تحلیل متن فارسی
چالش‌ها و تجربه‌ها

دانشگاه اصفهان

گروه پژوهشی تجارت الکترونیکی شناختی

محمدمهدی بخشی

مهر ۱۴۰۲

پایان‌نامه‌ی من چه بود؟

- بخش‌بندی فروشندگان با استفاده از روند احساسات نظرات برخط در بازارهای الکترونیکی
  - تحلیل احساسات
    - پیش‌پردازش متن
    - مدل‌سازی تحلیلگر احساسات
  - خوشه‌بندی روندهای احساسات

دیتاست

- دیتا از کجا بیارم؟
  - جمع‌کردن دستی <- روش من

- https://nlpdataset.ir/
- https://github.com/MEgooneh/awesome-Iran-datasets
- http://dataheart.ir/
- https://github.com/persiannlp/parsinlu

پیش‌پردازش متن

- زدودن متون از عبارت‌ها، کلمه‌ها و سایر مواردی که باعث ایجاد کج‌فهمی در مدل تشخیص احساسات می‌شود یا تغییر آن‌ها به نحوی که متن مطلوب به دست آید.
- چه روشی برای پیش‌پردازش متن‌ها خوبه؟
- PERSIAN SENTIMENT ANALYSIS: FEATURE ENGINEERING, DATASETS, AND CHALLENGES (Asgarnezhad, Monadjemi)
- چه‌جوری بفهمم خوبن یا نه؟

پیش‌پردازش متن

- نرمالسازی و عملیات‌های مشابه مثل تبدیل متن محاوره‌ای به رسمی
- با چه ابزاری؟

- https://github.com/ICTRC/Parsivar
- https://www.roshan-ai.ir/hazm/docs/index.html
- https://text-mining.ir

- تبدیل متن محاوره‌ای به رسمی

پیش‌پردازش متن

- پاکسازی
- چه چیزهایی باید حذف بشن؟
  - شماره‌های تماس، آدرس پست الکترونیکی، پیوند به سایت‌ها، نماد ارزها، اعداد و شماره‌ها، علائم نگارشی به جز علامت تعجب، تگ‌های HTML و هشتگ‌ها
  - کلمات بی‌اثر -> از کجا پیدا کنیم؟
  - ایموجی‌ها -> کدوم ایموجی‌ها؟

پیش‌پردازش متن

- ریشه‌یابی
  - Stemmer
  - Lemmatizer

پیش‌پردازش متن

- نرمالسازی
- پاکسازی تگ‌ها، هشتگ‌ها ...
- کلمات بی‌اثر
- ریشه‌یابی

مدل‌سازی و آموزش تحلیلگر

- استخراج ویژگی‌ها -> کدوم روش؟
  - TF-IDF
  - Word embedding
- یادگیرنده‌ها
  - الگوریتم‌های یادگیری ماشینی مرسوم
  - الگوریتم‌های جدیدتر مثل XLNET، BERT، GPT و T5 -> نیاز به سرورهای گرافیکی
    - https://cloud.jarvislabs.ai/ همراه با شارژ رایگان در صورت عضویت در https://forums.fast.ai/
    - https://ferdowsi.cloud/ به همراه تخفیفات https://labsnet.ir/
    - https://lambdalabs.com/service/gpu-cloud

پایان

ممنون که مرا شنیدید!