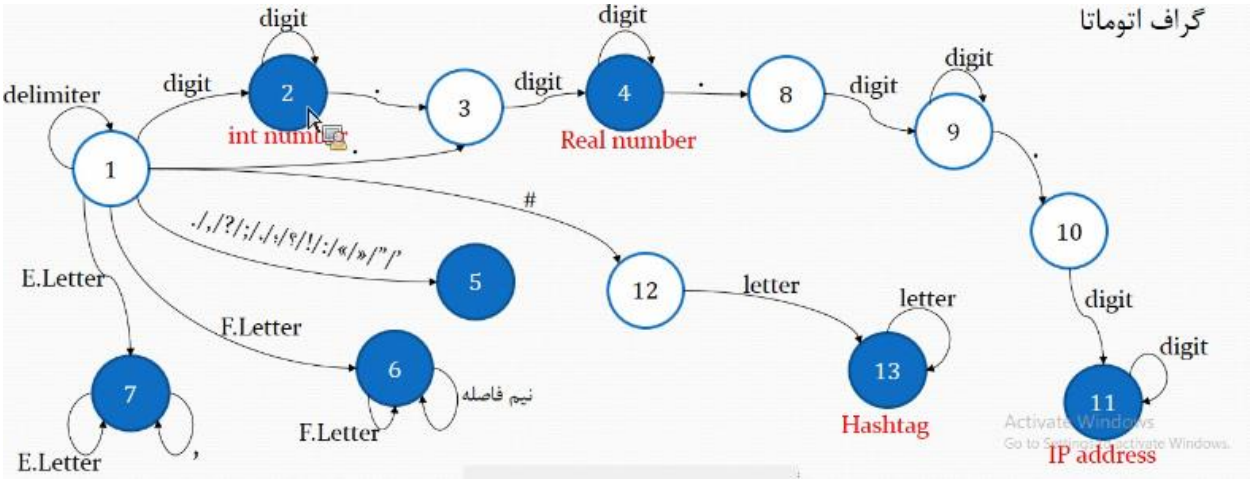
Tokenizer

شرح برنامه:

این برنامه هر نوع متنی را گرفته و توکن‌های آن را جداسازی و هر توکن را متناسب با نوعش چاپ میکند.

برای پیاده سازی این برنامه از گراف اتوماتای زیر استفاده شده است.



همانطور که مشاهده میکنید این برنامه از 13 حالت مختلف تشکیل شده است. برای پیاده سازی به ازای هر حالت یک کلاس در نظر گرفته شده است. به علاوه یک کلاس جهت نگهداری توکن ها و یک کلاس نیز برای نگهداری رفرنس های کلی برنامه در نظر گرفته شده است که جمعا تعداد کلاس ها به 15 عدد میرسد.

در این برنامه سعی شده است تا تمامی حالت های نشان داده شده در گراف اتوماتای بالا را به درستی پشتیبانی کند. برای مثال نیم فاصله ها تاثیری در جداسازی کلمات فارسی ندارند. یا زمانیکه کاربر به اشتباه space را وارد نکند برنامه تشخیص دهد و اشتباه کاربر را اصلاح کند. برای مثال I am20years old تمامی کلمات am و 20 و years از هم جدا میشوند. یا it is cold and#freezing در اینجا هشتگ #freezing به درستی تشخیص و جداسازی میشود.

تحلیل کد برنامه:

برنامه از کلاس main form شروع میشود. در هر فریم از برنامه یکی از state ها اجرا میشود. برای این منظور یک interface به کار برده شده است که همگی state ها از آن derive شده اند و در هر فریم تابع UpdateState آن رابط اجرا میشود.

در ابتدا یک بار کل متن به کد اسکی تبدیل میشود و در یک صف ریخته میشود تا برای تحلیل کاراکترها از آن بهره بگیریم.

current state در ابتدای کار که نمونه همان رابط است مقدار initial state را میگیرد. پس updateState کلاس initialState فراخوانی میشود. در اینجا بسته به خروجی صف کدها اسکی به یکی از state های digit یا digitDot یا Sharp یا specialToken یا FarsiLetter یا EnglishLetter میرود. در هر state اگر ورودی با روال آن مطابقت نداشت بسته به درست یا غلط بودن آن با عملیات مناسب (سیو کردن یا نکردن توکن) به حالت آغازین جهت تحلیل ادامه توکن ها برمیگردد.

در state هایی که آبی هستند نشانگر این است که توکن تا آنجا با موفقیت ثبت شده است پس در کلاس Token با توجه به نوع آن ثبت میشوند. اما در کلاس هایی که حالت نهایی نیستند برای مثال ورودی 192.168.1.. بدین معنی است که ورودی دارای فرمت اشتباه است پس در کلاس Token به عنوان unknown token ثبت میشوند.

تحلیل توکن ها تا زمانی ادامه میابد که صف کدهای اسکی به اتمام برسد که به این معنی است که به انتهای کاراکترها رسیدیم. اکنون زمان آن است که از لیست توکن های تهیه شده جهت چاپ خروجی استفاده کنیم. پس هر توکن را به علاوه نوع آن در خروجی چاپ میکنیم.

ضمنا کد برنامه کامنت گذاری شده است تا برای درک بهتر قابل استفاده باشد.