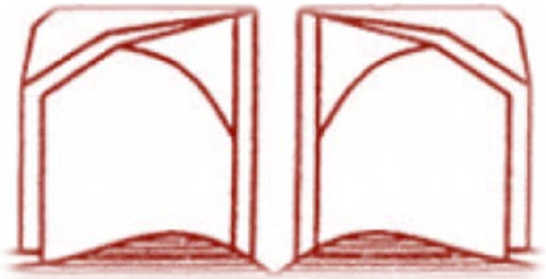


# روش‌های یادگیری ماشین در پردازش زبان طبیعی

## دسته‌بندی متون


هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران – دانشکده علوم و فنون نوین

# فهرست

# دسته‌بندی متون: کاربردها . . .

- تشخیص عنوان (موضوع)

Türkçe | Español | Русский | اردو ۰۲ آبان ۱۳۹۴

خبرگزاری جمهوری اسلامی ۱۳۱۳

صفحه اصلی | سیاسی | اقتصادی | اجتماعی | فرهنگی | علمی | ورزشی | بین الملل | استان ها | پژوهش | حوادث | عکس

تازه های خبری | هجوم نظامیان آل خلیفه به عزاداران عاشورایی - 52 دقیقه پیش

کد خبر: ۸۱۸۱۰۴۹۸ (۴۹۴۵۴۶۰) | تاریخ خبر: ۰۲/۰۸/۱۳۹۴ | ساعت: ۴:۴۳ | نسخه چاپی | ارسال به دوستان

## کتاب نذری برای دومین بار در اهواز توزیع شد

**اهواز- ایرنا- با همت فعالان اجتماعی و به منظور سوق دادن مردم به کتاب و کتابخوانی برنامه «نذر کتاب» برای دومین بار در شب های محرم در اهواز برگزار شد.**

رییس هیات مدیره کانون آموزش شهروندی برای توسعه (کاشت) در گفت وگو با ایرنا اظهار کرد: برنامه نذر کتاب در شب نهم ماه محرم برای دومین بار در طول دهه اول محرم در اهواز برگزار شد.

فهیمه صیادی افزود: این کانون دومین برنامه خود را در ساعت ۲۱ منطقه حصیر آباد اهواز برگزار کرد.

به گفته وی، اولین برنامه نذر کتاب در منطقه کوی نفت اهواز در شب هفتم محرم انجام شد.

وی هدف اولیه این کانون فرهنگی را ترویج نذر فرهنگی عنوان کرد و گفت: ترویج نذر فرهنگی برای این است که به جامعه آموزش داده شود که به جای غذای نذری می توانند کتاب یا کالاهای فرهنگی دیگر را جایگزین کنند.

صیادی با بیان اینکه این کانون پیش از این برنامه اهدای کتاب در روستاهای شادگان و ایذه را اجرا کرده است، گفت: نذر کتاب در محرم و در هیات های عزاداری برای نخستین بار انجام می شود.

وی اظهار کرد: از علاقه مندان به مشارکت در این برنامه ۲۴ میلیون ریال جمع آوری شد که بخشی از آن در برنامه های نذر کتاب توزیع شده و مابقی در برنامه های آتی این کانون توزیع خواهد شد.

- سیاسی
- مذهبی
- علمی
- هنری
- ورزشی
- ...

# دسته‌بندی متون: کاربردها . . .

## عقیده کاوی (دسته‌بندی نظرات کاربران)

### موافق/مخالف

**امید دلبر** (1393/1/24) : 1

فوق العادس

**محمود ابراهیمی جاویدی** (1393/1/22) : 0

من این لپ تاپ 5 ماه دارم و از همه نظر عالی مثل: کیفیت صفحه نمایش یا کیفیت لولاش برای لمسی بودنش و...

**مرتضی اکبری** (1393/1/17) : 1

یکی از مشکلاش ازنظر من کارت گرافیکشه.
منظورم باس کارت گرافیکه!

**هادی کریم** (1392/12/22) : 0

سلام..اصلا تو خریدنش شک نکنید.من خریدم خیلی عالیه.خیلی بیشتر از عکسش خوشگله.واقعا زیباست.اصلا لولایی دیده نمیشه واستحکامش خیلی بیشتره.به قیمتش می ارزه.صدای فنش هم اصلا اذیت نمیکنه.جون میده برای بازی

