به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۶

تاریخ تحویل:
۱۴۰۳/۱۰/۲۵

---

۱. (۲۰٪) [**پژوهش – مفاهیم شبکه‌های عصبی بازگشتی**] در درس با مفاهیم مختلف شبکه‌های بازگشتی آشنا شدید. در این بخش می‌خواهیم برخی مفاهیم دیگر در این حوزه را بررسی کنیم.

أ. پیش‌بینی ساختاریافته[۱] توضیح دهید و فرق آن با طبقه‌بندی[۲] بررسی کنید این تکنیک در چه کاربردهایی استفاده می‌شود؟

ب. هم‌ترازی دنباله‌ای[۳] چیست و چگونه می‌توان از RNN برای این کار استفاده کرد؟ کاربردهای آن را در حوزه‌هایی مانند بیوانفورماتیک[۴] بررسی کنید.

ت. مدل‌های ترکیبی که شبکه‌های بازگشتی را با معماری‌های دیگر مانند CNN یا Transformer ترکیب می‌کنند، چه مزایا و کاربردهایی دارند؟ توضیح دهید هرکدام از شبکه‌های عصبی چه نقشی در دسته‌بندی می‌توانند داشته باشند.

ث. روش تدریس اجباری[۵] را در شبکه‌های عصبی بازگشتی توضیح دهید.

---

۲. (۴۰٪) [**پیاده‌سازی –ترجمه ماشین عصبی با شبکه‌های مبدل**] در این تمرین، شما از شبکه‌های عصبی مبدل برای پیاده‌سازی یک سیستم ترجمه ماشین عصبی[۶] استفاده خواهید کرد. هدف شما این است که متنی را از یک زبان به زبان دیگر ترجمه کنید. برای این کار از مدل‌های مبدل و مجموعه داده‌ی این لینک استفاده خواهید کرد.

أ. پیش‌پردازش داده‌ها

۱. **تمیز کردن داده‌ها**: وظایفی شامل حذف نویزها، کلمات توقف، نشانه‌ها، و کاراکترهای غیرضروری را بر مجموعه داده اعمال کنید.

۲. **توکن‌سازی**: متن را به توکن‌ها (کلمات یا جملات) و تبدیل آن‌ها به فرمت‌های قابل استفاده برای مدل تقسیم کنید.

۳. **ایجاد دیکشنری‌ها**: یک دیکشنری از کلمات و توکن‌ها برای مدل‌های ورودی و خروجی بسازید

۴. **تقسیم داده‌ها**: در صورتی که حجم داده‌ها بزرگ است از ۲۰٪ داده‌ها استفاده کنید وداده‌ها را به سه دسته آموزش، آزمون، و ارزیابی تقسیم کنید.

۵. **Padding**: ازآنجایی که طول همه جملات یکسان نیستند، Padding مناسبی برای ورودی شبکه در نظر بگیرید. این Padding می‌تواند در سمت راست جمله و یا در سمت چپ آن باشد.

ب. آموزش مدل

۱. **مدل مبدل**: معماری شبکه‌ی نهایی می‌تواند مانند جدول زیر باشد اما در صورتی که معماری بهتری مد نظر دارید از آن نیز می‌توانید استفاده کنید:

---

[۱] Structured prediction
[۲] Classification
[۳] Sequence Alignment
[۴] Bioinformatics
[۵] Teacher forcing
[۶] Neural Machine Translation (NMT)

h.veisi@ut.ac.ir

Output Shape	Layer Type
(None,)	Input (encoder_input)
(None, sequence_len, embed_dim)	TokenAndPositionEmbedding (encoder)
(None, sequence_len, embed_dim)	TransformerEncoder
(None, None, embed_dim)	Input (encoded_seq_input)
(None,)	Input (decoder_input)
(None, sequence_len, embed_dim)	TokenAndPositionEmbedding (decoder)
(None, sequence_len, embed_dim)	TransformerDecoder
(None, sequence_len, embed_dim)	Dropout
(None, sequence_len, fr_vocab_size)	Dense (decoder_output)
(None, sequence_len, fr_vocab_size)	Output

جدول ۱ معماری شبکه

ت. ارزیابی مدل

۱. معیارهای ارزیابی:

- دقت ترجمه: میزان دقت ترجمه‌های مدل با استفاده از معیارهای همچون BLEU و ROUGE ارزیابی کنید.
- نمودار دقت و خطا در طول آموزش را رسم کنید.

