به نام خدا

دانشکده علوم و فنون نوین

# آشنایی با زبان‌شناسی رایانشی (۱۲۱-۰۵-۸۳)

## نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تاریخ تحویل: ۱۴۰۲/۰۲/۱۳

تمرین شماره ۲

---

۱. (۱۰٪) [**پژوهش: غلط‌های املایی و دستوری**] هدف از این تمرین آشنایی با روش‌های اخیر در اصلاح املا و دستور زبان است. این فرصت مناسبی است تا پیشرفت‌های جدید در پردازش زبان طبیعی را بررسی کنید و بینش‌هایی را در مورد اینکه چگونه هوش مصنوعی حوزه اصلاح املا و دستور زبان را تغییر داده است، کسب کنید.

**۱.۱** در پردازش زبان طبیعی، چه تکنیک‌های پیشرفته‌ای برای اصلاح املا و دستور زبان وجود دارد و چگونه تحولات اخیر در هوش مصنوعی، به ویژه در **مدل‌های زبانی بزرگ (LLMs**[1])، به این روش‌ها کمک کرده‌اند؟

**۲.۱** یک مقاله تحقیقاتی را که در **دو سال گذشته** منتشر شده و روش نوآورانه‌ای را برای اصلاح املا یا دستور زبان معرفی کرده است، انتخاب کنید و خلاصه‌ای کوتاه از یافته‌های مهم آن را ارائه دهید.

---

۲. (۳۵٪) [**پیاده‌سازی: عبارات با قاعده**] هدف از این تمرین آشنایی و تمرین با عبارات منظم[2] است.

۲-۱ مجموعه داده textCorpus.txt در اختیار شما قرار گرفته است. برای هر یک از قسمت‌های زیر عبارت منظم و الگو مورد نیاز را ایجاد و نتایج خواسته شده هر قسمت را در فایل خروجی مرتبط با هر قسمت ارائه کنید.

۱. سطرهایی که به علامت‌های «نقطه، سوال، علامت تعجب» ختم نشده‌اند را به سطر بعدی بچسبانید.
۲. فضای خالی ابتدا و انتهای همه سطرها را حذف کنید.
۳. فضاهای خالی که دارای بیش از یک کاراکتر فاصله هستند را به یک فاصله تبدیل کنید.
۴. اگر بیش از یک نیم‌فاصله در جایی استفاده شده، به یک نیم‌فاصله تبدیل کنید.
۵. اگر پیش از علائم نگارشی «»، }، ]، :، ،، ؛، ؟، !» فضای خالی وجود دارد، حذف کرده و اگر پس از آن فضای خالی نیست یک فاصله اضافه کنید.

---

[1] Large Language Models
[2] Regular Expression



h.veisi@ut.ac.ir

دانشکده علوم و فنون نوین

**آشنایی با زبان‌شناسی رایانشی (۱۲۱-۰۵-۸۳)**

**نیم‌سال دوم ۱۴۰۲-۱۴۰۱**

**تاریخ تحویل: ۱۴۰۲/۰۲/۱۳**

**تمرین شماره ۲**

۶. اگر پیش از علائم نگارشی «)، }، ]» فضای خالی نیست یک فاصله اضافه کنید و اگر پس از این علائم فضای خالی هست، آن را حذف کنید.

۷. کاراکترهای غیر استاندارد فارسی را به کاراکترهای استاندارد فارسی تبدیل کنید (برای دسترسی به لیست کامل آن به صفحه ویکی الفبای فارسی مراجعه کنید).

۲-۲ به کمک کتابخانه‌های هضم و پارسی‌ور به صورت جدا تغییرات بخش قبل را اعمال و تعداد توکن و کاراکترهای متن نهایی را در قالب جدولی گزارش کنید.

۲-۳ حال با توجه به پیکره نهایی که در بخش قبل نرمال‌سازی کردید، برای موارد خواسته شده تابعی جدا بنویسید که الگوی خواسته شده را دریافت کرده و نتایج مورد نظر را برگرداند.

۱. به کمک عبارت منظم عددهای موجود به صورت حروف (یک، دو، سه و ...) را به عدد استاندارد فارسی تبدیل کنید.

۲. به کمک عبارت منظم کلمات فارسی که دارای دو حرف بی‌صدا متوالی هستند را تشخیص دهید. به عنوان مثال، کلمه «کتاب» دارای دو حرف بی‌صدا متوالی «ک» و «ت» است.

۳. به کمک عبارت منظم کلمات فارسی که شامل یک دنباله خاص از حروف هستند را تشخیص دهید. به عنوان مثال، کلمه «خورشید» شامل دنباله «شی» است.

۴. به کمک عبارت منظم کلمات فارسی را که با یک حرف خاص شروع می‌شوند، تشخیص دهید. به عنوان مثال، کلماتی که با حرف «ب» شروع می‌شوند.

