دانشگاه صنعتی شریف

پردازش زبان طبیعی مدرس: دکتر عسگری

گزارش تمرین سری یک

نام و نام‌خانوادگی : امیر ملک حسینی *شماره دانشجویی: ۴۰۱۱۰۰۵۲۸*

چکیده

این گزارش، فرآیند ایجاد یک مجموعه داده از دستور پخت غذاهای محلی ایرانی را تشریح می‌کند. هدف اصلی، جمع‌آوری دستور پخت‌ها از منابع آنلاین مختلف، ساختارمند کردن داده‌ها، حاشیه‌نویسی (An-notation) آن‌ها با فراداده‌های مرتبط (نوع وعده، مناسبت، کیفیت) و انجام ارزیابی کیفی بود. داده‌ها از چندین وب‌سایت پوشش‌دهنده استان‌های مختلف با استفاده از کتابخانه‌های پایتون مانند requests و BeautifulSoup استخراج (Scrape) شدند. داده‌های جمع‌آوری‌شده تجمیع و سپس با استفاده از پلتفرم LabelStudio حاشیه‌نویسی شدند. ارزیابی کیفی بر اساس قضاوت حاشیه‌نویس در مورد وضوح و کامل بودن دستور پخت انجام شد. ورودی‌هایی که کیفیت پایینی داشتند، فیلتر شدند. مجموعه داده نهایی شامل ۱۷۰ دستور پخت با کیفیت بالا است. تحلیل‌ها نشان داد که تعداد کل کلمات در فیلدهای مرتبط در مجموعه داده نهایی ۲۶,۵۲۹ کلمه است و میانگین طول دستورالعمل‌ها تقریباً ۱۱۹ کلمه به ازای هر دستور پخت می‌باشد. این مجموعه داده، منبعی ساختاریافته برای وظایف پردازش زبان طبیعی (NLP) مرتبط با آشپزی ایرانی فراهم می‌کند.

۱ مقدمه

هدف این پروژه، گردآوری یک مجموعه داده ساختاریافته از دستور پخت غذاهای سنتی ایرانی از استان‌های مختلف بود. چنین مجموعه داده‌ای می‌تواند برای تحقیقات پردازش زبان طبیعی، حفظ میراث فرهنگی و کاربردهای آشپزی ارزشمند باشد. به دلیل کمبود مجموعه داده‌های ساختاریافته و آماده در این حوزه، از رویکرد استخراج وب برای جمع‌آوری اطلاعات از منابع آنلاین متفاوت استفاده شد. داده‌های جمع‌آوری‌شده سپس پردازش، برای ویژگی‌های خاص حاشیه‌نویسی و از نظر کیفی فیلتر شدند.
استان‌های هدف شامل موارد زیر بودند:

- اصفهان
- فارس (شیراز)
- هرمزگان
- چهارمحال و بختیاری
- خوزستان
- بوشهر
- کهگیلویه و بویراحمد

این گزارش، روش‌شناسی مورد استفاده برای جمع‌آوری داده‌ها، تجمیع، حاشیه‌نویسی، پاک‌سازی و تحلیلی از مجموعه داده نهایی را ارائه می‌دهد.

۲ روش‌شناسی

این فرآیند شامل چندین مرحله مجزا بود: جمع‌آوری داده از طریق استخراج وب، تجمیع داده‌ها از منابع مختلف، حاشیه‌نویسی با استفاده از LabelStudio و در نهایت، پاک‌سازی و فیلتر کردن داده‌های حاشیه‌نویسی‌شده.

۱.۲ جمع‌آوری داده‌ها

استخراج وب با استفاده از پایتون انجام شد و عمدتاً از کتابخانه requests برای دریافت صفحات وب و BeautifulSoup4 برای تجزیه (Parse) محتوای HTML استفاده گردید. به دلیل تفاوت در ساختار و نحوه ارائه محتوا، اسکرپرهای (Scrapers) مخصوصی برای هر وب‌سایت هدف توسعه داده شد.
چالش‌ها و تکنیک‌های کلیدی شامل موارد زیر بودند:

