سند ۴٫۹: مستندات مجموعه داده (Data Sheet + Dictionary) — ShahnamehMap

نسخه: ۱٫۰

تاریخ: ۰۵/۰۹/۱۴۰۳

مدیر ارشد محتوا (CCO) و مدیر فنی (CTO) تهیه‌کننده:

وضعیت: فعال (محصول/ML برای تیم)

---

۱۰. مقدمه: نقش داده و ML در ShahnamehMap

به صورت (ML) از یادگیری ماشین ShahnamehMapدر حال حاضر، محدود و متمرکز استفاده می‌کند. تمرکز اصلی بر ساخت پایگاه داده و استفاده از قوانین (Knowledge Graph) دانش ساختاریافته است. با این حال، ما از روز (Rule-based Systems) مبتنی بر دانش مانند) ML اول زیرساخت جمع‌آوری داده را برای کاربردهای آینده توصیه‌گر محتوا، تشخیص محتوای نامناسب، بالانس خودکار بازی) طراحی کرده‌ایم.

این سند **سه مجموعه داده اصلی** را مستند می‌کند: **۱) داده هسته ساختاریافته شاهنامه، ۲) داده رفتاری کاربران، و ۳) داده محتوای کاربرساخته (UGC)**.

---

۲. Data Sheet: مجموعه داده هسته شاهنامه (Shahnameh Core Knowledge Graph)

۲,۱. انگیزه (Motivation)

* **منبع حقیقت** ایجاد یک **(Single Source of Truth)**، معتبر، ساختاریافته و قابل ماشین‌خوانی برای تمام موجودیت‌ها و روابط در شاهنامه فردوسی.

۲,۲. ترکیب داده (Composition)

* گراف** داده‌ها به صورت **(Graph) در Neo4j ذخیره می‌شوند.

* **نودها (Nodes/Entities):** شخصیت‌ها، مکان‌ها، موجودات افسانه‌ای، اشیاء، رویدادها.

* **یال‌ها (Edges/Relationships):** روابط بین نودها (مثلاً: `سفر_کرد_به`، `کشته_شد_توسط`، `فرزند_است_از`، `متولد_شده_در`).

* **ویژگی‌ها (Properties):** ویژگی‌های هر نود (مانند نام، نام پدر، لقب، جنس، وابستگی قومی).

۲,۳. **فرآیند جمع‌آوری (Collection Process)**

* منبع اولیه:** متن تصحیح‌شده شاهنامه (چاپ مسکو و خالقی مطلق) به عنوان **منبع مرجع**.

* منابع ثانویه:** کتب و مقالات دانشگاهی معتبر برای **تکمیل و تصحیح** روابط جغرافیایی و شجره‌نامه‌ای (مانند "جغرافیای شاهنامه" از احمدی).

* روش:** **استخراج دستی و نیمه‌خودکار** توسط **تیم متخصص محتوا** (با تحصیلات ادبیات فارسی) با استفاده از ابزارهای annotating و سپس تبدیل به گراف.

۲,۴. **پیش‌پردازش/تمیزکاری (Preprocessing/Cleaning)**

* **یکسان‌سازی نام‌ها:** همه variant (مثلاً "رستم"، "رستم زال") به یک نود canonical مپ می‌شوند.

* **حل ابهام (Disambiguation):** برای نام‌های مشترک (مثلاً چند "گردآفرید")، از زمینه (Context) و منابع برای ایجاد نودهای جداگانه استفاده شده است.

* **اعتبارسنجی:** هر رکورد توسط حداقل دو عضو تیم و نهایتاً توسط یک استاد مشاور خارجی بررسی و تأیید شده است.

۲,۵. توزیع (Distribution)**

* **دسترسی عمومی (Read-only):** یک نسخه محدود و فقط-خواندنی از گراف (حدود ۲۰۰ موجودیت کلیدی) از طریق API** عمومی** در آینده قابل دسترس خواهد بود.

* **دسترسی داخلی:** گراف کامل به عنوان سرویس داخلی برای محصول ما استفاده می‌شود.

۲,۶. نگهداشت (Maintenance)**

* **بروزرسانی:** دستی و بر اساس پروژه. فعلاً برنامه‌ای برای بروزرسانی خودکار نیست.

