Assignment NO.3 Solutions

NLP | Fall 1401 | Dr.Minayi

Student name : **Amin Fathi**

Student id : **400722102**

Problem 1

a)

در زبان چینی و ژاپنی، blank ندارند و خودشان بدون blank می‌توانند بخوانند. علاوه بر این، در ژاپن چهار نوع الفا نیز وجود دارد و به‌صورت قاطی از آن‌ها استفاده می‌شود که این کار را سخت‌تر هم می‌کند. در زبان چینی، برای Tokenization از الگوریتم Maximum Matching استفاده می‌شود که در زبان انگلیسی و فارسی کاربرد ندارد. درست است که در زبان چینی blank وجود ندارد، اما کلماتشان یک syllable دارد و از روی آن می‌فهمند. این مورد در زبان فارسی وجود ندارد. مثلاً کلمه‌ی "کاشان"، یک syllable است اما کلمه‌ی "امیرکبیر"، ۴ تا syllable است (اَ، میر، کَ، بیر). همچنین، کلمه‌ی گفتم نیز ۲ syllable دارد. به همین دلیل برخلاف زبان چینی که کلمات دارای یک syllable است و راحتی می‌توان از این الگوریتم استفاده کرد، در زبان فارسی و انگلیسی نمی‌توان.

این الگوریتم به‌صورت greedy عمل می‌کند، به این صورت که میگه برو فهرست کلمات چینی را لیستش را ببین و شروع کن کاراکتر کاراکتر، بزرگترین کلمه‌ای که داخل دیکشنری پیدا میشه را پیدا کن. یک string میشه. اگر این در دیکشنری پیدا شد و بعد از آن چیزی نبود که موازاتش در دیکشنری باشد، پس این چیزی که تا الان یافت شد، یک کلمه می‌شود. این الگوریتم حتی با دقت ۱۰۰٪ هم جواب داده است.

این الگوریتم در زبان انگلیسی کار نمی‌کند. این مسئله در مثال زیر نشان داده شده است.

Thetabledownthere	The table down there
	Theta bled own there

b)

- ✓ **Algorithm Byte Pair Encoding (BPE)**: این الگوریتم در ابتدا برای فشرده‌سازی طراحی شده است؛ به این شکل که به‌صورت مکرر تعداد دوبایتی‌ها را می‌شمارد و سپس به‌جای پرتکرارترین بایت یک بایت استفاده نشده جایگزین می‌کرد. در پردازش زبان به جای بایت از واحد زبانی استفاده می‌شود. ابتدا کاراکترهای موجود در مجموعه متن را وارد vocabulary کرده و سپس جفت کاراکترها را می‌شماریم و آن جفتی که بیشتر از همه تکرار شده را به vocabulary اضافه کرده و آن دو را از این به بعد به‌عنوان یک کاراکتر درنظر می‌گیرد. این کار آنقدر تکرار می‌شود تا اندازه‌ی vocabulary به مقدار مشخصی که در ابرپارامتر مسئله مشخص شده برسد.
- ✓ **WordPiece Algorithm**: روش word piece که در مدل‌هایی مانند BERT، DistilBERT و Electra استفاده می‌شود. این الگوریتم در جستجوی صوتی ژاپنی و کره‌ای تشریح شده و بسیار مشابه BPE است. WordPiece ابتدا واژگان را برای گنجاندن هر کاراکتر موجود در داده‌های آموزشی مقداردهی می‌کند و به تدریج تعداد معینی از قوانین ادغام را یاد می‌گیرد. برخلاف BPE، WordPiece متداول‌ترین جفت نماد را انتخاب نمی‌کند، بلکه یکی را انتخاب کرده که احتمال داده‌های آموزشی را پس از افزودن به واژگان به حداکثر می‌رساند (یعنی سعی می‌کند آن جفتی را انتخاب کند که likelihood داده آموزش را بیشینه کند).

c)

