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.

نکاتی درباره Naive bayes:

1. سریع، پیچیدگی حافظه پایین

2. در دامنه هایی که فیچر های آن ها از اهمیت یکسانی برخوردارند،کارایی خوبی دارد در حالی که درختان تصمیم اینگونه نیستند.

3. اگر فرض مستقل بودن فیچر ها درست باشد، این روش بهینه است.

4. برای Text Classification یک Baseline محسوب می گردد.

Precision به درصد انتخاب هایی که به درستی از تمام آیتم های انتخاب شده، انتخاب شده است گفته می شود.

Recall به درصد آیتم های به درستی انتخاب شده از میان تمام آیتم های درست، گفته می شود.

طبق جدول زیر هر یک از انتخاب ها بر اساس ویژگی ها متفاوت نام گذاری می شود.

|  |  |  |
| --- | --- | --- |
| **Not Correct** | **Correct** |  |
| False Positive | True Positive | **Selected** |
| True Negative | False Negative | **Not Selected** |

اگر انتخاب کردیم(Positive) و انتخاب درست بود(True)، پس True positive محاسبه می گردد.

اگر انتخاب کردیم(Positive) و انتخاب درست نبود(False)، پس False positive محاسبه می گردد.

اگر انتخاب نکردیم(Negative) و انتخاب درست بود(True)، پس True Negative محاسبه می گردد.

اگر انتخاب نکردیم(Negative) و انتخاب درست بود(False)، پس False Negative محاسبه می گردد.

برای محاسبه Percision باید تعداد انتخاب های True Positive را تقسیم بر تعداد کل انتخاب ها در آن کلاس کنیم و برای محاسبه Recallباید تعداد True Positive را تقسیم بر تعداد کل انتخاب ها در آن ویژگی کنیم.

برای محاسبه دقت یا Accuricy باید تعداد True Positive را با تعداد True Negative جمع کرده و حاصل را بر مجموع کل حالت ها کنیم.

Acc = (True Positive + True Negative) / (True Positive + True Nengative + False Positive + False Negative)

تا اینجا روش های ارائه شده برای دو کلاس بود ولی اگر بیش از دوکلاس داشتیم چه کنیم؟کل کلاس ها را تبدیل به دوکلاس میکنیم به عنوان مثال کلاس x و کلاس سایر.در حقیقت کلاس سایر مجموعه ای از کلاس های مورد نظر می باشد و کلاس x یکی از کلاس های درخواست شده. به همین ترتیب ما دوکلاس داریم که عملیات مورد نظر را روی آن ها انجام دهیم. سپس همین فرایند را برای تک تک کلاس های خواسته شده انجام می دهیم.کلاس با بیشترین ضریب احتمال، کلاس نهایی و منتخب است.

در روش میانگین گیری ماکرو(Macro)، کارایی هر کلاس را جداگانه اندازه گرفته و میانگین آن ها را محاسبه می کنیم.

در روش میکرو(Micro)، تمام داده های کلاس ها را با هم ترکیب کرده و تنها یک بار ارزیابی را انجام می دهیم.

مثال



برای جلوگیری از overfitting از روش cross validation استفاده می کنیم

داده ها را به k قیمت مساوی تقسیم شده و هر بار k-1 قسمت برای آموزش و 1 قسمت برای تست استفاده می شود. به این روش k-fold cross validation نیز می گویند.

در این جزوه در احتمالات برای تبدیل ضرب احتمالات به جمع آن ها با استفاده از لگاریتم صورت می گیرد.

در spell correction ما دو تسک داریم یکی تشخیص غلط املایی و دیگری اصلاح آن.

دو نوع خطا در spell correction داریم:

1. خطا هایی که منجر به ایجاد کلمات بی معنی می شوند مثل graffe => giraffe

2. خطا هایی که منجر به ایجاد کلمات با معنی می شوند که این دسته خطا ها به دو دسته تقسیم می شوند:

1. Typographical errors: خطا هایی که ظاهر نسبتا شبیه ولی تلفظ متفاوت دارند. مثل

Three => There

2. Conitive errors (Homophones): خطا هایی که ظاهری متفاوت ولی تلفظ شبیه دارند.مثل

Piece => Peace Too => Two

ایجاد کلماتی که در دیکشنری موجود نیست خطا محسوب می شود.

فرایند تصحیح خطا:

1. لیستی از کلمات شبیه به کلمه خطا دار تولید می شود.

