بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۴-۰۳

مدرس: احسان‌الدین عسگری

تمرین اول | **داده** | مهلت ارسال: ۲ اردیبهشت

- **مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.**
- **انتخاب چالش در کوئرا** : در تمرین‌هایی که چند چالش دارند، فقط یک نفر از هر گروه در کوئرا باید چالش مورد نظر گروه را انتخاب کند. امکان تغییر چالش تا قبل از زمان ددلاین انتخاب چالش وجود دارد. البته ذکر این نکته ضروری است که هر چالش محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند دارد. بنابراین در اسرع وقت برای انتخاب چالش اقدام کنید.
- **امکان ارسال با تأخیر و قوانین آن:** در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز (با سقف ۵ روز برای هر تمرین) وجود دارد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- **قابلیت بازاجرای کامل نوت‌بوک‌ها:** توجه داشته‌باشید که نوت‌بوک‌های شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در همان نوت‌بوک قرار داده شود.
- **نحوه آپلود فایل‌های پروژه و مدل‌ها:** تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند، باید به شکل فایل فشرده در **CW** آپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره‌شده یک مدل در صورتی که بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی (نظیر Google Drive) آپلود کرده و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
  کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتماً در گزارش کار، نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها ذکر شود.
- **اهمیت گزارش کار:** بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری که کرده‌اید را توضیح دهید؛ بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. چند نمونه از خروجی‌های مساله را در گزارش بیاورید و بر اساس آن، رفتار برنامه‌تان را تحلیل کنید. در صورتی که پارامتری مانند دقت، صحت یا هر معیار دیگری خواسته شده باشد، آن‌ها را محاسبه کرده و در گزارش خود ارائه دهید.
- **تأثیر موارد امتیازی:** دقت داشته باشید، موارد امتیازی که در این تمرین آمده است، صرفاً بر روی امتیاز همین تمرین اثر دارد و روی نمرات تمرین‌های دیگر تأثیرگذار نخواهد بود.
- **نحوه پرسش سؤال و رفع ابهام:** در صورت وجود هرگونه ابهام یا مشکل، در کوئرا مطرح کنید و از ارسال پیام مستقیم به تیم تدریس خودداری نمایید.

توضیحات کلی

یکی از مهم‌ترین موضوعات پردازش زبان طبیعی کار بر روی داده، به خصوص دادگان مخصوص هر زبان یا دامنه‌های خاص می‌باشد. امروزه بحث دادگان مخصوص هر زبان و همچنین دادگان با کیفیت مهم شده است و با استفاده از این دادگان می‌توان LLM مطابق با ترجیح کاربران و دنبال کننده دستورات کاربران در عین رعایت امنیت تولید کرد. در این تمرین به این موضوع می‌پردازیم. هدف از این تمرین جمع‌آوری و کراولینگ دادگان فارسی با تمرکز بر مباحث فرهنگی می‌باشد به این شکل که بعد از به اتمام رسیدن تمرین یک مجموعه داده با کیفیت شامل دادگان متنی و مولتی‌مودال مرتبط

با فرهنگ‌های مختلف ایران و استان‌های مختلف خواهیم داشت که گامی مهم در جهت اضافه کردن ترجیحات فرهنگی ایرانی به مدل‌های زبانی می‌باشد. این تمرین به شکل **گروهی** می‌باشد و تیم شما باید ابتدا یکی از موضوعات مانند غذاها، تاریخ و غیره که توضیحات آن‌ها در ادامه آورده شده است را انتخاب کند، سپس برای هر کدام از موضوعات باید یکی از زیر بخش‌های آن را انتخاب کنید به عنوان مثال یک تیم ممکن است موضوع غذا را علاقمند و آشنا باشد همچنین به استان‌های گروه ۲ آگاهی بیشتری داشته باشد و به منابع آن دسترسی داشته باشد بنابراین موضوع تمرین آن‌ها غذاها در گروه ۲ استانی می‌شود و باید به حل این مسئله در دامنه استان‌های گروه ۲ بپردازند. توجه داشته باشید که برای انتخاب موضوع تمرین یک فرم منتشر می‌شود که باید در آن علاقمندی خود را وارد کنید و براساس ظرفیت هر موضوع (حدود ۲ تا ۳ تیم) یکی از اولویت‌های شما اعلام می‌شود. پس از مشخص کردن موضوع باید منابع مختلف برای آن را گردآوری کنید که تعدادی از آن‌ها در متن تمرین به عنوان راهنمایی آورده شده است، سپس با استفاده از مطالب آموخته شده در ورکشاپ کراولینگ باید یک کد برای جمع‌آوری دادگان مورد نظر خود پیاده‌سازی کنید سپس با استفاده از موارد آموزش داده شده در درس برای استخراج اطلاعات و پیش‌پردازش، کد فریمورک خود را به شکلی گسترش دهید که اطلاعات مورد نظر را تمیز و استخراج کند. لازم به ذکر است برای هر موضوع ساختار دقیق اطلاعاتی که باید استخراج شوند آورده شده است و خروجی نهایی شما باید طبق این ساختار برای هر موضوع باشد. در نهایت باید دادگان خود را در یک چارچوب برچسب‌گذاری **labelstudio** قرار دهید و از نظر کیفیت برچسب بزنید. برای اینکار نیاز به یک سیاست‌برچسب‌زنی دارید که باید درون تیم مشخص کنید و این سیاست را در گزارش تمرین خود وارد کنید، سپس شروع به برچسب‌زنی دادگان از نظر کیفی کنید به این شکل که حداقل دو نفر کیفیت هر نمونه را بسنجند و در صورتی که نظرات مخالف داشتند نفر سوم داده را برچسب بزند و در صورتی که داده کیفیت مناسبی نداشت به شکل دستی درست شود. در نهایت بعد از تولید دیتاست خود باید آن را در هاگینگ‌فیس که در اختیارتان قرار می‌گیرد آپلود کنید و کدها و گزارش کار خود را در CW به همراه گزارش آماری از دادگان (مشابه کاری که در نوت‌بوک‌های درس انجام شد) آپلود کنید.
لطفا به موارد زیر توجه فرمایید:

