Sindokht

Compiler Project Report – Final Phase - Spring 2017

Benyamin Delshad - Mohammad Mahzoun

Under supervision of Dr. Sajedi

07.14.2017

COMPUTER SCIENCE DEPARTMENT

UNIVERSITY OF TEHRAN

TEHRAN, IRAN

مقدمه

پروژه‌ی درس کامپایلر در ترم جاری(پاییز ۹۵ و بهار ۹۶) عبارت است از ساختن یک قصه‌گو که بتواند داستان‌های نسبتا معقولی را تولید کند و بخواند. ما نام این پروژه خود را "سیندخت" نهادیم و دلیل این نام‌گذاری آن است که در فرهنگ پارسی زبان‌ها، عموما این مادربزرگ‌ها هستند قصه برایمان می‌گویند و سیندخت مادربزرگ رستم است و رستم بزرگ‌ترین پهلوان اساطیری ایرانیان.

در فاز نخست سیندخت، او باید بتواند بر حسب داستان هایی که در ورودی می‌گیرد، در خروجی داستانی را تولید نماید که در ادامه الگوریتم این کار، شرح داده می‌شود.

در فاز دوم، داستان‌هایی که سیندخت تولید می‌نماید تحت بررسی‌های واژگانی قرار می‌گیرد تا جملات نادرست را تا آن جا که می‌توانیم شناسایی کنیم و حذف کنیم تا داستان نهایی بهبود یابد.

در فاز نهایی، وقت آن است که سیندخت جانی دوباره گیرد. برای آن جسمی می‌سازیم و این‌بار او می‌تواند داستان تولیدی خود را بخواند تا هرچه بیشتر به آن چه هدف بود نزدیک شود.

در ادامه مفصل‌تر به هر فاز و چالش‌هایش پرداخته می‌شود. زبان مورد استفاده‌ی ما در سیندخت، پایتون ۲,۷ است.

ابزار ها و کتابخانه‌های به کار رفته

- HAZM 0.5

ازین کتابخانه برای نرمالایز کردن داستان های ورودی و برخی ویژگی‌های خوب دیگر استفاده میشود.

- Re

ازین کتابخانه برای ویرایش‌های ساده‌ای در انتها که متن نهایی ساخته می‌شود، استفاده می‌شود.

- Trello

برای پروژه‌های بزرگ و تیمی، نیاز به تجزیه کردن پروژه به تسک‌های کوچک‌تر بودیم و مدیریت تسک‌ها و اولیت‌بندی و ... که ازین رو از Trello بهره جستیم.

- Git

وقتی تقسیم وظایف صورت می‌گیرد، در هر لحظه پروژه در یک مرحله‌ای قرار دارد و هرتغییری توسط عضوی از تیم، لازم است با توضیح آن تغییر، به اطلاع عضو دیگر تیم برسد و هر عضو لازم است بتواند به راحتی به آخرین نسخه از پروژه دسترسی یابد و آن را ادامه دهد. لذا از git استفاده کردیم.

- Flex

برای به دست آوردن نقش کلمات ازین ابزار استفاده کردیم.

- Bison

با استفاده ازین ابزار، گرامر های متن آزادی (Context Free Grammars) نوشتیم که با استفاده از نقش‌ها، درستی هر زبان را بررسی نماید.

- برخی کتابخانه‌های کوچک که برای مواردی خاص در پروژه به کار آمد

فاز نخست: شرح الگوریتم تولید محتوا

روش کار ما، بسیار مشابه با آن چه در کلاس حل تمرین یا در فایل توضیح فاز نخست پروژه گفته شد، می‌باشد. در ابتدا، متن ورودی را از آدرس داده شده، استخراج می‌کنیم و در ابتدا آن را نرمالایز می‌کنیم. کار هایی از قبیل: حذف فاصله های تکراری اضافی، یکی کردن انواع مختلف حروفی از فارسی که کد های مختلفی در Unicode دارند مثل انواع 'ک' ها و 'ی' ها، بعضی فاصله ها که میان یک کلمه هستند را تبدیل کردن به نیم‌فاصله تا در جداسازی توکن‌ها از هم جدا نگردند مثل 'می باشد' که یک کلمه است ولی اگر با فاصله توکنایز کنیم به دو کلمه تبدیل می‌شود پس در عملیات نرمالایز به 'می‌باشد' تبدیل می‌گردد. و .... یکی از دلایل استفاده از کتابخانه HAZM نیز هندل‌کردن راحت این موارد بود. سپس به توکنایز کردن آن می‌پردازیم. این کار دیگر به راحتی با دلیمیتر فاصله قابل انجام است.

در مرحله بعدی، n–گرام ها را می‌سازیم و به جای احتمال‌ها از فرکانس یا همان تعداد تکرار استفاده می کنیم. این اعداد صحیح مثبت، بعد تر وزن های ما می‌شوند در رندوم وزن‌داری که می‌زنیم. در واقع به ازای هر n-1 توکن متوالی وزن هر توکن برای بعد این n-1 تا بودن را داریم و با رندوم وزن‌دار کلمه بعدی را می‌سازیم و به پیش می‌رویم. برای شروع هم یک n-1 تایی رندوم اول کار می‌گذاریم که قبل‌تر به همین صورت متوالی در رشته ورودی ظاهر شده است.

