محمدرضا قادری

9627057

زمستان 99

در این پروژه قصد داریم تا به وسیله پردازش زبان طبیعی متون فارسی را دسته بندی کنیم

در این پروژه از مدل های زبانی استفاده میکنیم که شامل unigram, bigram میباشد و برای هموار سازی این مدل ها از روش backoff Model استفاده میکنیم

در ابتدا برای هر یک از شعرا فایل های تمرینی داده شده که در پوشه train-set هست خوانده شده و واژه نامه مربوط به هر شاعر را اماده کردیم که در ان واژه گانی که کمتر از دوبار تکرار شده اند حذف شده با پردازش این خطوط توانستیم تعداد کلمات و زوج کلمات را استخراج کنیم (هر کدام را در فایل که نام شاعر و سپس نوع مدل زبانی نوشته شده)

در مرحله بعد با توجه به این که بعضی از کلمات تنها یک بار تکرار شده اند را UNK و کلمات ناآشنا میگیرم و در فایل اصلی به جای این کلمات کلمه NONE آورده شده تا تخمین بهتری زده شود.

از نکات قابل توجه میتوان به این اشاره کرد که برای بالا بردن دقت علامت Start را به ابتدای هر جمله اضافه کردم تا در زوج کلمات، کلمات اول جمله نیز حساب شود.

(با توجه به اینکه ضرب ها مقادیر کوچکی بودن بهتر بود از جمع log استفاده میکردیم)

برای هموار سازی از ترکیب خطی که در سوال امده بود استفاده میکنیم استفاده کردیم و دو مقدار مختلف برای landa و دو مقدار برای epsilon محاسبه شده است. در این قسمت اگر واژه در مدل ما نبود به اندازه epsilon احتمال به وجود آن میدهیم.

برای backoff

مقادیر یک بار (epsilon= 0.0001 ,landa3=0.002 ,landa2 = 0.008 ,landa1 = 0.99) که در این حالت بیشترن دقت را داشت (0.7694545454545455) حدودا 77 درصد

مقادیر بار دیگر (epsilon= 0.0001 ,landa3=0.10 ,landa2 = 0.20 ,landa1 = 0.70) که در این حالت دقت ان (0.7261818181818182) بود

با تغیر دادن epsilon اگر بیشتر شود ممکن هست اشتباهاتی رخ دهد و اگر کمتر از این هم کنیم تاثیر زیادی نخواهد داشت