- **مدیریت ساختارهای متنوع:** وب‌سایت‌ها از طرح‌بندی‌های HTML متفاوتی (مانند جداول، لیست‌ها، تگ‌های عنوان خاص مثل <h2>، <h3>) برای ارائه اطلاعات دستور پخت (عنوان، مواد لازم، دستورالعمل‌ها، تصاویر) استفاده می‌کردند. اسکرپرها برای شناسایی و استخراج داده‌ها از این ساختارهای خاص، سفارشی‌سازی شدند.
- **پردازش متن فارسی:** توابع کمکی (convert_persian_numbers، clean_title) برای مدیریت اعداد فارسی (تبدیل '۰'-'۹' به '۰'-'۹') و کلمات عددی فارسی (مثلاً 'پنج' به ۵) ایجاد شدند. عناوین برای حذف شماره‌گذاری یا خط تیره ابتدایی، پاک‌سازی شدند.
- **تجزیه مواد لازم:** تابعی (parse_amount_unit) برای تجزیه خطوط مواد لازم توسعه داده شد که سعی می‌کرد نام ماده، مقدار عددی و واحد (مانند «گرم»، «پیمانه»، «قاشق») را جدا کند. این تابع فرمت‌های مختلفی از جمله کسرها ('۱/۲')، عبارات خاص ('مقدار لازم') و کلمات عددی فارسی را مدیریت می‌کرد.

- **استخراج دستورالعمل‌ها:** دستورالعمل‌ها اغلب درون پاراگراف‌ها (<p>) یا آیتم‌های لیست (<li>) پس از عناوین خاص (مانند «طرز تهیه») یافت می‌شدند. گاهی از عبارات منظم (Regular expressions) برای شناسایی مراحل شماره‌گذاری‌شده استفاده می‌شد.
- **استخراج تصاویر:** URL تصاویر، به‌ویژه تصویر اصلی دستور پخت، از تگ‌های <img> استخراج می‌شد که اغلب نیازمند شناسایی بر اساس نزدیکی به عنوان دستور پخت یا کلاس‌های CSS خاص بود. از های URL پایه برای تبدیل مسیرهای نسبی تصاویر استفاده شد.

هر اسکرپیر یک فایل JSON حاوی لیستی از اشیاء دستور پخت برای استان مربوطه خود تولید کرد.

۲.۲ تجمیع داده‌ها

فایل‌های JSON منفرد تولید شده از فرآیند استخراج برای هر استان (isfahan_foods.json، shiraz_foods.json و غیره) در یک فایل اصلی به نام Local_Foods.json ترکیب شدند. اسکریپت تجمیع به سادگی هر فایل JSON ورودی را خوانده و محتویات آن (با فرض اینکه هر فایل حاوی لیستی از دستور پخت‌ها بود) را به یک لیست اصلی اضافه می‌کرد که سپس ذخیره می‌شد. این امر منجر به یک مجموعه داده اولیه با ۱۸۵ ورودی دستور پخت شد.
نمونه‌ای از ساختار یک ورودی در فایل تجمیع‌شده اولیه در ادامه نشان داده شده است:

```
}
  "title": "بندری سوسیس",
  "location": }
   "province": "خورستان",
   "city": "اهواز",
   "coordinates": }
    "latitude": 318327.31,
    "longitude": 67062.48
   {
  {,
  "ingredients": ][,
  "instructions": ]
   "طرز در که است خوزستان به‌خصوص ایران، جنوبی مناطق غذاهای معروف‌ترین از بندری سوسیس
    ↪ این اولیه مواد که آنجا از. می‌شود استفاده گوجه‌فرنگی رب و ادویه پیاز، سوسیس، از آن تهیه
    ↪ و نیمه‌آماده غذای یک می‌شود، سرو ساندویچی نان با معمولا و است دسترس در و کم غذا
    ↪ .دارید غذا پخت برای کمی وقت که است زمان‌هایی مناسب و می‌آید حساب به فست‌فودی",
   "مخصوص بندری سوسیس تهیه طرز"
  [,
  "meal_type": ]
   "اصلی",
   "دسر",
   "غذا پیش"
  [,
  "occasion": ]
   "شام",
   "ناهار",
   "صبحانه"
  [,
  "images": }
   "نهایی تصویر": "https//:api2.kojaro.com/media2021/2-5440-f3474-e6546-ec-bc19-
    ↪ ddce96cbf82667-c46109c1067c5ba76732f9?w&1920=q80="
  {
```

```
{,

}
 "title": "ماسوا آش",
 "location": }
  "province": "خورستان",
  "city": "اهواز",
  "coordinates": }
   "latitude": 318327.31,
   "longitude": 67062.48
  {
 {,
 "ingredients": ][,
 "instructions": ]
  "منبع: عکس اینستاگرام @maria.saadat62",
  "آش ماسوا با ماست، گشنیز، (اسفناج، سبزی غلیظ، سابیده کشک پیاز، گندم، بلغور نخود، گوشت، با ماسوا آش
   ↪ خشک پونه یا داغ نعنا با معمولا و شده درست زردچوبه و قرمز فلفل نمک، روغن، ،)جعفری
   ↪ .", کرد استفاده می‌توان نیز سفید لوبیا مانند دیگری حبوبات از آش این طبخ در .می‌شود تزیین
  "آش ماسوا محلی در دزفول شهر در ماسوا محلی آش
   ↪ .است «آش به‌معنای اینجا در «با و است «با «ماست معنی به
  "ماسوا آش تهیه طرز"
 [,
 "meal_type": ]
  "اصلی",
  "دسر",
  "غذا پیش"
 [,
 "occasion": ]
  "شام",
  "ناهار",
  "صبحانه"
 [,
 "images": }
  "نهایی تصویر": "https//:api2.kojaro.com/media2021/3--c421dbd97-e534-b19-bb05261-
   ↪ a2d45b07667-c46109c1067c5ba7673ebe?w&1920=q80="
 {
{,
```


