بسمه تعالی

پروژه HMM درس مقدمه‌ای بر بیوانفورماتیک

نیم‌سال دوم سال ۱۳۹۹–۱۳۹۸

مهلت تحویل ساعت ۲۳:۵۵ روز جمعه ۱۳۹۹/۴/۶

توجه کنید که توضیحات پروژه بروز شده است.

۴. پیاده‌سازی روش مدل مخفی مارکوف گسسته برای ساخت پروفایل

هدف از این پروژه پیاده‌سازی روش HMM برای ساخت پروفایل از روی یک هم‌ترازی چندگانه است. برای این منظور شما باید با استفاده از پایتون برای حل مسائل سه‌گانه HMM کدنویسی کنید. برنامه شما باید بتواند ابتدا از روی یک هم‌ترازی ساختار مدل را پیدا کند که در عمل پیدا کردن تعداد حالات M در اسلایدها (match states) خواهد بود. سپس با توجه به نمونه‌ها به صورت اتوماتیک تشخیص دهد که دنباله‌ها مربوط به DNA هستند و یا پروتئین (کاراکترها را بررسی و بر اساس آن تصمیم‌گیری کنید). پس از این پارامترهای مدل با استفاده از هم‌ترازی ورودی یادگرفته خواهند شد.

در مرحله ارزیابی یک و فقط یک دنباله جدید به شما داده خواهد شد. برای این دنباله باید محتمل‌ترین هم‌ترازی (دنباله حالات) را با استفاده از الگوریتم ویتربی بدست بیاورید.

از آنجایی که زمان اجرای کدهای قبلی بیشتر از محدودیت زمانی کوئرا است و خیلی وقت‌ها به این علت مشخص نمی‌شود خروجی کد برای آن تست کیس چیست ارزیابی کدها به صورت آفلاین تغییر داده شد. این ارزیابی دو مرحله خواهد داشت که مرحله اول آن را با ۶ تست کیس مختلف شما انجام خواهید داد و در مرحله دوم روی تعداد دیگری تست کیس ما آن را ارزیابی خواهیم کرد. از این رو نیاز هست شرح مختصری از ارزیابی خود بر روی تست کیس‌ها بنویسید و به صورت فایل PDF در کوئرا آپلود کنید.

برای اینکه در زمان ارزیابی همه کدها به یک صورت قابل اجرا باشند باید به این صورت برنامه را بنویسید که برنامه ابتدا چک می‌کند که آیا پارامتری از کنسول به آن پاس داده شده یا نه. اگر بله، آن پارامتر آدرس پوشه تست‌ها خواهد بود. اگر نه با اجرای دستور زیر آدرس پوشه را از کاربر می‌گیرد:

پایتون ۲: raw_input('Enter dir of test cases: ')

پایتون ۳: input('Enter dir of test cases: ')

ساختار پوشه تست‌کیس‌ها به صورت زیر است:

```
main_dir:
    in:
        input1.txt
        input2.txt
        …
    out:
        output1.txt
        output2.txt
        …
```

که در آن in و out دو پوشه هستند که هر کدام به تعداد برابری ولی نامشخص فایل در آنها وجود دارد. ساختار فایل‌های ورودی به این صورت زیر هستند:

```
4 1
TTAGATTGATGC
TTAGGTTGGTGC
---GGTTAGTGC
-T-GGTTGGTGC
TAGGTTGGTGC
```

