Data Mining

ACECR

ACECR, Shahid Beheshti University of Medical Sciences, Medical Data Mining Group

بخش اول:
مفاهیم اولیه

پیشگفتار

- انقلاب دیجیتال
- انفجار اطلاعات
- عصر رایانه
- عصر اطلاعات و ارتباطات
- شبکه جهانی وب
- سیستم‌های یکپارچه اطلاعاتی
- سیستم‌های یکپارچه بانکی
- تجارت الکترونیک

افزایش لحظه به لحظه به حجم داده ها در پایگاه داده

پیشگفتار

- افزایش روزافزون حجم داده ها در پایگاه داده ها

  هر دو سال دو برابر

- لازمه موفق بودن سازمان ها: تحلیل حداقل ۷٪ داده هایشان
  - **در عمل تحلیل کمتر از ۱درصد داده ها در سازمان ها**

- غرق در داده ها و تشنه دانش

DATA MINING

- Mine به معنای استخراج از منابع نهفته و با ارزش زمین
- ادغام این کلمه با Data به معنی جستجویی عمیق از داده های قابل دسترس با حجم زیاد

هدف : یافتن اطلاعات نهفته مفید

پیشگفتار

داده = طلا

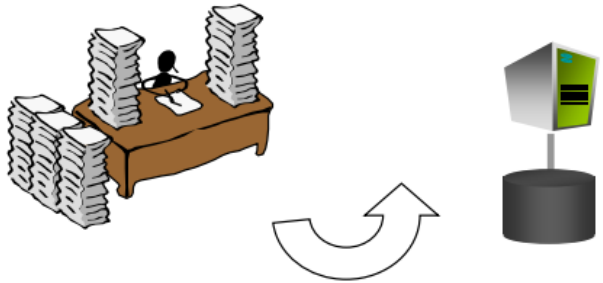
هرچه داده های موجود در پایگاه داده شرکت افزایش پیدا کند،افزایش سرمایه

داده کاوی:

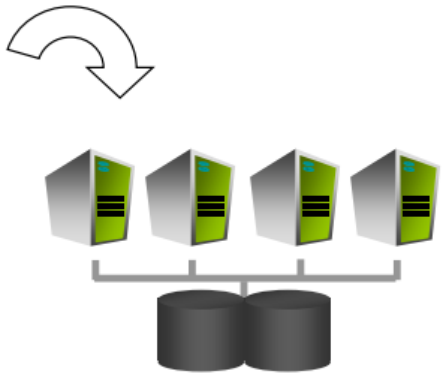
تبدیل داده های حجیم و تمیز نشده به اطلاعات مفید و گزارش های مناسب برای تصمیم گیری

پیشگفتار

ضرورت کشف و استخراج سریع و دقیق دانش از پایگاه داده ها



داده ها در اکثر سازمان‌ها به سرعت درحال جمع آوری و ذخیره شدن



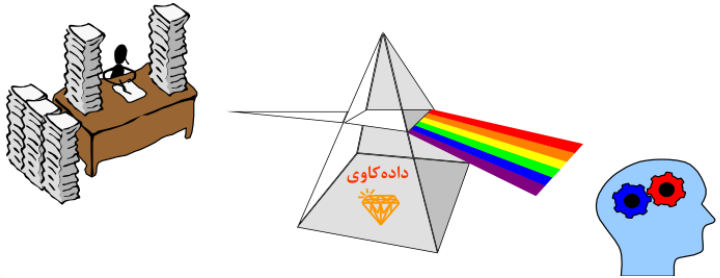
علاقه سازمان‌ها به استفاده از این منبع بزرگ و ارزشمند برای رشد و توسعه خود

پیشگفتار

- نیاز به طراحی سیستم های قادر به اکتشاف سریع اطلاعات مورد علاقه کاربران
  با تاکید بر حداقل مداخله انسانی
- روی آوردن به روش های تحلیل متناسب با حجم داده های حجیم

- **داده کاوی مهم ترین فناوری برای بهره وری موثر ، صحیح و سریع از داده های حجیم**