- **استفاده از چارچوب برچسب‌گذاری**: در این پروژه لازم است داده‌ها در یک چارچوب استاندارد برچسب زده شوند. سیستم پیشنهادی برای اینکار labelstudio می‌باشد. این قسمت باعث می‌شود که با فرآیند و ابزارهای برچسب‌زنی داده آشنا شوید که موضوع مهمی در صنعت و تحقیقات پردازش زبان طبیعی و به شکل کلی هوش مصنوعی می‌باشد.
- **مدل‌های زبانی بزرگ**: در صورت نیاز و زمانی که استخراج اطلاعات به شکل قاعده‌محور ممکن نبود، شما می‌توانید و حتی توصیه می‌شود که از مدل‌های زبانی بزرگ استفاده کنید، همچنین می‌توانید از این مدل‌ها در فرآیند برچسب‌زنی خود استفاده کنید. در صورتی که نیاز به استفاده از مدل‌های منبع‌باز LLM داشتید می‌توانید از مدل Gemma۳ استفاده کنید که نسخه‌های کم پارامتر آن قابل اجرا بر روی منابع موجود مانند کولب یا ککل می‌باشد. همچنین اگر نیاز به مدل OCR برای جمع‌آوری داده داشتید می‌توانید از مدل Olmocr استفاده کنید. در صورتی که مشکل در اجرای این مدل‌ها داشتید از تیم تدریس کمک بگیرید.
- **کنترل کیفیت و گزارش متریک‌های برچسب‌گذاری:** پس از گردآوری داده و انجام برچسب‌گذاری، باید فرایند کیفیت‌سنجی روی داده‌ها انجام شود. در این فرآیند باید معیارهای برچسب‌زنی و توافق برچسب‌زن‌ها را با متریک ارزیابی مانند امتیاز کاپا گزارش کنید. همچنین جزئیات فرآیند برچسب‌زنی مانند برچسب‌هایی که هر داده گرفته است و همچنین داده قبل و بعد از اصلاح را در CW آپلود کنید.
- **تحویل و بارگذاری داده‌ها روی مخازن هاگینگ‌فیس**: تمامی داده‌ها و مدل‌های نهایی (در صورت نیاز) باید روی مخازن هاگینگ‌فیس بارگذاری شوند. دسترسی به هاگینگ‌فیس به مسئول گروه داده خواهد شد.
- یک فعالیت امتیازی برای این تمرین در نظر گرفته شده است که تاثیر مثبت بالایی خواهد داشت. این فعالیت به این صورت هست که به عنوان مثال تیم x موضوع مشاهیر هنری را انتخاب کرده است و یکی از اعضای این تیم منابع یا دادگان خوبی در خصوص غذاهای استان خاصی دارد در اینصورت می‌تواند با پرکردن فرم مخصوص این فعالیت، اطلاعات و دادگان خود را در اختیار تیم‌هایی که روی موضوع غذا کار می‌کنند قرار دهد و باعث بهبود

و گستردگی دادگان شود. هدف از این فعالیت استفاده از اطلاعات خاص فرهنگی ساکنان شهرها و استان‌های مختلف می‌باشد که حضور دانشجویان از شهرهای مختلف این امکان را فراهم می‌کند و دادگانی از این جنس که مستقیم توسط منابع هر استان معرفی شود کمک به غنای دادگان تیم‌های دیگر خواهد کرد.