نسخه‌بندی:** هر تغییر اساسی، یک نسخه جدید از مجموعه داده ** *
.(`shahnameh-core-v1.2` مثلاً) ایجاد می‌کند

---

۳. Data Sheet: مجموعه داده رفتاری کاربران (User Behavioral Event Data)

۳,۱. انگیزه

* درک تعامل کاربران با محصول برای **بهینه‌سازی تجربه کاربری (UX)**، **شخصی‌سازی (Personalization)** و **تحلیل رشد (Growth Analytics)**.

۳,۲. ترکیب داده

* ذخیره داده‌ها به صورت **رویداد (Event)** در **ClickHouse** می‌شوند. هر رکورد یک رویداد است.
* **جدول: `user_events`**

۳,۳. فرآیند جمع‌آوری

* **منبع:** تعاملات کاربران با فرانت‌اند و بک‌اند پلتفرم.

* **روش:** ردیابی رویداد (Event Tracking) با **کدهای Instrumentation** در فرانت‌اند (با رعایت حریم خصوصی) و **لاگ‌های سرور** در بک‌اند.

۳,۴. پیش‌پردازش/تمیزکاری

* **ناشناس‌سازی (Anonymization):** `user_id` واقعی به یک `pseudo_user_id` پس از ۲۴ ساعت حذف می‌شود. **آدرس IP** یکطرفه تبدیل می‌شود.

* **اعتبارسنجی Schema:** رویدادها در برابر یک Schema (با Apache Avro یا **Protobuf**) ثبت‌شده اعتبارسنجی می‌شوند تا از انسجام داده اطمینان حاصل شود.

۳,۵. توزیع

* **دسترسی داخلی:** توسط تیم‌های محصول، تحلیل داده و ML مهندسی (Need-to-Know) با کنترل‌های دسترسی **نیاز-به-دانستن** آینده) استفاده می‌شود.

۳,۶. نگهداشت

* **حذف خودکار:** داده‌های رویداد پس از ۱۴ ماه به طور خودکار حذف می‌شوند (مطابق با خط‌مشی حریم خصوصی و GDPR).
* **نظارت بر Drift:** اسکریپت‌های مانیتورینگ، تغییرات ناگهانی در توزیع رویدادها را گزارش می‌کنند (مثلاً اگر ناگهان رویداد `character_created` ۵۰٪ کاهش یابد).

---

۴. Data Sheet: مجموعه داده محتوای کاربرساخته (UGC)

۴,۱. انگیزه

* **توسعه سیستم توصیه‌گر (Recommender)** برای کمپین‌های کاربری، **تشخیص محتوای نامناسب (Content Moderation)** و **تحلیل خلاقیت جامعه**.

۴,۲. ترکیب داده

* داده‌ها در **Neo4j** (ساختار کمپین) و **MinIO** (فایل‌های آپلودشده) ذخیره می‌شوند.

* ** جدول/ساختار:** شامل متادیتای کمپین (عنوان، توضیح، سازنده، برچسب‌ها) و گراف صحنه‌ها.

۴,۳.** فرآیند جمع‌آوری**

* ** منبع:** کاربران خالق که از ابزار Campaign Builder استفاده می‌کنند.

* ** رضایت و مالکیت (ToS):** کاربران با پذیرش **قوانین خدمات** و **خط‌مشی محتوا**، صراحتاً:

۱. تأیید می‌کنند که **مالک محتوای تولیدشده** هستند.

۲. مجوز غیرانحصاری، جهانی و رایگان** ShahnamehMap به **ذخیره‌سازی، نمایش، توزیع و ایجاد آثار مشتق (برای بهبود الگوریتم‌ها)** می‌دهند.

۳. موافقت می‌کنند که محتوای آن‌ها ممکن است برای **تمرین داخلی** مورد استفاده قرار گیرد ML مدل‌های.

۴,۴.** پیش‌پردازش/تمیزکاری**

* **سانیتایز کردن متن:** ورودی‌های متنی برای جلوگیری از XSS سانیتایز می‌شوند.

* **بررسی اولیه محتوای نامناسب:** یک **فیلتر کلمه‌کلیدی ساده (Keyword Filter)** برای مسدود کردن آشکارترین محتواهای توهین‌آمیز در لحظه آپلود اجرا می‌شود.

۴,۵. **توزیع**

* **عمومی:** کمپین‌های منتشرشده توسط کاربران برای همه کاربران قابل مشاهده است.