**ابودر هومن** (1392/12/17) : 5

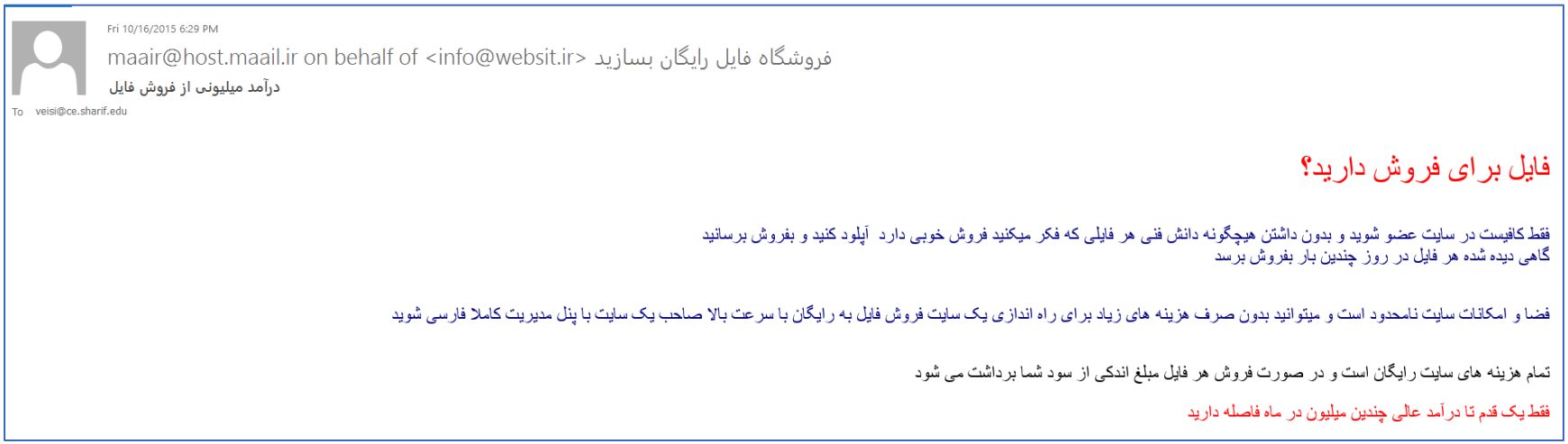
با سلام
بعد از 4ماه کار مداوم
از هر نظر عالی
قدرت خوبی داره.بسیار سریع
تاچ خیلی خوبی داره.کیفیت واقعا عالی.صفحه عجیبی داره که اصلا خسته نمیشین موقع کار باهاش
وزنش هم که عالیه.یه همراه همیشگی میشه براتون
مرسی

**هادی کریم** (1392/12/1) : 0

سلام..اصلا تو خریدنش شک نکنید.من خریدم خیلی عالیه.خیلی بیشتر از عکسش خوشگله.واقعا زیباست.اصلا لولایی دیده نمیشه واستحکامش خیلی بیشتره.به قیمتش می ارزه.صدای فنش هم اصلا اذیت نمیکنه.جون میده برای بازی

# دسته‌بندی متون: کاربردها . . .

- تشخیص هرزنامه‌ها
  - ایمیل دریافتی هرزنامه است یا نه؟



Fri 10/16/2015 6:29 PM

maair@host.maail.ir on behalf of <info@websit.ir> فروشگاه فایل رایگان بسازید

درآمد میلیونی از فروش فایل

To veisi@ce.sharif.edu

فایل برای فروش دارید؟

فقط کافیست در سایت عضو شوید و بدون داشتن هیچگونه دانش فنی هر فایلی که فکر میکنید فروش خوبی دارد آپلود کنید و بفروش برسانید
گاهی دیده شده هر فایل در روز چندین بار بفروش برسد

فضا و امکانات سایت نامحدود است و میتوانید بدون صرف هزینه های زیاد برای راه اندازی یک سایت فروش فایل به رایگان با سرعت بالا صاحب یک سایت با پنل مدیریت کاملا فارسی شوید

تمام هزینه های سایت رایگان است و در صورت فروش هر فایل مبلغ اندکی از سود شما برداشت می شود

فقط یک قدم تا درآمد عالی چندین میلیون در ماه فاصله دارید

# دسته‌بندی متون: کاربردها . . .

- تشخیص نویسنده/جنسیت نویسنده
  - نویسنده متن کیست؟ زن است یا مرد؟

چراغهای امامزاده، از دور در تاریکی؛ مثل چراغ خانه ای بود که تو را میخواهد. گرم، روشن و منتظر. سرم را به ضریح چسباندم. سلام آقا. دوسش دارم از بین این همه آدم ، فقط اون! شاید بچه گیام فقط برای ظاهرش بود ،اما روزی که به خاطر من ، دعوا کرد،دیدم جوونمرده. مثل قهرمونای قصه. وقتی منو سر مزار دوستش برد و گریه کرد، دیدم مهربونه.دل داره؛ و پاکه، مگه آدم چند بار میتونه دلشو هدیه بده؟ من هیچوقت روم نشده از خدا چیزی بخوام.اما این بار میخوام ! عمر در برابر عمر! از من نگیرش خدا! چیزی ندارم بت بدم،جز عشقی که خودت تو قلبم گذاشتی. پیرزن بخش زنانه گفت: تو اتاقشه. اما گفته کسی رو نمیبینه! گفتم:بگو دخترت اومده ! پشت در اتاقش بودم.از اتاقهای کوچک اجاره ای

