به نام خدا

داده کاوی

تمرین امتیازی

دکتر شاکری

محمد امین عرب خراسانی

۸۱۰۱۰۲۲۰۵

بهار ۱۴۰۳

بخش عملی

۱) در مرحله‌ی اول می‌بایست دیتاست موردنظر لود شود تا با نوع دیتای آن آشنا شد. به همین منظور دیتاست در گوگل کولب آپلود می‌شود. بعد از لود دیتاست، ۵ سطر اول آن مشاهده می‌شود که نتیجه‌ی آن در زیر آورده شده است.

	ID	TITLE	ABSTRACT	Computer Science	Physics	Mathematics	Statistics	Quantitative Biology	Quantitative Finance
0	1	Reconstructing Subject-Specific Effect Maps	Predictive models allow subject-specific inf...	1	0	0	0	0	0
1	2	Rotation Invariance Neural Network	Rotation invariance and translation invarian...	1	0	0	0	0	0
2	3	Spherical polyharmonics and Poisson kernels fo...	We introduce and develop the notion of spher...	0	0	1	0	0	0
3	4	A finite element approximation for the stochas...	The stochastic Landau--Lifshitz--Gilbert (LL...	0	0	1	0	0	0
4	5	Comparative study of Discrete Wavelet Transfor...	Fourier-transform infra-red (FTIR) spectra o...	1	0	0	1	0	0

