Or منطقی در عبارت منظم در [ ] قرار میگیرد مثلا

[wW] oodchuck این کلمه یا با دبلیو بزرگ یا کوچک نوشته شده است که می توان از pipe یا | هم استفاده کرد

Negation در عبارت برای بیان نفی، کاراکتر ^ باید داخل [ ] و ابتدای آن قرار گیرد.مثلا

[^A-Z] غیر حروف بزرگ

علامت سوال در وسط کلمه به معنای این است که حرف ما قبل علامت سوال میتواند در کلمه موجود باشد یا نباشد.مثلا colou?r

علامت ستاره در کلمه به صفر یا بیش از صفر عدد از کاراکتر ماقبل آن اشاره دارد مثلا (oo\*h!) oh! ooh! oooh!

علامت مثبت در کلمه به یک یا بیش از یک عدد از کاراکتر ماقبل آن اشاره دارد مثلا (oo+h!) ooh! oooh! ooooh!

علامت نقطه در کلمه به معنای این است که هر کاراکتر ممکنی قابل قبول است مثلا beg.n که میتواند begin یا begun یا هر چیز دیگری باشد.

علامت ^ اگر در ابتدای براکت باشد به این معنی است که کلمه باید با حروف داخل براکت شروع شود مثلا ^[A-Z] و اگر بخواهیم حرف در انتهای کلمه آیند از علامت $. استفاده می کنیم(اگر بخواهیم از خود نقطه استفاده کنیم از علامت \.$ استفاده می کنیم)

|  |  |
| --- | --- |
| **Expansion** | **RE** |
| [0-9] | \d |
| [^0-9] | \D |
| [a-zA-Z0-9\_] | \w |
| [^\w] | \W |
| [ \r\t\n\f] | \s |
| [^\s] | \S |

{n} به n عدد از کاراکتر قبلی یا عبارت قبلی اشره دارد مثلا Hel\.{2}o/ که به معنی Hello

{n,m} حداقل n و حداکثر m عدد از کاراکتر یا عبارت قبلی وجود دارد مثلا /Hel\.{2,4}o/ که به معنی Hello,Helllo,Hellllo است.

{n, } حداقل n عدد از کاراکتر یا عبارت قبلی وجود دارد مثلا /Hel\.{3, }o/

دو نوع خطا داریم. نوع 1 که شامل استخراج رشته هایی که نباید استخراج می شدند است که اصطلاحا به آن False Positives گفته می شود.نوع 2 که شامل عدم استخراج رشته هایی که باید استخراج می شدند است و به آن False negative گفته می شود.

کاهش تعداد رشته هایی که نباید استخراج می شدند باعث افزایش percision و افزایش تعداد رشته هایی که باید استخراج شوند باعث افزایش recall خواهد شد.

برای پردازش بعضی از مدل خهای زبانی ممکن است از روش های بادگیری ماشین استفاده کنیم که در این حالت، عبارات منظم نقش ویژگی ها(features) را بازی می کنند.

نرمال سازی متن به دو بخش تقسیم می شود. Lemmatization به معنی اینکه فرم های مختلف کلمه تبدیل شود به هدوورد(کلمه ای در دیکشنری که حالت پایه آن کلمه محسوب می شود) آن شود. بخش دوم Stemming در این حالت فقط شاخ و برگ کلمه زده شده و ریشه کلمه نگه داشته می شود.

مورفولوژی به معنی استخراج یک کلمه و استخراج پسوندو پیشوند های آن است

Morphemes یا مورفیم ها جز کوچکی از زبان است که جزئی از کلمه را می سازد.

Stems: جزء اصلی یک کلمه که با معنی می باشد.

Affixes: اجزای دیگر کلمه که به stem می چسبند.

Stemming یا ریشه یابی از کلمات Stem را استخراج می کند.به عبارت دیگر، affix ها را از کلمه جدا می کند.

الگوریتم porter برای Stemming: برای اجرای این الگوریتم باید مراحل زیر طی شوند:

1. اگر کلمه دارای sses بود تبدیل به ss شود مانند caresses که تبدیل به caress شده است.

2. اگر کلمه دارای ies بود تبدیل به i شود مانند ponies که تبدیل به poni شده است.

3. اگر کلمه دارای ss بود تبدیل به ss شود مانند caress که تبدیل به caress شده است.

4. اگر کلمه دارای s بود حذف شود مانند cats که تبدیل به cat شده است.