- ✓ **Lemmatization:** یک تکنیک عادی‌سازی متن است که در پردازش زبان طبیعی شده و هر نوع کلمه را به حالت ریشه اصلی آن تغییر می‌دهد. Lemmatization وظیفه‌ی گروه‌بندی اشکال مختلف عطف کلمات را به شکل ریشه‌ای دارد که معنای یکسانی دارند. Lemmatization معمولاً شامل استفاده از واژگان و تجزیه‌وتحلیل صرفی کلمات، حذف پایان‌های عطفی و برگرداندن فرم فرهنگ لغت یک کلمه (لم) است. به عبارتی می‌توان گفت که Lemmatization عبارت است از گروه‌بندی شکل‌های مختلف یک کلمه باهم. در پرس‌وجوهای جستجو، واژه‌سازی به کاربران نهایی اجازه می‌دهد تا هر نسخه‌ای از یک کلمه‌ی پایه را جستجو کنند و نتایج مرتبط را دریافت کنند. به‌عنوان مثال، رایانه می‌تواند کلماتی را که ریشه یکسانی ندارند، اما دارای همان معنای عطفی هستند، در کنار هم قرار دهد. گروه‌بندی کلمه "خوب" با کلماتی مانند "بهتر" و "بهترین" نمونه‌ای از واژه‌سازی است.

  Lemmatization یکی از بهترین راه‌ها برای کمک به چت‌بات‌ها برای درک بهتر سوالات مشتریان است. از آنجایی که این شامل تجزیه‌وتحلیل مورفولوژیکی کلمات است، ربات چت می‌تواند درک بهتری از معنای کلی جمله‌ای که در حال اصطلاح‌سازی است به دست آورد. همچنین Lemmatization برای فعال‌کردن ربات‌ها برای صحبت و مکالمه استفاده می‌شود. این موضوع باعث می‌شود واژه‌سازی بخش نسبتاً مهمی از درک زبان طبیعی و پردازش زبان طبیعی در هوش مصنوعی باشد.
- ✓ **Stemming:** فرآیند کاهش یک کلمه به ریشه کلمه آن است که به پسوندها و پیشوندها یا به ریشه کلمات معروف به لم می‌چسبد. ریشه در درک زبان طبیعی و پردازش زبان طبیعی مهم است. الگوریتم‌های مختلفی جهت انجام عمل Stemming وجود دارد. در زبان انگلیسی الگوریتم Porter بسیار معروف است. این الگوریتم طبق یک سری قاعده‌ی منظم (مثلاً حذف s در آخر کلمات جمع) می‌تواند ریشه‌ی کلمات را با دقت خوبی به دست آورد. همچنین در زبان فارسی، الگوریتم کاظم تقوی، این کار را با دقت بالایی (برای کلمات فارسی) انجام می‌دهد.

به‌صورت کلی می‌توان اینگونه بیان کرد که Stemming فرایندی است که چند کاراکتر آخر یک کلمه را منشا می‌گیرد یا حذف می‌کند، که اغلب منجر به معانی و املای نادرست می‌شود. Lemmatization، زمینه را درنظر گرفته و کلمه را به شکل پایه معنی‌دار خود تبدیل می‌کند. به‌عنوان مثال، برای Lemmatization، مجموعه کلمات {فرماندهان، فرماندهی، فرمانده، فرمانده‌ای، فرمانده‌ی، فرماند} را "فرمانده" نتیجه می‌دهد. برای Stemming، رفتن به رفت، گفتن به گفت، آمدن به آمد، خوردن به خورد که در تمام این‌ها، -َن در آخر کلمات حذف شده است.