۳.۲ برچسب‌زنی داده‌ها

مجموعه داده تجمیع‌شده (Local_Foods.json) برای حاشیه‌نویسی به LabelStudio (https://labelstud.io/) وارد شد. فرآیند حاشیه‌نویسی بر افزودن اطلاعات معنایی و ارزیابی کیفیت داده‌ها متمرکز بود.

وظایف حاشیه‌نویسی:

۱. **طبقه‌بندی نوع وعده (meal_type):** تخصیص برچسب‌هایی مانند 'پیش غذا'، 'غذای اصلی' یا 'دسر'.

۲. **طبقه‌بندی مناسبت (occasion):** تخصیص برچسب‌هایی که مناسبت‌های مناسب را نشان می‌دهند، مانند 'صبحانه'، 'ناهار' یا 'شام'. انتخاب چند گزینه مجاز بود.

۳. **ارزیابی کیفیت (quality):** تخصیص برچسب 'خوب' یا 'بد' بر اساس کیفیت و کامل بودن درک‌شده

دستور پخت، به‌ویژه دستورالعمل‌ها و لیست مواد لازم. عواملی مانند وضوح، جریان منطقی و کامل بودن ظاهری مراحل در نظر گرفته شدند.

سیاست حاشیه‌نویسی: برچسب‌ها برای meal_type و occasion بر اساس مواد لازم دستور پخت، روش تهیه، تصاویر مرتبط (در صورت وجود) و دانش عمومی آشپزی ایرانی تخصیص داده شدند. برچسب quality ذهنی بود اما هدف آن شناسایی دستور پخت‌هایی با دستورالعمل‌های نامشخص، ناقص یا بی‌معنی بود.
در خصوص استان هایی که اطلاعات کمی از آنها در دسترس بود(مانند استان مهگیلویه و بویراحمد) این سیاست ها با سختگیری کمتری همراه بودند. **توافق بین حاشیه‌نویس‌ها:** حاشیه‌نویسی توسط یک نفر انجام شد. بنابراین، محاسبه معیارهای استاندارد توافق بین حاشیه‌نویس‌ها مانند کاپای کوهن یا کاپای فلایس، که سازگاری بین چندین ارزیاب را اندازه‌گیری می‌کنند، در این مرحله امکان‌پذیر نبود. قابلیت اطمینان برچسب‌ها به سازگاری و پایبندی حاشیه‌نویس واحد به سیاست حاشیه‌نویسی بستگی دارد.
خروجی از LabelStudio (LabelStudio_Output.json) شامل داده‌های اصلی به همراه حاشیه‌نویسی‌های جدید و فراداده‌های مربوط به فرآیند حاشیه‌نویسی (شناسه حاشیه‌نویس، مُهرهای زمانی و غیره) بود. نمونه‌ای از چنین ورودی قبل از پاک‌سازی در زیر نشان داده شده است:


```
}
  "title": "بندری سوسیس",
  "location": }
   "province": "خورستان",
   "city": "اهواز",
   "coordinates": }
    "latitude": 318327.31,
    "longitude": 67062.48
   {
  {,
  "ingredients": ][,
  "instructions": ]
   "طرز در که است خوزستان به‌خصوص ایران، جنوبی مناطق غذاهای معروف‌ترین از بندری سوسیس
    ↪ این اولیه مواد که آنجا از .می‌شود استفاده گوجه‌فرنگی رب و ادویه پیاز، سوسیس، از آن تهیه
    ↪ و نیمه‌آماده غذای یک می‌شود، سرو ساندویچی نان با معمولا و است دسترس در و کم غذا
    ↪ .دارید غذا پخت برای کمی وقت که است زمان‌هایی مناسب و می‌آید حساب به فست‌فودی ",
   "مخصوص بندری سوسیس تهیه طرز"
  [,
