به نام آنکه آموخت انسان را آنچه نمی دانست

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس پردازش زبان‌های طبیعی

CA1

m.ghoroobi@ut.ac.ir

محمد غروبی

بهمن ۱۳۹۸

فهرست

مقدمه

هدف از انجام این تمرین، آشنایی با مباحث ان‌گرام‌ها[1] و معیار سرگشتگی[2] در زبان فارسی است. انتظار می‌رود در طی این تمرین، با توجه به ان‌گرام‌ها، مدل‌های زبانی[3] مربوطه در زبان فارسی ایجاد نمایید، پیش‌بینی‌های خود را با استفاده از معیار سرگشتگی انجام دهید و در نهایت ارزیابی نتایج را با استفاده از معیار F۱ گزارش کنید.

نکته: هدف از تمرین، ایجاد تفکر علمی در شما می‌باشد، پس فارغ از نتایج، بخشی اعظمی از نمره به تفسیر نتایج تعلق می‌گیرد.

---

[1] N-grams

[2] Perplexity

[3] Language Models

دادگان

دادگان فارسی از اخبار روزنامه همشهری بین سال‌های ۱۳۷۵ تا ۱۳۸۶ با استفاده از خزشگرهای وب ایجاد شده است. با توجه به حجم بالای دادگان اصلی، نمونه‌ای از آن ایجاد و در اختیار شما قرار گرفته است.

تعداد ۲۳۸۱ متن خبری در ۶ کلاس (تکنولوژی، ورزشی، اجتماعی، سیاسی، اقتصادی و فرهنگی) در بخش دادگان آموزشی[۴] قرار گرفته است. هم چنین ۶۰۰ متن خبری بدون کلاس در بخش دادگان تست[۵] قرار گرفته است.

[۴] Train data

[۵] Test data

سوال ۱ – پیش پردازش (۱۰ نمره)

با توجه به کتابخانه‌های آماده در زبان فارسی و کتابخانه‌های آماده دیگر که در اختیار شما قرار گرفته، می‌توانید عملیاتی همچون نشانه‌گذاری، ریشه‌یابی و ... را انجام دهید.

دقت کنید که پیش‌پردازش یکی از مهمترین مراحل پروژه‌های پردازش زبان طبیعی است که کیفیت آن بر روی نتیجه، تاثیر مستقیم دارد. همینطور به پیش پردازشهایی که خاص زبان فارسی است دقت کنید.

راهنمایی:

منظور از نشانه‌گذاری مشخص‌نمودن ابتدا و انتهای تمامی جملات است.

علی به مدرسه رفت. بابک آمد.

<s>علی به مدرسه رفت</s> <s>بابک آمد</s>

سوال ۲- ایجاد مدل‌های زبانی (۳۰ نمره)

ابتدا دادگان آموزشی را به دو بخشی آموزشی (٪۸۰) و اعتبارسنجی[۶] (٪۲۰) تقسیم کنید.

با استفاده از دادگان آموزشی و با استفاده از کتابخانه‌های موجود برای هر کلاس مدل‌های زبانی یکتایی[۷] و دوتایی[۸] برای ان‌گرام‌های کلمه و حرف ایجاد کنید.

راهنمایی:

مدل یکتایی کلمه: هر کلمه در متن را جداگانه در نظر بگیرید.

مدل دوتایی کلمه: دو کلمه پشت سر هم در متن را در نظر بگیرید.

مدل یکتایی حرف: هر حرف در متن را در نظر بگیرید. (به فاصله دقت شود)

مدل دوتایی حرف: هر دو حرف پشت سر هم در متن را در نظر بگیرید. (به فاصله دقت شود)

۴ مدل که هر یک برای۶ کلاس ساخته شده است، پس نهایتا ۲۴ مدل‌زبانی متفاوت ایجاد می‌گردد.

---

[۶] Validation

[۷] Unigram

[۸] Bigram

سوال ۳ – معیار سرگشتگی (۴۵ نمره)

الف)

در سوال ۲ مدلهای زبانی را ایجاد کردید. حال برای هریک از متن‌های خبری در دادگان اعتبارسنجی کلاس مورد نظر را با توجه به حداقل میزان سرگشتکی که از مدل زبانی مربوط بدست می‌آید تعیین کنید.

