


با نام دوست

گزارش تمرین اول درس پردازش زبان طبیعی - گروه ۹








اعضای گروه: علیرضا حسین خانی، صالح شجاعی و محمدجواد سعیدی

لازم به ذکر است تمامی کدها و دیتاهای استخراج شده و همچنین دیتای قبل و بعد از اصلاح (طبق هیستوری گیت) در ریپازیتوری گیت‌هاب زیر موجود می‌باشد. در تصویر زیر نمایی از ساختار ریپازیتوری قابل مشاهده است.

https://github.com/mohammadjavadsaeidi/NLP_HW1

 **NLP_HW1** Public Watch 1

main 3 Branches 0 Tags t Add file Code

 mohammadjavadsaeidi	add statistical_report.	f44e9e3 · 6 hours ago	15 Commits
 crawler	codes and data has been added.		4 days ago
 llm	codes and data has been added.		4 days ago
 output	[Update]		12 hours ago
 statistical_report	add statistical_report.		6 hours ago
 README.md	Initial commit		4 days ago
 Report.pdf	add Report.pdf.		12 hours ago

پوشه کراولر مربوط به فایل‌های کراولر سایت‌های مختلف، پوشه llm مربوط به کد تعامل با llm در فرایند حل تمرین، پوشه output مربوط به فایل‌های خام اولیه هر استان و همچنین دیتای اگرگیت شده و نسخه‌های قبل و بعد اصلاح دیتا (طبق هیستوری گیت) و پوشه گزارش‌های آماری در تصویر قابل مشاهده می‌باشد.

بخش اول - استخراج دادگان

در این تمرین به سراغ استخراج دادگان مرتبط با غذاهای استان‌های گروه ۳ رفتیم که شامل ۶ استان گیلان، آذربایجان غربی و شرقی، اردبیل، کردستان و زنجان می‌شود.

پس از بررسی سایت‌ها و منابع موجود اقدام به خزش کردن داده‌های ۳ سایت <https://roostanet.ir>، <https://fa.wikibooks.org> و <https://ghazaland.com> نمودیم و برای هرسایت خزش‌گر مربوطه‌ش را نوشتیم.

علت انتخاب این سایت‌ها هم اطلاعات کامل و با جزئیاتشان بود تا بتوانیم مواد مورد و طرز تهیه غنی‌ای را خزش کرده‌باشیم.

در فایل‌های پیوست این گزارش ۳ فایل مربوط به خزش‌گر به نام‌های:

`crawler_roostanet.py`

`crawler_ghazaland.py`

`crawler_wiki_book.py`

مشاهده می‌کنید که شامل کدهای خزش‌گرهای این ۳ سایت است (فایل کرالرها در گیت‌هاب ذکر شده موجود است)

باتوجه به اینکه فرمت نمایش هرکدام از غذاها و جزئیاتی که بیان می‌کردند در هر یک از سایت‌های خزش‌شده متفاوت بود و ما نیز می‌خواستیم به فرمت خواسته شده در صورت تمرین غذاها را ذخیره‌سازی نیاز به یک عملیات transformation میان jsonهای ذخیره شده از هر سایت و قالب مدنظر داشتیم که برای انجام این کار از مدل‌های زبانی بزرگ کمک گرفتیم.

باتوجه به اینکه چند صد غذای خزش شده داشتیم و استفاده از ال‌های LLMهای بزرگ دشواری‌هایی داشت اقدام به استفاده از APIهای مدل‌های زبانی بزرگ کردیم.

با استفاده از <https://metisai.ir> که به امکان استفاده از APIهای مدل‌های زبانی بزرگ را می‌داد یک توکن خریداری کردیم و برای ساختاردهی داده‌های خام از آن استفاده کردیم که کد آن را در فایل:

`llm.py`

مشاهده می‌کنید که ابتدا از پکیج openai استفاده کرده و یک کلاینت نوشتیم و سپس با کمک پرامپتی که به LLM خواسته‌ها و محدودیت‌های ما را منتقل می‌کرد به LLMها درخواست زدیم و jsonهای خروجی را ذخیره کردیم که در بخش outputs می‌توانید به تفکیک هر استان خروجی‌های موردنظر را مشاهده کنید.

همچنین مواردی نظیر meal_type و occasion در سایت‌هایی که خزش کرده بودیم مشخص نشده بود و این فیلدها را نیز به کمک LLM و پرامپتی که به آن داده بودیم مشخص کردیم.