۵. به کمک عبارت منظم در پیکره اولیه عددهایی را که در هر دو نوشتار فارسی و عربی نوشته شده‌اند را تشخیص دهید، به عنوان مثال، عدد «۴۵» می‌تواند به صورت «۴۵» (نوشتار فارسی) یا «۴۵» (نوشتار عربی) نوشته شود.

۶. به کمک عبارت منظم فعل به اصطلاح مجهول که در فارسی از ساختار «ریشه + (ته/ده) + فعل شدن» ساخته می‌شود را تشخیص دهید.



h.veisi@ut.ac.ir

به نام خدا



۳. (۱۵٪) **[نظری: مبدل با حالت محدود]** هدف از این تمرین آشنایی با مبدل حالت متناهی است. به طور کلی، یک مبدل حالت متناهی یک مدل ریاضی برای توصیف رفتار یک سیستم است که در هر لحظه فقط در یک حالت از مجموعه‌ای متناهی از حالات قابل قبول قرار دارد و با دریافت ورودی‌های مختلف، بین این حالات تغییر می‌کند. برای طراحی یک مبدل حالت متناهی، باید ابتدا حالات، ورودی‌ها و خروجی‌های ممکن را شناسایی کنید. سپس باید تابع انتقال و تابع خروجی را تعریف کنید که نشان دهند با دریافت هر ورودی در هر حالت، مبدل به کدام حالت جدید منتقل می‌شود و چه خروجی را تولید می‌کند.

می‌خواهیم یک FST ساده طراحی و از آن استفاده کنیم.

۱. FST را با استفاده از نمادهای خروجی {N, PREP, V, ADJ} طراحی کنید که همه جملات زیر را بپذیرد. FST شما دنباله کلمات زیر را گرفته و دنباله گرامری آن را تولید می‌کند.

- لباس زیبایی پوشید.
- باز می‌گردیم.
- من امروز کلاس دارم.
- امروز هوا سرد است.
- به دانشگاه می‌رویم.
- هوای تهران آلوده است.

| N | V | ADJ | PREP |
|---|---|---|---|
| امروز | است | آلوده | از |
| تهران | باز می‌گردیم | زیبا | به |
| خانه | پوشید | سرد | در |
| دانشگاه | دارم | | |
| کلاس | می‌رویم | | |
| لباس | می‌مانم | | |
| من | | | |
| هوا | | | |




h.veisi@ut.ac.ir

۲. برای FST قسمت «۱»، هفت‌تایی مرتبط با تعریف FST $(Q, \Sigma, \Delta, q_0\ F, \delta(q, i), \sigma(q, w))$ را تعیین کنید.

۳. با پیمایش جملات زیر در FST طراحی شده قسمت «۱»، خروجی مرتبط با پذیرش یا عدم پذیرش آن‌ها را تعیین نمایید.

- از تهران باز می‌گردیم.
- کلاس زیبا در دانشگاه
- من امروز در خانه می‌مانم.

---

**۴. (۴۰٪) [پیاده‌سازی: ویراستار ساده]** هدف از این تمرین آموختن نحوه استفاده از الگوریتم فاصله لونشتاین است، که یک روش برای اندازه‌گیری شباهت بین دو رشته با شمارش حداقل تعداد ویرایش‌های (درج، حذف یا جایگزینی) لازم برای تبدیل یک رشته به رشته دیگر است. این الگوریتم می‌تواند برای بررسی املای کلمات با مقایسه یک کلمه غلط با لیستی از کلمات درست املایی و پیشنهاد کلمه‌ای با کمترین فاصله ویرایش به عنوان اصلاح، استفاده شود.

برای پیاده‌سازی، شما به دو فایل نیاز دارید: فایل story.txt شامل یک داستان با برخی کلمات غلط و فایل word_list.csv یک لیست از کلمه‌ها با املای درست برای کلمات غلط آمده است. وظیفه شما این است که یک برنامه پیاده‌سازی کنید که برای هر کلمه غلط در فایل داستان، کلمات اصلاح شده را با استفاده از الگوریتم فاصله لونشتاین پیشنهاد دهد. سپس شما می‌بایست کلمه‌ی با املای غلط را با کلمه‌ی با املای درستی که دارای کمترین فاصله ویرایش است جایگزین کنید. اگر پیشنهادهای چندگانه‌ای با همان فاصله ویرایش وجود داشته باشد، شما می‌توانید یکی را به صورت تصادفی انتخاب یا از معیارهای دیگری برای تعیین بهترین تطبیق استفاده و معیار استفاده خود را در گزارش شرح دهید.

