Training Statistical NLP Model - Bigrams & Unigrams

تاثیر حذف کلمات پر تکرار و کم تکرار در دقت دست آمده

قطعا موثر است، این کار به در چند روش پیاده شده:

- با Lemmatize کردن کلمات میزان نا خالصی های دیکشنری ها را کاهش دادیم.
- با حذف 10 کلمه اول و 10 کلمه آخر از هر دیکشنری زبان کلمات بیش از حد تکرار شده را حذف کردیم.

اگر این کار را نکنیم دقت به شدت پایین می آید، به عبارتی مدل general تر عمل میکند و قادر به تشخیص دقیق نمی باشد.

تاثیر ضرایب در دقت بدست آمده

ضریب lambda3 زمانی مورد استفاده است که کلمه یا کلماتی داشته باشیم که در دیکشنری ها نباشند (مدل آنها را ندیده باشد) و به عنوان یک bias بکار میرود تا از صفر شدن کل احتمال جلوگیری شود.

ضرایب lambda1 و lambda2 تاثیر هر کدام از بخش ها را مشخص میکند، به عبارتی اگر lambda1 را زیاد کنیم تاثیر bigrams را را بیشتر کردیم و بلعکس.

بهترین دقت بدست آمده و تحلیل تاثیر پارامتر ها

بهترین دقت زمانی است که ضریب تاثیر bigram بیشتر از unigram باشد.