انبوه داده ها و فقر دانش



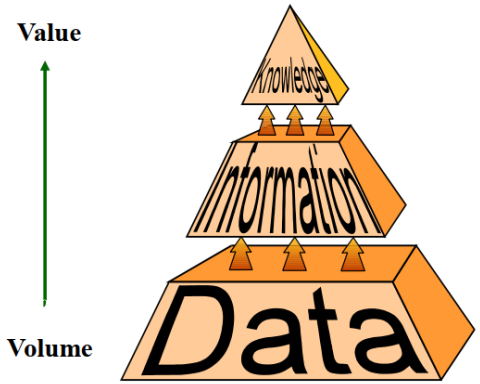
داده های خام و اعداد و ارقام به تنهایی هیچ کمکی نمی کنند.

علی رغم حجم انبوه داده ها سازمان ها با فقر دانش روبرو هستند.

راه حل :

علم داده کاوی

داده کاوی: استخراج اطلاعات و دانش و کشف الگوهای پنهان از پایگاه داده های بسیار بزرگ



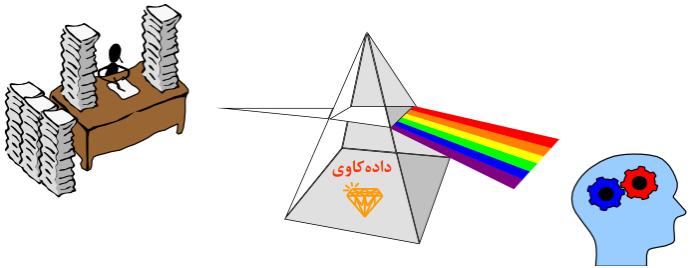
- تبدیل داده ها به اطلاعات با پردازش بسیار جزئی
- دانش سطح بالاتری نسبت به داده و اطلاعات
- دانش یک مفهوم ذهنی در مورد یک مسئله.

پیشگفتار

- **داده کاوی**

  - تحلیل و آنالیز مجموعه داده های بزرگ
  - استخراج اطلاعات معتبر ، یافتن الگوهای مفید و الگوهایی میان داده ها
  - یافتن ارتباطات غیر قابل انتظار از پیش ناشناخته ، قابل فهم و قابل اعتماد برای صاحب داده

مدل : ارتباطات پیدا شده



پیشگفتار

- جریان پایدار اطلاعات داده
- ارائه اطلاعات مبتنی بر حقیقت
- امکان بررسی یک قسمت از داده
- انجام آسان تست ها

پیشگفتار

- بنابر اعلام دانشگاه MIT دانش نوین داده کاوی (Data mining) یکی از ده دانش در حال توسعه
- انقلاب تکنولوژیکی در دهه آینده

- عدم وجود حد و مرزی برای کاربرد این دانش
- زمینه های کاری این دانش از ذرات کف اقیانوس‌ها تا اعماق فضا

سیر تاریخی

- اواخر دهه ۸۰ میلادی ، شروع تلاش برای استخراج و استفاده از اطلاعات پایگاه های داده ای
- آغاز دهه ۹۰ فرآیند داده کاوی عنوان یک علم مطرح شد
- مطرح شدن اصطلاح داده کاوی به طور رسمی اولین بار توسط « فیاض »
- در اولیـن کنفرانس بین المللی « کشف دانش و داده کاوی » در سال ۱۹۹۵
- از سال ۱۹۹۵ ورود جدی به مباحث آمار

- داده کاوی: نگرشی نو ، به مسئله استخراج اطلاعـــات از پایگـــاه داده ها

داده کاوی چیست

- استفاده از عبارات گوناگون برای یک فرایند واحد

سال	نام	استفاده کنندگان
۱۹۶۰	لایروبی داده، ماهی گیری داده	کارشناسان آمار
۱۹۸۹	کشف دانش در پایگاه های داده	هوش مصنوعی،جامعه کاربری یادگیری ماشین
۱۹۹۰	داده کاوی	جامعه پایگاه داده،کسب و کار
سایر اسامی	باستان شناسی داده،هرس اطلاعات،کشف اطلاعات،استخراج دانش	

داده کاوی علم بین رشته ای

داده کـاوی پل ارتباطی میان علوم مختلف

داده کاوی درختی است با ریشه در تکنولوژی های دیگر

مرزهای این رشته ها در داده کاوی مبهم ودارای اشتراک های فراوان

ریشه‌های داده کاوی در

آمار

هوش مصنوعی

یادگیری ماشین

و...