چهارشنبه آخر ماه پیش وقتی از مسیر دورتر می رفت، سر یک کوچه ترمز کرد نگاهی به این طرف و آن طرف انداخت، بعد گفت: «ببخشید الان برمی گردم» و از ماشین پیاده شد. دوباره کمی این طرف و آن طرف را نگاه کرد، یک کوچه را تا نیمه رفت و برگشت بعد سوار شد و رفتیم. به دست هایش نگاه کردم، فرمان را آنقدر محکم گرفته بود که ترسیدم از جا کنده شود، اما لرزش دست هایش پیدا بود، پرسیدم «حالتون خوبه؟» گفت «نه.» نگاهش کردم و بعد برایم تعریف کرد. چهل و شش سال پیش عاشق دختر جوانی می شود. چهارشنبه آخر یک ماه دختر جوان به او می گوید خانواده اش اجازه نمی دهند با او ازدواج کند. راننده از دختر جوان می خواهد لااقل ماهی یک بار او را از دور ببیند. دختر جوان قول می دهد تا آخر عمر چهارشنبه آخر هر ماه سر این کوچه بیاید. چهل و شش سال دختر جوان چهارشنبه آخر هر ماه سر کوچه آمده، راننده او را از دور دیده و رفته است. از راننده پرسیدم «دختر جوان ازدواج کرد؟» نمی دانست. پرسیدم «آدرسشو دارین؟» نداشت. در این چهل و شش سال با او حتی یک کلمه هم حرف نزده بود فقط چهارشنبه های آخر هر ماه دختر جوان را دیده بود و رفته بود. راننده گفت «چهل و شش سال چهارشنبه آخر هر ماه اومد، ولی

# دسته‌بندی متون: کاربردها

## تشخیص زبان متن

- فارسی

زبان‌شناسی علمی است که به مطالعه و بررسی روشمند زبان می‌پردازد. در واقع، زبان‌شناسی می‌کوشد تا به پرسش‌هایی بنیادین همچون «زبان چیست؟»، «زبان چگونه عمل می‌کند و از چه ساخت‌هایی تشکیل شده‌است؟»، «انسان‌ها چگونه با یکدیگر ارتباط برقرار می‌کنند؟»، «زبان آدمی با سامانه ارتباطی دیگر جانوران چه تفاوتی دارد؟»، «کودک چگونه سخن گفتن می‌آموزد؟»، «زبان بشر چگونه تکامل یافته‌است؟»، «زبان‌ها چه قرابتی با یکدیگر دارند؟»، «ویژگی‌های مشترک زبان‌های جهان کدامند؟»، «انسان چگونه می‌نویسد و از چه راهی زبان نانوشتاری را واکاوی

- کردی

زمانەوانی ، زمانناسی یا (بە ئینگلیسی Linguistics ) بریتییە لە زانستی لێکۆڵینەوەی ورد و هەمەلایەنەی زمان. بە واتایەکی تر زمانەوانی بە دوای دۆزینەوەی ئەم پرسیارانەدایە: «زمان چیە؟»، «چۆن کار دەکا و لە چی پێکهاتووە؟»، «مرۆڤەکان چۆن پەیوەندی بە یەکەوە دەکەن؟»، «زمانی مرۆڤ لە چیدا لە گەڵ سیستەمە پێوەندیەکانی گیانداران جیاوازە؟ »، «منداڵ چۆن زمان دەگرێ؟»، «زمانی مرۆڤ چۆن گەشەی سەندووە؟»، «خزمایەتی نێوان زمانەکان چۆنە؟»، «تایبەتمەندیە هاوبەشەکانی زمانەکانی دنیا کامانەن؟»، «مرۆڤ چۆن دەنووسێ و چلۆن زمانی نووسین لێک

- عربی

اللغويات أو اللسانيات أو الألسن(هامش) هي العلم الذي يهتم بدراسة اللغات الإنسانية ودراسة خصائصها وتراكيبها ودرجات التشابه والتباين فيما بينها. أما اللغوي فهو الشخص الذي يقوم بهذه الدراسة. ظهرت في القرن ١٩م وهي متعلقة بدراسة اللغة.جاءت بفكرة رئيسة مع العالم دي سوسير فمع علمنة الثورة الصناعية اراد علمنة اللغة أيضا في كتابه /محاضرات في اللغويات العامة/فاللغة عنده تحمل هويات من القيم الدين.المحيط.الثقافة.الفكر الفلسفي.