در پایان این بخش حدود ۳۱۳ غذای مربوطه استخراج و جمع‌آوری و تمیز شد که به تفکیک هر استان در زیر مشاهده می‌کنید:

گیلان: ۱۰۱ غذا

زنجان: ۲۲ غذا

کردستان: ۴۵ غذا

اردیبل: ۳۵ غذا

آذربایجان شرقی: ۸۰ غذا

آذربایجان غربی: ۳۰ غذا

فایل‌های خام مربوطه در پوشه‌ی output ریپازیتوری موجود هستند که طبق نام استان تفکیک شده‌اند. همچنین فایل تجمیعی نیز در همان پوشه موجود است. دقت شود که طبق هیستوری گیت فایل تجمیع شده می‌توان نسخه‌های قبل و بعد اصلاح را مشاهده کرد.

NLP_HW1 / output / aggregated_with_id.json

```

Code Blame 19942 lines (19942 loc) · 660 KB  Code

76 "مقداری نمک مزه دار کرده و به صورت کشت قهقی سرخ می‌کنیم"
77 حیویات و مایعی مواد را اضافه کرده و اجازه می‌دهیم آش جا بپزد"
78 را در ظرف مناسب ریخته و با نعنا داغ و زرد داغ ترکیب می‌کنیم"
79 ایده ۳: رول خوش کرده، سپس در هنگام استفاده از صافی رد می‌کنیم"
80 "بورگ زردالی، آلوجه و لواشک را ۴۴ ساعت خیس می‌کنیم"
81 },
82 "meal_type": [
83 "غذای اصلی"
84 ],
85 "occasion": [
86 "ناهار",
87 "شام"
88 ],
89 "images": {
90 "final_image": "https://encrypted-tbn0.gstatic.com"
91 }
92 },
93 {
94 "id": 1,
95 "title": "ساج کباب",
96 "location": {
97 "province": "اردبیل",
98 "city": "اردبیل",
99 "coordinates": {
100 "latitude": "38.2432",
101 "longitude": "48.2976"
102 }
103 },
104 "ingredients": [
105 {
106 "name": "گوشت گوسفند",

```

نمایی از فایل aggregated_with_id.json که عملاً محل قرارگیری دیتا قبل و بعد اصلاح (طبق هیستوری گیت) می‌باشد. برای سهولت برچسب‌زنی به هر رکورد غذا یک آیدی اختصاص داده شده است.

بخش دوم - برچسب‌گذاری دادگان

برای برچسب‌گذاری با توجه به محدودیت‌های لیبل استودیو برای انجام کار به صورت جمعی تحت وب و با مشورت تیمی تصمیم گرفتیم تا از گوگل شیت برای این کار استفاده کنیم. سیاست برچسب‌گذاری به این صورت بود که هر نفر یکی از سه نوع لیبل زیر را بسته به شرایط می‌توانست برای هر رکورد دیتا (هر غذا) ثبت کند:

- صحیح (Correct) به معنی اینکه داده تماماً سالم و طبق ساختار مدنظر می‌باشد
- نیاز به اصلاح (Needs Edit) به معنی اینکه داده ایراداتی دارد اما قابل اصلاح است
- غلط (Wrong) برای داده‌هایی که تکراری هستند (غذای تکراری) یا اینکه قابل اصلاح نیستند.

نکته قابل ذکر این است که تشخیص دیتای سالم و یا نیازمند اصلاح طبق قوانین زیر که داخل تیم تعریف شدند انجام گرفته است:

- عنوان غذا باید دقیقاً نام غذا باشد، نه جمله یا عبارت توصیفی.
- مواد اولیه (اینگریدینتس) باید کامل، دقیق و قابل مصرف باشند. در صورت ناقص بودن، نادرستی، یا استفاده از مواد غیربهداشتی یا نامتناسب، باید اصلاح شوند.
- بررسی تکراری بودن غذا باید بر اساس عنوان انجام شود و غذاهای تکراری حذف یا ادغام گردند.
- دستور پخت باید یک متن معنی‌دار، ساختارمند و دارای آغاز و پایان مشخص باشد. از ترکیب‌های بی‌معنی یا ناقص پرهیز شود. در صورت ادغام اشتباه دو دستور یا حذف بخشی از آن، لازم است اصلاح صورت گیرد.
- نوع وعده (Meal Type) در صورت خالی بودن، باید از میان چهار گزینه‌ی زیر انتخاب و درج شود:
 - ◆ غذای اصلی، غذای دریایی، دسر، پیش‌غذا
- مناسبت مصرف (Occasion) در صورت عدم وجود، باید یکی یا چند مورد از گزینه‌های زیر بر اساس نوع غذا انتخاب شود:
 - ◆ صبحانه، میان‌وعده، ناهار، شام
 - ◆ توجه شود که ناهار و شام معمولاً با هم در نظر گرفته می‌شوند.

