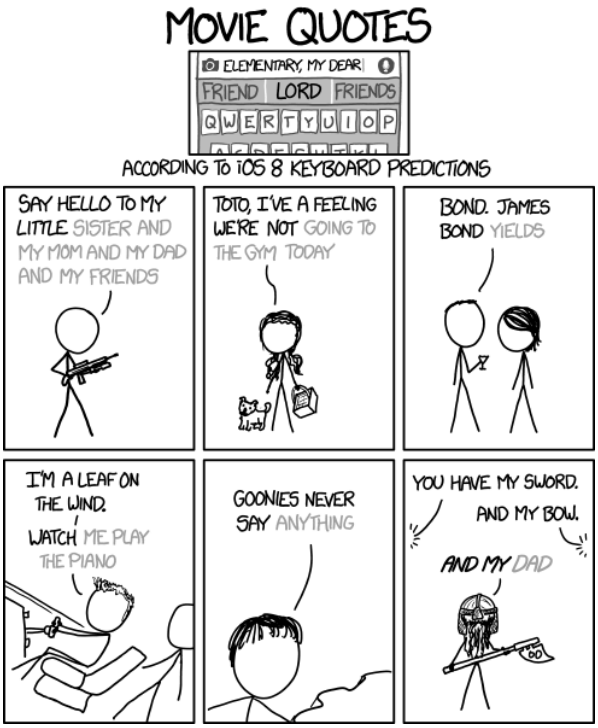
$$p(w_1 \dots w_n) = \prod_{i=1}^{n} q(w_i|w_{i-2}, w_{i-1})$$

و $w_0 = w_{-1} = *$ و $w_n = <del>$ است. همچنین پارامترها را به این صورت تخمین می‌زنیم

$$q(w|u,v) = \frac{\text{Count}(u,v,w)}{\text{Count}(u,v)}$$

یک مجموعه آموزش و آزمون بنویسید به طوری که سرگشتگی مدل زبانی آموزش دیده شده روی مجموعه آزمون **دقیقا** **۲** باشد.

h.veisi@ut.ac.ir



شکل ۱: صفحه‌کلید IOS [۲]

مرجع‌ها

[1] D. Jurafsky, “Speech and Language Processing.” [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/. [Accessed: 05-May-2023].

[2] iOS Keyboard, *xkcd*. [Online]. Available: https://xkcd.com/1427/. [Accessed: 05-May-2023].

h.veisi@ut.ac.ir

به نام خدا

راهنمای تحویل

- انجام تمرین به صورت **تک نفره** است.
- جهت تحویل تمرین می‌بایست پیاده‌سازی‌ها از پایه و بدون استفاده از کتابخانه‌های موجود انجام شود، اما در شرایط قید کتابخانه در صورت سوال، منعی ندارد.
- زمان مناسبی را برای نوشتن و آماده‌سازی گزارش در نظر بگیرید، چرا که نیمی از نمره هر سوال، بر اساس گزارش تحویلی است.
- در صورت استفاده از **کدهای آماده**، دلیل استفاده، کامنت‌گذاری و توضیحات کافی، ضروری است، در غیر اینصورت تقلب تلقی می‌گردد.
- گزارش‌های تمرین خود را در مسیر documentation و اطلاعات مرتبط با پیاده‌سازی‌ها نیز در مسیر source قرار داده شوند.
- لطفا گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت CL_YourFamilyName_YourStNo_HW#.zip به ایمیل h.veisi@ut.ac.ir ارسال فرمائید.
- برای اطلاعات بیشتر به صفحه درس به آدرس https://dsp.ut.ac.ir/courses/y1401/introduction-to-computational-linguistics مراجعه کنید.

در صورت وجود سوال، ابهام و درخواست راهنمایی در گروه اسکایپی یا تلگرامی و یا از طریق ایمیل با دستیار آموزشی در ارتباط باشید.