راهنمایی:

در مدل یکتایی کلمه، برای هر متن خبری، ۶ مدل زبانی دارید، پس ۶ مقدار سرگشتگی بدست می‌آورید، کمترین مقدار سرگشتگی متعلق به هر کلاسی باشد، کلاس متن خبری همان کلاس می‌باشد. حال برای هر متن خبری در داده‌های اعتبار سنجی یک کلاس حقیقی (از پیش تعیین شده) دارید و یک کلاس تخمینی که با توجه به معیار سرگشتگی بدست آورده‌اید، (پس براحتی به معیارهای دقت، صحت و درنتیجه **f۱** دسترسی دارید)

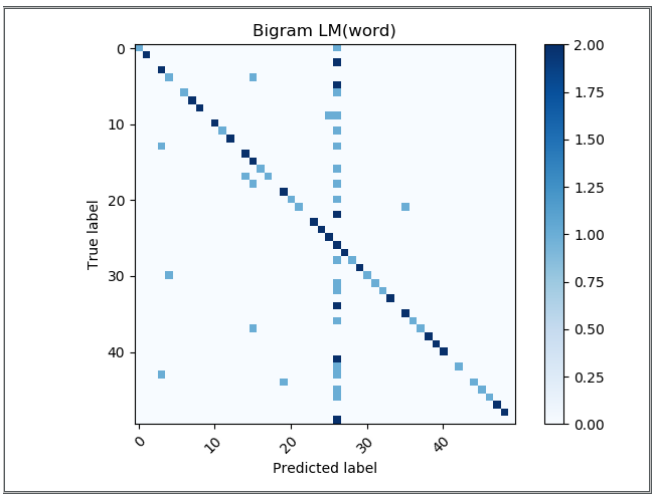
برای بدست آوردن سرگشتگی بر روی مدل‌های زبانی می‌توانید از کتابخانه‌های موجود کمک بگیرید.

ب)

در نهایت ۴ مدل زبانی داریم، که برای هر یک، یک ماتریس در هم‌ریختگی[۹] ایجاد کنید و برای هر نمودار تنها معیار میانگین f۱ تمامی کلاس‌ها را گزارش کنید.

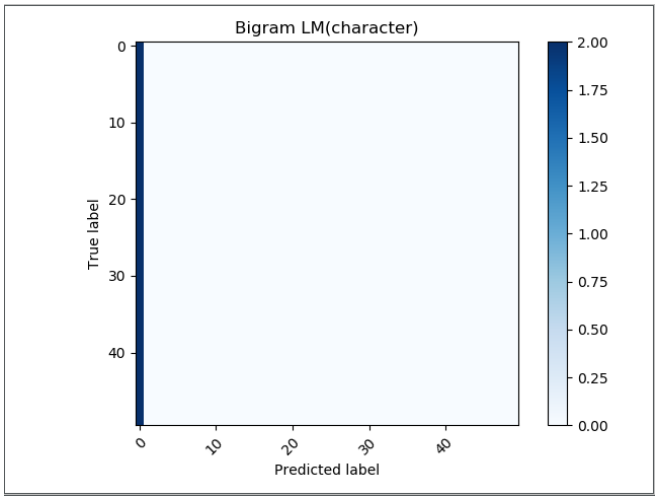
راهنمایی:

نمونه‌ای از ماتریس درهم ریختگی با معیار **f۱** بالا



[۹] Confusion Matrix

نمونه‌ای از ماتریس درهم ریختگی با معیار **۱f** پایین



سوال ۴- دادگان تست (۱۵ نمره)

جهت ارزیابی نتایج شما، دادگان تست که فاقد کلاس می‌باشند در اختیار شما قرار گرفته. با توجه به بهترین مدل‌زبانی که در سوال ۳ بدست آورید. لطفا برچسب هر متن به فرمت زیر در فایلی تحت عنوان Result.csv ارسال کنید.

Filename ,Class

۱۰۰۳۸.txt ,finance

به قواعد حل تمرین در سایت حتما توجه گردد.

موفق باشید.