جایگاه داده کاوی در میان علوم مختلف

- **ریشه های داده کاوی میان سه خانواده از علوم**

- آمارکلاسیک مهمترین این خانواده ها
- بدون آمار هیچ داده کاوی وجود نخواهد داشت بطوریکه آمار اساس اغلب تکنولوژیی می باشد که داده کاوی بر روی آنها بنا می‌باشد
- **هدف آمار:** تحلیل و آنالیز دیتای موجود و نتیجه‌گیری بر اساس همین دیتا

- هوش مصنوعی دومین خانواده‌ای که داده کاوی به آن تعلق دارد
- هوش مصنوعی که برپایه روش های ابتکاری می باشد و با آمار ضدیت دارد
- تلاش هوش مصنوعی در جهت ایجاد فرآیندی مانند فکر انسان برای حل مسائل آماری
- **هدف یادگیری ماشین:**پیش بینی بر اساس الگوی استخراج شده از دیتا

- یادگیری ماشین سومین خانواده ی داده کاوی:
- یادگیری ماشین مفهوم دقیق تر اجتماع آمار و هوش مصنوعی می باشد.
- یادگیری ماشین مخلوطی از روش های ابتکاری هوش مصنوعی به همراه تحلیل آماری پیشرفته
- یادگیری ماشین داده‌کاوی را قادر به دریافت نتیجه می‌نماید

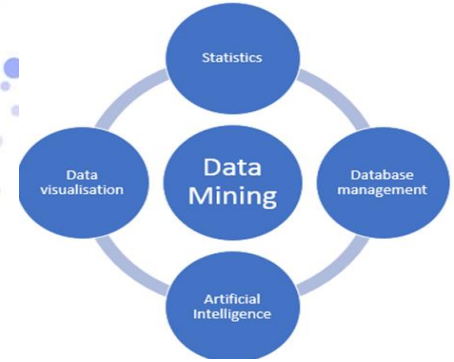
کاربرد یادگیری ماشین و آمار در داده کاوی:

- داده کاوی، یادگیری ماشین و سایر مفاهیم مرتبط آن، برآمده از علم آمار و احتمال

- داده کاوی جانشین تکنیک های آماری سابق نیست بلکه وارث آنهاست
- داده کاوی تغییر و گسترش تکنیک های سابق برای متناسب سازی آنها با حجم داده ها و مسائل امروزی
- داده کاوی ترکیب تکنیک های کلاسیک با الگوریتم های جدید مثل شبکه های عصبی و درخت تصمیم گیری.

- داده کاوی راهکاری برای مسائل تجاری امروز به کمک تکنیک های آماری و هوش مصنوعی
  برای افراد حرفه ای با هدف ایجاد یک مدل پیش بینی

شباهت ها و تفاوت های داده کاوی و آمار



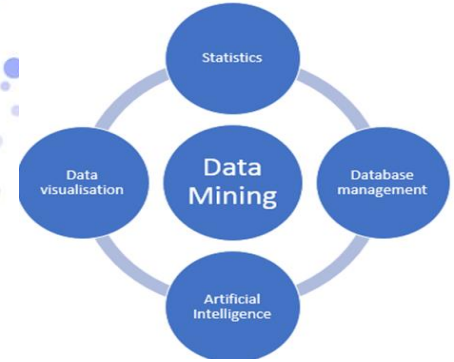
- تفاوت ها:

- علم آمار و احتمال مبتنی بر نمونه گیری / داده کاوی و یادگیری ماشین، مبتنی بر تحلیل کل
- علم آمار و احتمال: نمونه گیری از جامعه، استنتاج های به حد کافی خوب و سپس تعمیم آن

- در علم آمار و احتمال، ابتدا فرضیه مطرح می شود و سپس برای اثبات یا رد فرضیه داده پیرامون فرضیه جمع آوری می شود
- در داده کاوی بر اساس الگوی ذاتی نهفته در داده ها، دانش و الگوی کسب شده تفسیر می شود.

