پروژه پایانی طراحی الگوریتم

در این پروژه می‌خواهیم برنامه‌ای بنویسیم که قادر باشد یک متن از زبان انگلیسی را به زبان فارسی ترجمه نماید. این ترجمه به صورت کلمه‌به‌کلمه و بی‌توجه به ترتیب کلمات انجام خواهد شد. برای ترجمه می‌خواهیم از یک optimal BST استفاده نماییم. به این ترتیب که هر نود آن شامل یک زوج کلید و مقدار است که کلید، کلمه انگلیسی و مقدار، معادل فارسی آن است. الگوریتم ساخت BST بهینه در ویدئوها به صورت کامل توصیف شده است. از همان الگوریتم استفاده نمایید و پس از ساخت درخت چند متن کوتاه انگلیسی را به صورت اتوماتیک و با استفاده از درخت ترجمه کنید. برای محاسبه احتمالات و لیست کلیدها هم از فایل dictionary در پوشه پروژه استفاده کنید.

فایل dictionary و محاسبه احتمالات:

یک فایل متنی است حاوی 31805 کلمه انگلیسی که بر اساس تکرارشان و به صورت نزولی مرتب شده‌اند. هر خط از این فایل حاوی سه رشته‌ی به ترتیب کلمه انگلیسی، معادل فارسی و احتمال رخداد آن کلمه انگلیسی است که با یک فاصله از هم جدا شده‌اند مثلا:

downloaded بارگیری 8.40E-06

که نشان می‌دهد کلمه download با احتمال 8.40E-06 تکرار می‌شود و معادل فارسی آن "بارگیری" است. البته همین فایل با فرمت excel هم در پوشه پروژه قرار داده شده است.

طبیعتا تعداد کلمات ممکن در زبان انگلیسی بیشتر از 31805 عدد است و بنابراین بسیاری از کلماتی را که در یک متن وجود دارد، با درختی که برمبنای این فایل ساخته شود، نمی‌توانید ترجمه کنید. در چنین مواردی به جای کلماتی که درخت شما نمی‌داند، یک علامت خاص (مثلا ؟) برگردانید.

احتمال کلیدهای نماینده چنین کلماتی (q) را با تقسیم باقی احتمالات بر تعداد این کلیدها (n) بدست آورید یعنی:

$$\frac{\left(1-\sum_{i=1}^{31805} p_i\right)}{n}$$

که در آن، p_i احتمال رخداد کلمه موجود iام و n تعداد کلمات (کلیدهای) ناموجود است. تعداد کلیدهای ناموجود (برگ‌ها در درخت دودویی) چندتا است؟

نکاتی که در پیاده‌سازی این پروژه باید در نظر داشته باشید:

- ✓ برنامه شما باید BST بهینه را فقط یک بار بسازد و در موارد تست از همان درخت ساخته‌شده استفاده نماید.
- ✓ باید قادر باشد هر متنی را در زمان تست گرفته و ترجمه‌شده آن را تولید کند.
- ✓ باید بتواند زمان اجرای ساخت BST بهینه و ترجمه یک متن را محاسبه نماید (بر اساس واحد زمان)
- ✓ شما مجاز هستید با توجه به اندازه RAM خود، از بخشی از دیکشنری استفاده نمایید.

عوامل موثر در نمره پروژه(100 نمره):

- ✓ به درستی اجرا شدن برنامه. (وتویی: سایر نمرات در صورت برقراری این شرط، به پروژه مربوطه تعلق خواهد گرفت)
- ✓ خوانایی کد (20 نمره):
  - استفاده از comment، نامگذاری مناسب متغیرها، رعایت indent و فاصله‌گذاری‌های مناسب
- ✓ عملکرد درست برنامه (30 نمره)
- ✓ سرعت اجری مناسب (10 نمره)
- ✓ محاسبه زمان اجرا (10 نمره)
- ✓ ذخیره مناسب BST (20 نمره)
- ✓ محاسبه درست احتمالات (10 نمره)

نمره اضافه (تا 60 نمره):

- ✓ نمایش گرافیکی درخت بهینه (بخشی از درخت کافیست)
- ✓ استفاده از واسط گرافیکی برای تست برنامه
- ✓ استفاده از کل فایل dictionary به صورت بهینه