  "meal_type": "اصلی غذای",
  "occasion": }
   "choices": ]
    "ناهار",
    "شام"
   [
  {,
  "images": }
   "نهایی تصویر": "https//:api2.kojaro.com/media2021/2-5440-f3474-e6546-ec-bc19-
    ↪ ddce96cbf82667-c46109c1067c5ba76732f9?w&1920=q80="
  {,
  "id": 80,
  "quality": "بد",
  "annotator": 1,
  "annotation_id": 80,
  "created_at": "202505-01-T18473482.44:05:Z",
  "updated_at": "202505-01-T21324598.15:59:Z",
  "lead_time": 091.12
```

```
{,
}
 "title": "ماسوا آش",
 "location": }
  "province": "خورستان",
  "city": "اهواز",
  "coordinates": }
   "latitude": 318327.31,
   "longitude": 67062.48
  {
 {,
 "ingredients": ][,
 "instructions": ]
  "ترە، گشنیز، (اسفناج، سبزی غلیظ، سابیدە کشک پیاز، گندم، بلغور نخود، گوشت، با ماسوا آش
  ↪ خشک پونە یا داغ نعنا با معمولا و شدە درست زردچوبە و قرمز فلفل نمک، روغن، )جعفری
  ↪ .",کرد استفادە می توان نیز سفید لوبیا مانند دیگری حبوبات از آش این طبخ در .می شود تزیین
  "محلی زبان در «» ماسوا .دارد زیادی طرفداران خوزستان، استان در دزفول شهر در ماسوا محلی آش
  ↪ ."است «» آش بەمعنای اینجا در «» با و است با «ماست معنی بە
 [,
 "meal_type": "غذا پیش",
 "occasion": "صبحانه",
 "images": }
  "نهایی تصویر": "https//:api2.kojaro.com/media2021/3--c421dbd97-e534-b19-bb05261-
  ↪ a2d45b07667-c46109c1067c5ba7673ebe?w&1920=q80="
 {,
 "id": 81,
 "quality": "بد",
 "annotator": 1,
 "annotation_id": 81,
 "created_at": "202505-01-T18349729.44:11:Z",
 "updated_at": "202505-01-T21878880.16:06:Z",
 "lead_time": 509.7
{,
```


۴.۲ پاک‌سازی و فیلتر کردن داده‌ها

مرحله نهایی شامل پاک‌سازی داده‌های حاشیه‌نویسی‌شده از LabelStudio بود. یک اسکریپت پایتون برای موارد زیر استفاده شد:

۱. **فیلتر بر اساس کیفیت:** پیمایش ورودی‌ها در LabelStudio_Output.json و حذف هر ورودی که فیلد quality آن 'بد' علامت‌گذاری شده بود.

۲. **حذف فراداده‌ها:** حذف فراداده‌های خاص حاشیه‌نویسی که توسط LabelStudio اضافه شده بودند (مانند id، quality، annotator، annotation_id، created_at، updated_at، lead_time) از ورودی‌های باقی‌مانده با کیفیت 'خوب'.

مجموعه داده پاک‌سازی و فیلتر شده حاصل، که فقط شامل دستور پخت‌های با کیفیت بالا با اطلاعات ضروری و

برچسب‌های اضافه‌شده meal_type و occasion بود، با نام Filered_Data.json ذخیره شد. این فرآیند اندازه مجموعه داده را از ۱۸۵ به ۱۷۰ ورودی کاهش داد.

۳ نتایج و تحلیل

مجموعه داده نهایی (Filered_Data.json) شامل ۱۷۰ ورودی دستور پخت است که کیفیت خوبی دارند.