5. اگر کلمه دارای حروف صدا دارد و ing بود حذف شود شود مانند walking که تبدیل به walk شده است.

6. اگر کلمه دارای حروف صدا داردو ed بود حذف شود مانند plastered که تبدیل به plaster شده است.

7. اگر کلمه دارای ational بود تبدیل به ate شود مانند relational که تبدیل به relate شده است.

8. اگر کلمه دارای izer بود تبدیل به ize شود مانند digitizer که تبدیل به digitize شده است.

9. اگر کلمه دارای ator بود تبدیل به ate شود مانند operator که تبدیل به operate شده است.

10. اگر کلمه دارای al بود حذف شود مانند revival که تبدیل به reviv شده است.

11. اگر کلمه دارای able بود حذف شود مانند adjustableکه تبدیل به adjust شده است.

12. اگر کلمه دارای ate بود حذف شود مانند activate که تبدیل به activ شده است. که باگ الگوریتم است

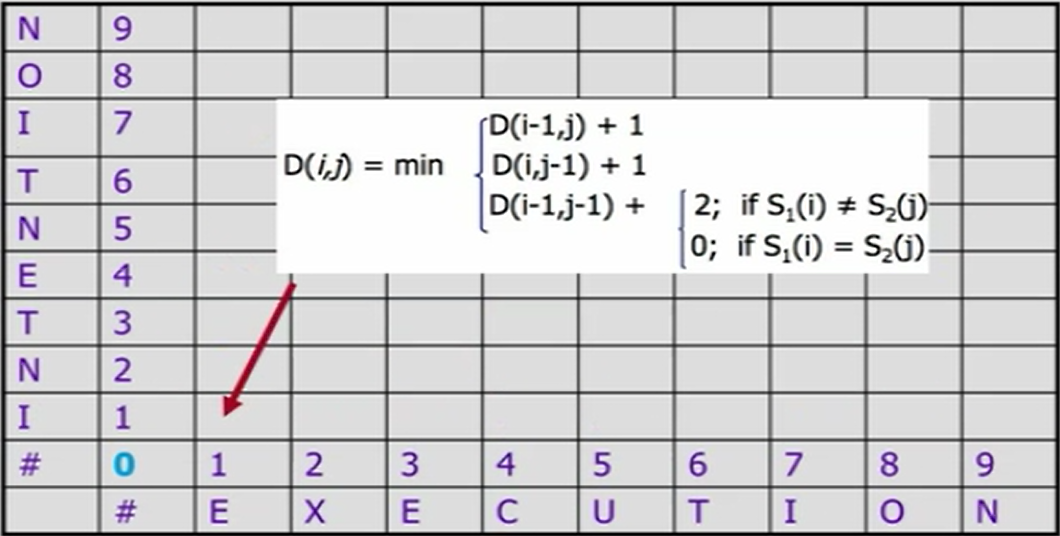
توکن به مجموعه یا یک تک کاراکتر که یه کلمه یا حرف یا یک نشانه را در جمله تشکیل می دهند گویند.مثلا جمله "من یک دانشجو هستم!" دارای 5 توکن است. به عمل استخراج این توکن ها توکنایز(Tokenize) کردن گفته می شود.

MED یا Minimum Edit Distance بین دو رشته معادل است با حداقل تعداد اعمال زیر که یک رشته را به رشته دیگری تبدیل می کند.1)Insertion 2)ِDeletion 3)Substitution