- داده کاوی نتایج غیرمنتظره ایجاد می کند
- در آمار نتایج کسب شده یا در محدوده آزمون فرض هست یا نیست.

شباهت ها و تفاوت های داده کاوی و آمار



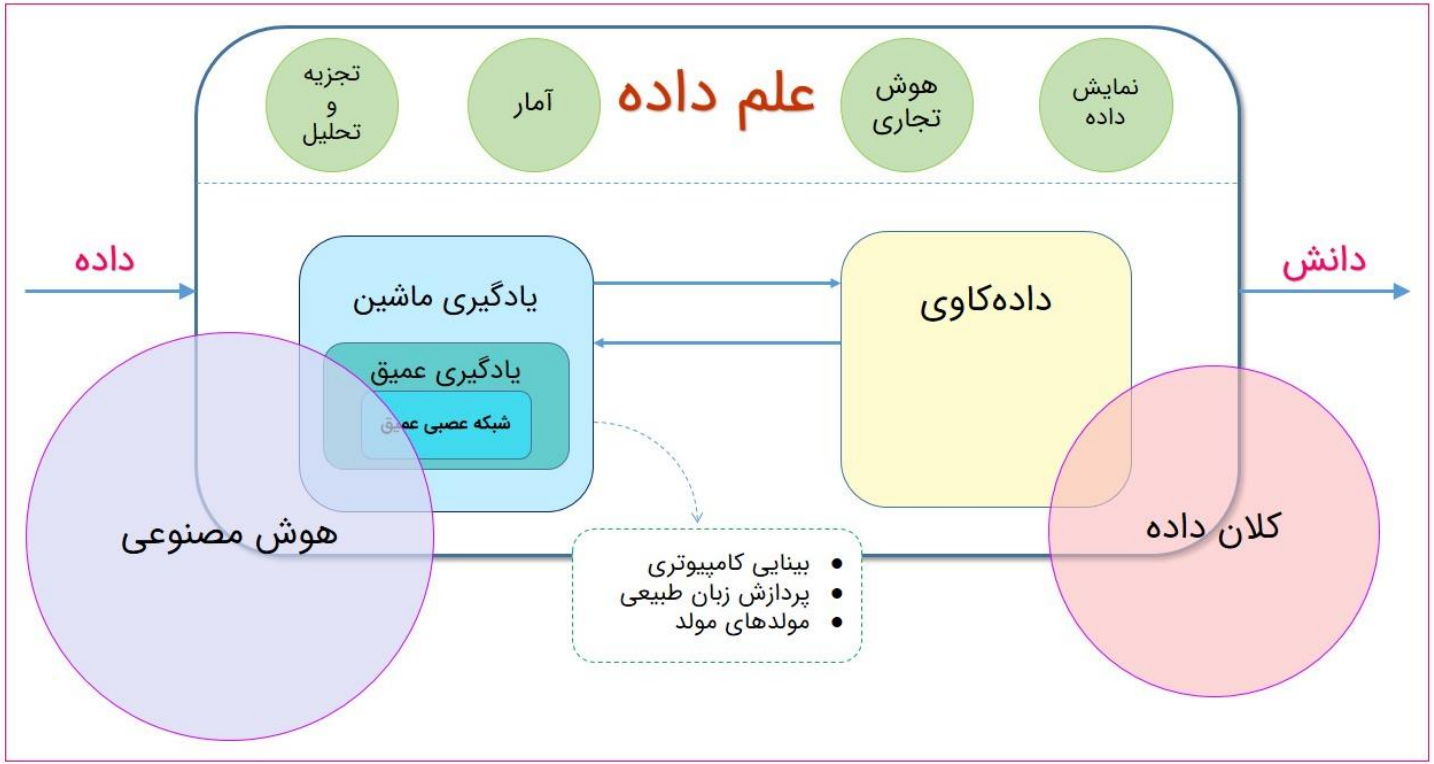
- تعداد ویژگی ها در آمار محدود بوده
- ولی در داده کاوی امکان پردازش داده ها با ویژگی های بزرگ تر

- فرض اولیه اساسی در آمار، نرمال بودن داده هاست، در داده کاوی، چنین فرضی وجود ندارد.
- هر چند در مباحث پیشرفته آماری، چنین فرض هایی وجود ندارد ولی شاکله اساسی مفاهیم آماری مبتنی بر توزیع نرمال است
- اساسا چنین فرضی در داده کاوی مطرح نیست.

