به نام خدا

درس پردازش زبان طبیعی

تکلیف برنامه نویسی: مدل سازی آماری زبان

دانشگاه تهران

گروه علوم کامپیوتر

یک متن فارسی دلخواه در یک حوزه خاص (مثلاً ورزشی یا پزشکی) با حداقل 20000 کلمه تهیه کنید. از این متن مدل‌های زبانی unigram، bigram و trigram را استخراج کنید و با روش‌های مختلف هموارسازی (چه آنهایی که در این درس آموخته اید و چه سایر روش‌ها) عمل هموارسازی را انجام دهید. سپس سرگشتگی (perplexity) حاصل از هر یک از این مدل‌ها را برروی یک متن جدید از همان حوزه (با حداقل 2000 کلمه) (به ازای مدل‌های زبانی مختلف و روش‌های مختلف هموارسازی) به دست آورید و در قالب جدول و یا نمودار نمایش دهید. مراحل کار را شرح دهید و نتایج را تفسیر کنید.

توجه: استفاده از ابزارهای موجود مدل‌سازی زبانی (CMU SLM Toolkit, SRLLM Toolkit, IRSTLM Toolkit) مجاز است. فقط تمام مراحل کار (از نحوه تهیه داده تا روش استخراج مدل زبانی و محاسبه سرگشتگی و ...) را همراه با نتایج و تفسیر آنها بنویسید.

لطفا کد به همراه گزارش را در قالب یک فایل فشرده شده تا قبل از موعد اعلام شده به آدرس ایمیل ut.cs.exam@gmail.com ارسال نمایید.

Format : FirstName.LastName.HW1

EX: Bagher.BabaAli.HW1

با آرزوی سربلندی