به نام خدا

دانشکده علوم و فنون نوین

مبانی محاسبات (رایانش) نرم (۰۱-۰۳۹-۰۵-۸۳)
نیم‌سال اول ۱۴۰۳-۱۴۰۲

تمرین شماره ۵

تاریخ تحویل: ۱۴۰۲/۱۰/۰۸

---

۱. (۱۰٪) [**پژوهش: شبکه‌های عصبی بزرگ**] هدف این سوال آشنایی با شبکه‌های عصبی پیش‌آموزش داده بسیار بزرگ مانند GPT-3 است که طی آن این شبکه و شبکه‌های مقیاس بزرگ مشابه را مرور کرده و مواردی مانند روش یادگیری، دادگان و کاربردهای آنها را گزارش کنید.

---

۲. (۲۰٪) [**پژوهش: روش‌های دیگر یادگیری**] روش‌های یادگیری زیر را مرور کرده و علاوه بر بیان ایده آنها، روش کار و نمونه کاربردهایی از هر کدام را بیان کنید.

الف) Progressive Learning

ب) Meta Learning

ج) Zero Shot Learning

د) One Shot Learning

---

۳. (۷۰٪) [**پیاده‌سازی: رمزگشایی و رمزگذاری با استفاده از الگوریتم ژنتیک**] هدف از رمزنگاری، تبدیل متن خام (پیام) به متن رمزشده است تا هیچکس جز مقصد پیام آن را نفهمد. هر چه بازگرداندن متن رمزشده به پیام اصلی از نظر زمانی، پیچیده‌تر باشد و بیشتر طول بکشد، رمزنگاری ارزشمندتر است. یکی از روش‌های مرسوم رمزنگاری، رمزنگار به روش جایگزینی است. در این روش، هر حرف به یک حرف دیگر نگاشت داده می‌شود و در پیام اصلی جایگزین آن قرار داده می‌شود تا رمز موردنظر به دست آید. به این نگاشت ۲۶ حرفی کلید گفته می‌شود. جدول زیر نمونه‌ای از این کلید است.

Alphabet	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**	**o**	**p**	**q**	**r**	**s**	**t**	**u**	**v**	**w**	**x**	**y**	**z**
Key	h	i	j	a	z	d	w	v	u	p	q	r	b	c	e	g	f	m	n	o	k	l	t	s	x	**y**

هنگام رمزنگاری هر حرف ردیف بالا با حرف متناظر از ردیف پایین جایگزین می‌شود و در فرآیند رمزگشایی، برعکس این اتفاق می‌افتد. به عنوان مثال، عملیات رمزنگاری با استفاده از کلید بالا به صورت زیر است. در فرآیند رمزنگاری حروف بزرگ به حروف بزرگ متناظر خود می‌روند.

h.veisi@ut.ac.ir

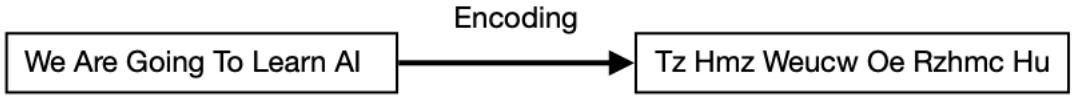
به نام خدا

دانشکده علوم و فنون نوین

مبانی محاسبات (رایانش) نرم (۰۱-۰۳۹-۰۵-۸۳)
نیم‌سال اول ۱۴۰۳-۱۴۰۲

تمرین شماره ۵

تاریخ تحویل: ۱۴۰۲/۱۰/۰۸



یکی از روش‌های برگرداندن پیام اصلی امتحان کردن تمام نگاشت‌های ممکن است. اما این کار ۲۶! حالت دارد و انجام آن به روش آزمون و خطا با کامپیوتر شخصی امکان پذیر نیست. در مسائلی به این شکل که فضای حالت بسیار بزرگی وجود دارد، استفاده از الگوریتم ژنتیک بسیار موثر است. با کمی دقت، می‌توان دید که این الگوریتم رمزنگاری آسیب‌پذیر است چراکه با داشتن لغت‌نامه‌ای از حروف می‌توان میزان مفید بودن یک کلید را سنجید و این کلید را آنقدر عوض کرد که به متن اصلی رسید. هنگامی که می‌خواهیم بررسی کنیم که رمزگشایی تا چه حد مفید بوده است، به دنبال کلمات آشنا و معنادار در آن می‌گردیم. هر چقدر طول یک کلمه معنادار بلندتر باشد، رمزگشایی ما مفیدتر بوده است. این یکی از روش‌های سنجش میزان مفید بودن رمزگشایی است و روش‌های دیگر نیز می‌تواند مفید باشد. در این پروژه یک پیام رمزشده به شما داده می‌شود و شما باید با استفاده از الگوریتم ژنتیک، کلید موردنظر را پیدا کنید و پیام رمزگشایی شده را بازگردانید.