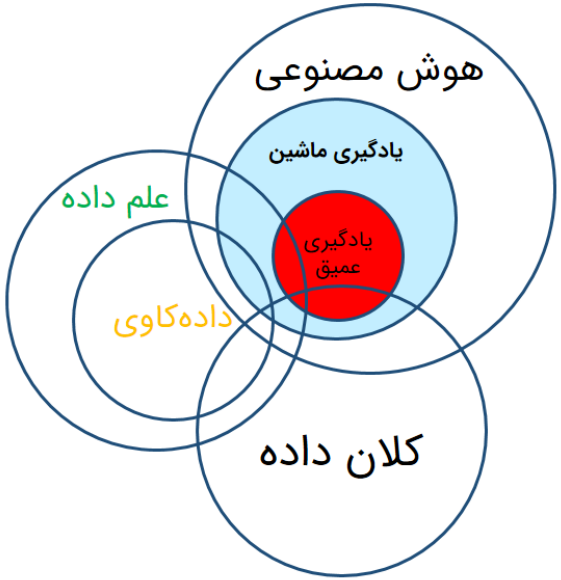
- داده های آماری به جز در مواردی که خطای انسانی دخیل است، نیاز به پیش پردازش سنگینی ندارند
- ولی در داده کاوی معمولا بیشتر زمان داده کاوی صرف پیش پردازش داده ها می شود.

داده‌کاوی

تبدیل داده‌ی ورودی به بینش



اجزای علم داده در یک قاب



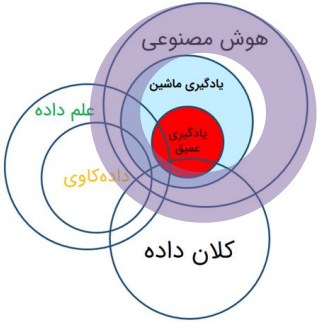
علم داده:
علم استفاده از روش‌های کمی آمار و ریاضیات در بستر تکنولوژی

اهداف:
- توسعه الگوریتم‌های طراحی شده
- کشف الگوها
- پیش‌بینی نتایج
- و یافتن راه‌حل‌های بهینه برای مسائل پیچیده

داده کاوی:تبدیل اطلاعات موجود از منابع به الگوهای برجسته
ارایه الگوها به عنوان پایه‌ی کار به هوش مصنوعی و یادگیری ماشین