برای موضوعات غذاهای محلی، جغرافیا و سایت‌های گردشگری و آداب و رسوم محلی، تقسیم‌بندی به شکل استانی می‌باشد بنابراین باید هر تیم یک گروه از استان‌ها را انتخاب کند. انتخاب دسته‌بندی استان‌ها توسط گروه‌ها در فرم مربوطه به همراه انتخاب چالش انجام خواهد شد. دسته‌بندی استان‌ها عبارتند از:

- **گروه ۱:** تهران، قزوین، مازندران، سمنان، گلستان، البرز، قم
- **گروه ۲:** اصفهان، فارس، بوشهر، چهارمحال و بختیاری، هرمزگان، کهگیلویه و بویراحمد
- **گروه ۳:** آذربایجان شرقی، کردستان، آذربایجان غربی، زنجان، گیلان، اردبیل
- **گروه ۴:** لرستان، ایلام، کرمانشاه، همدان، مرکزی، خوزستان
- **گروه ۵:** خراسان رضوی، خراسان جنوبی، خراسان شمالی، کرمان، یزد، سیستان و بلوچستان

برای موضوع تاریخ دسته‌بندی وجود ندارد و هدف جمع‌آوری وقایع تاریخی مانند جنگ‌ها می‌باشد.
برای موضوع مشاهیر دسته‌بندی براساس موارد زیر است و نیاز هست هر تیم که موضوع مشاهیر را انتخاب می‌کند یکی از دسته‌های زیر را انتخاب کند.

- مشاهیر علمی و فناوری و فلسفی
- مشاهیر ادبی
- مشاهیر هنری و موسیقی
- مشاهیر سیاسی
- مشاهیر ورزشی

غذاهای محلی ایران

در این بخش، هدف ما جمع‌آوری و **ساختاردهی اطلاعات** مربوط به غذاهای محلی ایران از منابع متنی مختلف در سطح وب است. برای این منظور، داده‌های متنی مرتبط با غذاهای استان‌ها را از اینترنت استخراج می‌کنیم (دسته‌بندی استان‌ها در بخش ابتدایی تمرین مشخص شده است). منابع پیشنهادی برای این کار شامل وبلاگ‌های آشپزی، سایت‌های آموزش آشپزی، مقالات مرتبط با گردشگری غذایی و **ویکی‌پدیا** است. این منابع اطلاعاتی درباره **نام غذاها، مواد اولیه و مراحل تهیه** ارائه می‌دهند که می‌توانند در ساختاردهی داده‌های مرتبط مفید باشند.
برای استخراج اطلاعات از این منابع، می‌توان از کتابخانه‌های BeautifulSoup، Scrapy و Selenium استفاده کرد.
این ابزارها امکان پردازش و تحلیل داده‌های متنی را فراهم می‌کنند. بسته به منبع داده، ممکن است لازم باشد که **مراحل پیش‌پردازش متن** شامل حذف نویزهای متنی، یکسان‌سازی فرمت‌ها و استانداردسازی اطلاعات انجام شود. این کار باعث افزایش کیفیت داده‌های استخراج‌شده خواهد شد.
پس از اتمام مراحل پردازش داده، اطلاعات مرتبط با هر غذای محلی در یک **ساختار استاندارد** ذخیره می‌شوند. این ساختار شامل **نام غذا، استان و شهر مربوطه، مواد اولیه همراه با مقادیر آن‌ها، دستور تهیه و نوع وعده غذایی** خواهد بود. نمونه‌ای از این ساختار در قالب JSON به شکل زیر ارائه شده است:

```
{
  "title": "قلیه ماهی",
  "location": {
    "province": "خوزستان",
    "city": "آبادان",
    "coordinates": {
      "latitude": 30.3473,
      "longitude": 48.2934
    }
  },
  "ingredients": [
    {
      "name": "ماهی",
      "amount": 500,
      "unit": "گرم"
    },
    {
      "name": "سبزی (گشنیز و شنبلیله)",
      "amount": 250,
      "unit": "گرم"
    }
  ],
  "instructions": [
    "پیاز را خرد کرده و در روغن تفت دهید تا طلایی شود.",
    "تمبر هندی را در آب حل کرده و از صافی رد کنید، سپس به قابلمه اضافه کنید.",
    "ماهی را به قطعات متوسط برش داده و به خورش اضافه کنید."
  ],
  "meal_type": [
    "غذای اصلی",
    "دریایی"
  ],
  "occasion": [
```

```
    "ناهار",
    "شام"
  ],
  "images": {
    "تصویر نهایی": "https://example.com/images/ghalieh_mahi.jpg",
    "مرحله ۱": "https://example.com/images/step1.jpg",
    "مرحله ۲": "https://example.com/images/step2.jpg"
  }
}
```

این ساختار امکان پردازش و تحلیل داده‌ها را برای کاربردهای مختلف، از جمله **توسعه سامانه‌های معرفی غذاهای محلی** و **تحلیل داده‌های تغذیه‌ای** فراهم می‌کند.

