**Tokenizing**

با استفاده از آن میتوان کلمات و یا جملات یک متن را جدا کرد.

Sent\_tokenize جملات یک متن را جدا کرده و انها را داخل لیست قرار میدهد.

Word\_tokenize کلمات یک متن را جدا کرده و آنها را داخل لیست قرار میدهد.

باگها و خطاها: پرانتز و کاما و نقطه را به عنوان یک کلمه میشناسد.

جفت کلمات معنی دار را نمیتواند تشخیص دهد و آنهارا جدا میکند.

sents=sent\_tokenize(example\_string)

words=word\_tokenize(example\_string)

**Stop Words**

کلماتی مانند an وis وin وغیره که در اغلب جملات کاربرد دارند داخل stopwords هستند و میتوان از آن برای فیلتر کردن کلمات استفاده شده در متن استفاده کرد.ابتدا باید یک لیست از کلمات تشکیل دهیم سپس با استاپ وردز آنها را فیلتر کنیم. که اینکار میتواند با بیرون کشیدن تمام کلمات از متن باشد یا با بیرون کشیدن مرحله به مرحله کلمات از هر جمله‌ی متن باشد.

با استفاده از متد casefold() میتوان کلمات با حروف بزرگ یا کوچک را همسان در نظر گرفت.

باگ ها و خطاها:

اگرکوتیشین و یا دبل کوتیشن یا اکولاد و غیره در متن وجود دارد بهتر است حذف شود و سپس پری پراسسینگ شود.

stop\_words = set(stopwords.words("english"))

filtered\_list = [word for word in words if word.casefold() not in stop\_words]

**Stemming**

برای کاهش کلمات به ریشه هایشان کابرد دارد. مثلا برای تبدیل containing و container به contain استفاده میشود. به این صورت میتوان وزن وجود این keyword در داخل متن را مشخص کرد و لازم به مشخص کردن دستی نخواهد بود.

باگها و خطاها:

در برخی موارد کلمات با ریشه ی مشخص، روتهای مختلفی برگردانده میشود. مثلا برای discovery ، discoveri ولی برای discovering ، discov برمیگردد که منطقی نیست.

stemmer=PorterStemmer()

stemed\_word=[stemmer.stem(word) for word in words]

**Tagging**

برای تعیین نقش کلمات در جمله میتوان از این ترم استفاده کرد. مثلا فعل، اسم و یا صفت بودن یک کلمه به صورت یک تاپل در جلوی آن کلمه مشخص میشود.

w\_tag=nltk.pos\_tag(words)