در ادامه، اطلاعات مربوط به این دیتاست با اجرای کد df.info() مشخص می‌شود که در زیر آورده شده است.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20972 entries, 0 to 20971
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ID                    20972 non-null  int64
 1   TITLE                 20972 non-null  object
 2   ABSTRACT              20972 non-null  object
 3   Computer Science      20972 non-null  int64
 4   Physics               20972 non-null  int64
 5   Mathematics           20972 non-null  int64
 6   Statistics            20972 non-null  int64
 7   Quantitative Biology  20972 non-null  int64
 8   Quantitative Finance  20972 non-null  int64
dtypes: int64(7), object(2)
memory usage: 1.4+ MB
```

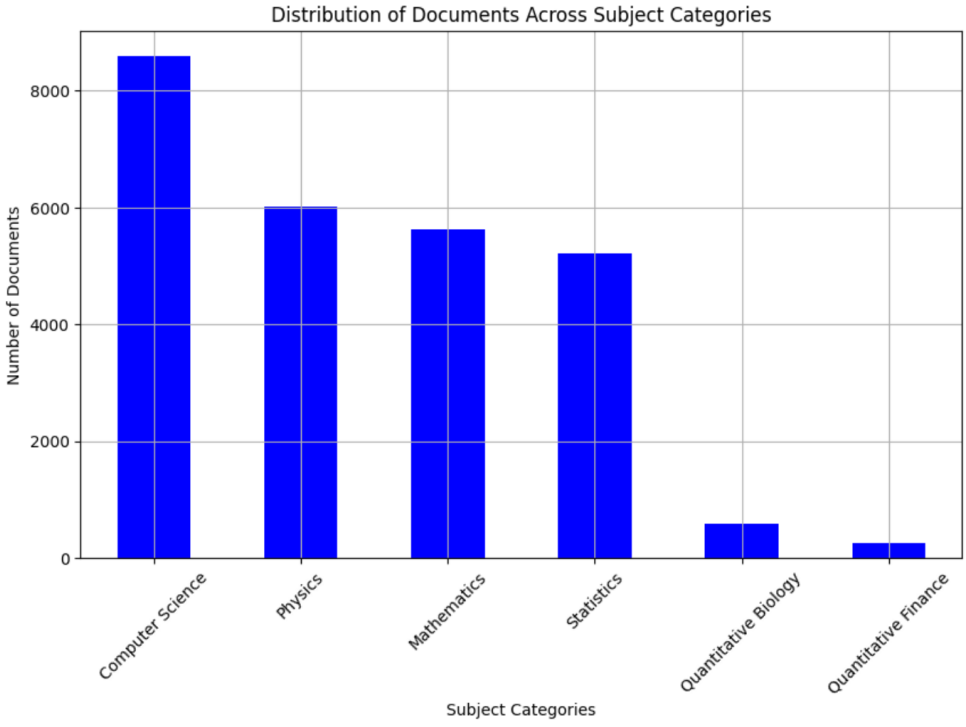
در ادامه برای آن که به جواب مشخص و درستی از دسته‌بند برسیم می‌بایست قبل از تقسیم دیتاست و آموزش دادن مدل بر اساس بخش ترین دیتاست، دیتاست تمیز شود.

از آن‌جایی که دیتای موجود در دیتاست به شکل متن می‌باشد می‌بایست چک شود که آیا تمامی کاراکترهای موجود در ABSTRACT و TITLE استاندارد هستند یا خیر. بنابراین این کاراکترها شناسایی شده و در نهایت حذف

می‌شوند. همچنین تمامی حروفی که در ABSTRACT و TITLE وجود دارند به صورت کوچک نوشته می‌شوند تا نرمالیزیشن انجام شود. نتیجه برای ۵ سطر ابتدایی به شکل زیر خواهد بود.

	TITLE	ABSTRACT
0	reconstruct subjectspecif effect map	predict model allow subjectspecif infer when a...
1	rotat invari neural network	rotat invari and translat invari have great va...
2	spheric polyharmon and poisson kernel for poly...	we introduc and develop the notion of spheric ...
3	a finit element approxim for the stochast maxw...	the stochast landaulifshitzgilbert llg equat c...
4	compar studi of discret wavelet transform and ...	fouriertransform infrar ftir spectra of sampl ...

در این مرحله، توزیع دیتای تمیز شده رسم می‌شود. نتیجه به شکل زیر خواهد بود.



۲) برای multi label classification دو حالت کلی وجود دارد. یکی از این حالات problem transformation methods و دیگری algorithm adaptation methods می‌باشد.

در مورد اول، هر مسئله‌ی multi label classification به چند مسئله‌ی single label classification تبدیل می‌شود. این اعمال به کمک دو روش binary relevance و classifier chains انجام می‌شود.

از سمت دیگر برای مورد دوم، الگوریتم به گونه‌ای تغییر پیدا می‌کند که مسئله‌ی multi label classification می‌تواند انجام شود. دو تکنیک مورد استفاده در این بخش multi lael k-nearest neighbor می‌باشد که مشابه الگوریتم k-nearest می‌باشد. تکنیک دوم random k-labelstes می‌باشد.

در این روش، بر اساس الگوریتم اصلی، هر دیتا به کلاسی مربوط می‌شود که بیشترین همسایگی را داشته باشد.

۳) در این بخش از binary relevance استفاده شده است. توضیحات مربوط به این تکنیک در سوال ۲ آورده شده است. برای تعریف مدل از tensorflow استفاده شده است.

در این مرحله به نسبت ۸۰ و ۱۰ و ۱۰، داده‌ها را به ۳ دسته‌ی اصلی تقسیم می‌شوند. در ادامه، با تنظیم پارامترهای مدلی که منجر به بهترین و بیشترین دقت می‌شود انتخاب می‌شود. بعد از ساخت مدل، آموزش لازم روی آن آنجا می‌شود.

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 300, 128)	640000
lstm (LSTM)	(None, 300, 64)	49408
dropout (Dropout)	(None, 300, 64)	0
lstm_1 (LSTM)	(None, 32)	12416
dense (Dense)	(None, 6)	198

Total params: 702022 (2.68 MB)
Trainable params: 702022 (2.68 MB)
Non-trainable params: 0 (0.00 Byte)

این پارامترها برای مدل به شکل زیر آورده شده است.

```
history = model.fit(X_train, y_train, epochs=100, batch_size=64,
validation_split=0.1)
```

۴) پس از اعمال تست روی مدل ساخته شده نتایج زیر حاصل می‌شود.