Problem 2

a)

مدل زبان[1]، توزیع احتمال بر روی توالی کلمات است. در عمل، یک مدل زبانی احتمال "معتبر" بودن یک دنباله کلمه خاص را می دهد. اعتبار در این زمینه به معتبر بودن از لحاظ گرامری اشاره نمی کند بلکه به این معنی است که شبیه نحوه صحبت کردن (یا به عبارت دقیق

[1] Language model

تر، نوشتن) مردم باشد که همان چیزی است که مدل زبان یاد می گیرد. در واقع مدل زبانی فقط ابزاری است برای ترکیب اطلاعات فراوان به شیوه‌ای مختصر که قابل استفاده مجدد در متنی که خارج از موارد نمونه است، باشد.

مدل‌های آماری شامل توسعه مدل‌های احتمالی است که قادر به پیش‌بینی کلمه بعدی در دنباله با توجه به کلمات قبل از آن هستند. برخی از مدل‌های زبان عبارتند از:

- ✓ N-Gram: یکی از ساده‌ترین رویکردها برای مدل‌سازی زبان است. در اینجا، توزیع احتمال برای دنباله‌ای از n ایجاد می‌شود که در آن n می‌تواند هر عددی باشد و اندازه gram را مشخص می‌کند. اگر n=4 باشد، یک gram ممکن است به این صورت باشد: "can you help me". اساساً n مقدار زمینه‌ای است که مدل برای درنظر گرفتن آن آموزش دیده است. مدل‌های N-Gram انواع مختلفی دارند مانند unigrams، bigrams و trigrams.
- ✓ Unigram: ساده‌ترین نوع مدل زبان است. در محاسبات خود به هیچ زمینه‌ی شرطی نگاه نمی‌کند. هر کلمه یا اصطلاح را به‌طور مستقل ارزیابی می‌کند. مدل‌های Unigram معمولاً وظایف پردازش زبان مانند بازیابی اطلاعات را انجام می‌دهند. Unigram پایه یک مدل خاص‌تری به نام مدل احتمال پرس‌وجو[2] است که از بازیابی اطلاعات برای بررسی مجموعه‌ای از اسناد و تطبیق مرتبط‌ترین آن‌ها با یک پرس‌وجو خاص استفاده می‌کند.
- ✓ Bidirectional: برخلاف مدل‌های n-gram که متن را در یک جهت تجزیه‌وتحلیل می‌کنند، مدل‌های دوطرفه متن را در هر دو جهت، عقب و جلو تجزیه‌وتحلیل می‌کنند. این مدل‌ها می‌توانند هر کلمه‌ای را در یک جمله یا بدنه متن با استفاده از هر کلمه دیگری در متن پیش‌بینی کنند. بررسی متن به‌صورت دو طرفه دقت نتیجه را افزایش می‌دهد. این نوع اغلب در یادگیری ماشین و برنامه‌های تولید گفتار استفاده می‌شود. به‌عنوان مثال، گوگل از یک مدل دو جهته برای پردازش پرس‌وجوهای جستجو استفاده می‌کند.
- ✓ Exponential: این نوع مدل آماری متن را با استفاده از معادله‌ای که ترکیبی از n-gram و توابع ویژگی است ارزیابی می‌کند. در اینجا ویژگی‌ها و پارامترهای نتایج مورد نظر از قبل مشخص شده است. این مدل بر اساس اصل آنتروپی است که می‌گوید توزیع احتمال با بیشترین آنتروپی بهترین انتخاب است. مدل‌های نمایی دارای مفروضات آماری کمتری هستند که به این معنی است که شانس داشتن نتایج دقیق بیشتر است.
- ✓ Continuous Space: در این نوع مدل آماری، کلمات به صورت ترکیبی غیرخطی از وزن‌ها در یک شبکه عصبی مرتب می‌شوند. فرآیند تعیین وزن به یک کلمه به‌عنوان جاسازی کلمه شناخته می‌شود. این نوع مدل در سناریوهایی که مجموعه داده‌های کلمات همچنان بزرگ می‌شوند و شامل کلمات منحصربه‌فرد می‌شوند، مفید است. در مواردی که مجموعه داده بزرگ است و از کلمات کم استفاده یا منحصر به فرد تشکیل شده است، مدل‌های خطی مانند n-gram کار نمی‌کنند. این به این دلیل است که با افزایش کلمات، توالی کلمات ممکن افزایش می‌یابد و در نتیجه الگوهای پیش‌بینی‌کننده کلمه بعدی ضعیف‌تر می‌شوند.