تاریخ

در این بخش، هدف ما جمع‌آوری و **ساختاردهی اطلاعات** مربوط به مهمترین رویدادهای تاریخی ایران از منابع متنی مختلف در سطح وب است که می‌تواند شامل دسته‌های زیر باشد:

- جنگ‌ها
- قراردادها
- انقلاب‌ها
- مهمترین دستاوردهای دوره‌ها و پادشاهان
- ...

و اطلاعات مربوط به آن‌ها را با ساختار ارائه شده جمع‌آوری کنید. منابع پیشنهادی برای این کار:

- ویکی پدیا
- بلاگ های سایت کجارو و بیتوته
- سایت تاریخ ما
- https://tarikh.inoor.ir/
- https://tarikhirani.ir/
- https://historydocuments.ir/
- https://tarikhirani.ir/
- کتب تاریخی
- ...

برای استخراج اطلاعات از این منابع، می‌توان از کتابخانه‌های BeautifulSoup، Scrapy، Selenium برای کراول استفاده کرد. بسته به منبع داده، ممکن است لازم باشد که مراحل پیش‌پردازش متن شامل حذف نویزهای متنی، یکسان‌سازی فرمت‌ها و استانداردسازی اطلاعات انجام شود. این کار باعث افزایش کیفیت داده‌های استخراج‌شده خواهد شد.
پس از اتمام مراحل پردازش داده، اطلاعات مرتبط با هر رویداد در یک **ساختار استاندارد JSON** ذخیره می‌شوند. این ساختار شامل **منبع، نام رویداد، زمان رخداد، مکان رویداد، نتیجه، علت، طرفین، اهمیت تاریخی و توضیحات** خواهد بود. نمونه‌ای از این ساختار در قالب JSON به شکل زیر ارائه شده است:

```
{
  "title": "نبرد چالدران: برخورد ایران و عثمانی",
  "description": "نبرد چالدران میان صفویان ایران و امپراتوری عثمانی رخ داد. این جنگ به دلیل اختلافات مذهبی و سرزمینی به وقوع پیوست و عثمانی‌ها به علت برتری تسلیحاتی، پیروز شدند، اما موفق به اشغال ایران نشدند.",
  "period": {
    "start_year": "۱۵۱۴",
    "end_year": "۱۵۱۴"
  },
  "location": {
```

```
    "province": "آذربایجان غربی",
    "city": "چالدران",
    "coordinates": {
        "latitude": 36.2695,
        "longitude": 59.5863
    }
  },
  "causes": [
    "رقابت مذهبی بین شیعه و سنی",
    "اختلافات سرزمینی و توسعه‌طلبی عثمانی",
  ],
  "belligerents": [
    {
      "name": "صفویان",
      "leader": "شاه اسماعیل اول"
    },
    {
      "name": "امپراتوری عثمانی",
      "leader": "سلطان سلیم اول"
    }
  ],
  "result": "پیروزی عثمانی",
  "casualties": {
    "تلفات صفویان": "۵۰۰۰ نفر",
    "تلفات عثمانی‌ها": "۲۰۰۰ نفر"
  },
  "impact": [
    "کاهش نفوذ صفویان در آناتولی",
    "اثبات برتری سلاح‌های گرم عثمانی"
  ],
  "historical_significance": "این جنگ نقطه عطفی در تاریخ ایران بود که نشان داد برتری سلاح‌های گرم می‌تواند نتیجه نبردها را تغییر دهد.",
  "references": [
    {
      "title": "تاریخ ایران در دوران صفوی",
      "author": "دکتر حسن نصر",
      "year": "۱۹۹۸"
    },
    {
      "title": "تاریخ نظامی عثمانی",
      "author": "الیزابت جونز",
      "year": "۲۰۰۵"
    }
  ],
  "source": {
    "title": "نبرد چالدران: برخورد ایران و عثمانی",
    "author": "محمد رضایی",
    "publication_date": "۱۵ مه ۲۰۲۳",
    "url": "https://example.com/article"
  }
}
```

