سوالات تمرین RNN [redacted]		گروه یادگیری عمیق دکتر قادری

در این تکلیف با اســـتفاده از کتابخانه های موجود برای پیش پردازش متن و مدل های RNN یک مدل زبانی متن طراحی خواهید کرد. بسیاری از بخش ها پیاده سازی شده اند اما توجه داشته باشید که این یک کد پیشنهادی است و شما میتوانید آن را تغییر دهید. سوالات مطرح شده را با توجه به مدل پیاده سازی خود پاسخ دهید.

عنوان تمرین:

توئیتر یک کانال ارتباطی است که به مردم این امکان را می دهد تا پیام ها، نظرات و حتی وضعیت های اضطراری را مشاهده کنند. حال ما یک دیتاست از توئیت های توئیتر داریم که میخواهیم با استفاده از یک مدل یادگیری ماشین، پیش بینی کنید کدام توئیت ها درباره بلایا واقعی هستند و کدام نیستند.

دیتاست دراختیار شما قرار گرفته است.

۱- در پردازش زبان طبیعی، تمام داده ها به صورت متن یا رشته هستند. درحالی که مدل سازی با الگوریتم طبقه بندی کننده یادگیری ماشین، به یک بردار ویژگی عددی نیاز دارد، برای این مسئله چه راه حلی را پیشنهاد می کنید؟

۲- نمودار Heatmap را رسم کرده و آن را تشریح کنید.

۳- بیست کلمه کلیدی برتر را استخراج و ترسیم کنید و سپس با کلیدواژه فاجعه بار (disaster) و غیر فاجعه بار (non-disaster) مجددا این عمل را تکرار کنید.

۴- در مرحله پردازش متن، نیاز اســت که علائم نگارشــی را حذف کنیم، پس یک تابع به نام "Toclean_text" تعریف کنید و با استفاده از آن علائم نگارشی را از روی داده های آموزشی حذف کنید و یک ستون clean_text برای داده های آموزشی ایجاد کنید که دارای متن بدون نقطه گذاری باشد.

۵- نویز در متن را میتوان هرچیزی دانســت که به تعامل زبانی عادی انســان تعلق دارد. نویز در متن را به طور کلی به عنوان URL، اختصــارات، ایموجی ها، پیام داخل تگ HTML و ... میتوان در نظر گرفت. دلیل اصــلی قرار گرفتن اختصارات به عنوان نویز این است که برخی افراد برای تشکر thx مینویسند. اگر اختصارات با کلمه اصلی جایگزین نشوند، "thx" و "thankyou" به عنوان دو کلمه متفاوت در نظر گرفته می شوند. نویز ها را شناسایی و سپس یک تابع به نام "clean_tweet()" تعریف و تمام نویز های متن را حذف کنید.

۶- Stopword ها کلماتی هســتند که معمولاً مورد اســتفاده قرار می گیرند که هیچ ویژگی متمایزکننده ای مانند "a"، "an"، "the" و غیره ندارند و موتور جســتجو طوری برنامه ریزی شــده اســت که هنگام نمایه ســازی ورودی ها و در حین بازیابی نتایج جســتجوی جســتجو، آنها را نادیده بگیرد. میتوانید با اســتفاده از ابزار nltk پایتون این کلمات را از متن حذف کنید.

۷- عمل Tokenization و دلیل اصلی استفاده از آن چیست؟ میتوانید از Keras Tokenizer() استفاده کنید. (پیش فرض در keras سطح کلمه است)

- مدل را یکبار بدون محدود کردن تعداد توکن و یکبار با محدود کردن تعداد توکن بررسی کنید.
- آیا لازم به انجام عمل Lowercasing در روش پیش‌پردازش هست؟

۸- تکنیک Pad_sequence() چیست و چه موقع از آن استفاده می کنیم؟

۹- لایه Embedding، اولین لایه شبکه عصبی است و دارای سه پارامتر هست:

- ✓ Input_dim: Number of distinct token vector
- ✓ Output_dim: Dimension of embedding vector
- ✓ Input_length: Size of input layer

سایز لایه Embedding چقدر است؟ دلیل استفاده از آن چیست؟

۱۰- برای بررسی نتایج از Confusion matrix استفاده کنید.

۱۱-از چه Optimizer میتوان استفاده کرد و اندازه Learning Rate چه اهمیتی دارد؟ با استفاده از صورت مسئله چه Activation function در لایه Dense و چه Loss Function مناسب است؟

۱۲-مدل پیشنهادی را با LSTM، GRU و ترکیب هر دو بررسی کنید. و نمودار یادگیری و نتایج را تحلیل کنید. (تعداد لایه ها و نحوه ترکیب آنها به اختیار خودتان است.)

۱۳-به نظر شما کدام یک از شبکه های عصبی زیر از مسئله vanishing gradient رنج میبرند؟ (علت انتخاب خود را توضیح دهید.)

- شبکه عصبی feed forward تک لایه
- شبکه عصبی feed forward عمیق
- شبکه عصبی Recurrent
- شبکه عصبی Recursive

۱۴-به نظر شما با چه تکنیک هایی میتوان مسئله Overfit شدن در مدل را برطرف کرد؟ شما از چه تکنیک هایی در پیاده سازی خود استفاده کرده اید؟ میزان تاثیر هر یک از روش های پیشنهادی شما چقدر است؟

۱۵-آیا اضافه کردن تعداد لایه های بیشتر در مدل شما، باعث حل مسئله Vanishing gradient می شود؟ (در مدل ارائه شده توسط شما)

۱۶-همانطور که میدانیم در زیر معادله تعریف LSTM را داریم، که در آن o به element-wise multiplication اشاره دارد و σ که sigmoid function است.

$$
\begin{aligned}
i_t &= \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1}\right) \\
f_t &= \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1}\right) \\
o_t &= \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1}\right) \\
\widetilde{c_t} &= \tanh\left(W^{(c)}x_t + U^{(c)}h_{t-1}\right) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \widetilde{c_t} \\
h_t &= o_t \circ \tanh\left(c_t\right)
\end{aligned}
$$

موارد زیر را بررسی کنید.

- اگر $x_t = 0$ درنتیجه $h_t = h_{t-1}$ خواهد بود.
- اگر f_t بسیار کوچک یا مساوی صفر باشد، خطا (error)، back-propagated به مراحل زمان قبلی منتقل یا منتشر نمی شود.
- ورودی های f_t, i_t, o_t ، غیرمنفی (non-negative) هستند.
- پارامتر های f_t, i_t, o_t را میتوانند به عنوان توزیع احتمال (Probability distributions) مشاهده شوند.

چند نکته مهم:

- موارد خواسته شده را اجرا و میتوانید تکنیک های BERT، GloVe، EDA و TF IDF را در کنار جزئیات خواسته شده میتوانید اعمال کنید.
- گزارش شما باید شامل روش های بکار گرفته شده و روش انجام محاسبات و کد نتایج و جمع بندی باشد. یک فایل ipynb و یک فایل pdf شرح نتایج و سوالات را در نهایت آپلود کنید.
- برنامه شما باید با پایتون نوشته شود و سایر زبان ها مورد قبول نیست. حجم گزارش به هیچ عنوان معیار نمره دهی نیست . دقت کنید که گزارش شما تعیین کننده میزان یادگیری و تفاوت کد شما با منابع دیگر را می رساند.
- [redacted]