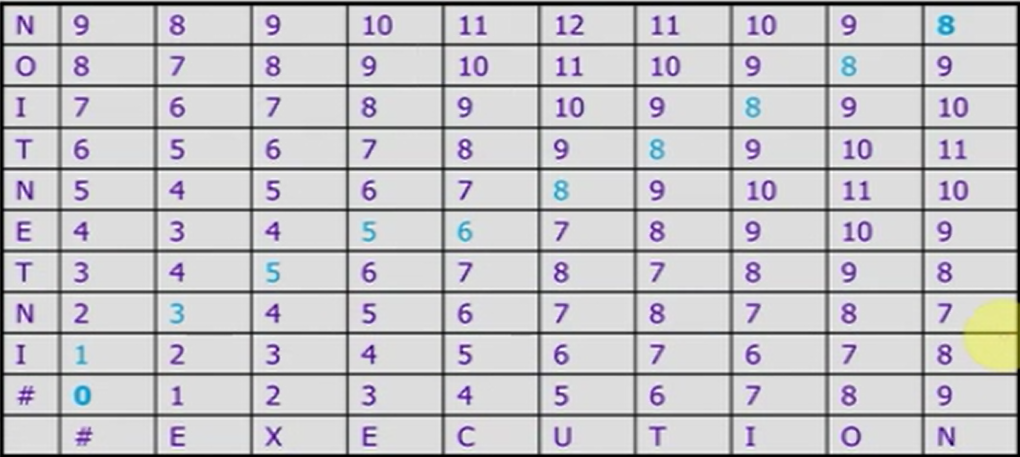
به عنوان مثال در تبدیل کلمه Intention به Execution در حرف I عمل Deletion در حرف n عمل substitution در حرف T نیز همینطور و برای حرف بعد از e نیز یک عمل Insertion و برای n دوم نیز عمل substitution خواهیم داشت. در این اعمال اگر هزینه هر عمل 1 واحد در نظر بگیریم، در کل 5 واحد هزینه خواهیم داشت،ولی برای عمل substitution که متشکل از یک عمل Deletion و یک عمل Insertion است، دو واحد هزینه در نظر میگیریم.درنتیجه در فرایند تبدیل کلمات گفته شده،8 واحد هزینه خواهیم داشت.

برای محاسبه Med از الگوریتم برنامه نویسی پویا استفاده می گردد برای اینکه این الگوریتم همیشه بهینه ترین راه ها را به ما پیشنهاد می دهد.

خلاصه استفاده این الگوریتم در این مسئله به این گونه است که پس از نشست هر دو کلمه یا متن در جدول به صورت افقی و عمودی و نسبت یک عدد به هر کدام از حروف حل مسئله پیدا کردن کوتاهترین و بهینه ترین راه برای تبدیل دو کلمه به یکدیگر شروع خواهد شد.



در حروف اول کلمات به دلیل یکسان نبودن حروف،چه در تبدیل رشته اول به دوم و چه در تبدیل رشته دوم به اول عملیاتی خواهیم داشت. درتبدیل رشته اول به دوم،در رشته اول یک عملیات Insertion خواهیم داشت و در رشته دوم عملیات ایجاد یک Gap و در تبدیل رشته دوم به اول نیز به همین گونه.همر دو عملیات هزینه های یکسان و برابر 1 دارند که جمعا هزینه دو برای آن ها در نظر گرفته خواهد شد.به عبارت دیگر هزینه عملیات Substitution 2 است و در این جدول اگر بخواهیم به صورت اریب حرکت کنیم 0 + 2 برابر با هزینه 2 خواهد بود.اگر حروف با هم یکسان بودند به صورت اریب هزینه 0 برای آن ها در نظر گرفته خواهد شد. جدول تکمیل شده به صورت زیر می باشد که بهینه ترین مسیر با رنگ آبی مشخص شده است.

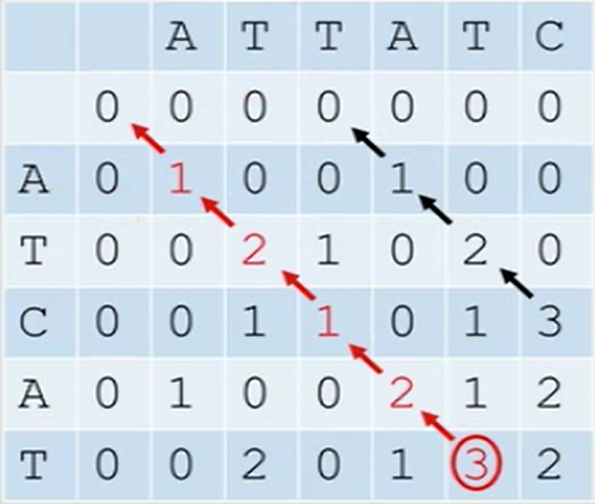


پیچیدگی زمانی این الگوریتم O(nm) و پیچیدگی فضای آن O(nm)

الگوریتم عقبگرد برای پیدا کردن بهینه ترین مسیر استفاده می شود و طر کار آن به این صورت است که اعداد پایین و سمت چپ را مقایسه می کند و وارد خانه ای از جدول می شود که عدد کوچکتری دارد.پیچیدگی زمانی این الگوریتم O(n+m) می باشد.