برای این منظور، ابتدا باید مفهوم کروموزوم را در پروژه تعریف کنید. سپس، جمعیت[1] اولیه‌ای از کروموزوم‌های خود تعریف کنید. سپس، میزان مفید بودن هر کروموزوم را بسنجید. در این مسئله نیاز به تابع تناسب[2] دارید که بتواند به هر کروموزوم یک عدد به عنوان میزان مناسب بودن نسبت دهد. برای محاسبه امتیاز تناسب کروموزوم، نیاز به یک لغت نامه[3] داریم. یکی از کارهای اساسی و اولیه در پروژه‌ها فراهم کردن مجموعه‌ی داده[4] است. لغت‌نامه در این پروژه مجموعه داده‌ی ماست. برای درست کردن آن متنی را که در پیوست آمده است، باید پردازش کنید. تضمین می‌شود که تمامی لغات پیام‌ها در متن پیوست آمده‌اند. (در پروژه‌های واقعی خودتان باید لغت‌نامه خود را ایجاد و گسترش دهید تا پوشش مناسب را ایجاد کنید.) یکی از مراحل جمع‌آوری

[1] Population
[2] Fitting Function
[3] Dictionary
[4] Dataset

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

مبانی محاسبات (رایانش) نرم (۰۱-۰۳۹-۰۵-۸۳)
نیم‌سال اول ۱۴۰۳-۱۴۰۲

تمرین شماره ۵

تاریخ تحویل: ۱۴۰۲/۱۰/۰۸

مجموعه داده، تمیز کردن داده‌هاست. تمیزکردن در قالب این پروژه چگونه باید انجام شود؟ روند تولید لغت نامه از متن را توضیح دهید (به کاراکترهای غیر کارآمد و ایست واژه‌ها[5] فکر کنید.)

الف)پس از انتخاب کروموزوم‌های برتر، عملیات cross over و mutation انجام می‌شود تا جمعیت جدید به دست آید. روند ایجاد جمعیت جدید ادامه پیدا می‌کند تا با بهبود مرحله به مرحله کلید، کلید اصلی و متن خام اولیه حاصل شود. اگر تعداد جمعیتی را که در هر دوره نگه می‌داریم را افزایش دهیم، چه تاثیری بر روی سرعت و دقت می‌گذارد؟

ب) به نظر شما تاثیر mutation چیست؟ اگه فقط از crossover استفاده بشود، چه مشکلی پیش می‌آید؟

پ)Mutation موثرتر است یا cross over؟ کدام باعث سریعتر بالا رفتن دقت می‌شود؟

ت) با استفاده از این روش، باز ممکن است کروموزوم‌هایتان بعد از چند مرحله دیگر تغییر نکنند. چرا این اتفاق رخ می‌دهد؟ این سوگیری کروموزوم‌ها چه مشکلی پیش می‌آورد؟ برای حل آن چه راه حلی پیشنهاد می‌دهید؟ راه‌حل های مختلف خود را بیان کنید و آن‌ها را امتحان کنید. بهترین آنها را بر روی پروژه خود پیاده‌سازی کنید. (به میزان استفاده از انتخاب تصادفی فکر کنید)

نکات:

- تمام مراحل کار خود (تعریف کروموزوم، تولید جمعیت اولیه، تابع تناسب، عملیات cross over، mutation و جلوگیری از سوگیری کروموزوم‌ها، پردازش داده اولیه و تولید لغت نامه) را در گزارش کار خود بیاورید.
- برخی از پرسش‌های مطرح شده در متن از جمله مشکلاتی هستند که باید راهکاری برای آنها ارائه دهید. این دست از مشکلات و راه‌حل پیشنهادی باید در گزارش کار تفسیر شوند. برخی دیگر سوالاتی هستند که فکر کردن به آن‌ها در روند یادگیری و کیفیت

[5] Stop Words

h.veisi@ut.ac.ir

پروژه شما تاثیرگذار است. بر روی آن‌ها فکر کنید و مبنی بر پیاده‌سازی خود به آن‌ها پاسخ دهید.