مشاهیر

در این بخش، هدف ما جمع‌آوری و **ساختاردهی اطلاعات** مربوط به مشاهیر ایران از منابع متنی مختلف در سطح وب است. برای این منظور، ابتدا یک دسته از مشاهیر ادبی، علمی، هنری و ... را انتخاب کنید و اطلاعات مربوط به آن‌ها را با ساختار ارائه شده جمع‌آوری کنید. منابع پیشنهادی برای این کار وبسایت ویکی‌پدیا، وبسایت گنجور و سایر وبسایت‌ها و وبلاگ‌های مرتبط با مشاهیر و آثار ادبی و علمی است.
برای استخراج اطلاعات از این منابع، می‌توان از کتابخانه‌های BeautifulSoup، Scrapy و Selenium استفاده کرد. این ابزارها امکان پردازش و تحلیل داده‌های متنی را فراهم می‌کنند. بسته به منبع داده، ممکن است لازم باشد که مراحل پیش‌پردازش متن شامل حذف نویزهای متنی، یکسان‌سازی فرمت‌ها و استانداردسازی اطلاعات انجام شود. این کار باعث افزایش کیفیت داده‌های استخراج‌شده خواهد شد.
پس از اتمام مراحل پردازش داده، اطلاعات مرتبط با هر شخص در یک **ساختار استاندارد JSON** ذخیره می‌شوند. این ساختار شامل **نام فرد، جنسیت، محل و تاریخ تولد و فوت و همچنین مقبره آن، دوره زندگی، مشاغل، آثار، رویدادهای مهم در طول زندگی آن فرد و تصاویر مربوطه** خواهد بود. نمونه‌ای از این ساختار در قالب JSON به شکل زیر ارائه شده است:

```
{
  "name": "ابوالقاسم فردوسی توسی",
  "sex": "مرد",
  "nick-names": [
    "حکیم توس",
    "حکیم سخن"
  ],
  "birth": {
    "date": "940 AD",
    "location": {
      "province": "خراسان",
      "city": "توس",
      "coordinates": {
        "latitude": 36.2695,
        "longitude": 59.5863
      }
    }
  },
  "death": {
    "date": "1025 AD",
    "location": {
      "province": "خراسان",
      "city": "توس",
      "coordinates": {
        "latitude": 36.2695,
        "longitude": 59.5863
      }
    },
    "tomb_location": {
      "province": "خراسان رضوی",
      "city": "توس",
      "coordinates": {
        "latitude": 36.2695,
```

```
        "longitude": 59.5863
      }
    }
  },
  "era": "سامانیان، غزنویان",
  "occupation": [
    "شاعر",
    "دهقان"
  ],
  "works": [
    "شاهنامه"
  ],
  "events": [
    {
      "title": "سرودن شاهنامه",
      "start_date": "974 AD",
      "end_date": "1004 AD",
      "location": {
        "province": "خراسان",
        "city": "توس",
        "coordinates": {
          "latitude": 36.2695,
          "longitude": 59.5863
        }
      },
      "related_people": [
        "منثورِ ابومنصوری"
      ],
      "description": "فردوسی در جوانی به مطالعه تاریخ ایران علاقه‌مند شد. وقتی دید که شاهنامه‌ی منثورِ ..."
    }
  ],
  "image": {
    "young": "https://example.com/images/ferdowsi_young.jpg",
    "adult": "https://example.com/images/ferdowsi_adult.jpg",
    "tomb": "https://example.com/images/ferdowsi_tomb.jpg"
  }
}
```

تغییرات اقلیمی و منابع طبیعی و سایت‌های گردشگری در ایران

در این بخش، هدف ما جمع‌آوری و **ساختاردهی اطلاعات** مربوط به **منابع طبیعی، جاذبه‌های گردشگری و تأثیرات تغییرات اقلیمی** در استان‌های ایران از منابع متنی مختلف در سطح وب است. برای این منظور، ابتدا یک گروه از استان‌ها را مطابق توضیحات ابتدای تمرین انتخاب کرده و داده‌های متنی مرتبط با این موضوعات را از منابع مختلف اینترنتی استخراج کنید. منابع پیشنهادی **وب‌سایت‌های گردشگری، کتاب‌های جغرافیای استانی، پایگاه‌های علمی، مقالات محیط‌زیستی، گزارش‌های تغییرات اقلیمی و ویکی‌پدیا** می‌باشد. این منابع اطلاعاتی درباره **ویژگی‌های طبیعی، جاذبه‌های گردشگری و تغییرات اقلیمی** ارائه می‌دهند که می‌توانند در ساختاردهی داده‌های مرتبط مفید باشند. استخراج اطلاعات و پیش‌پردازش از کتاب‌های جغرافیای استانی باید بخشی از سیستم شما باشد.