۱.۳ مرور کلی مجموعه داده

مجموعه داده نهایی ساختار اصلی (عنوان، مکان، مواد لازم، دستورالعمل‌ها، تصاویر) را حفظ کرده و شامل فیلدهای حاشیه‌نویسی‌شده دستی meal_type و occasion است. نمونه‌ای از یک ورودی نهایی و پاک‌سازی‌شده در زیر نشان داده شده است:


```
}
  "title": "باقالی‌شوید‌کوفته",
  "location": }
   "province": "اصفهان",
   "city": "اصفهان",
   "coordinates": }
    "latitude": 6539.32,
    "longitude": 666.51
   {
  {,
  "ingredients": ]
   }
    "name": "ریحان‌و‌مرزه‌ترخون،‌شوید،‌گشنیز،‌جعفری،‌(تره،‌کوفته‌سبزی",
    "amount": 0.1,
    "unit": "گرم‌کیلو"
   {,
   }
    "name": "باقلا",
    "amount": 0.1,
    "unit": "گرم‌کیلو"
   {,
   }
    "name": "چربی‌بدون‌گوشت",
    "amount": 0.500,
    "unit": "گرم"
   {,
   }
    "name": "برنج",
    "amount": 0.2,
    "unit": "پیمانه"
   {,
   }
    "name": "جو‌یا‌جو‌بلغور",
    "amount": 0.2,
    "unit": "غذاخوری‌قاشق"
   {,
   }
```

```
      "name": "مرغ تخم",
      "amount": 0.1,
      "unit": "عدد"
    },
    {
      "name": "دارچین و زردچوبه و فلفل و نمک",
      "amount": "لازم مقدار",
      "unit": "لازم مقدار"
    {
  [,
  "instructions": ]
    "و کنید آبکش را آن درآمد پز نیم حالت به وقتی. بپزید آن داخل را باقالا و ریخته آب قابلمه در
     ↪ بگذارید کنار.",
    "شوند خرد کاملا تا ریخته میکسر در را جو و برنج سبزی، گوشت، با برنج پیمانه یک.",
    "بگیرید را باقالاها پوست.",
    "و بریزید را دارچین و زردچوبه فلفل، نمک، باقالا، برنج، بقیه شدند، ترکیب خوب مواد که زمانی
     ↪ دهید ورز دست با.",
    "شوند یکدست تا کنید مخلوط دوباره و بشکنید را مرغ تخم.",
    "بجوشد بگذارید و گذاشته ملایم حرارت روی روغن پیمانه نصف و آب پیمانه دو با را قابلمه یک.",
    "کنید گرد دست با و بردارید مواد از پرتقال یک اندازه به.",
    "بپزند تا بگذارید آن کف در را کوفته ها و کنید کم را شعله آمد جوش آب که آن از بعد."
  [,
  "meal_type": "غذای اصلی",
  "occasion": }
    "choices": ]
      "ناهار",
      "شام"
    [
  {,
  "images": }
    "نهایی تصویر": "https//:storage.jainjas.com/storage/blog/1/original1/
     ↪ c5152a1fb064c1b95aaa85648894194.jpg"
  {
{,


}
  "title": "شیرازی فالوده",
  "location": }
    "province": "فارس",
    "city": "شیراز",
    "coordinates": }
      "latitude": 5926.29,
      "longitude": 5836.52
    {
  {,
  "ingredients": ]
    }
      "name": "گندم نشاسته",
      "amount": 0.300,
      "unit": "گرم"
```

```
    {,
    }
      "name": "شکر",
      "amount": 0.1,
      "unit": "دسته‌دار فرانسوی لیوان"
    {,
    }
      "name": "آب",
      "amount": 0.4,
      "unit": "دسته‌دار فرانسوی لیوان"
    {,
    }
      "name": "گلاب",
      "amount": 0.4,
      "unit": "غذاخوری قاشق"
    {,
    }
      "name": "یاآبلیمو شیرازی ترش لیمو",
      "amount": 0.1,
      "unit": "عدد"
    {,
    }
      "name": "آلبالو شربت",
      "amount": "لازم مقدار",
      "unit": "لازم مقدار"
    {
  [,
  "instructions": ]
    "کنید درست را شیرازی فالوده شربت :اول مرحله",
    "کنید اضافه را گلاب :دوم مرحله",
    "بزنید هم را فالوده شربت :سوم مرحله",
    "بجوشانید را آب و نشاسته :چهارم مرحله",
    "دهید تفت را آب و نشاسته :پنجم مرحله",
    "کنید آماده را نشاسته دادن فرم وسایل :ششم مرحله",
    "دهید فرم را فالوده نشاسته :هفتم مرحله",
    "بریزید شربت در را نشاسته‌ها :هشتم مرحله",
    "کنید سرو را شیرازی فالوده :آخر مرحله"
  [,
  "meal_type": "دسر",
  "occasion": }
    "choices": ]
      "ناهار",
      "شام",
      "صبحانه"
    [
  {,
  "images": }
    "نهایی تصویر": "https//:blog.okcs.com/wp-content/uploadsفالوده./2024/07/jpg"
  {
{,
```