- اردو

لسانیات (Linguistics)ایک ایسا مضمون ہے جس میں انسانی زبانوں کا ، زبانوں کی موجودہ صورت کا اور زبانوں میں وقت کے ساتھ ساتھ ہونے والی تبدیلیوں کا مطالعہ کیا جاتا ہے۔ اس علم میں مختلف زبانوں کی آپس میں مشابہت کے بارے میں مطالعہ کے ساتھ ساتھ اس چیز کا بھی مطالعہ کیا جاتا ہے کہ زبانوں کا اس دنیا کی دیگر چیزوں کے ساتھ کیا تعلق ہے۔ گویا لسانیات در اصل وہ علم ہے جس میں صرف انسانی زبان پر بحث کی جاتی ہے اور اس کے علاوہ کسی دوسری نظام کا مطالعہ نہیں کیا جاتا۔ ۔اردومیں اسم" لسان" کے ساتھ "یات" بطور لاحقۂ نسبت لگانے سے لسانیات کی اصطلاح بنتی ہی ۔ یہ اصطلاح اردو میں بطور اسم استعمال ہوتی ہے اور علم اللسان یا علم زبان

- پشتو

انسانان د زمانې په لمن کې پیښې زیږوي او مورخین د هغو پیښو په پرله پسې ډولو تاریخ لیکنه کوي . تر ټولو ښه تاریخ ، هغو مورخینو لیکلی چې پیښې یې په هغه ډول چې واقع شویدي ، د زمانې په قید کې راټولې کړي او د پوهیدو وړ یې ګرځولي دي . تاریخ لیکنه د هر هیواد په ملت جوړونه کې ستر رول لري ، هغومره چې د ټولنې سیاست پوهان ، واکمنان ، ټولواکان د ټولنې په شتون ، وده او پراختیا او یا هم بدحالی کې ونډه لري د مورخینو ونډه هم په هماغه پیمانه ده . تاریخ د ټولنې له پاره هغه هنداره ده چې هر لوستونکی کولای شي چې د هغې ټولنې ځیره ، قد



Hadi Veisi (h.veisi@ut.ac.ir)

# دسته‌بندی متون . . .

- **مساله**
  - **ورودی**
    - یک سند d
    - تعداد مشخص (و ثابت) دسته $C = \{c_1, c_2, \ldots, c_J\}$
  - **خروجی**
    - پیش‌بینی یکی از دسته‌ها $c \in C$ برای سند d
- **راه حل**
  - **استفاده از تعدادی قانون برای تعیین هر دسته بر اساس ویژگی‌های مشخصی از متن**
    - استخراج مجموعه قوانین مناسب زمان بر، پرهزینه و گاهی مشکل است
  - **استفاده از یادگیری ماشین**
    - یادگیری قوانین از روی داده‌ها

# دسته‌بندی متون . . .

- استفاده از یادگیری ماشین (با ناظر) . . .
  - فاز آموزش
    - تعداد مشخص (و ثابت) دسته $C = \{c_1, c_2, \ldots, c_J\}$
    - ورودی: داده آموزش = تعداد $M$ نمونه سند دارای برچسپ (دسته سندها مشخص است)
      - $(d_1,c_1),\ldots,(d_m,c_m)$
    - خروجی: مدل‌های دسته‌ها/تابع جداکننده دسته‌ها
  - فاز استفاده (آزمون)
    - پیش‌بینی یکی از دسته‌ها $c \in C$ برای سند $d$
    - ورودی: یک سند $d$ با دسته نامشخص + مدل‌های دسته‌ها/تابع جداکننده دسته‌ها
    - خروجی: دسته $c \in C$ پیش‌بینی شده برای سند $d$

# دسته‌بندی متون . . .

- روش‌های یادگیری ماشین (با ناظر)
  - بیز ساده (Naïve Bayes)
  - رگرسیون (Regression)
  - نزدیک‌ترین همسایه (k-Nearest Neighbors)
  - شبکه عصبی مصنوعی (Artificiel Neural Network)
  - ماشین بردار پشتیبان (SVM: Support Vector Machines)
  - مدل مخفی مارکوف (HMM: Hidden Markov Model)
  - ...