---

۳. (۴۰٪) **[پیاده‌سازی – تشخیص احساسات با شبکه‌های عصبی بازگشتی]** در این تمرین، شما با استفاده از چهار روش دسته‌بندی شامل شبکه عصبی MLP، RNN، GRU و LSTM، تشخیص احساسات را تحلیل می‌کنید و نتایج این مدل‌ها را با یکدیگر مقایسه می‌کنید.

۱. پیش‌پردازش داده‌ها

أ. تمیز کردن متن: حذف نویزها، کلمات توقف[۱]، نشانه‌ها و کاراکترهای غیرضروری.

ب. توکن‌سازی[۲]: تقسیم متن به کلمات و توکن‌های مجزا.

ت. داده‌ها به سه دسته آموزش، آزمون و ارزیابی تقسیم کنید.

۲. استخراج ویژگی‌ها با روش‌های مختلف:

- Word2Vec:

[۱] Stop words
[۲] Tokenization

h.veisi@ut.ac.ir

بردارهای از پیش آموزش‌داده‌شده را از این لینک دریافت کنید. برای هر سند، میانگین و جمع بردارهای کلمات آن را محاسبه کرده و به‌عنوان نمایش برداری سند استفاده کنید.

- TF-IDF:

با استفاده از کتابخانه‌های موجود، ماتریس TF-IDF را برای اسناد ایجاد کنید.

- Bag of Words:

ماتریس حضور کلمات را برای اسناد ایجاد کنید.

- FastText:

مدل FastText را بر روی دادگان آموزش داده و بردارهای کلمات را استخراج کنید. مانند Word2Vec، میانگین و جمع بردارهای کلمات هر سند را محاسبه کنید

۳. آموزش مدل

- **شبکه عصبی ساده:**

أ. مدل شبکه‌ی عصبی

a. از TF-IDF برای استخراج ویژگی‌ها استفاده کنید.
b. یک لایه Dense با تعداد مناسب نرون‌ها برای یادگیری روابط ترتیبی در داده‌ها
a. لایه Dense با یک نورون و تابع فعال‌سازی Sigmoid برای پیش‌بینی کلاس‌ها.

- **شبکه‌های عصبی بازگشتی:**

ب. مدل RNN: یک مدل RNN با معماری زیر طراحی کنید:

a. لایه Embedding برای تبدیل کلمات به بردارهای عددی.
b. یک لایه SimpleRNN با تعداد مناسب نرون‌ها برای یادگیری روابط ترتیبی در داده‌ها.
c. لایه Dense با یک نورون و تابع فعال‌سازی Sigmoid برای پیش‌بینی کلاس‌ها.

ت. مدل LSTM: یک مدل LSTM با معماری زیر طراحی کنید:

a. لایه Embedding برای تبدیل کلمات به بردارهای عددی.
b. یک لایه LSTM برای یادگیری روابط طولانی‌مدت بین کلمات.
c. لایه Dense با یک نورون و تابع فعال‌سازی Sigmoid برای پیش‌بینی کلاس‌ها.

ث. مدل GRU: یک مدل GRU با معماری زیر طراحی کنید:

a. لایه Embedding برای تبدیل کلمات به بردارهای عددی.
b. یک لایه GRU برای یادگیری روابط بین کلمات.
c. لایه Dense با یک نورون و تابع فعال‌سازی Sigmoid برای پیش‌بینی کلاس‌ها.

۴. ارزیابی مدل

به نام خدا

أ. معیارهای ارزیابی
- چهار مدل را با استفاده از معیارهای زیر ارزیابی کنید:
  - دقت
  - AUC
  - F-1 socre

ب. مقایسه مدل‌ها
- عملکرد سه مدل را با استفاده از جدول یا نمودار مقایسه کنید.
- تحلیل کنید که کدام مدل عملکرد بهتری داشته و چرا.