* **داخلی (برای ML):** مجموعه داده کامل UGC (شامل کمپین‌های منتشرنشده) ممکن است برای **تمرین مدل‌های داخلی تشخیص محتوا یا سیستم توصیه‌گر** استفاده شود.

۴,۶. **نگهداشت**

* **حذف به درخواست کاربر:** کاربران می‌توانند محتوای خود را حذف کنند. در این صورت، داده از مجموعه‌های آموزشی ML **حذف** خواهد شد.

حذف به دلیل تخلف:** محتوای حذف‌شده توسط مادها، در یک ** *
برای `violation` **مجموعه داده جداگانه "منتشرنشده"** با برچسب
.**آموزش مدل تشخیص تخلف** آینده نگهداری می‌شود

ویژگی‌های کلیدی - (Data Dictionary) فرهنگ داده .۵ ##

:از مجموعه داده هسته شاهنامه .۵,۱ ###

| **نام ویژگی** | **نوع داده** | **توضیح** | **مقدار نمونه** |
Missingness |
| --- | --- | --- | --- | --- |
| `entity_id` | String (UUID) | .شناسه یکتای موجودیت |
`char_7f3a...` | ۰٪ |
| `name` | String | ۰٪ | `"رستم"` | نام اصلی فارسی. |
| `type` | Enum | نوع موجودیت: `CHARACTER`, `LOCATION`,
`CREATURE`, `OBJECT`, `EVENT`. | `CHARACTER` | ۰٪ |

| `description` | Text | `پهلوان بزرگ | توضیح متنی از منبع. سیستان، فرزند زال و رودابه...` | ~۱۰٪ (برای موجودات کم‌اهمیت)

| `canonical_source` | String | | منبع اصلی (شماره بیت). `"شاهنامه، جلد ۲، بیت ۱۲۳۴"` | ۵٪ (برای اطلاعات استنتاج‌شده)

| `gender` | Enum (`MALE`, `FEMALE`, `UNKNOWN`) | `MALE` | ~۳۰٪ (برای موجودات افسانه‌ای) | جنسیت.

۵,۲. داده رفتاری کاربران مجموعه از:

| **ویژگی‌های کلیدی (Properties)** | **نام رویداد** | **توضیح**

| --- | --- | :---: |

| `game_session_started` | `session_id`, `campaign_id`, `pseudo_user_id`, `character_id` | شروع یک جلسه بازی.

| `player_action` | `session_id`, `action_type` (`ATTACK`, `DIALOG_CHOICE`, `MOVE`), `target_id`, `outcome` | هر عمل بازیکن در بازی.

| `campaign_published` | `campaign_id`, `pseudo_user_id`, `category`, `word_count` | انتشار یک کمپین کاربرساخته. |

| `payment_succeeded` | `pseudo_user_id`, `amount`, `plan_type` | **پرداخت موفق کاربر. **(حساس)

۵,۳. از مجموعه داده UGC:

نام ویژگی**	**نوع**	**توضیح**	**مسائل کیفیت (Quality Concerns)**
`campaign_narrative_text`	Text	متن اصلی داستان/دیالوگ نوشته‌شده توسط کاربر.	**Bias:** بازتاب دیدگاه فرهنگی/سیاسی کاربر.
`creator_style_embedding`	Vector (Float[])	امبدینگ از سبک نوشتاری کاربر (آینده).	نیاز به مدل برای تولید.
`user_rating_avg`	Float	میانگین امتیاز کاربران دیگر داستان‌های ناتمام.	**Incompleteness:**
`pseudo_user_id`	String	شناسه کاربر.	**Missingness** در ابتدا بالا. **Bias:** ممکن است تحت تأثیر دوستان سازنده باشد.

۶.** ملاحظات اخلاقی، سوگیری (Bias) و انحراف (Drift)**

۶,۱.** سوگیری‌های شناخته‌شده (Known Biases):**

* ** سوگیری جنسیتی در داده هسته:** شاهنامه جهان‌سازی مردانه دارد. تعداد شخصیت‌های مرد بسیار بیشتر از زن است. **هر مدلی که روی این داده آموزش ببیند، این سوگیری را به ارث می‌برد.**

* ** اقدام کاهش:** در کاربردهای آموزشی، این سوگیری به وضوح توضیح داده می‌شود. در کاربردهای بازی (مثلاً تولید NPC)، سعی می‌شود از شخصیت‌های زن موجود به صورت متعادل‌تر استفاده شود یا شخصیت‌های جدید با حفظ اصالت افزوده شوند.