2. بهترین کلمه کاندید انتخاب می گردد.(منظور از کلمه کاندید کلمه ای است که در کمترین فاصله ویرایشی را با کلمه ی خطا دار دارد یا کلمه ای که بیشترین مقدار احتمال را در روش Noisy Channel بدست آورد. لیست کلمات کاندید میتواند کلمات با تلفظ یکسان یا با املای مشابه باشد.

بهترین کلمه ها با یکی از روش های Noisy Channel یا Classification به دست خواهد آمد.

اگر کلمه درستw باشد و به صورت اشتباه به صورت x تایپ شود:

ArgMax P(x | w) P(w)

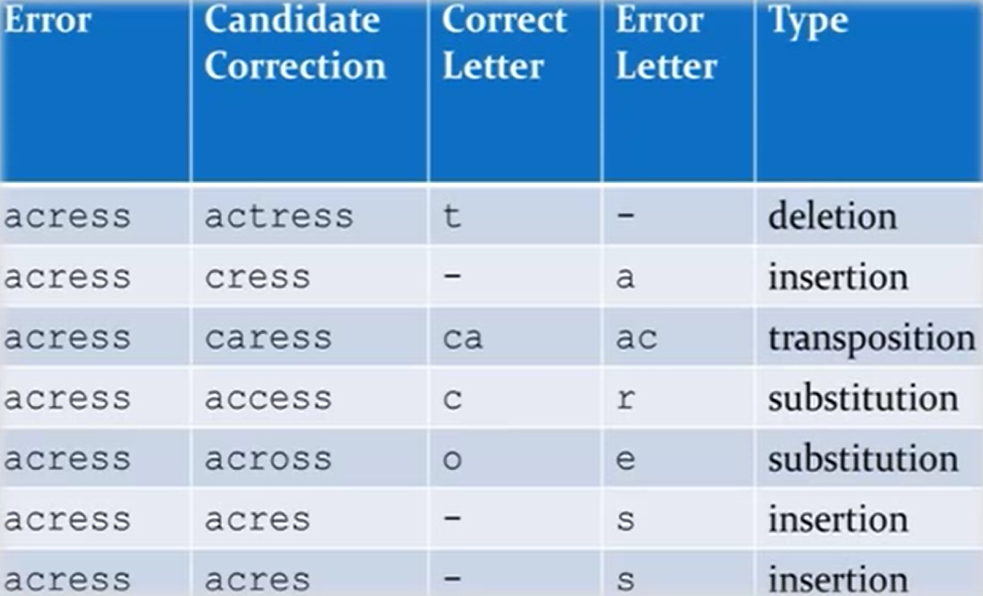
فاصله ویرایشی با یک یا چند تا از اعمال زیر به وجود می آید.

Insertion یا افزودن

Deletion یا حذف

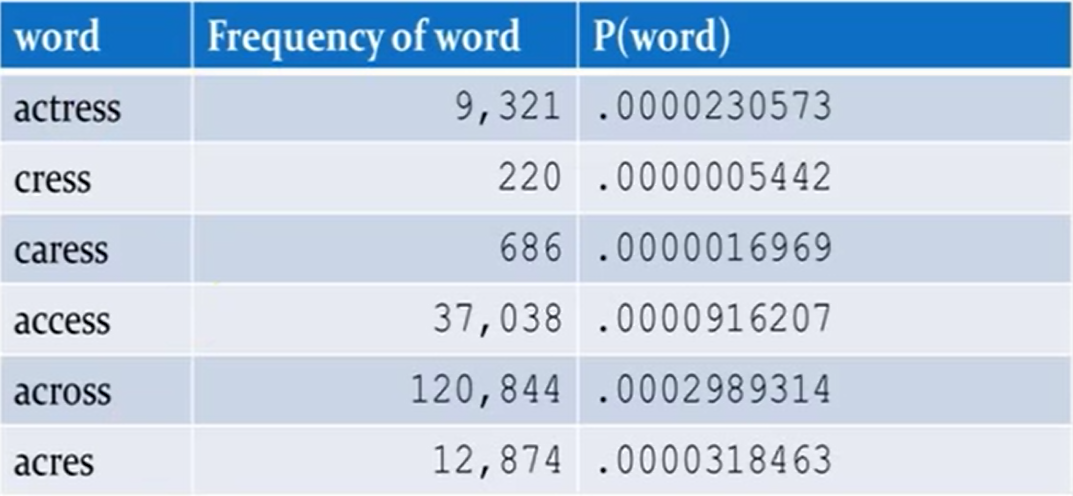
Transposition of two adjacent letters یا تغییر مکان دو حرف مجاور

مثال: فردی سعی در تایپ کلمه acress دارد میخواهیم ببینیم اشتباه کجا است.