طبق قوانین و ساختار تعریف شده برای برچسب زنی ابتدا دو نفر کار برچسب زنی را انجام دادند (علیرضا و جواد که در شیت‌های جداگانه در لینک زیر در دسترس می‌باشد) و سپس نفر سوم (صالح) به بررسی نهایی، رفع مغایرت برچسب‌ها و همچنین اصلاح داده‌ها در صورت نیاز پرداخت که همه‌ی موارد ذکر شده در گوگل شیت زیر در دسترس می‌باشد. **دقت شود که دیتای اصلاحی در گیت‌هاب روی فایل aggregated_with_id.json در پوشه output قابل مشاهده است.**

<https://docs.google.com/spreadsheets/d/1FJU9xQALP8bGr1k1Xpbm3JR9S141lgiEiqUG78nK8CQ/edit?usp=sharing>

در بخش زیر اطلاعاتی از وضعیت کلی برچسب‌های زده شده توسط نفرات قرار گرفته است.

- تعداد رکوردهای حذف شده: 51 مورد
- تعداد رکوردهای اصلاح شده: 207 مورد
- تعداد رکوردهای صحیح بدون نیاز به تغییر: 55 مورد
- همچنین تعداد 235 لیبل یکسان توسط دو نفر برچسب زننده وجود داشت
- نفر اول 59 مورد صحیح، 205 مورد نیازمند ویرایش و 48 مورد غلط
- نفر دوم 118 مورد صحیح، 146 مورد نیازمند ویرایش و 49 مورد غلط
- اشتراک بین نفر اول و دوم شامل 55 مورد صحیح، 142 مورد نیازمند ویرایش و 38 مورد غلط

سنجش دقت برچسب‌گذاری

برای سنجش دقت برچسب‌گذاری، ضریب کاپای کوهن محاسبه شد. ۳۱۲ نمونه وجود داشت و برچسب‌زن اول به‌ترتیب ۵۹ «Needs Edit»، ۲۰۵ «Correct»، و ۴۸ «Wrong» و برچسب‌زن دوم ۱۱۸ «Needs Edit»، ۱۴۶ «Correct»، و ۴۸ «Wrong» داده بود. مواردی که دو برچسب‌زن دقیقاً هم‌نظر بودند ۵۵ «Needs Edit»، ۱۴۲ «Correct»، و ۳۸ «Wrong» بود؛ بنابراین احتمال توافق مشاهده‌شده $P_o = 235/312 \approx 0.753$ شد. احتمال توافقی تصادفی نیز با حاصل‌ضرب فراوانی نسبی کلاس‌ها محاسبه و $P_e \approx 0.403$ به‌دست آمد. در نهایت با فرمول مربوط به کاپا مقدار $K \approx 0.59$ به‌دست آمد که طبق معیار Landis & Koch **نشان‌دهنده توافق متوسط رو به قوی است و نشان می‌دهد کیفیت برچسب‌گذاری قابل قبول است.**

فایل اطلاعات مربوط به گزارش‌های آماری نیز در مسیر اصلی ریپازیتوری گیت‌هاب موجود است که تصویر آن نیز در زیر آمده است

```
Code Blame 31 lines (30 loc)... Code 55% faster

1 Total number of records: 261
2 Total word count (including numbers): 46653
3
4 Average lengths per field:
5 - ingredients.amount:
6   • Average characters: 2.63
7   • Average words: 0.99
8 - ingredients.name:
9   • Average characters: 9.57
10  • Average words: 2.16
11 - ingredients.unit:
12  • Average characters: 6.61
13  • Average words: 1.48
14 - instructions:
15  • Average characters: 79.78
16  • Average words: 18.28
17 - location.city:
18  • Average characters: 6.86
19  • Average words: 1.16
20 - location.province:
21  • Average characters: 8.49
22  • Average words: 1.35
23 - meal_type:
24  • Average characters: 7.80
25  • Average words: 1.74
26 - occasion:
27  • Average characters: 4.33
28  • Average words: 1.02
29 - title:
30  • Average characters: 11.85
31  • Average words: 2.43
```