تطبیق محلی یا Local Alignment به یکسان بودنبخشی از رشته اول با رشته دوم گویند.برای این کار از جدول به صورت زیر استفاده می شود. فرض کنید دو رشته X = ATCAT و Y = ATTATC داریم و می خواهیم زیر رشته های یکسان دو رشته را پیدا کنیم. به صورت زیر عمل می کنیم.

در جدول اگر حروف عمودی و افقی با هم یکسان بودند عدد اریب را با 1 جمع و اگر یکسان نبودند منهای 1 می کنیم.برای راحتی کار در جدول می توانیم اعداد منفی را صفر در نظر بگیریم.در جدول زیر این الگوریتم پیاده سازی شده است.



در ترجمه ماشین یا تصحیح لغت و یا تشخیص گفتار ما نیاز داریم برای انجام جمله بندی از درست ما نیاز داریم که احتمال وقوع ساختار کلمه x بعد از کلمه y و یا عبارت x بعد از عبارت y را داشته باشیم. به هر کدام از این احتمال ها یک Language model و یا یک Grammer گفته می شود.مثلا در تشخیص گفتار:

P(I saw a van) > P(Eyes awe of an)

است.مثال:

This is transparent => P(This, is, transparent)

با توجه به فرمول احتمال شرطی:

P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)

داریم:

P(x1,x2,x3,...,xn) = P(x1)P(x2|x1)P(x3|x1,x2)....P(xn|x1,...,xn-1))

پس برای محاسبه احتمال وقوع یک جمله یا Chain Rule داریم:

P(“Its water is so transparent”) = P(Its) × P(water | Its) × P(is | its water) × P(so | Its water is) × P(transparent | Its water is so)

فرضیه Markov عملیات بالا را ساده می کند ولی در این حال باعث پایین آمدن دقت آن می شود.فرضیه به این صورت است که به جای محاسبه و مقایسه یک کلمه با کل جمله بهتر است کلمه مورد نظر با یک کلمه یا دو کلمه و یا با n کلمه اخر هر جمله مقایسه و محاسبه گردد.مثال جمله قبل:

P(so | Its water is) P(so | is)

N-Gram به معنای این است که کلمه مورد نظر با چند کلمه از جمله مورد نظر محاسبه شود که پر واضح است که هر چه قدر N بیشتر شود قدت در محاسبات نیز بیشتر خواهد شد.از gram های معروف و پرکاربرد می توان به Unigram که کلمه با یک کلمه دیگر از جمله محاسبه میگردد.bigram در اسن حالت کلمه مورد نظر با دو کلمه از جمله محاسبه میگردد.در trigram با سه کلمه و در Quadgram با چهار کلمه و همین طور الی آخر.

تخمین احتمال در بایگرم به این صورت انجام می شود که تعداد تکرار کلمه پیشین در متن در مخرج کسر و تعداد رخداد ساختار کلمه پسین بعد از کلمه پیشین در صورت کسر قرار می گیرد که فرمول آن به صورت زیر می باشد:

P(wi | wi-1) =c(wi-1, wi) / c(wi-1)

برای محاسبات خود از لگاریتم استفاده می کنیم برای جلوگیری از پاریز یا under flow و اینکه جمع کردن سریعتر از ضرب کردن است.

log(p1 × p2 × p3 × p4) = log p1 + log p2 + log p3 + log p4

احتمال بالاتری باید به جملات معمول تر و واقعی تر انتساب گردد.با استفاده از یک Traingin set پارامتر های مدمان را تخمین می زنیم،سپس مدل ارائه شدخ را روی Test set آزمایش می کنیم.

روش نخست روش Extrinsic Evaluation است. در این روش دو مدل A و B در تعداد کلمات به درستی اصلاح و ترجمه شده را با هم مقایسه می کنند.که روشی پر هزینه و زمان بر خواهد بود.از طرفی تا زمانی که Test Set کاملا شبیه به Trainging set نباشد تخمین خوبی نخواهد زد.

روش دوم Intrinsic Evaluation است که در این جزوه با راه حلی با نام Perplexity پیاده خواهد شد.Perplexity در معنای کاربردی،به قدرت انتخاب و حدس یک کلمه در یک عبارت و یا یک جمله اشاره دارد ودر معنای تئوری به معکوس احتمال وقوع کلمات در Test Set اشاره دارد که با تعداد کلمات به صورت توان معکوس نرمالسازی شده است.فرمول به صورت زیر می باشد:

PP(W) =