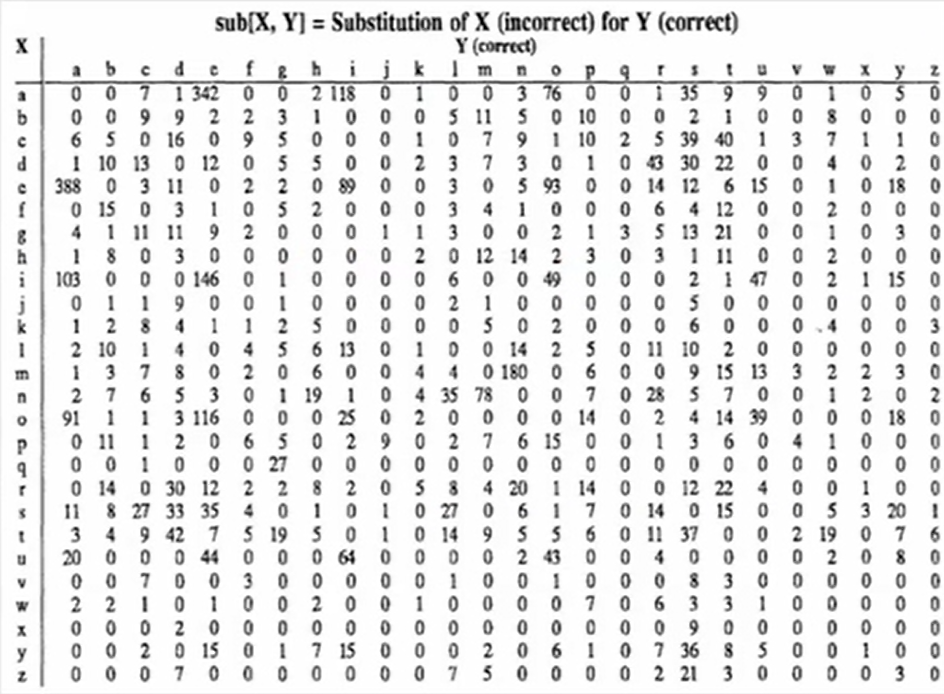
برای خطا یابی در وب نیز میتوان از الگوریتم Stupid backoff استفاده کرد.

احتمال کلمات درست توسط یک Corpus معتبر به نام COCAجدول زیر نتیجه شده است.



افزودن یک کاراکتر یا حذف کردن این کاراکتر که احتمال وقوع آن در پیکره، نسبت به کاراکتر قبلی محاسبه خواهد شد.

جدول زیر گزارش جابجایی و تعویض کلمات در متون را طبق یک منبع معتبر می دهد.



طبق گفته های بالا روش محاسبه P(x | w) در چهار عملیات گفته شده به شورت زیر خواهد بود.

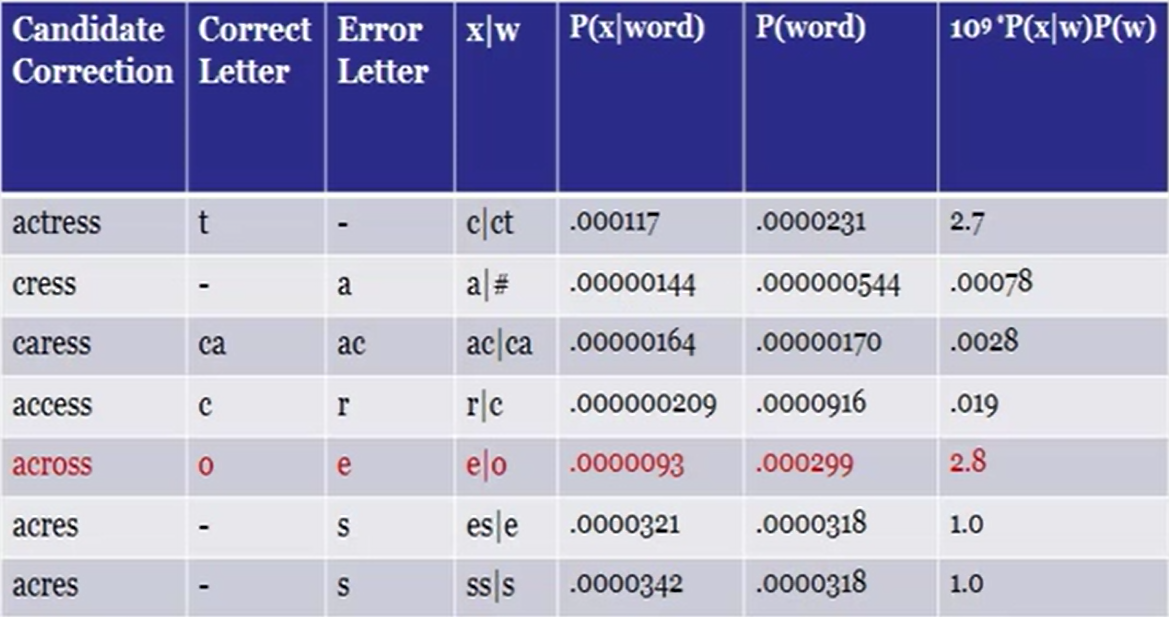
Del[wi - 1,wi] / count[(wi – 1) wi] if deletion

