دسته‌بندی متون

پروژه مبانی برنامه‌سازی

پاییز ۹۵

مقدمه

به احتمال زیاد تا کنون چیزهای زیادی راجع به هوش‌مصنوعی و یادگیری‌ماشین و تحولی که در اثر پیشرفت دانش بشر در این حوزه در حال وقوع است شنیده‌اید. با پیشرفت این علوم، ماشین‌ها روزبه‌روز به انسان‌ها نزدیک‌تر می‌شوند و با یادگیری بیشتر، توان انجام کارهایی که هیچ‌وقت انتظار آن نمی‌رفت را پیدا می‌کنند.
یادگیری ماشین انواع مختلفی دارد که مسئله دسته‌بندی متون از نوع یادگیری تحت نظارت است. هدف از این پروژه آشنایی ابتدایی شما با این دست از مسائل و پیاده‌سازی چند الگوریتم ابتدایی یادگیری روی متون دریافت شده از یک سایت خبری و استفاده از آن برای دسته‌بندی متون با دسته‌بندی نامشخص است.
در حین انجام این پروژه شما با چیزهای دیگری مانند پروتکل‌های ردوبدل کردن اطلاعات روی شبکه، برخی زبان‌های نشانه‌گذاری و ابزارهای مدیریت نسخه نیز آشنا خواهید شد. پروژه فازبندی شده است و برآورده کردن انتظارات هر فاز در صورت برنامه‌ریزی منطقی کار سختی نخواهد بود.
شما در فاز صفر به وسیله‌ی ردوبدل کردن اطلاعات با سرور پروژه و استفاده از `Git` کمی با مفاهیم شبکه و مدیریت نسخه آشنا خواهید شد و برای استفاده از آن‌ها در فازهای بعدی پروژه آماده می‌شوید.
در فاز یک با پردازش ابتدایی متن‌هایی که از سرور پروژه دریافت می‌کنید، کمی با چالش‌های پردازش متن و ذخیره‌سازی اطلاعات آشنا می‌شوید و در فاز بعدی به صورت جدی‌تر وارد این چالش شده و زمینه‌ی استفاده از متون برای یادگیری ماشین در فاز آخر پروژه را فراهم می‌کنید. در فاز آخر نیز با الگوریتم‌های ساده برای دسته‌بندی متون آشنا می‌شوید و به وسیله‌ی آن‌ها و خلاقیت خود متون را دسته‌بندی می‌کنید.
امیدواریم هر آن چه در این پروژه می‌آموزید، به تسلط بیش‌تر شما به زبان `C` و پیدا کردن دید بهتر به مفاهیم مختلف برنامه‌نویسی کمک کند.

با آرزوی موفقیت
تیم پروژه‌ی مبانی برنامه‌سازی
پاییز ۹۵

نکاتی راجع‌به استفاده از Git

یکی از اهداف جدی ما از طراحی این پروژه آشنایی شما با ابزار مدیریت نسخه Git و استفاده اصولی از آن است به همین دلیل از شما نیز انتظار داریم که نکات زیر را در رابطه با استفاده از Git در نظر داشته باشید:

- همان طور که می‌دانید یکی از مهم‌ترین اهداف استفاده از ابزارهای مدیریت نسخه نگهداری سوابق تغییرات در پروژه و توانایی بازگشت به وضعیت‌های قبلی است. در همین راستا commit هایی که شما انجام می‌دهید نیز بایستی به خوبی تغییرات ایجاد شده در پروژه را مشخص کنند. به صورت کلی توصیه می‌شود که پس از ایجاد هر تغییر کوچک تغییرات را commit کنید تا هم نگرانی از دست رفتن تغییرات به دلایل مختلف نباشید و هم کارتان را برای بازگشت به وضعیت‌های قبلی راحت‌تر کنید.
- اصولا در هر سیستمی مکانیزم‌هایی برای خروج از وضعیت‌های بحرانی وجود دارد که استفاده از آن‌ها در حالت‌های عادی توصیه نمی‌شود. یکی از مثال‌های چنین مکانیزم هایی در گیت Push -f است که تغییرات شاخه remote را overwrite می‌کند. از جایی که ما به استفاده اصولی از گیت اهمیت زیادی می‌دهیم برای این دسته از اعمال خشونت آمیز(!) با Git جریمه‌های خاصی در نظر گرفته ایم که در صورت پرهیز نکردن شما بر نمره نهایی شما تاثیر می‌گذارد. برای مثال هربار استفاده از این دستور موجب کسر ۰.۵ نمره از نمره نهایی شما می‌شود
- یکی از اهداف استفاده از Git در این پروژه همکاری متعادل شما عزیزان در انجام پروژه است. خوش بختانه سایت‌های ارائه دهنده سرویس Git نیز ابزارهای مناسبی برای پایش میزان و نحوه همکاری افراد دخیل در پروژه در آن دارند (مانند تعداد خط کد زده شده توسط هر شخص یا زمان فرستاده شدن هاcommit روی سایت). تقسیم کار و برنامه‌ریزی مناسب شما در پروژه نیز از چشمان تیزبین دستیاران آموزشی پروژه دور نخواهد ماند! لذا به این موضوع نیز توجه لازم را داشته باشید.
- از جایی که هنگام تحویل حضوری پروژه فازهای پروژه از ابتدا اجرا و بررسی می‌شوند باید انتهای هر فاز را با یک tag مربوط به آن فاز مشخص شده باشد.