برای استخراج اطلاعات از این منابع، می‌توان از کتابخانه‌های BeautifulSoup، Scrapy و Selenium استفاده کرد. این ابزارها امکان پردازش و تحلیل داده‌های متنی را فراهم می‌کنند. بسته به منبع داده، ممکن است لازم باشد که **مراحل پیش‌پردازش متن** شامل حذف نویزهای متنی، یکسان‌سازی فرمت‌ها و استانداردسازی اطلاعات انجام شود. این کار باعث افزایش کیفیت داده‌های استخراج‌شده خواهد شد.

پس از اتمام مراحل پردازش داده، اطلاعات مرتبط با هر استان در یک **ساختار استاندارد** ذخیره می‌شوند. این ساختار شامل **نام استان، منابع طبیعی، جاذبه‌های گردشگری، و تأثیرات تغییرات اقلیمی به همراه تصاویر مختلف از آنها در طول سال‌های مختلف** خواهد بود. نمونه‌ای از این ساختار در قالب JSON به شکل زیر ارائه شده است:

```
{
  "title": "ویژگی‌های جغرافیایی اصفهان",
  "location": {
    "province": "اصفهان",
    "city": "اصفهان",
  },
  "geographical_features": [
    {
      "name": "رودخانه‌ها",
      "description": [
        {"name": "زاینده‌رود", "images": []},
        {"name": "رودخانه کوهپایه", "images": []},
        {"name": "رودخانه چم‌آباد", "images": []},
        {"name": "رودخانه دالانکوه", "images": []},
        {"name": "رودخانه‌های غرب اصفهان", "images": []}
      ]
    },
    {
      "name": "کوه‌ها",
      "description": [
        {"name": "کوه صفه", "images": []},
        {"name": "کوه‌های کرکس", "images": []},
        {"name": "کوه دالانکوه", "images": []},
        {"name": "کوه‌های زیاران", "images": []},
        {"name": "کوه‌های آران و بیدگل", "images": []}
      ]
    },
    {
      "name": "دریاچه‌ها",
```

```
      "description": [
        {"name": "تالاب گاوخونی", "images": []},
        {"name": "دریاچه نمک", "images": []}
      ]
    },
    {
      "name": "پوشش گیاهی",
      "description": [
        "بیابان‌های مرکزی ایران",
        "گونه‌های مقاوم به کم‌آبی",
        "گیاگان نادر مناطق بیابانی",
        "پوشش گیاهی مراتع کوهستانی",
        "گونه‌های گیاهی منطقه‌ای (پالیز و بادام وحشی)"
      ]
    }
  ],
  "topography": [
    {"name": "منطقه اقلیمی", "description": ["بیابانی", "نیمه‌بیابانی"]},
    {"name": "ارتفاع", "description": ["کوه‌های کرکس", "کوه صفه"]},
    {"name": "توپوگرافی", "description": ["ترکیب کوه‌های مرتفع", "دشت‌های وسیع",
    "مناطق بیابانی و کویری"]},
    {"name": "ویژگی‌های زمین‌شناسی", "description": ["گسل‌های فعال مانند گسل اصفهان-کاشان",
    "وجود گسل‌های زمین‌شناسی مهم", "منطقه کوهستانی"]}
  ],
  "natural_resources": [
    {"name": "منابع آبی", "description": ["آب‌های زیرزمینی", "چاه‌های کشاورزی",
    "رودخانه‌های فصلی", "چاه‌های عمیق", "سد زاینده‌رود"]},
    {"name": "منابع معدنی", "description": ["سنگ آهن", "نمک", "گچ",
    "گاز طبیعی", "معدن مس", "معدن سنگ مرمر"]}
  ],
  "tourist_attractions": [
    {"name": "میدان نقش جهان",
    "images": [], "year_built": "1598",
    "architect": "علی‌اکبر اصفهانی - محمدرضا ابن حسین بنای", "Constructor": "شاه عباس اول" ,
    "description": "یکی از بزرگ‌ترین میدان‌های جهان و از آثار ثبت‌شده در یونسکو."},
    {"name": "پل خواجو", "images": [],
    "year_built": "1650", "architect": "نامشخص",
    "Constructor": "شاه عباس دوم",
    "description": "یکی از زیباترین پل‌های تاریخی اصفهان با معماری بی‌نظیر."},
    {"name": "چهل ستون", "images": [], "year_built": "1647",
    "architect": "شیخ بهایی", "Constructor": "شاه عباس دوم",
    "description": "کاخی باشکوه با نقاشی‌های تاریخی و معماری صفوی."},
    {"name": "کاخ عالی‌قاپو", "images": [], "year_built": "1597",
    "architect": "علی‌اکبر اصفهانی - محمدرضا ابن حسین بنای اصفهانی",
    "Constructor": "شاه عباس اول",
    "description": "کاخی تاریخی با تزیینات زیبا و نمایی فوق‌العاده از میدان نقش جهان."},
    {"name": "مسجد شیخ لطف‌الله",
    "images": [], "year_built": "1619", "architect": "محمدرضا اصفهانی",
    "Constructor": "شاه عباس اول",
    "description": "مسجدی منحصربه‌فرد با کاشی‌کاری‌های بی‌نظیر در میدان نقش جهان."}
```

```
  ],
"additional_info": {
  "books_source": "http://chap.sch.ir/category/%D8%AF%D9%88%D8%B1%D9%87/%D8%AF%D9
   %88%D8%B1%D9%87-%D8%A2%D9%85%D9%88%D8%B2%D8%B4-%D9%85%D8%AA%D9%88%D8%B3%D8%B7
   %D9%87/%D8%AC%D8%BA%D8%B1%D8%A7%D9%81%D9%8A%D8%A7%DB%8C-%D8%A7%D8%B3%D8%AA%D8
   %A7%D9%86-%D9%87%D8%A7"
   }

}
```