ت. نتیجه‌گیری
- مشخص کنید که کدام روش برای این مسئله مناسب‌تر است.
- پیشنهادهایی برای بهبود مدل‌ها (مانند استفاده از داده‌های بیشتر یا مدل‌های پیشرفته‌تر) ارائه دهید.

---

۴. (۳۰٪ **نمره اضافی**) **[پیاده‌سازی – آشنایی با بستر ساخت چت‌بات RASA]** در این پرسش با ربات پاسخگو به پرسش‌های پرتکرار[۱] آشنا خواهید شد و یک نمونه از آن را با ابزار Rasa پیاده‌سازی خواهید کرد. پرسش و پاسخ‌های پرتکراری که در نظر گرفته شده، پرسش و پاسخ‌های پرتکرار شرکت ارتباطات سیار ایران، همراه اول در زمینه اینترنت است.

با توجه به آموزشی که در قالب فیلم به پیوست فرستاده شده، پس از نصب Rasa ، یک پروژه ابتدایی از آن ایجاد خواهید کرد. سپس اقدام به شخصی سازی بخش های مختلف پروژه و تغییر آن به شکلی که تبدیل به ربات پاسخگو به پرسش های پرتکرار شود خواهید نمود و در نهایت آن را آموزش خواهید داد تا ربات ساخته شود. برای ساخت ربات پاسخگو به پرسش های پرتکرار با ابزار Rasa می توان دو رویکرد کلی داشت:

۱- هر جفت پرسش و پاسخ، یک کلاس[۲] یا اندیشه‌ی[۳] جداگانه است.

۲- یک کلاس یا اندیشه واحد به نام پرسش و پاسخهای پرتکرار داریم و هر جفت پرسش و پاسخ، یک کلاس یا زیراندیشه از کلاس یا اندیشه مادر، یعنی اندیشه پرسش و پاسخهای پرتکرار هستند

در این تمرین فقط رویکرد اول را پیاده‌سازی خواهید کرد.

A. آماده‌سازی داده‌ها

داده ها در قالب جدولی با نام TrainData_Internet_MCI و با پسوند .xlsx به پیوست فرستاده شده اند. هر سطر در این جدول مربوط به یک اندیشه است که در آن هر اندیشه یک پاسخ و تعدادی نمونه پرسش ۵ دارد. مشابه با آن چه در یادگیری ماشین به آن عمل طبقه‌بندی می گویند، در این جا نیز ابتدا ربات با نمونه پرسشهای هر اندیشه آموزش می بیند و یک مدل شبکه عصبی تولید می کند. پس از آن، در عمل و هنگام کار و اجرا وقتی کاربر چیزی از ربات می پرسد، ربات اندیشه او را بر اساس مدل آموزش دیده مشخص می کند، سپس پاسخ مربوط به همان اندیشه را به کاربر برمی گرداند .در این بخش باید داده ها را در پرونده های مخصوص به خود در پروژه Rasa قرار

---

[۱] Frequently Asked Questions (FAQ)
[۲] Class
[۳] Intent

h.veisi@ut.ac.ir

دهید. پاسخها در پرونده domain.ymlو پرسشهای نمونه در پرونده nlu.yml (که در پوشه data قرار دارد) قرار می گیرند.

صورت دستی به صورت دستی تک تک داده ها را در قالبی مشابه با آنچه در پرونده های nlu.yml و domain.yml مشاهده می کنید قرار دهید.