ساختار بندی فازها

برای ساده تر کار شما در ساختار بندی فایل‌های پروژه برای هر فاز از پروژه ساختار معینی از فایل‌ها در نظر گرفته‌شده که شما باید کدهای خود را در آن قالب قرار دهید. هم‌چنین پس از اتمام هر فاز برای تشخیص موارد تقلب فایل‌های خود را در سامانه کوئرا در همان قالب بارگذاری می‌کنید.

فاز صفر

هدف از این فاز یادگیری نحوه رد و بدل کردن اطلاعات با سرور پروژه و استفاده ابتدایی از ابزار مدیریت نسخه Git است. تمامی کد این فاز بایستی در یک فایل نوشته و در کوئرا آپلود شود.

زمان بندی فاز

مهلت انجام فاز از روز شنبه ۲۷ آذر تا ساعت ۲۳:۵۹ جمعه ۳ دی است. ضوابط ارسال با تاخیر همانند تمرین‌ها است.

گام اول

ابتدا کتابخانه libcurl را با استفاده از مستندات تهیه‌شده روی سیستم عامل خود نصب کنید. سپس برای اطمینان از کار کردن این کتابخانه به لینک `http://fop-project.ir/welcome` یک درخواست GET با Header Authrorization نام‌کاربری و کلمه عبور تیم خود بفرستید(نام کاربری تیمها به شما داده شده است و از طریف کانال تلگرام یا کوئرا قابل دسترسی است) تا پیام خوشامدگویی را که شامل نام اعضای تیم است از سرور دریافت کنید!

گام دوم

Repository اختصاص‌یافته به تیمتان روی GitHub را روی سیستم خود Clone کنید. سپس متن آماده شده روی سرور را به وسیله ارسال یک درخواست از نوع GET به `http://fop-project.ir/phase0` دریافت کنید. متنی که دریافت می‌کنید یه متن نسبتا طولانی است که کلمات به وسیله space و enter و tab از هم جدا شده‌اند.

گام سوم

در این مرحله از این فاز هرکدام از اعضای تیم باید یکی از توابع زیر که ورودی آن‌ها متن دریافت شده از سرور است، پیاده‌سازی کند .(طبعا عضو اول باید تابع اول را پیاده سازی کند و عضو دوم تابع دوم را) در ضمن می‌توانید برای سهولت کار خود متن دریافت‌شده را در یک فایل ذخیره کنید.

- تابع اول باید کوچک‌ترین کلمه لغت‌نامه‌ای (کلمه ای که پیش از سایر کلمات در لغت‌نامه می‌آید) در بین کلماتی که بیش‌ترین تکرار را دارند پیدا کند.
- تابع دوم باید تعداد کلمات متمایز با طول بیشینه را پیدا کند.

گام چهارم

برای بررسی برنامه، خود خروجی‌ای که از توابع بالا دریافت می‌کنید را به صورت یک درخواست از نوع POST به آدرس `http://fop-project.ir/phase0` (همان آدرسی که متن را گرفتید) بفرستید. باید در بدنه Request جواب و این که کدام عضو تیم هستید را مشخص کنید؛ مثلا به این صورت:

`member=1&answer=salam` یا `member=2&answer=234`.

سرور هم در پاسخ اعلام میکند که جواب یافت شده درست است یا غلط.

اینکه شما عضو چندم تیم هستید در همان پیام خوش‌آمد گویی که دریافت میکنید مشخص شده (و به ترتیب اسامی اعضای گروه در آن پیام است)

آپلود در کوئرا و کار با گیت

درنهایت لازم است برنامه خود را در کوئرا آپلود کنید. فراموش نکنید که آخرین کامیت مربوط به این فاز را با تگ phase0 در Git مشخص کنید زیرا در هنگام تحویل حضوری باید به این کامیت برگردید و کد خود را مجدد اجرا کنید و اینبار با یک متن جدید دریافتی از سرور کد شما تست می‌شود.
دقت کنید که استفاده درست شما از Git در این فاز و توانایی همکاری شما در این بستر اهمیت فراوانی دارد.در واقع این فاز تمرینی است برای آشنایی و تسلط نسبی شما بر گیت و کتابخانه cURL که در فازهای بعدی نقش‌های اساسی ایفا میکنند بنابراین لازم است هرکس کامیت مربوط به تابعش را خودش انجام دهد، و ما توصیه میکنیم هر کس تابعش را خودش و به تنهایی بزند تا با جزییات و ریزه کاری‌های کار با گیت و merge کردن آشنا شوید.(البته این به این معنا نیست که کل فاز را جدا از هم بزنید)

نکته مهم

دقت داشته باشید که در نهایت آخرین کامیت شما در این فاز باید تابع هر دو نفر را پشتیبانی کند یعنی مثلا در ابتدای کار شماره یک عضو گروه را میگیرد(۱ یا ۲) سپس با توجه به این شماره متن را از سرور دریافت کرده و پردازش مربوط به آن عضو را روی آن انجام میدهد و برای سرور میفرستد.