پیشنهاد دیتاست اصلی پروژه — فاز ۱

پروژه: ساخت گراف دانش از متن غیرساختاریافته **دیتاست پیشنهادی:** **Re-DocRED** (نسخه بازبینی‌شده و اصلاح‌شده‌ی DocRED) **زبان:** انگلیسی · **لایسنس:** MIT (استفاده و انتشار آزاد) **منبع رسمی:** https://github.com/tonytan48/Re-DocRED **مقاله مرجع دیتاست پایه:** Yao et al., *DocRED: A Large-Scale Document-Level Relation Extraction Dataset*, ACL 2019 — https://aclanthology.org/P19-1074 / **مقاله نسخه اصلاح‌شده:** Tan et al., *Revisiting DocRED — Addressing the False Negative Problem*, EMNLP 2022

---

۱. چرا این دیتاست

مسئله‌ی پروژه، **استخراج رابطه در سطح سند** است: ورودی یک متن چندجمله‌ای است و خروجی مجموعه‌ای از سه‌تایی‌های (موجودیت، رابطه، موجودیت) همراه با شاهد متنی.

Re-DocRED دقیقاً همین مسئله را با برچسب طلایی انسانی می‌سنجد:

- برچسب طلایی دارد، پس **دقت، بازخوانی و F1 واقعی** قابل محاسبه است. دیتاست شهرداری برچسب طلایی ندارد و فقط ارزیابی دستی می‌دهد.
- **بنچمارک استاندارد جهانی** است. نتیجه‌ی ما با نتایج منتشرشده در ACL و EMNLP قابل مقایسه است.
- هر سه‌تایی **شماره‌ی جمله‌ی شاهد** را ذخیره می‌کند. این دقیقاً با طراحی فعلی سامانه‌ی ما (ذخیره‌ی evidence برای هر یال) هم‌راستا است.
- لایسنس MIT است و بدون درخواست مجوز قابل دانلود و ارسال است.

---

۲. جدول ۱ — ابعاد کلی دیتاست (اعداد از روی فایل‌های واقعی محاسبه شد)

بخش	تعداد سند	تعداد جمله	تعداد توکن	تعداد موجودیت	تعداد ذکر (mention)	تعداد سه‌تایی
Train	۳,۰۵۳	۲۴,۲۵۶	۶۰۳,۴۶۸	۵۹,۳۵۹	۷۹,۴۸۱	۸۵,۹۳۲
Dev	۵۰۰	۴,۱۱۰	۱۰۱,۹۷۰	۹,۶۸۴	۱۳,۱۸۹	۱۷,۲۸۴
Test	۵۰۰	۳,۹۶۶	۹۸,۷۳۴	۹,۷۷۹	۱۳,۰۱۸	۱۷,۴۴۸
جمع	**۴,۰۵۳**	**۳۲,۳۳۲**	**۸۰۴,۱۷۲**	**۷۸,۸۲۲**	**۱۰۵,۶۸۸**	**۱۲۰,۶۶۴**

شاخص	مقدار
تعداد نوع رابطه	۹۶
تعداد نوع موجودیت	۶
میانگین جمله در هر سند	۸٫۰
میانگین توکن در هر سند	۱۹۸٫۴
میانگین موجودیت در هر سند	۱۹٫۴
میانگین سه‌تایی در هر سند	۲۹٫۸
حجم خام فایل‌ها	۲۴ مگابایت (JSON)

---

۳. جدول ۲ — توزیع نوع موجودیت

نوع موجودیت	معنی	تعداد	سهم
LOC	مکان	۲۳,۱۷۴	۲۹٫۴٪
TIME	زمان	۱۵,۱۴۰	۱۹٫۲٪
MISC	سایر	۱۲,۴۵۹	۱۵٫۸٪
PER	شخص	۱۲,۲۷۷	۱۵٫۶٪
ORG	سازمان	۱۰,۷۳۵	۱۳٫۶٪
NUM	عدد	۵,۰۳۷	۶٫۴٪
جمع		**۷۸,۸۲۲**	**۱۰۰٪**

---

۴. جدول ۳ — ده رابطه‌ی پرتکرار (از ۹۶ رابطه)

شناسه‌ها، شناسه‌ی خصیصه‌ی Wikidata هستند.