Ins[wi - 1,xi] / count[wi - 1] if insertion

Sub[xi,wi] / count[wi] if substitution

Trans[wi,wi + 1] / count[(wi + 1) wi] if transposition

در جدول زیر اختلافات بدست آمده با استفاده از مدل Noisy Channel برای کلمه acress نمایش داده می شود که کلمه درست across بیشترین احتمال را دارد ولی با دخالت مدل bigram نتیجه عوض می شود. بدلیل اینکه عدد خیلی کوچک می شد،ضرب در 109 می شود.



اما جمله زیر را در نظر بگیرید

A stellar and versatile acress whose combination of sass and glamour.....

با توجه به امار های بدست آمده از Contemporary american english

P(actress | versatile) = 0.000021 P(whose | actress) = 0.0010

P(across | versatile) = 0.000021 P(whose | across) = 0.000006

P(versatile actress whose) = 0.000021 0.0010 = 210 10-10

P(versatile across whose) = 0.000021 0.000006 = 1 10-10

میبینیم که کلمه actress احتمال بیشتری برای درست بودن دارد و با ضرب 210 در احتمال کلمه actress در جدول این احتمال نیز بیشتر می گردد.

منظور از lemma حالت ریشه ای و بدون شاخ و برگ کلمه در متن است و منظور از Wordform حالت عادی کلمه بدون تغییر است.

از مشکلات دیگری که در NLP رخ میدهد وجود کلمات با تلفظ و فرم یکسان اما با معانی مختلف در متن است که به آن ها Homograph گفته می شود.

کلماتی که بیش از یک معنی دارند ولی مرتبط به یکدیگرند را polysemy گویند مانند:

The bank was constructed in 1875

بانک (ساختمان بانک) در سال 1875 ساخته شده است.

I withdrew money from the bank

پولم را از بانک (حساب بانکی) بیرون کشیدم.

به این نوع ارتباط ها Systematic Polysemy یا Metonymy نیز گفته می شود.

چگونه بدانیم یک کلمه بیتر از یک معنی در دو جمله دارد؟ با استفاده از تست Zeugma

این تست اینگونه عمل می کند که دو جمله مورد نظر با یک دیگر ترکیب می شود، در صورتی که جمله ترکیب شده از نظر ساختاری و معنایی مشکلی نداشت پس کلمه یک معنی بیشتر ندارد. در غیر این صورت بیشتر از یک معنی دارد. مانند:

1. Which flights serve breakfast?
2. Does Lufthansa serve Philadelphia?

جمله ترکیب شده Does Lufthansa serve breakfast and Philadelphia?

جمله ترکیب شده از لحاظ معنایی اشتباه است پس کلمه serve دو معنی دارد.

اگر یک کلمه از کلمه بعدی خاص تر باشد یا به عبارتی از لحاظ رابطه پدر فرزندی، کلمه اول فرزند کلمه دوم باشد به این رابطه Hyponomyگفته می شود. نام دیگر این رابطه IS-A نیز می باشد. به عکس این رابطه Hypernym یا Superordinate گفته می شود.مانند:

Car is hyponym of vehicle (Car IS-A vehicle)

Vehicle is hypernym of car

Instance نیز مانند Hypernym است اما به یک شی خاص یا یک فرد خاص اشاره دارد.مانند:

San Francisco یک نمونه از مفهوم City است.در حالی که خود مفهوم City یک کلاس است.

تزاروس به رابطه سلسله مراتبی و یا رابطه درختی یا پدر فرزندی را گویند به طور خلاصه.

WordNet به مجموعه از اماده به زبان های مختلف است که هم به صورت آنلاین و هم به صورت آفلاین و نصبی می توان استفاده نمود.

از مجموعه ای از synset ها ساخته شده است. Synset ها مجموعه ای از کلمه ها هستند که یک معنی را ارائه می دهند.

در جدول زیر خلاصه ای از روابط نشان داده شده است.