همان طور که مشخص است اولین خط فایل ورودی شامل دو عدد است که عدد اول (n) نشان دهنده تعداد دنباله‌ها در هم‌ترازی خواهد بود. دومین عدد در این خط (m) نشان‌دهنده مقدار آستانه برای تعداد گپ‌های مجاز در هر ستون است. **ستون‌هایی که تعداد گپ‌های آنها بیشتر از این مقدار باشد باید برای آموزش در نظر گرفته نشوند**. سپس به تعداد مشخص شده (n) هم‌ترازی چندگانه در خطوط جداگانه در فایل وجود دارد. در نهایت خط آخر برای ارزیابی خواهد بود. برای این خط باید محتمل‌ترین دنباله حالات برای تولید آن را با الگوریتم ویتربی پیدا کرده و بر اساس آن یک خط خروجی تولید کنید. بعد از بدست آوردن هم‌ترازی ویتربی، خروجی دنباله‌ای از کاراکترها است که به حالات M هم‌تراز شده‌اند و برای هر حالت D یک گپ در محل مربوطه به خروجی اضافه خواهد شد. کاراکترهایی که به حالات I تراز می‌شوند را در خروجی لحاظ نکنید.

فرمت فایل خروجی (output) ساده است و فقط شامل یک خط می‌شود که دنباله درستی که باید تولید شود را نشان می‌دهد. به عنوان مثال برای فایل ورودی بالا فرمت فایل خروجی به صورت زیر است:

```
TGGTTGGTGC
```

برنامه شما باید برای پردازش هر جفت فایل ورودی و خروجی یک و فقط یک خط خروجی به فرمت زیرچاپ کند:

```
input_sequance correct_output_sequence your_output_sequnce (True|False)
```

به عنوان مثال اگر فرض کنیم کد ما به درستی خروجی را تولید کرده است خط زیر باید در کنسول برای تست بالا چاپ شود.

```
TAGGTTGGTGC TGGTTGGTGC TGGTTGGTGC True
```

برای اینکه مشکل ضرب اعداد کوچک در هم پیش نیاید دو روش کلی Scaled و کار در حوزه لگاریتم ارائه شده‌اند. با توجه به مشکلاتی که برای طراحی تست‌ها وجود داشت و همچنین محدودیت‌های پردازشی کوئرا طول تست‌کیس‌ها و تعداد دنباله در آنها کم در نظر گرفته شده است. به همین خاطر مشکل محاسباتی احتمالاً به وجود نخواهد آمد. به همین خاطر استفاده از روش‌های مذکور در اینجا اختیاری است، امّا توجه کنید برای برنامه‌های واقعی حتماً باید از این روش‌ها استفاده شود. برای جزئیات بیشتر از پیاده‌سازی HMM می‌توانید به لینک‌های زیر مراجعه کنید. به عنوان مثال روابط موجود در صفحه ۶ فایل دوم به عنوان روابط کامل‌تر از چیزی که در اسلایدها درس دادیم و شامل بیش از یک رشته می‌شود است.

توجه: این تمرین به عنوان پروژه‌ای کوچک در نظر گرفته شده است و از این رو نمره آن نسبت به تمرین‌های قبلی بیشتر است.

https://www.diva-portal.org/smash/get/diva2:833697/FULLTEXT01.pdf

https://www.ece.ucsb.edu/Faculty/Rabiner/ece259/speech%20recognition%20course/term%20projects/HMM%20Project.pdf

توجه مهم: برای این پروژه باید کدهای خود را به صورت جداگانه به همراه یک توضیح مختصر در رابطه با تست‌کیس‌ها و نتایج کدهای خود در کوئرا آپلود کنید.

لطفاً به نکات زیر توجه کنید.

---

۱. لطفاً پروژه را به صورت انفرادی انجام دهید.
۲. pdf توضیحات خود را با کیفیت مناسب و خوانا اسکن کرده و یا تایپ شده به صورت یک فایل درآورید و با نام "HW3_StudentNumber.pdf" روی quera آپلود کنید.
۳. در صورتی پیداشدن هرگونه کپی در پروژه نمره‌ی هر دو نفر ۱۰۰- در نظر گرفته خواهد شد.
۵. اشکالات خود را می توانید از طریق ایمیل و یا آی‌دی تلگرام @deepmine_admin بپرسید.
hsn.zeinali@gmail.com