رتبه	شناسه	معنی	تعداد
۱	P131	واقع در واحد تقسیمات کشوری	۲۸,۲۹۶
۲	P17	کشور	۱۹,۹۰۶
۳	P27	تابعیت	۶,۲۶۹
۴	P150	شامل واحد تقسیمات کشوری	۴,۷۶۴
۵	P800	اثر شاخص	۴,۲۵۷
۶	P527	دارای جزء	۳,۳۰۷
۷	P361	جزئی از	۳,۰۱۲
۸	P175	اجراکننده	۲,۳۸۶
۹	P577	تاریخ انتشار	۲,۲۵۸
۱۰	P463	عضو	۱,۸۷۷

---

۵. جدول ۴ — ویژگی‌های هر رکورد (Schema)

فیلد	سطح	نوع	توضیح
`title`	سند	متن	عنوان سند
`sents`	سند	آرایه‌ی آرایه	جمله‌های سند، توکن‌شده
`vertexSet`	سند	آرایه	فهرست موجودیت‌ها
`vertexSet[i][j].name`	ذکر	متن	شکل ظاهری موجودیت در متن
`vertexSet[i][j].type`	ذکر	برچسب	یکی از ۶ نوع موجودیت
`vertexSet[i][j].sent_id`	ذکر	عدد	شماره‌ی جمله
`vertexSet[i][j].pos`	ذکر	بازه	موقعیت توکنی شروع و پایان
`labels[k].h`	رابطه	اندیس	موجودیت مبدأ
`labels[k].t`	رابطه	اندیس	موجودیت مقصد
`labels[k].r`	رابطه	برچسب	یکی از ۹۶ رابطه
`labels[k].evidence`	رابطه	آرایه	شماره‌ی جمله‌های شاهد

تعداد ویژگی متمایز: ۱۱ (۳ ویژگی در سطح سند، ۴ در سطح ذکر موجودیت، ۴ در سطح رابطه).

---

۶. چرا این دیتاست به‌اندازه‌ی کافی بزرگ و چالشی است

۶.۱ استدلال چندجمله‌ای اجباری. ۶۴,۵۲۴ سه‌تایی، یعنی ۵۳٫۵٪ کل رابطه‌ها، بین دو موجودیتی است که هرگز در یک جمله با هم ظاهر نمی‌شوند. مدل باید از چند جمله و از زنجیره‌ی هم‌مرجعی عبور کند. یک مدل جمله‌محور روی نیمی از داده صفر می‌گیرد.

۶.۲ نامتوازنی شدید کلاس. تعداد کل جفت‌موجودیت‌های نامزد ۱,۵۸۴,۹۹۴ است، اما فقط ۱۲۰,۶۶۴ جفت رابطه دارند. **نرخ مثبت ۷٫۶۱٪** است. یعنی مدل باید ۹۲٫۴٪ موارد را درست رد کند. این همان مسئله‌ی «توهم رابطه» است که در پروژه‌ی شهرداری هم با اعتبارسنجی سخت‌گیرانه کنترل کردیم.

۶.۳ فضای برچسب بزرگ و دنباله‌دار. ۹۶ نوع رابطه وجود دارد، اما ۵ رابطه‌ی نخست ۵۲٫۶٪ کل نمونه‌ها را می‌گیرند و ۷ رابطه کمتر از ۱۰۰ نمونه دارند. مدل باید روی کلاس‌های کم‌نمونه هم کار کند.

۶.۴ چندبرچسبی بودن. ۲۴,۶۳۶ جفت‌موجودیت بیش از یک رابطه‌ی هم‌زمان دارند. پس مسئله دسته‌بندی تک‌برچسبی نیست و ارزیابی باید چندبرچسبی باشد.

