جداسازی بخش‌های نام در اسامی بهم‌چسبیده

(تمام اسامی دنیا نوشته شده با حروف انگلیسی)

دانشجو :نیکی بیات

استاد راهنما :دکتر مسعود اسدپور

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

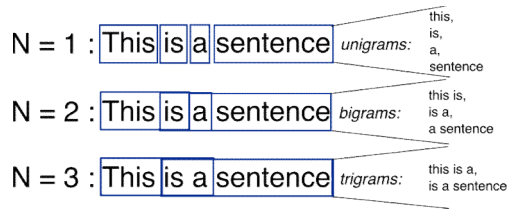
مقدمه

از جمله خطاهای رایج در نام، حذف فاصله میان بخش‌های سازنده همچون نام کوچک، نام میانی و یا نام خانوادگی می‌باشد. در این پژوهش هدف ما جدا کردن بخش های نام از جمله نام کوچک، نام میانی و نام خانوادگی در نام بهم چسبیده است. نام بهم‌چسبیده به نامی که فاصله میان بخش‌های مختلف آن حذف شده باشد گفته می‌شود.

شکل ۱- جداسازی نام بهم چسبیده

روش پیشنهادی

روش پیشنهادی برای جداسازی بخش‌های نام در این پژوهش استفاده از مدل‌های زبانی است.



شکل ۲- نمونه دسته بندی کلمات در مدل های ۱حرفی، ۲حرفی و ۳حرفی

مدل ۱حرفی:

در مدل زبانی ۱حرفی، احتمال توکن‌ها مستقل از یکدیگر محاسبه می‌شود .برای جداسازی توسط این مدل ابتدا تمامی نامزدهای جداسازی ممکن برای رشته ورودی را پیدا کرده و سپس احتمال رخداد هر تکه جداسازی را در متن مرجع یافته و در یکدیگر ضرب می‌کنیم تا امتیاز جداسازی به دست آید.

مدل ۲حرفی:

در مدل زبانی ۲حرفی، احتمال کلمات مشروط به کلمه پیشین محاسبه می‌شود .در این روش ترتیب رخداد بخش‌های نام حائز اهمیت است.

$$P(W_i|W_0 \dots W_{i-1}) \approx P(W_i|W_{i-1})$$

رابطه ۱- نحوه محاسبه احتمال توکن در مدل ۲حرفی

روش هموارسازی:

در ۱حرفی و ۲حرفی تولید شده، فرکانس بسیاری از توکن‌ها صفر گزارش می‌شود، چرا که در دادگان آموزش وجود نداشته‌اند. در این پژوهش از روش‌های هموارسازی افزایشی، روش معرفی شده توسط آقای peter norvig، بازگشت به عقب و تعامل برای اختصاص احتمال غیر صفر به این توکن ها، استفاده شده است.

$$probability = \frac{۱۰}{N \times ۱۰^{len(word)}}$$

رابطه ۲- روش هموارسازی peter norvig

تولید مجموعه داده:

دادگان این پژوهش از نام‌های موجود در مجموعه داده‌های متعدد که به صورت رایگان در اینترنت در اختیار محققین قرار گرفته است، استخراج شده است. در مجموع ۱۲۲۷۱۵۱۷۱ نام از زبان‌های مختلف برای این پژوهش جمع‌آوری شده‌اند.

نتایج

با استفاده از شبکه اجتماعی آکادمیک ۶۱۸۱۹۰ نام تحت عنوان دادگان آزمون به منظور ارزیابی عملکرد الگوریتم جداسازی جمع‌آوری شده است.

جداسازی توسط مدل ۱حرفی:

در این آزمایش اسامی موجود در داده آزمون را با استفاده از مدل ۱حرفی جدا می‌کنیم. دقت جداسازی برای ۳ حالت بدون هموارسازی، با هموارسازی افزایشی و با هموارسازی معرفی‌شده توسط آقای peter norvig، مقایسه خواهد شد.

جدول ۱ -مقایسه دقت هموارسازی‌های مختلف در مدل ۱حرفی

روش هموارسازی	دقت جداسازی
بدون هموارسازی	٪۷۵
هموارسازی افزایشی	٪۷۷.۶
هموارسازی norvig	٪۷۹.۴۸

جداسازی توسط مدل ۲حرفی:

روش هموارسازی تعامل، همانطور که انتظار می‌رفت، عملکرد ضعیفی داشت. دقت هموارسازی بازگشت به عقب، ٪۷۲.۷ است. هموارسازی بازگشت به عقب برای مواردی که اندازه داده بزرگ باشد بهتر عمل می‌کند این درحالی است که روش‌های تعامل برای داده‌های کوچک بهینه هستند و این موضوع در نتایج آزمایش نیز کاملا مشهود است.

جمع بندی

- حذف فاصله میان بخش‌های مختلف نام یک خطای رایج است .در این پژوهش هدف ما جداسازی بخش‌های نام در نام های بهم‌چسبیده بود که از جمله کاربردهای آن تطبیق موجودیت است.
- در این پژوهش، روش پیشنهادی برای جداسازی استفاده از مدل‌های زبانی ۱حرفی و ۲حرفی است.
- مجموعه داده آموزش و آزمون از دادگان دیتاست‌های معتبر جمع‌آوری شده و در مجموع بیش از ۱۲۰ میلیون نام برای به دست آوردن فرکانس توکن‌ها جمع‌آوری شده است.
- جداسازی توسط مدل ۱حرفی و با استفاده از روش هموارسازی آقای peter norvig با دقت ٪۷۹.۴۸ بهترین روش برای جداسازی بخش‌های نام در اسامی بهم‌چسبیده است.

مراجع اصلی

1. P. Norvig, “Natural Language Corpus Data: Beautiful Data” 2008-2009. [Online]. Available: https://norvig.com/ngrams/ [Accessed: JAN 28, 2019].
2. K. Wang, C. Thrasher, E. Viegas, X. Li, and B. Hsu. 2010. “An overview of Microsoft web n-gram corpus and applications”. In *Proc. NAACL/HLT-2010*, Los Angeles, CA.
3. K. Wang, C. Thrasher, and B. Hsu. “Web scale nlp: a case study on url word breaking”. in Proceedings of the 20th international conference on World wide web. 2011. ACM

شماره پوستر :۵۲