Stop Word به کلمات و حروف هایی در جمله که به تنهایی معنای خاصی ندارند و باید برای پردازش از جمله حذف گردند.

از کاربرد های Text Classification میتوان به مواردی مانندتشخیص موضوع متن،شناسایی Spam، شناسایی سن یا جنس نویسنده، شناسایی زبان، تحلیل احساسات و ..... مورد استفاده قرار گیرد.در این جزوه از روش های یادگیری ماشین(با نظارت) استفاده می کنیم.از classifier های مختلفی می توان استفاده کرد که 4 نمونه از آن ها عبارت اند از:

Naive Bayes, Logistic regression, support-vector machines, k-nearest Neighbors

روش Laplace smoothing میگوید که هنگام محاسبه احتمال یک کلمه یا یک عبارت در جمله امکان صفر شدن احتمال وجود داشته باشد که در پی آن صفر شدن کل احتمال نتیجه می شود. در این روش به فرمول احتمال در صورت کسر به علاوه 1 و مخرج کسر به علاوه سایز و تعداد لغات می گردد.

P(w | c) = count(w, c) + 1 / count(c) + | V |

به بررسی Naive Bayes می پردازیم.در قانون بیز برای کلاس c و داکیومنت d داریم:

P(c | d) = P(d | c) P(c) / P(d)

در نهایت با توجه به قانون بیز برای Naive Bayes داریم:

Argmax P(x1, x2,...,xn | c) P(c)

کار درستی نیست ولی برای سادگی کار هر یک از احتمالات P(x1) P(xn) را مستقل از هم میدانیم.

در نتیجه

P(x1,...,xn | c) = P(x1 | c) × P(x2 | c) × P(x3 | c) ×....× P(xn | c)

به طور ساده از تعداد وقوع یک کلمه در Training Set بهره می گیریم.

P(cj) = doccount(C = cj) / Ndoc

P(wi | cj) = count(wi,cj) / count(w,cj)

که با استفاده از Laplace Smoothing خواهیم داشت

P(w | c) = count(w,c) + 1 / count(c) + | V |

در مثال زیر استفاده نکردن از Laplace Smoothing نتیجه ای برابر صفر خواهد داشت.

مثلا کلمه fantastic با اینکه مثبت است(از نظر حس و احساسی) اما شاید در بین کلمات موجود در trainging set مشاهده نشده باشد.

P(“fantastic” | positive) = count(“fantastic”, positive) / count(w,positive) = 0



جدول بالا از ما میخواهد با توجه به Training set که از شماره 1 تا 3 متعلق به کلاس C و شماره 4 متعلق به کلاس j است. شماره 5 را که یک نمونه از Test set است را تشخیص دهیم که متعلق به کدام کلاس می باشد.

با توجه به فرمول P(c) = Nc / N

داریم P(c) = ¾ و P(j) = ¼

و با توجه به فرمول P(w | c) = count(w,c) + 1 / count(c) + | V |

برای تک تک توکن های جمله خواسته شده در هر یک از کلاس ها داریم

P(Chinese | c) = (5 + 1) / (8 + 6) = 6/14 = 3.7

P(Tokyo | c) = (0 + 1) / (8 + 6) = 6/14 = 1/14

P(Japan | c) = (0 + 1) / (8 + 6) = 6/14 = 1/14

P(Chinese | j) = (1 + 1) / (3 + 6) = 2/9

P(Tokyo | j) = (1 + 1) / (3 + 6) = 2/9

P(Japan | j) = (1 + 1) / (3 + 6) = 2/9

حال با مقادیر بدست آمده شروع به تشخیص کلاس میکنیم.برای این کار باید احتمال بودن جمله پنجم در کلاس های c و j بدست آورده شده و با هم مقایسه شوند.جمله 5 مطعلق به کلاسی است که مقدار احتمال بالاتری را دارا باشداحتمال هر دو به صورت زیر محاسبه خواهند شد.

P(c | d5) = 3/4 × (3/7)3 × 1/14 × 1/14 0.0003

P(j | d5) = 1/4 × (2/9)3 × 2/9 × 2/9 0.0001

در مقایسه دو احتمال بالا نتیجه محاسبات بزرگتر بودن احتمال تعلق جمله 5 به کلاس c است.پس، جمله 5 در کلاس c قرار دارد.