- Decoder خود را در قالب یک کلاس پیاده‌سازی کنید. این کلاس باید تمام ویژگی‌های لازم برای رمزگشایی را داشته باشد تا بتوان به راحتی آن را تست کرد. سازنده این کلاس باید صرفا یک ورودی متن رمزشده را بگیرد .همچنین این کلاس، باید یک تابع decode داشته باشد که هیچ ورودی ندارد و خروجی آن متن رمزگشایی شده است.

کد شما باید بتواند به صورت زیر تست شود.

```
encoded_text = open("encoded_text.txt").read()

from code import Decoder
d = Decoder(encoded_text)
decoded_text = d.decode()

print(decoded_text)
```

---

۴. (۴۰٪ **اختیاری- نمره اضافی**) [**پیاده‌سازی: تحلیل احساس در متن با شبکه LSTM**] در این سوال قرار است با استفاده از شبکه‌ی حافظه‌ی کوتاه مدت ماندگار (LSTM) مدلی را آموزش دهیم که بتواند احساس مثبت یا منفی مربوط به هر نظر در داده‌های IMDB را تشخیص دهد. بدین منظور در ابتدا نیاز به آماده‌سازی داده و سپس طراحی مدل و آموزش و ارزیابی می‌باشد.

مجموعه داده نقد فیلم که معمولا با نام مجموعه داده IMDB شناخته می‌شود، شامل ۲۵۰۰۰ نقد مثبت و منفی برای آموزش شبکه و تعدادی مشابه برای ارزیابی می‌باشد. مسئله طراحی شده تشخیص حس مثبت و منفی در هر نظر نوشته شده است. داده مورد نظر از لینک زیر قابل دسترس است.

/http://ai.stanford.edu/~amaas/data/sentiment

قبل از شروع کار بر روی و طراحی مدل نیاز به آماده سازی و آنالیز داده می‌باشد. بدین منظور لازم است تمام داده به حروف کوچک تبدیل شوند. علائم نگارشی را حذف کنید. لیستی از نظرات را بسازید. لازم است متن را به واحدهای مجزا تجزیه کنید (واحدسازی) که برای این کار می‌توانید از کاراکتر

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

مبانی محاسبات (رایانش) نرم (۰۱-۰۳۹-۰۵-۸۳)
نیم‌سال اول ۱۴۰۳-۱۴۰۲

تمرین شماره ۵

تاریخ تحویل: ۱۴۰۲/۱۰/۰۸

فاصله و علائم سجاوندی به عنوان مرز بین کلمات استفاده کنید.

لازم است هر واحد (کلمه) را به یک بردار عددی تبدیل کنید. برای این کار از مدل‌ها و بردارهای آماده موجود مانند Word2Vec، FastText و یا Bert استفاده کنید. در این کار از بردارهای Bert استفاده کنید. مثلا در لینک زیر دریافت کنید:

https://github.com/google-research/bert

حال نوبت به آموزش شبکه می‌رسد. برای این کار از شبکه‌ی LSTM استفاده کنید و مدلی بسازید که ورودی آن یک نظر باشد و خروجی آن، بیانگر موافق یا مخالف بودن باشد. برای این منظور لازم است الگوریتم LSTM را به صورت کلی (با امکان تغییر تعداد لایه‌های مخفی، تعداد نرون‌های هر لایه، تنظیم تابع فعالساز، تغییر اندازه دسته (Batch) و...) پیاده‌سازی کنید. برای این تمرین می‌توانید از کتابخانه‌ها و ابزارهای موجود مانند تنسورفلو استفاده کنید.

الف) یک شبکه LSTM یک طرفه با batch size = 50، یک لایه مخفی با ۶۴ واحد حافظه و تابع فعال‌ساز سیگموید ایجاد کنید و یادگیری آن‌را با نرخ ۰.۰۱ انجام دهید. شبکه را با بخش Train دادگان آموزش داده و نموداری از روند تغییرات خطای MSE شبکه بر روی این داده‌ها رسم کنید. درستی (Accuracy) مدل آموزش یافته را روی داده آزمون گزارش کنید.

ب) شبکه بخش الف را دوطرفه کنید و نتایج را در این حالت بدست آورد و با حالت یک طرفه مقایسه کنید.

ج) در شبکه قسمت ب، تعداد لایه‌ها را به دو لایه افزایش داده و تعداد واحدهای حافظه هر لایه را با انجام چند آزمایش برای مقادیر مختلف بدست آورید تا بر اساس آن ساختار بهتر شناسایی شود. برای هر کدام از مقادیر بررسی شده، تعداد تکرار آموزش و مقدار درستی را روی دادگان آزمون گزارش کنید.

h.veisi@ut.ac.ir