به نام خدا

پروژه ی سوم هوش مصنوعی:

براي اينجام اين پروژه بايد ابتدا فايل هاي ترين داده شده را بخوانيم و در صورت نياز روي آن ها تغييراتي اعمال كنيم.

ابتدا فايل هاي مورد نظر را باز ميكنيم. براي اينكه دشكنري درستي از كلمات داشته باشيم نياز است كه علايم نگارشي را از فايل حذف كنيم و اين كار را انجام ميدهيم و براي اينكه هر خط را مشخص كنيم از علامت <s></s> استفاده ميكنيم و هر مصرع را معادل يك خط در نظر ميگيريم.

در مرحله ي بعد بايد يك ليست از همه ي كلمات در نظر بگيريم كه ان را با parse كردن در هر خط انجام ميدهيم و كلمات آن را به ليست اضافه كنيم.

براي يافتن تعداد تكرار هر كلمه بايد ابتدا يك ديكشنري از كل كلمات درست كنيم و مقدار value براي هر كلمه تعداد تكرار آن جمله را ست كنيم.(اين كار براي روش يونيگرام لازم است.)

براي روش بايگرام چون بايد هر كلمه با توجه به كلمه ي آن يعني به صورت جفت كلمه سنجيده شود نياز به ديكشنري ديگري است كه كليد آن جفت كلمات كنار هم باشد و مقدار آن تعداد تكرار آن ها باشد كه اين كار مشابه حالت قبلي ابتدا كليد ها را يافته در ليستي قرار ميدهيم سئس ديكشنري را با مقدار هاي تكرار هر كدام تشكيل ميدهيم.

سپس براي مدل يونيگرام و بايگرام بايد احتمال هر كدام را حساب كنيم و آن مقدار را براي كليد هاي به دست آمده در ديكشنري جديد قرار ميدهيم.

با توجه به اينكه ممكن است احتمال تكرار بعضب جفت ها و كلمات صفر باشد بنابراين بايد از مدل backoffmodel استفاده كنيم. كه براي اينكار با توجه به ضرايبي كه تعيين ميكنيم از نتايج يونيگرام و بايگرام استفاده ميكنيم.

براي نسبت دادن مقادير به ضرايب بايد درنظر بگيريم كه مقدار اپسيلون عدد بسيار كمی باشد زيرا كه آن عدد برای وقت هايی است كه آن زوج در داده های آموزشی وجود نداشته باشد. از طرفی وجود يک لغت به صورت بايگرام نيز اهميت زيادی دارد پس لاندا2 مهم تر از لاندا3 می باشد.

براي استفاده از فايل تست بايد نتيجه ي مدل را به ازا هر سه شاعر به دست بياوريم و شاعري كه احتمال بالاتري داشته باشد را به عنوان جواب بازميگردانيم.