نتایج بدست آمده را در قالب جدولی مشابه جدول زیر گزارش کنید.



h.veisi@ut.ac.ir

| کلمه‌های مرجع / کلمه‌های آزمون | کلمه درست اول | کلمه درست دوم | ... | آخرین کلمه درست |
|---|---|---|---|---|
| کلمه نادرست اول | | | | |
| کلمه نادرست دوم | | | | |
| ... | | | | |
| آخرین کلمه نادرست | | | | |

- همچنین بیان کنید؛ چه محدودیت‌هایی در استفاده از الگوریتم فاصله لونشتاین برای بررسی املا وجود دارد و چگونه می‌توان این محدودیت‌ها را رفع کرد؟

---

**۵.** (نمره اضافی ٪۵) **[پژوهش: چت جی‌پی‌تی[3]]** هدف از این تمرین آشنایی با سیستم‌های گفتگو مبتنی بر هوش مصنوعی است. سیستم چت جی‌پی‌تی شرکت OpenAI توانایی انجام گفتگوهای متنی به طور مشخص با زبان انگلیسی و حتی دیگر زبان‌ها را در طیف گسترده‌ای از موضوعات، پشتیبانی می‌کند. می‌توانید از سیستم، اطلاعاتی در مورد موضوع مشخصی بخواهید و درخواست نوشتن آن را با سبک‌های خاصی بدهید (مانند؛ مقاله، شعر و ...). این سیستم قادر به تولید کد نیز است. در دسترس بودن این سیستم (از اواخر سال ۲۰۲۲) هم شور و شوق و هم نگرانی‌هایی را برانگیخته است. به طوری که متنی که تولید می‌کند، کاملا روان است، به حدی که تشخیص آن از متن نوشته شده توسط یک انسان اغلب دشوار است. از جمله نگرانی‌ها، می‌توان موارد زیر را یاد کرد:

- ممکن است مردم توانایی‌های آن را بیش از حد ارزیابی کنند و تشخیص ندهند که چه زمانی چیزی نادرست، و یا گمراه‌کننده ایجاد می‌کند.
- از آنجایی که برای تقلید و ایجاد متنی مشابه با متنی که دیده، آموزش دیده است، ممکن است محتوای مغرضانه[4] یا مضر[5] تولید کند.

[3] ChatGPT
[4] Biased


h.veisi@ut.ac.ir

- این سیستم مرجع‌های خود را ذکر نمی‌کند، بنابراین ممکن است تا حد زیادی محتوایی تولیدی، سرقت ادبی باشد که توسط انسان نوشته شده است.
- ممکن است افراد به نادرست ادعای مالکیت بر خروجی آن کنند (به عنوان مثال، دانش‌آموزان/دانش‌جویان ممکن است از آن برای تقلب استفاده کنند).

۵-۱ خودتان ChatGPT را آزمایش کنید (این کار مستلزم ایجاد یک حساب کاربری در وب‌سایت OpenAI یا استفاده از اپلیکیشن **zigap** است) به این سوال پاسخ دهید: آیا فکر می‌کنید می‌تواند برای یادگیری شما مفید باشد (نه برای تقلب، بلکه برای بهبود درک شما از مفاهیم مورد مطالعه)؟ چرا و چرا نه؟

۵-۲ چند نمونه از سوالات/پاسخ‌های مکالمه‌تان را به عنوان شواهدی از اینکه چگونه می‌تواند مفید (یا نامفید) باشد، ضمیمه کنید. می‌توانید صفحه نمایش عکس‌برداری کنید یا قسمت‌های مربوطه را به صورت متن الصاق کنید. فقط مطمئن شوید که واضح است کدام قسمت‌ها، پرسش‌های شما و کدام قسمت‌ها، خروجی سامانه هستند. بدون در نظر گرفتن این نمونه عکس‌ها، پاسخ خود را مختصر (به عنوان مثال در ۱-۲ پاراگراف) نگه دارید.

[5] Harmful


h.veisi@ut.ac.ir

به نام خدا



## راهنمای تحویل

- انجام تمرین به صورت **تک نفره** است.
- جهت تحویل تمرین می‌بایست پیاده‌سازی‌ها از پایه و بدون استفاده از کتابخانه‌های موجود انجام شود، اما در شرابط قید کتابخانه در صورت سوال، منعی ندارد.
- زمان مناسبی را برای نوشتن و آماده‌سازی گزارش در نظر بگیرید، چرا که نیمی از نمره هر سوال، بر اساس گزارش تحویلی است.
- در صورت استفاده از **کدهای آماده**، دلیل استفاده، کامنت‌گذاری و توضیحات کافی، ضروری است، در غیر اینصورت تقلب تلقی می‌گردد.
- گزارش‌های تمرین خود را در مسیر documentation و اطلاعات مرتبط با پیاده‌سازی‌ها نیز در مسیر source قرار داده شوند.
- لطفا گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت CL_YourFamilyName_YourStNo_HW#.zip به ایمیل h.veisi@ut.ac.ir ارسال فرمائید.
- برای اطلاعات بیشتر به صفحه درس به آدرس https://dsp.ut.ac.ir/courses/y1401/introduction-to-computational-linguistics مراجعه کنید.

در صورت وجود سوال، ابهام و درخواست راهنمایی در گروه اسکایپی یا تلگرامی و یا از طریق ایمیل با دستیار آموزشی در ارتباط باشید.