تشخيص موجوديت­هاي اسمي (Named Entity Recognition) يكي از مسائل پردازش زبان طبيعي است كه كاربردهاي عمده آن در سيستم­هاي خلاصه ساز متون، استخراج اطلاعات، بازيابي اطلاعات، پرسش و پاسخ، مترجم ماشيني و دسته بندي اسناد است. وظیفه یک سامانه تشخیص موجودیت اسمي، علاوه بر تعیین مرز هر یک از موجودیت­های اسمي، تشخیص نوع آن موجودیت و قرار دادن آن در دسته­های از پیش تعیین شده است. براي اين منظور قطعه كدهاي زير مورد استفاده قرار مي گيرند:

قطعه كد feature.py براي استخراج ويژگي هاي اصلي برنامه مي باشد. مسأله استخراج و انتخاب ويژگي، يکي از مسائلي است که در مبحث يادگيري ماشين و همچنين شناسائي آماري الگو مطرح است. اين مساله در بسياري از کاربردها (مانند طبقه بندي) اهميت به سزائي دارد، زيرا در اين کاربردها تعداد زيادي ويژگي وجود دارد، که بسياري از آن**­**ها يا بلااستفاده هستند و يا اينکه بار اطلاعاتي چنداني ندارند.

ویژگی­های Part of Speech و BaseForm با استفاده از نرم افزار GeniyaTagger تحت لینوکس استخراج شده اند، و براي استفاده در قطعه كد اصلي feature.py نياز به تغييراتي دارد، اين تغييرات با استفاده از قطعه كد editgenia.py اعمال مي شوند.

پس از اعمال مراحل فوق بر روی یک متن ساختار نیافته­ی شیمیایی، متن بصورت برداری از همه­ی ویژگی­های استخراج شده، در مي آيد که این متن جدید به الگوریتم یادگیری داده شده تا مدل هاي طبقه بند پايه را با استفاده از ابزار Mallet (كه يك نوع ابزار مدلسازي موضوعي است) و از الگوريتم CRF استفاده مي كند، تولید و با استفاده از قطعه كد evalIOB2.pl كه به زبان پرل نوشته شده است نتایج طبقه بندها مورد ارزيابي قرار مي گيرند.

پس از تولید مدل طبقه بند، دسته های مربوط به متن های جدید را می توان تشخیص داد. در واقع می توان گفت كه در اين مرحله طبقه بندها از روي متن هاي پيش پردازش شده اقدام به يادگيري مي كنند.

جهت توليد داده هاي آموزشي مختلف براي ايجاد طبقه بند و نيز انتخاب ويژگي هاي كارا جهت استفاده در فرآيند ايجاد مخزن طبقه بندها، اقدام به ساده سازي داده ي آموزشي با استفاده از قطعه كد skipLabel.py نموده، به صورتي كه جملاتي از داده هاي آموزشي كه برچسب هاي تمامي نشانه هاي آن جمله ها از نوع OUT بوده اند را حذف و سپس با استفاده از روش جايگذاري جملات، بصورت تصادفي داده هاي آموزشي متفاوت توليد مي گردد. و در آخر با استفاده از روش انتخاب و تركيب طبقه بند ها با روش شبه هيوريستيك از طريق قطعه كد eval.py بهترين نتيجه بدست مي آيد.

قطعه كد MajorityVoting.py براي راي گيري اكثريت در تمام مراحل فوق مورد استفاده قرار مي گيرد. همچنين قطعه كد main.py كد اصلي اجراي برنامه مي باشد، كه تمامي قطعه كدهاي موجود در آن فراخواني مي شوند.