B. انجام تنظیمات و آموزش ربات

ابزار Rasa برای ساخت ربات های گوناگونی مورد استفاده قرار می گیرد و هر ربات نیز تنظیماتی مخصوص به خود دارد. در این بخش شما باید با تغییر در پرونده config.yml تنظیمات مورد نیاز برای ساخت ربات پاسخگو به پرسشهای پرتکرار را متناسب با رویکردی که در حال اجرای آن هستید برگزینید . در این تمرین قصد داریم با عملکرد مدل‌های زبانی بر پایه Bert آشنا شوید. بنابراین باید ابتدا مدل‌های ParsBert و LaBSE را بیازمایید و عملکرد هر یک را گزارش کنید. سپس هر یک از این مدل ها که بهتر بود را انتخاب کنید و تاثیر تعداد دورهای[۱] آموزش را روی آن بررسی کنید. به این صورت که یک بار با ۵۰ دور، یک بار با ۱۰۰ دور و یک بار با ۲۰۰ دور آموزش را انجام دهید و در نهایت حالت بهینه را گزارش کنید. هر بار پس از انجام تغییرات، از محتوای پر وندهی Config.ymlعکس بگیرید و در گزارش خود بیاورید. سپس با توجه به عکس، تغییرات انجام شده را به طور مختصر توضیح دهید و کاربرد و دلیل هر یک از آن ها را بنویسید .)

C. ارزیابی ربات

با استفاده از دادگان آزمون که در پرونده TestData_Internet_MCIقرار گرفته، ربات و مقادیر دقت را ارزیابی کنید و سپس ماتریس آشفتگی و امتیاز F-1 را برای هر اندیشه گزارش کنید. (چگونگی انجام ارزیابی با استفاده از دادگانی غیر از دادگان آموزشی را از پایگاه اینترنتی مستندات Rasaمطالعه نمایید).

D. ابزار گفتوگوی تحت وب (بدون نمره و دلخواه)

به پیوست تمرین یک پوشه به نام WebChat قرار داده شده. این پوشه را به صورت کامل به پوشه پروژه ربات خود انتقال دهید. پس از آموزش و تکمیل ربات خود، برای گفت وگو با آن دو راه پیش رو دارید:

۱- استفاده از خط فرمان[۲]

حالت ساده و ابتدایی است، با اجرای دستور زیر در خط فرمان امکان پذیر است

```
rasa shell
```

۲- استفاده از ابزار گفت وگوی تحت وب

حالت دوم استفاده از یک رابط تصویری تحت وب است. برای استفاده از آن کافی است پس از ساخت ربات، در خط فرمان دستور زیر را اجرا کنید

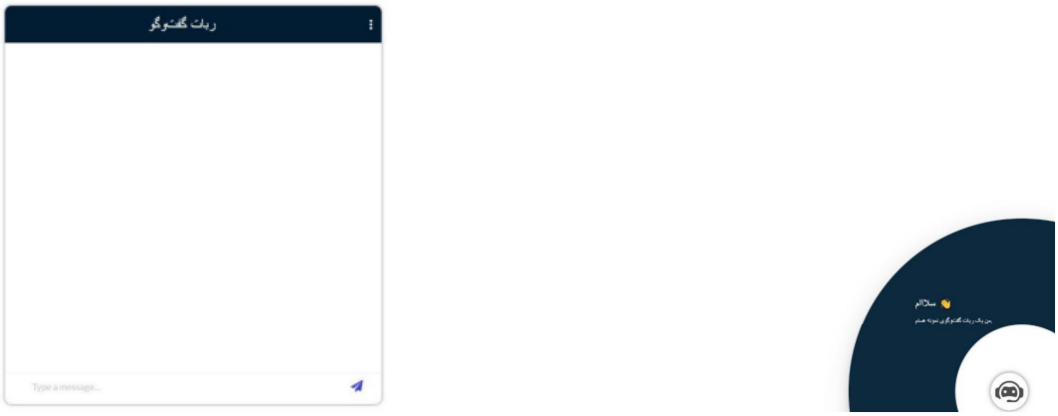
```
rasa run -m models --enable-api --cors "*" –debug
```

سپس وارد پوشه WebChatشده و پرونده html.indexرا باز کنید. سپس همان طور که در شکل زیر نشان داده

[۱] Epochs
[۲] Command Line

h.veisi@ut.ac.ir

شده، گوشه پایین-راست تصویر را لمس کنید تا صفحه گفت وگو با ربات نمایان شود.



شکل1 چت‌بات

h.veisi@ut.ac.ir