آداب و رسوم محلی ایران

در این بخش، هدف ما جمع‌آوری و **ساختاردهی اطلاعات** مربوط به آداب و رسوم مختص هر استان ایران، شامل مراسم مخصوص به آن منطقه در مناسبت‌های مختلف، صنایع دستی، لباس محلی و دیگر ویژگی‌های فرهنگی است.
برای این منظور، ابتدا یک گروه از استان‌ها را طبق توضیحات کلی تمرین انتخاب کرده و داده‌های متنی مرتبط با این موضوعات را از منابع مختلف اینترنتی استخراج کنید. منابع پیشنهادی شامل مقالات گردشگری، وب‌سایت‌های فرهنگی، پژوهش‌های مردم‌شناسی و **ویکی‌پدیا** است. این منابع اطلاعاتی درباره **مراسم سنتی، لباس‌های محلی، صنایع دستی** و دیگر شاخص‌های فرهنگی ارائه می‌دهند که می‌توانند در ساختاردهی داده‌های مرتبط مفید باشند.
برای استخراج اطلاعات از این منابع، می‌توان از کتابخانه‌های BeautifulSoup، Scrapy و Selenium استفاده کرد.
این ابزارها امکان پردازش و تحلیل داده‌های متنی را فراهم می‌کنند. بسته به منبع داده، ممکن است لازم باشد که **مراحل پیش‌پردازش متن** شامل حذف نویزهای متنی، یکسان‌سازی فرمت‌ها و استانداردسازی اطلاعات انجام شود. این کار باعث افزایش کیفیت داده‌های استخراج‌شده خواهد شد.
پس از اتمام مراحل پردازش داده، اطلاعات مرتبط با هر استان در یک **ساختار استاندارد** ذخیره می‌شوند. این ساختار شامل **نام استان، شهرهای مرتبط، مراسم خاص، صنایع دستی، لباس‌های محلی و سایر عناصر فرهنگی** خواهد بود.نمونه‌ای از این ساختار در قالب JSON به شکل زیر ارائه شده است:
برای مراسم‌های محلی می‌توانید از فرمت زیر استفاده کنید:

```
{
  "title": "نوروز",
  "category": "مراسم",
  "description": "نوروز یکی از بزرگ‌ترین و قدیمی‌ترین جشن‌های ایرانی است که در آغاز بهار جشن گرفته می‌شود.",
  "location": {
    "province": "ایران باستان",
    "city": "ایران باستان",
    "coordinates": {
      "latitude": 30.3473,
      "longitude": 48.2934
    }
  },
  "history": "نوروز از زمان‌های قدیم تا به امروز در ایران و برخی کشورهای همسایه جشن گرفته می‌شود.",
  "elements": [
    {
      "name": "سبزه",
      "description": "سبزه یکی از نمادهای نوروز است که در سفره هفت‌سین قرار می‌گیرد."
    },
    {
      "name": "هفت‌سین",
      "description": "هفت‌سین به مجموعه‌ای از هفت شیء با نام‌هایی که با حرف سین شروع می‌شود گفته می‌شود."
    }
  ],
  "occasion": [
    "نوروز",
    "سال نو"
  ],
  "attire": [
    {
      "name": "لباس سنتی",
```

```
      "description": "در نوروز مردم ایران معمولاً لباس‌های نو و سنتی می‌پوشند."
    }
  ],
  "music": [
    {
      "name": "موسیقی نوروزی",
      "instrument": "سازهای سنتی مانند تار و سه‌تار",
      "description": "موسیقی سنتی ایرانی که در مراسم نوروز نواخته می‌شود.",
      "audio": "https://example.com/audio/nowruz_music.mp3"
    }
  ],
  "special_customs": [
    "پذیرایی از مهمانان با شیرینی‌های مخصوص نوروز",
    "دید و بازدید خانواده‌ها"
  ],
  "activities": [
    "موسیقی زنده",
    "رقص‌های محلی",
    "نمایش‌های فرهنگی"
  ],
  "date_time": "۱ فروردین هر سال",
  "ceremony_type": "جشنواره فرهنگی",

  ],
  "images": {
    "تصویر ۱": "https://example.com/images/nowruz1.jpg",
    "تصویر ۲": "https://example.com/images/nowruz2.jpg",
    "تصویر ۳": "https://example.com/images/nowruz3.jpg"

  }
}
```