۶.۵ هم‌مرجعی و حل موجودیت. ۱۰۵,۶۸۸ ذکر به ۷۸,۸۲۲ موجودیت نگاشت می‌شود و ۱۸٫۶٪ موجودیت‌ها بیش از یک ذکر دارند. نام‌های «Willi Schneider» و «Schneider» باید به یک گره برسند. این همان مشکلی است که در دیتاست شهرداری برای ردیف‌های بودجه دیدیم و با کلید یکتاسازی حل کردیم — اینجا در مقیاس ۷۹ هزار موجودیت.

۶.۶ ارزیابی سخت‌گیرانه و قابل مقایسه. دیتاست دو معیار دارد: F1 و **Ign F1** (که سه‌تایی‌های دیده‌شده در آموزش را حذف می‌کند تا حفظ‌کردن جواب امتیاز نگیرد). این معیار جلوی تقلب مدل را می‌گیرد.

۶.۷ مقیاس محاسباتی. ۸۰۴,۱۷۲ توکن و ۱٫۵۸ میلیون جفت نامزد یعنی خط لوله باید دسته‌ای، قابل‌ازسرگیری و بدون تکرار کار باشد. معماری فعلی ما (کش در سطح chunk و بارگذاری idempotent در Neo4j) دقیقاً برای همین ساخته شده است.

---

۷. مقایسه با دیتاست شهرداری (که به‌عنوان مطالعه موردی می‌ماند)

شاخص	دیتاست شهرداری	Re-DocRED	نسبت
تعداد سند	۷	۴,۰۵۳	۵۷۹ برابر
تعداد توکن	۸۱۱	۸۰۴,۱۷۲	۹۹۲ برابر
تعداد موجودیت	۳۸	۷۸,۸۲۲	۲,۰۷۴ برابر
تعداد سه‌تایی	۳۹	۱۲۰,۶۶۴	۳,۰۹۴ برابر
تعداد نوع رابطه	۵	۹۶	۱۹ برابر
برچسب طلایی	ندارد	دارد	—
رابطه‌ی بین‌جمله‌ای	کم	۵۳٫۵٪	—

---

۸. گزینه‌های فارسی که بررسی شد

دیتاست	اندازه	چرا انتخاب نشد
PERLEX	۱۰,۶۹۰ جمله، ۹ رابطه	سطح جمله است، نه سطح سند. استدلال چندجمله‌ای ندارد. اندازه‌ی برچسب کوچک است.
FarsBase-KBP	۲۲,۰۱۵ سه‌تایی فارسی	فایل داده به‌صورت عمومی و آزاد در دسترس نیست و نیاز به مکاتبه با نویسندگان دارد.
ARMAN / PEYMA	فقط NER	رابطه ندارد، پس برای گراف دانش ناقص است.

اگر استاد محترم اصرار بر دیتاست فارسی داشته باشند، **FarsBase-KBP** گزینه‌ی دوم است و برای دریافت آن با نویسندگان مکاتبه می‌کنیم.

---

۹. برنامه‌ی اجرا مطابق نظر استاد

فاز	کار	خروجی
فاز ۱ (انجام شد)	انتخاب و دریافت دیتاست بزرگ و چالشی	همین سند + فایل داده
فاز ۲	نگاشت هستان‌شناسی Re-DocRED به خط لوله‌ی فعلی و اجرا روی زیرمجموعه‌ی Dev/Test	سه‌تایی‌های استخراج‌شده
فاز ۳	ارزیابی کمی با F1 و Ign F1 و مقایسه با نتایج منتشرشده	جدول نتایج
فاز ۴	فصل آخر: اجرای همان مدل روی دیتاست شهرداری به‌عنوان مطالعه‌ی موردی ایرانی	گزارش مطالعه موردی

---

۱۰. فایل‌های پیوست

فایل	حجم	محتوا
`dataset_re_docred.zip`	۴٫۷ مگابایت	کل دیتاست: `train_revised.json`، `dev_revised.json`، `test_revised.json` به‌همراه لایسنس
`re_docred_sample.json`	۳۷ کیلوبایت	دو سند نمونه با قالب خوانا، برای بررسی سریع ساختار
`DATASET_PROPOSAL.md`	—	همین سند