b)

$$\frac{m}{M} < \frac{m+\alpha}{M+V\alpha}$$

[2] Query likelihood model

$$mM + mV\alpha < Mm + M\alpha$$

$$m < \frac{M}{V}$$

c)

کلمات منحصر به فرد = {ایران، سرا، من، است، امید، بوستان، بانو، آبی، آسمان، دختر، علی، دایی، اسطوره، <s>، </s>}

- ✓ **جمله اول**: <s> آسمان آبی است </s>

حالت unigram: احتمال وقوع هر کلمه به تنهایی بررسی می‌شود و به کلمات قبل وابستگی ندارد.

P(s) = p(<s>) p(آسمان) p(آبی) p(است) p(</s>)

برای محاسبه‌ی احتمال تکی هر کدام از کلمات با استفاده از laplace smoothing از فرمول زیر استفاده می‌شود:

$$p(x) = \frac{n_x + k}{N + kv}$$

در این مسئله k = 1 درنظر گرفته شده است.

$$p(s) = \frac{5+1}{15+28} \times \frac{1+1}{15+28} \times \frac{1+1}{15+28} \times \frac{3+1}{15+28} \times \frac{5+1}{15+28}$$

حالت bigram: احتمال وقوع هر کلمه به کلمه‌ی قبل از خود بستگی دارد.

P(s) = p(<s>) p(| آسمان >s<) p(آبی | آسمان) p(است | آبی) p(</s> | است)

$$= \frac{6}{43} \times \frac{0+1}{5+15} \times \frac{0+1}{1+15} \times \frac{0+1}{1+15} \times \frac{3+1}{3+15}$$

- ✓ **جمله دوم**: <s> دختر دایی امید ایرانی است </s>

حالت اول:

P(s) = p(<s>) p(دختر) p(دایی) p(امید) p(ایرانی) p(است) p(</s>)

$$= \frac{6}{43} \times \frac{2}{43} \times \frac{2}{43} \times \frac{2}{43} \times \frac{4}{43} \times \frac{4}{43} \times \frac{6}{43}$$

حالت دوم:

P(s) = p(<s>) p(دختر| <s>) p(دایی|دختر) p(امید|دایی) p(ایرانی|امید) p(است|ایرانی) p(</s>|است)

$$= \frac{6}{43} \times \frac{1}{20} \times \frac{1}{16} \times \frac{1}{16} \times \frac{1}{16} \times \frac{2}{18} \times \frac{4}{18}$$

- ✓ **جمله سوم**: <s> ایران بانو در سرای بوستان است </s>

حالت اول:

$$p(S) = \frac{6}{43} \times \frac{4}{43} \times \frac{2}{43} \times \frac{1}{43} \times \frac{3}{43} \times \frac{2}{43} \times \frac{4}{43} \times \frac{6}{43}$$

حالت دوم:

$$p(s) = \frac{6}{43} \times \frac{3}{20} \times \frac{1}{18} \times \frac{1}{16} \times \frac{1}{15} \times \frac{1}{17} \times \frac{1}{16} \times \frac{4}{18}$$

- ✓ **جمله چهارم**: <s> سرای من ایران است </s>

حالت اول:

$$p(s) = \frac{6}{43} \times \frac{3}{43} \times \frac{2}{43} \times \frac{4}{43} \times \frac{4}{43} \times \frac{6}{43}$$

حالت دوم:

$$p(s) = \frac{6}{43} \times \frac{1}{20} \times \frac{2}{17} \times \frac{1}{16} \times \frac{2}{18} \times \frac{4}{18}$$

Problem 3

کد به پیوست ارسال شده است.