به نام خدا

مهسا میمری

تمرین اول پردازش زبان طبیعی:

سوال 1:ابتدا از روی فایل csv داده شده دیکشنری ساخته شد به طوری که کلید نحوه نوشتاری ومقدار , لیست انواع روش های خواندن آن است
سپس جمله ورودی از روی فایل خوانده شده و با استفاده از دیکشنری که در بالا توضیح داده شد لیستی از تمامی روش های که میتوان آن جمله را(به طوری که تمام کلمات متصل شده اند) خواند ساخت شد
سپس با استفاده از یک تابع بازگشتی روش های متفاوتی که میتوان این فونتیک ها را جدا کرد و جمله ساخت به دست آمد .
این تابع بدین گونه عمل میکند که جمله واینکه از کجا به بعد آن جمله به دنبال کلمه باشد را می گیرد به اولین فونتیکی که در واقع یک لغت بود رسید خود را صدا میکند و بقیه جمله را به آن ورودی میدهد در نهایت فلگی وجود دارد که در آخرین بار که به دنبال کلمه هستیم اگر چیزی پیدا نکرد آخرین جمله که در حال ساخت را حذف میکند زیرا این جمله پایانی ندارد در نهایت خروجی این فایل بازگشتی لیستی از تمتم جملاتی که می توان ساخت است
نکته:در فایل خروجی بعضی جواب ها به نظر شبیه هم هستند ولی از نظر تلفظ این ونه نیست مثلا برای فایل ورودی آخر کلمه (در) به دو صورت "dor" یا "dar" می تواند خوانده شود پس جواب هایی ایجاد کرده اند که از نظر نوشتاری کاملا مشابه ولی از نظر گفتاری متفاوتند

سوال 2:

^(Doctor|Dr\.) | (M\.D\.)$

با رج اکس بالا نمونه هایی که با Doctor یا Dr یا با M.D. تمام می شوند را می یابیم.

سوال 3:اگر سایز نمونه اول m و سایز نمونه دوم n باشد یک جدول m+1 در n+1 در نظر میگیریم بدین صورت که خانه i و j ان یعنی تا حرف i ام از نمونه 1 و حرف j ام از نمونه 2 minimum distance چند است در ردیف اول (ستون اول) این جدول مقدار خانه ها شماره ردیف(شماره ستون) ان است زیرا از نمونه دیگر چیزی نداریم و فقط باید insert کنیم در باقی خانه ها دو حالت داریم یا دو حرف از دو نمونه متفاوت یکی هستند که در این صورت همانمقدار هزینه تا قبل را داریم و این حرف جدید هزینه ای ندارد یعنی مقدار خانه i-1 و j-1 ولی اگر دو حرف متفاوت بود کمینه 3 خانه بالا و خانه چپ و خانه قطری بالا چپ را گرفته و به اضافه یک میکنیم که به ترتیب هزینه replace, insert , remove است.

سوال 4: ابتدا از روی کل کلمات در فایل داده شده لیستی ساخته شد سپس فایل ورودی خوانده شده و سپس بر روی لیست کلمات انگلیسی for زده شد و اگر حرف اول ان ها یکی بود (شرط سوال)

min edit distance ان کلمه با ورودی محاسبه شد و به صورت دیکشنری که کل مه کلید ان و md مقدار ان بود ذخیره شد سپس دیکشنری قید شده بر اساس عدد md (مقدار) به صورت صعودی مرتب شد و 3 کلید اول این دیکشنری به عنون خروجی در فایل نوشته شد