# دسته‌بندی متون: بیز ساده . . .

- **استفاده از قانون بیز**
  - **برای سند Doc و دسته Class داریم**

$$P(Class \mid Doc) = \frac{P(Doc \mid Class)P(Class)}{P(Doc)}$$

    - دسته‌ای انتخاب می شود که احتمال $p(Class \mid Doc)$ بالاتری داشته باشد

  - **تخمین بیشینه احتمال پسین (MAP: maximum a posteriori)**

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d) = \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)} = \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$

**عدم تاثیر مقدار مخرج بر تصمیم گیری**

  - **با فرض نمایش سند d به صورت بردار $[x_1, x_2, \ldots, x_n]$**

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c) = \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

# دسته‌بندی متون: بیز ساده . . .

- بیز ساده (Naïve Bayes) . . .
  - فرض اصلی: مستقل بودن ویژگی‌ها از همدیگر
    - استقلال شرطی: ویژگی به شرط دسته $P(x_i|c_j)$

$$P(x_1,\ldots,x_n \mid c) = P(x_1 \mid c)\times P(x_2 \mid c)\times P(x_3 \mid c)\times\ldots\times P(x_n \mid c)$$

- بنابراین

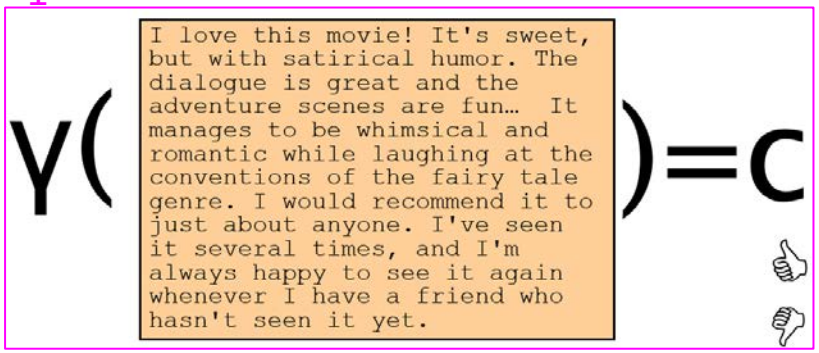
$$c_{MAP} = \underset{c\in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

$$c_{NB} = \underset{c_j\in C}{\operatorname{argmax}}\, P(c_j)\prod_{i=1}^{n} P(x_i \mid c_j)$$

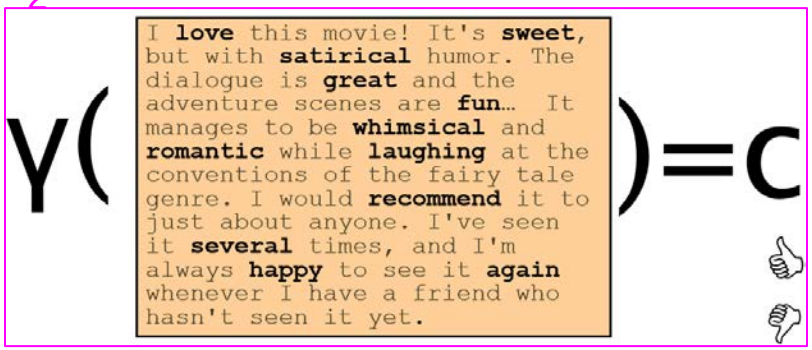
# دسته‌بندی متون: بیز ساده . . .

- بیز ساده (Naïve Bayes)
  - فرض کیسه لغات: محل کلمات در متن مهم نیست

1

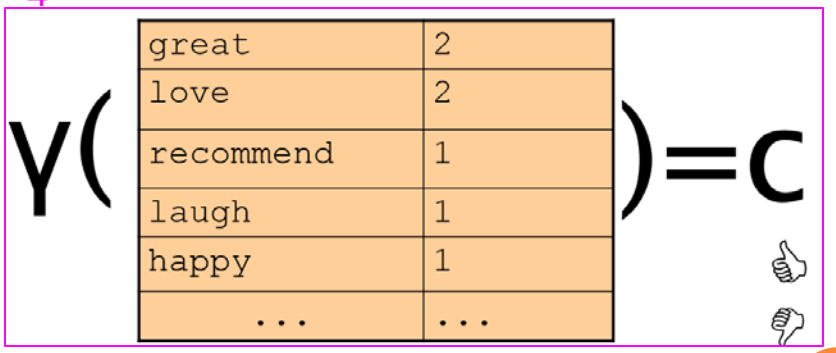



2




3


4



# دسته‌بندی متون: بیز ساده (آموزش) . . .

- آموزش: تخمین پارامترهای مدل
  - $P(c_j)$: احتمال دسته jام
  - $P(x_i|c_j)$: احتمال ویژگی (کلمه iام) به شرط دسته jام
  - تخمین شباهت بیشینه (MLE) برای احتمال دسته jام $P(c_j)$
    - تعداد مستنداتی که دسته jام هستند به تعداد کل مستندات
      - در بیشتر کاربردها از توزیع یکنواخت (برابر بودن احتمال برای همه دسته ها)

$$\hat{P}(c_j) = \frac{DocCount(c = c_j)}{N_{Doc}}$$

  - تخمین شباهت بیشینه (MLE) برای احتمال ویژگی (کلمه iام) به شرط دسته jام $P(x_i|c_j)$
    - تعداد تکرار کلمه $w_i$ در کل مستنداتی که دسته jام هستند، به تعداد کل کلمات موجود در مستندات دسته jام

$$\hat{P}(x_i = w_i \mid c_j) = \frac{Count(w_i, c_j)}{\sum_{w \in V} Count(w, c_j)}$$

