تمرین: پاک‌سازی و نرمال‌سازی متن

نرمال‌سازی متن یکی از مراحل کلیدی در پردازش زبان طبیعی است که دقت و کارایی مدل‌های NLP را به شدت افزایش می‌دهد.

چرا نرمال‌سازی مهم است؟

- یکسان‌سازی نوشتار (مثلاً تبدیل "میشه" به "می‌شود")
- حذف نویزها مانند علائم نگارشی اضافه یا فاصله‌های بی‌جا
- کاهش ابعاد داده و افزایش سرعت پردازش
- بهبود عملکرد مدل در درک متن و ارائه نتایج دقیق‌تر

هدف این تمرین، آشنایی با پاک‌سازی و نرمال‌سازی داده‌های فارسی است.

بخش۱:

- ابتدا با کتابخانه‌های hazm و dadmatools که در رودمپ اشاره شد را مطالعه کنید.
- فایل persian_text1.docx را باز کنید و متن آن را استخراج کنید.
- ایموجی‌ها، ایمیل، آدرس سایت، علائم نگارشی و تگ‌های html را با کمک ابزارهای معرفی شده حذف کنید.

بخش۲:

فایل paragraghs.docx را باز کنید و متن آن را استخراج کنید.

پیش پردازش‌های لازم (حذف فضاهای خالی اضافه، تبدیل فاصله به نیم فاصله، ایجاد فاصله مناسب بعد از ویرگول یا نقطه، و ..) را انجام دهید.

برای هر پاراگراف جدول مجزا ساخته و اطلاعات زیر را در جدول قرار دهید:

- تعداد جملات
- تعداد کل کلمات
- تعداد فعل ها
- تعداد اسم ها

۵ جمله دلخواه انتخاب کرده و:

- همه توکن ها را پیدا کنید.
- ریشه فعل ها و ریشه کلمات را پیدا کنید.
- لیست کلماتی که به درستی در خروجی مراحل قبل ظاهر شده اند را پیدا کنید.

نکات مهم در حل تمرین

- خروجی دو فایل ipynb. داده شود.
- از کامنت گذاری مناسب، سلول بندی، و Markdown استفاده کنید. (بخشی از معیار ارزیابی شما کد نویسی تمیز و خوانایی کد هست.
- یک فایل گزارش متنی جامع و کامل به صورت pdf آماده کنید. مسئله را در آن کامل و مرحله به مرحله شرح دهید، راه‌حل خود را نتایج کسب شده، و کدها را با جزئیات کامل توضیح دهید.
- گزارش مرتب و مناسبی ارائه دهید.