* **سوگیری در داده UGC:** محتوای تولیدشده توسط کاربران اولیه (بازی‌های RPG) ممکن است تم‌ها و سبک خاصی داشته باشد که نماینده کل جامعه ایران نباشد.

* ** اقدام کاهش:** تشویق کاربران متنوع (معلمان، دانش‌آموزان) به خلق محتوا.

۶,۲. انحراف (Drift) احتمالی:**

* **Concept Drift**: اگر ویژگی‌های جدیدی به بازی اضافه شود، الگوهای رفتار کاربران تغییر می‌کند.

* **مانیتورینگ**: نظارت بر توزیع رویدادهای کلیدی (`action_type`) برای شناسایی تغییرات ناگهانی.

۶,۳. ملاحظات اخلاقی و رضایت:**

* **رضایت برای ML**: همانطور که در بخش ۴,۳ آمده، مجوز استفاده از ML در ToS گنجانده شده است. ما از "انتخاب عدم مشارکت (Opt-out)" برای استفاده از محتوای کاربر برای آموزش مدل‌های جمعی **نمی‌کنیم**، زیرا این امر اثربخشی مدل را مختل می‌کند. در عوض، کاربر می‌تواند **محتوای خود را به طور کامل حذف کند**.

* **حریم خصوصی**: همه داده‌های رفتاری ناشناس‌شده هستند. هیچ مدلی سعی در شناسایی مجدد (Re-identify) کاربران ندارد.

---

۷. مالکیت، مجوز و مسئولیت‌پذیری (Ownership, Licensing, Accountability)

مجموعه داده**	**مالک اصلی**	**مجوز استفاده داخلی**
ShahnamehMap	**مجوز انتشار عمومی**	
:---	:---	:---
متون کهن	هسته شاهنامه (Public Domain)**	اما
CC BY-NC-SA 4.0 استفاده در محصول.	مجوز کامل برای **ساختار و گردآوری ما** متعلق به شرکت است.	بخشی تحت مجوز با ذکر منبع) منتشر می‌شود.
حقوق	**کاربران** (به عنوان موضوع داده)، اما **حقوق	داده رفتاری**
مجوز کامل برای تحلیل و بهبود محصول.	**هرگز** منتشر نمی‌شود.	جمع‌آوری و تحلیل** متعلق به شرکت است (طبق خط‌مشی حریم خصوصی).
**در گزارش‌ها ذکر شود	**تجمعی و ناشناس**	ممکن است به صورت
UGC داده	**	کاربر خالق** (مالکیت معنوی). کاربر به شرکت **
مجوز برای نمایش، ذخیره‌سازی و استفاده داخلی.	**مجوز استفاده** داده است.	فقط همان‌طور که کاربر انتخاب کرده است (معمولاً ML در **منتشرشده**).

مسئول نهایی کیفیت، **(CCO) مسئول:** **مدیر ارشد محتوا**
یکپارچگی و استفاده اخلاقی از همه مجموعه داده‌ها است. **مدیر فنی
.مسئول امنیت و زیرساخت داده است **(CTO)

نتیجه: داده به عنوان دارایی کنترل‌شده ##

یک **دارایی ShahnamehMap این مستندات نشان می‌دهد که داده در
:استراتژیک اما کنترل‌شده** است. ما

.منبع و مالکیت** هر داده را می‌دانیم** *

آن را مستند (Bias, Missingness) **کیفیت و محدودیت** *
.کرده‌ایم

.حقوق کاربران** را در قلب فرآیند داده قرار داده‌ایم** *

.ایجاد کرده‌ایم **ML چارچوبی برای استفاده مسئولانه و آینده‌نگر از** *

این رویکرد از تبدیل داده به یک **بدهی حقوقی، اخلاقی یا فنی** در آینده
جلوگیری می‌کند و آن را به موتور قابل اعتماد نوآوری محصول تبدیل می‌کند.

داده‌های شفاف و خوش‌تعریف، سنگ بنای هر سیستم هوش مصنوعی ارزشمندی هستند.