  - در عمل: تبدیل کلیه فایل‌های هر دسته به یک فایل برای محاسبه ساده‌تر

# دسته‌بندی متون: بیز ساده (آموزش) . . .

- **مشکل**
  - **طبق قانون بیز ساده، اگر یکی از احتمال‌های $P(x_i|c_j)$ صفر باشد، کل احتمال برای آن دسته صفر خواهد بود (مستقل از اینکه مابقی $P(x_i|c_j)$ها چه مقداری دارند)**
    - در داده‌های آموزش ویژگی (کلمه) $x_i$ در دسته $c_j$ رخ نداده است

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i=1}^{n} P(x_i \mid c_j)$$

- **مثال: عقیده کاوی (موافق/مخالف)**
  - صفر شدن احتمال برای کلمه «خفن» در دسته «موافق» به دلیل عدم وقوع این کلمه در داده آموزش

$$\hat{p}\left(\text{خفن} \middle| \text{موافق}\right) = \frac{Count(\text{"خفن"}, \text{موافق})}{\sum_{w \in V} Count\left(w, \text{موافق}\right)} = 0$$

  - باعث صفر شدن احتمال انتخاب دسته «موافق» در متن (نظر) زیر می‌شود: «طراحی زیبا و کیفیت صفحه خفن به همراه باطری قوی. این گوشی محبوب منه»

# دسته‌بندی متون: بیز ساده (آموزش) . . .

- رفع مشکل صفر شدن احتمال‌ها
  - اضافه کردن مقداری ثابت (مثلا ۱) به تعداد همه کلمات واژگان در محاسبه احتمال

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} \left( count(w, c_j) \right)}$$

$$\hat{P}(w_i \mid c_j) = \frac{Count(w_i, c_j) + 1}{\sum_{w \in V} \left( Count(w, c_j) + 1 \right)} = \frac{Count(w_i, c_j) + 1}{\left( \sum_{w \in V} Count(w, c_j) \right) + |V|}$$

اندازه واژگان= تعداد کلمات

  - می توان به جای ۱ مقدار ثابتی مانند α به تعداد اضافه کرد

$$\hat{P}(w_i \mid c_j) = \frac{Count(w_i, c_j) + \alpha}{\left( \sum_{w \in V} Count(w, c_j) \right) + \alpha |V|}$$

# دسته‌بندی متون: بیز ساده (الگوریتم آموزش) . . .

- گام ۱: در نظر گرفتن (استخراج) واژگان V و تعداد C دسته
  - اندازه واژگان=تعداد لغات = $|V|$
- گام ۲: محاسبه $P(c_j)$ها و $P(w_i | c_j)$
  - ۲-۱ برای هر دسته $c_j$
    - ۲-۱-۱ محاسبه احتمال $P(c_j)$
      - تقسیم تعداد سندهای دسته $c_j$ به کل سندها
        - $docs_j \leftarrow$ all docs with class $=c_j$

$$\hat{p}(c_j) = \frac{|Docs_j|}{|Total\ No\ of\ Docs|}$$

    - ۲-۱-۲ ایجاد یک فایل حاوی تمام متون آموزش دسته $c_j$: $Text_j \leftarrow$ single doc containing all $docs_j$
    - ۲-۱-۳ برای هر کلمه $w_i$ در واژگان $V$
      - ۲-۱-۳-۱ قرار بده $n_{ij} \leftarrow$ # of occurrences of $w_i$ in $Text_j$ و $n_j \leftarrow$ # of all words in $Text_j$
        - تعداد رخداد کلمه $w_i$ در متون دسته $c_j$
      - ۲-۱-۳-۲ محاسبه احتمال $P(w_i | c_j)$
        - تقسیم فراوانی کلمه $w_i$ در دسته $c_j$ بر کل کلمات این دسته

$$\hat{p}(w_i|c_j) = \frac{n_{ij} + \alpha}{n_j + \alpha|V|}$$

# دسته‌بندی متون: بیز ساده (مثال) . . .

- هدف: دسته‌بندی متون به دو دسته «ایران» و «آمریکا»
- داده آموزش و آزمون
  - تعداد واژه‌ها= $|V| = ۱۰$

| آموزش | | |
|---|---|---|
| شماره سند | متن | دسته |
| ۱ | تهران شهر بزرگ ایران | ایران |
| ۲ | ایران بزرگ آمریکا | ایران |
| ۳ | تهران فارسی زبان ایران | ایران |
| ۴ | واشنگتن شهر آمریکا | آمریکا |
| ۵ | زبان نیویورک انگلیسی | آمریکا |
| آزمون | | |
| ۱ | زبان فارسی انگلیسی تهران | ؟ |