۲.۳ تحلیل کمی

یک اسکریپت برای انجام تحلیل کمی پایه روی مجموعه داده نهایی و پاک‌سازی‌شده (Filered_Data.json) استفاده شد. معیارهای کلیدی به‌دست‌آمده عبارتند از:

- **تعداد کل ورودی‌ها:** ۱۷۰
- **تعداد کل کلمات (فیلدهای مرتبط):** ۲۶,۵۲۹ کلمه. این شمارش شامل کلمات فیلدهایی مانند عنوان، مکان (استان، شهر)، مواد لازم (نام، واحد)، دستورالعمل‌ها، نوع وعده و انتخاب‌های مناسبت می‌شود.
- **میانگین طول دستورالعمل‌ها:** ۱۱۹/۴۶ کلمه به ازای هر دستور پخت. این میانگین بر اساس تمام ۱۷۰ دستور پخت در مجموعه داده نهایی محاسبه شد، زیرا همه آن‌ها از فیلتر کیفیت عبور کرده و حاوی دستورالعمل بودند.

این معیارها یک نمای کلی از اندازه و حجم محتوای متنی مجموعه داده ارائه می‌دهند. میانگین طول دستورالعمل‌ها نشان می‌دهد که دستور پخت‌ها به طور کلی سطح جزئیات معقولی دارند.

۳.۳ تحلیل کیفی

مرحله فیلتر کردن بر اساس حاشیه‌نویسی 'کیفیت'، قابلیت اطمینان مجموعه داده را با حذف ورودی‌هایی با دستورالعمل‌های ضعیف توصیف‌شده یا ناقص، به طور قابل توجهی بهبود بخشید. ۱۷۰ ورودی باقی‌مانده به طور کلی دستور پخت‌هایی با مراحل منسجم را نشان می‌دهند. با این حال، به دلیل ماهیت استخراج وب از منابع متنوع با سطوح مختلف جزئیات و سبک‌های نوشتاری، ممکن است ناسازگاری‌های بالقوه‌ای همچنان وجود داشته باشد. تجزیه مواد لازم، اگرچه کاربردی است، ممکن است منجر به برخی تغییرات در نمایش واحد یا طبقه‌بندی نادرست گاه‌به‌گاه مقدار در مقابل واحد، به‌ویژه برای اندازه‌های کمتر رایج، شده باشد. حاشیه‌نویسی دستی meal_type و occasion اطلاعات معنایی ارزشمندی را اضافه می‌کند که همیشه به طور صریح در متون منبع وجود ندارد.

۴ نتیجه‌گیری

این پروژه با موفقیت یک مجموعه داده از ۱۷۰ دستور پخت غذای سنتی ایرانی از هفت استان را ایجاد کرد. این فرآیند شامل توسعه اسکرپرهای وب هدفمند، تجمیع داده‌های جمع‌آوری‌شده، حاشیه‌نویسی آن برای نوع وعده، مناسبت و کیفیت با استفاده از LabelStudio و فیلتر کردن ورودی‌های با کیفیت پایین بود. مجموعه داده نهایی اطلاعات ساختاریافته‌ای شامل عنوان، مکان، مواد لازم با مقادیر/واحدهای تجزیه‌شده، دستورالعمل‌های گام‌به‌گام، تصاویر و برچسب‌های معنایی را فراهم می‌کند. تحلیل کمی نشان داد که پایه متنی قابل توجهی (بیش از ۲۶,۰۰۰ کلمه) و میانگین طول دستورالعمل تقریباً ۱۱۹ کلمه وجود دارد. در حالی که رویکرد تک‌حاشیه‌نویس ارزیابی قابلیت اطمینان برچسب را از طریق معیارهای استاندارد مانند کاپا محدود می‌کند، مرحله فیلتر کردن کیفیت، قابلیت استفاده مجموعه داده را افزایش می‌دهد. این مجموعه داده به عنوان یک منبع ارزشمند برای وظایف آتی پردازش زبان طبیعی مرتبط با تحلیل دستور پخت، تولید متن یا انفورماتیک فرهنگی متمرکز بر آشپزی ایرانی عمل می‌کند.