برای لباس‌های محلی می‌توانید از فرمت زیر استفاده کنید:

```
{
  "title": "لباس محلی گیلانی",
  "category": "لباس محلی",
  "description": "لباس‌های سنتی مردم گیلان که در مراسم‌های مختلف، مانند عروسی‌ها پوشیده می‌شود.",
  "materials": [
    {
      "name": "پارچه ابریشمی",
      "description": "یک پارچه نرم و براق که در تولید لباس‌های محلی استفاده می‌شود."
    },
    {
      "name": "نوارهای تزئینی",
      "description": "نوارهای رنگارنگ که به عنوان تزئینات در لباس‌های گیلانی به کار می‌روند."
    }
  ],
  "history": "لباس‌های محلی گیلانی در مراسمات مهم زندگی همچون عروسی‌ها و جشن‌ها مورد استفاده قرار می‌گیرد.",
  "location": {
    "province": "گیلان",
    "city": "رشت",
```

```
    "coordinates": {
      "latitude": 30.3473,
      "longitude": 48.2934
    }
  },
  "occasion": [
    "عروسی",
    "جشن‌ها"
  ],
  "season": "تابستان و بهار",
  "cultural_significance": "نمادهایی چون رنگ‌های خاص نشان‌دهنده‌ی وضعیت اجتماعی افراد است.",
  "social_context": "در مراسم عروسی، لباس‌های گیلانی معمولاً نماد احترام و پذیرش مهمانان است.",
  "gender": "مردانه و زنانه",
  "images": {
    "تصویر ۱": "https://example.com/images/gilan_traditional_dress1.jpg",
    "تصویر ۲": "https://example.com/images/gilan_traditional_dress2.jpg",
    "تصویر ۳": "https://example.com/images/gilan_traditional_dress3.jpg",
  }
}
```

برای صنایع دستی محلی می‌توانید از فرمت زیر استفاده کنید:

```
{
  "title": "سفالگری لالجین",
  "category": "صنایع دستی",
  "description": "سفالگری یکی از هنرهای دستی معروف در شهر لالجین همدان است.",
  "materials": [
    {
      "name": "خاک رس",
      "description": "ماده اولیه برای ساخت سفال‌های لالجین که به دلیل کیفیت بالای آن استفاده می‌شود."
    },
    {
      "name": "لعاب",
      "description": "لعاب‌هایی که برای پوشش سطح سفال‌ها به کار می‌روند و باعث براق شدن آن می‌شوند."
    }
  ],
  "history": "سفالگری در لالجین از دیرباز رایج بوده و امروزه یکی از مهم‌ترین صنایع دستی ایران شناخته می‌شود.",
    "location": {
    "province": "همدان",
    "city": "همدان",
    "coordinates": {
      "latitude": 30.3473,
      "longitude": 48.2934
    }
  },
  "techniques": [
    "چرخ‌زنی",
    "کوزه‌گری"
  ],
  "production_steps": [
    "انتخاب خاک رس",
    "ساخت قطعات با استفاده از چرخ سفالگری",
```

```
    "خشک کردن قطعات",
    "لعاب زدن و پخت در کوره"
  ],

  "marketplaces": [
    "بازار همدان",
    "نمایشگاه‌های صنایع دستی"
  ],
  "cultural_factors": {
    "symbolism": "سفالگری لالجین به عنوان نمادی از هنر و ذوق مردم این منطقه شناخته می‌شود.",

  },
  "images": {
    "تصویر ۱": "https://example.com/images/lalejin_pottery1.jpg",
    "تصویر ۲": "https://example.com/images/lalejin_pottery2.jpg",
    "تصویر ۳": "https://example.com/images/lalejin_pottery3.jpg"  }
}
```