- آموزش . . .
  - احتمال دسته‌ها

تعداد سندهای «ایران»

$$p(c = \text{ایران}) = \frac{3}{5}$$

تعداد کل سندها

$$p(c = \text{آمریکا}) = \frac{2}{5}$$

# دسته‌بندی متون: بیز ساده (مثال) . . .

| آموزش | | |
|---|---|---|
| شماره سند | متن | دسته |
| ۱ | تهران شهر بزرگ ایران | ایران |
| ۲ | ایران بزرگ آمریکا | ایران |
| ۳ | تهران فارسی زبان ایران | ایران |
| ۴ | واشنگتن شهر آمریکا | آمریکا |
| ۵ | زبان نیویورک انگلیسی | آمریکا |

- آموزش
  - احتمال کلمات در دسته‌ها

$$p(w_i|c_j) = \frac{n_{ij} + 1}{n_j + |V|}$$

  - تعداد کلمات دسته «ایران» = ۱۱
  - تعداد کلمات دسته «آمریکا» = ۶

| کلمه | P(کلمه \| ایران) | P(کلمه \| آمریکا) |
|---|---|---|
| تهران | P(تهران \| ایران)=(2+1)/(11+10)=3/21 | P(تهران \| آمریکا)=(0+1)/(6+10)=1/16 |
| شهر | P(شهر \| ایران)=(1+1)/(11+10)=2/21 | P(شهر \| آمریکا)=(1+1)/(6+10)=2/16 |
| بزرگ | P(بزرگ \| ایران)=(2+1)/(11+10)=3/21 | P(بزرگ \| آمریکا)=(0+1)/(6+10)=1/16 |
| ایران | P(ایران \| ایران)=(3+1)/(11+10)=4/21 | P(ایران \| آمریکا)=(0+1)/(6+10)=1/16 |
| آمریکا | P(آمریکا \| ایران)=(1+1)/(11+10)=2/21 | P(آمریکا \| آمریکا)=(1+1)/(6+10)=2/16 |
| فارسی | P(فارسی \| ایران)=(1+1)/(11+10)=2/21 | P(فارسی \| آمریکا)=(0+1)/(6+10)=1/16 |
| زبان | P(زبان \| ایران)=(1+1)/(11+10)=2/21 | P(زبان \| آمریکا)=(1+1)/(6+10)=2/16 |
| واشنگتن | P( واشنگتن \| ایران)=(0+1)/(11+10)=1/21 | P( واشنگتن \| آمریکا)=(1+1)/(6+10)=2/16 |
| نیویورک | P(نیویورک \| ایران)=(0+1)/(11+10)=1/21 | P(نیویورک \| آمریکا)=(1+1)/(6+10)=2/16 |
| انگلیسی | P(انگلیسی \| ایران)=(0+1)/(11+10)=1/21 | P(انگلیسی \| آمریکا)=(1+1)/(6+10)=2/16 |

# دسته‌بندی متون: بیز ساده (مثال) . . .

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i=1}^{n} P(x_i \mid c_j)$$

- آزمون
  - محاسبه $p(\text{class} \mid \text{doc})$ برای doc=«زبان فارسی انگلیسی تهران»
  - محاسبه برای class=ایران

$$p\left(\text{ایران} \middle| \text{زبان فارسی انگلیسی تهران}\right) = \text{p}\left(\text{ایران}\right).\text{p}\left(\text{زبان}\middle|\text{ایران}\right).\text{p}(\text{فارسی}|\text{ایران}).\ \text{p}(\text{انگلیسی}|\text{ایران}).\ \text{p}\left(\text{تهران}\middle|\text{ایران}\right)$$

$$= \frac{3}{5} \cdot \frac{2}{21} \cdot \frac{2}{21} \cdot \frac{1}{21} \cdot \frac{3}{21} = 0.0004$$

  - محاسبه برای class=آمریکا

$$p\left(\text{آمریکا} \middle| \text{زبان فارسی انگلیسی تهران}\right) = \text{p}\left(\text{آمریکا}\right).\text{p}\left(\text{زبان}\middle|\text{آمریکا}\right).\text{p}(\text{فارسی}|\text{آمریکا}).\ \text{p}(\text{انگلیسی}|\text{آمریکا}).\ \text{p}\left(\text{تهران}\middle|\text{آمریکا}\right)$$

$$= \frac{2}{5} \cdot \frac{2}{16} \cdot \frac{1}{16} \cdot \frac{2}{16} \cdot \frac{1}{16} = 0.0002$$