در نهایت نیز یک متن حداکثر N جمله ای می‌سازیم که مقدار دیفالت N ده است و معیار هر جمله، نقطه، علامت سوال یا علامت تعجب است.

در انتها نیز متن خروجی را کمی می‌پردازیم. فاصله های زاید آن را می‌زداییم و در فایل خروجی می‌ریزیم. فایل Prob.txt نیز مقادیر فرکانس های صحیح ما را نگه می‌دارد.

برای بهترشدن متن تولیدی، نیاز داریم تا روابط معنایی بین جملات متوالی را نیز در نظر بگیریم. خوشبختانه با این روش، این موضوع نیز تا حدی پوشش داده می‌شود زیرا هنگامی که جمله‌ی جدید در حال شروع شدن است، واپسین توکن‌ها که مربوط به جمله پیشین‌اند در شانس انتخاب شروع جمله جدید تاثیر می‌گذارند و در نهایت موجب تولید محتوای بهتری می‌شوند. البته این وابستگی به وابسین توکن‌ها به طور کنترل شده‌ای است که جملاتی عینا یکسان با متن اصلی تولید نگردد.

فاز دوم: تحلیل واژگانی و چالش‌ها

همان‌طور که پیشتر گفته شد، هدف این فاز، حذف کردن برخی جملات نادرست است که در نهایت منجر به تولید متن بهتری گردد. ابزار Flex این قابلیت را به ما داد تا بتوانید نقش کلمات را تعیین کنیم. منتها این ابزار، از حروف فارسی پشتیبانی نمی‌کرد و این چالش نخست بود. برای حل این چالش، با استفاده از یک دیتابیس مناسب، کلمات فارسی را با حروف انگلیسی(یا به اصطلاح عامیانه به صورت فینگلیش) به دست آوردیم. مشکل این کار، این بود که حرکت حروف در فارسی نمایان نیست ولی در انگلیسی با استفاده از حروف صدادار نمایان میگردد. در واقع این‌جا بود که دیتابیس مناسب به داد ما رسید و توانستیم حرکت کلمات را تخمین زده تا فینگلیش بهتری داشته باشیم.

در ادامه باید انواع جملات صحیح را در زبان فارسی در نظر میگرفتیم و برای آن گرامر می‌ساختیم تا ابزار Bison به ما کمک کند جملات نادرست را از دست تمییز دهیم.

به طور مثال در زبان فارسی داریم:

علی آمد. = علی + آمد + .

در مرحله Flex به "علی" برچسب فاعل زده می‌شود و به "آمد" برچسب فعل و نقطه نیز یک نماد است.

گرامری که این جمله را صحیح انگارد، بدین صورت است:

$$\text{جمله} \leftarrow (\text{فاعل})(\text{فعل})(\text{نماد نقطه})$$

به نظر میاید همچین گرامری همواره جملات درستی را بسازد. ۸۰ مدل گرامر ازین دست داریم تا بتوانیم جملات پیچیده‌تر را نیز شناسایی کنیم.

فازسوم(نهایی): جان بخشیدن به سیندخت ...

همان‌طور که از تیتر بر میایید، خود این فاز شامل زیربخش هایی میشود.

- Text to Speech

در این بخش، باید نوشتار تبدیل به گفتار میشد که انجام آن از نخست، کاری بسیار پیچیده و زمانبر است و نیاز به دانش‌های بیشتری دارد. لذا از یک سیستم آماده استفاده کردیم وتا جای ممکن آن را بهبود دادیم تا صدایی نرم و واضح تولید شود.

- ساختن بدنه فیزیکی

درین بخش سعی کردیم مدل زیبایی را بسازیم با استفاده از وسایل ساده ای که موجود است.

- دکمه

برای این که بتوانیم با استفاده از تنها یک دکمه، داستان خوانی شروع گردد، کار های سخت افزاری و نرم افزاری لازم بود که در این فاز به آن‌ها پرداختیم. با استفاده از Breadboard و دکمه ساختاری طراحی کردیم که با فشردن دکمه، در رزبری پای اکشنی واقع گردد و سپس برنامه را در آنجا اجرا کردیم تا شروع به تولید و خواندن داستان نماید.

درباره‌ی ما

ما یک تیم دو نفره برای انجام این پروژه هستیم و متشکل از: بنیامین دلشاد ممقانی و محمد محزون با شماره های دانشجویی ۶۱۰۳۹۳۰۹۳ و ۶۱۰۳۹۳۱۳۵. هر دو از ورودی های ۹۳ علوم کامپیوتر دانشگاه تهران هستیم.

در این پروژه، سعی بر آن داشتیم که بهترین تلاشمان را انجام دهیم و علاوه بر چیزهایی که در این پروژه یاد گرفتیم، آموختیم که کار تیمی مناسب می‌تواند نتیجه را به آن چه بهتر از تصور است، تغییر دهد.