  - بنابراین داریم $p\left(\text{ایران}\middle|\text{doc}\right) > p\left(\text{آمریکا}\middle|\text{doc}\right)$
    - سند متعلق به دسته «ایران» است

# دسته‌بندی متون: بیز ساده . . .

- ویژگی‌ها
  - خیلی سریع است
  - عدم نیاز به حافظه زیاد
  - مقاوم به ویژگی‌های غیرمرتبط (به علت لغو اثر این ویژگی‌ها توسط همدیگر)
  - کارایی خوب در کاربردهایی که ویژگی‌هایی با اهمیت مشابه دارند
  - در صورت برقراری شرط استقلال، این روش تبدیل به دسته‌بند بهینه خواهد شد
  - درارای کارایی خوب برای دسته‌بندی متن
  - دارای ارتباط مفهومی با مدل زبانی
  - هنوز نیاز به روشی با کارایی بالاتر
    - روش‌های دیگر یادگیری ماشین

# دسته‌بندی متون: بیز ساده (ارتباط با مدل زبانی) . . .

- ویژگی‌های مورد استفاده بیز ساده در دسته‌بندی متون = کلمات
  - امکان استفاده از هر نوع ویژگی‌ای در این روش
  - اگر از همه کلمات متن در بیز ساده استفاده کنیم، آنگاه دسته‌بندی مشابه محاسبه احتمال در مدل زبانی است
- مدل زبانی یک تایی (Unigram)
  - محاسبه احتمال برای هر کلمه = $P(w_i)$
  - محاسبه احتمال برای هر جمله با ضرب احتمال تک تک کلمات در همدیگر

$$P(W) = P(w_1 w_2 \cdots w_m) \cong \prod_{i=1}^{m} P(w_i \mid w_{i-(k)} \cdots w_{i-1}) \cong \prod_{i=1}^{m} P(w_i) = P(w_1)P(w_2)P(w_3)\cdots P(w_m)$$

- در صورتی که مقدار احتمال هر کلمه $P(w_i)$ برای یک دسته محاسبه شود = $P(w_i \mid c_j)$

$$P(Sentence \mid c_j) = P(W \mid c_j) = \prod_{i=1}^{m} P(w_i \mid c_j)$$

# دسته‌بندی متون: بیز ساده (ارتباط با مدل زبانی) . . .

- مثال
  - محاسبه احتمال هر کلمه در هر دسته

| Model Positive | |
|---|---|
| 0.1 | I |
| 0.1 | love |
| 0.01 | this |
| 0.05 | fun |
| 0.1 | film |

| Model Negative | |
|---|---|
| 0.2 | I |
| 0.001 | love |
| 0.01 | this |
| 0.005 | fun |
| 0.1 | film |

| I | love | this | fun | film |
|---|---|---|---|---|
| 0.1 | 0.1 | 0.01 | 0.05 | 0.1 |
| 0.2 | 0.001 | 0.01 | 0.005 | 0.1 |

$$P(s|pos) > P(s|neg)$$

# دسته‌بندی متون: بیز ساده (نکات کاربردی) . . .

- رفع مشکل Underflow
  - ضرب تعداد زیادی مقدار احتمال (با مقادیر کوچک) = صفر شدن

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i=1}^{n} P(x_i \mid c_j)$$

  - استفاده از لگاریتم برای رفع مشکل

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

  - دسته‌ای با لگاریتم احتمال بیشتر (در صورت عدم نرمال کردن)، دسته برنده است

- دارای کارایی خوب با داده کم

# دسته‌بندی متون: بیز ساده

- مثال (فارسی)
  - تعداد ۳۱۹۳ سند
  - ۸ موضوع (عنوان): ادبی، مذهبی، اقتصادی، هنری، پزشکی، تاریخی، سیاسی و ورزشی
  - تعداد کلمه‌ها (ویژگی‌ها): ۲۹۰۰ کلمه

- کارایی روش‌های مختلف (بر حسب درصد)

| | HMM | ANN | SVM | Naïve Bayes | 50-NN | DT |
|---|---|---|---|---|---|---|
| میانگین Recall | ۸۳ | ۷۴ | ۶۹ | ۶۳ | ۶۴ | ۵۹ |
| میانگین Precision | ۷۶ | ۷۳ | ۷۳ | ۶۰ | ۷۰ | ۵۵ |
| میانگین معیار F | ۷۹ | ۷۳ | ۷۱ | ۶۱ | ۶۷ | ۵۷ |

Gharavi E., Veisi H., A Hidden Markov Model for Persian Text Classification, 3rd National Computational Linguistics Conference, Tehran-Iran, 2014