هوش مصنوعی

دانش شناخت و طراحی عامل‌های هوشمند

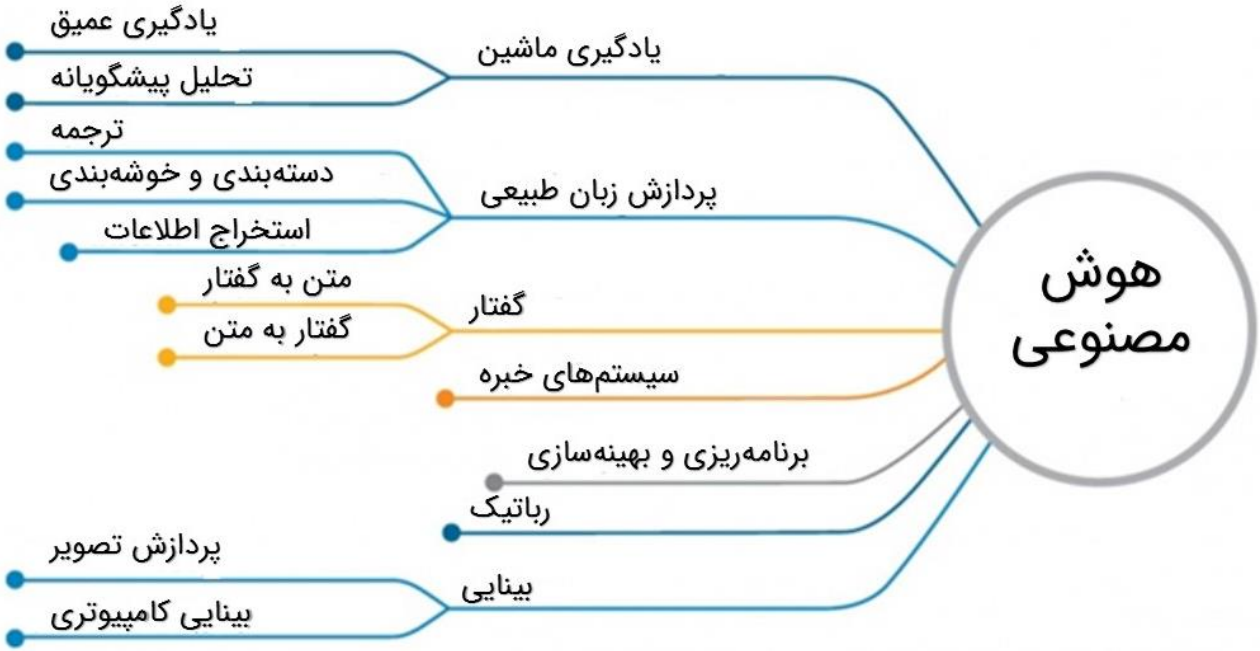


- سیستم‌هایی واکنش‌هایی مشابه رفتارهای هوشمند انسانی دارند از جمله
  - درک شرایط پیچیده
  - توانایی کسب دانش
  - یادگیری و استدلال برای حل مسائل
  - شبیه‌سازی فرایندهای تفکری
  - شیوه‌های استدلال مشابه انسان

- اشتراک هوش مصنوعی در مواردی چون یادگیری ماشین و یادگیری عمیق با علم داده
- هوش مصنوعی الهام‌بخش ابزارها و تکنیک‌های علم داده

- توسعه مفاهیم مانند یادگیری ماشین و یادگیری عمیق به‌گونه‌ای از طریق هوش مصنوعی

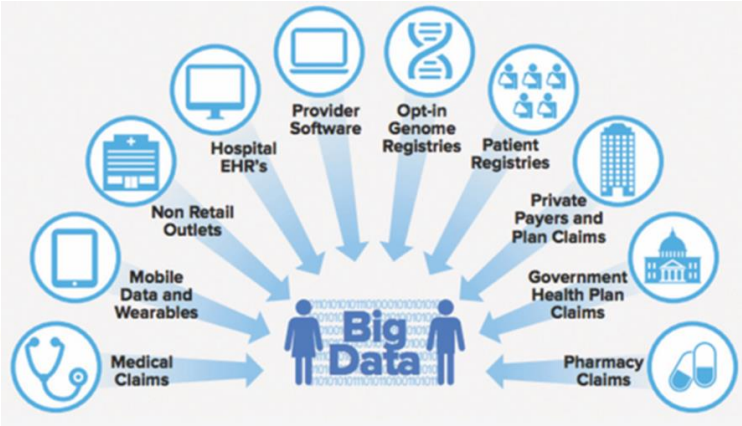
زیر شاخه‌های موجود در هوش مصنوعی



کلان داده

- **کلان داده عنصر کلیدی**
- کلان داده دارایی اصلی و داده کاوی مبدل این دارایی به نتایج سودمند
- داده‌کاوی فعالیت ورود به پایگاه داده‌های عظیم به منظور جستجو در اطلاعات مرتبط

- **کلان داده**
- راهی برای توصیف انبوهی از اطلاعات که سازمان‌ها به منظور تبدیل داده‌های خود به دانش با آن مواجه هستند
- داده‌هایی که بسیار انبوه، پرشتاب و گوناگون هستند
- نیاز به روش‌های پردازشی تازه
- نیاز تصمیم‌گیرنده‌گان به داده‌های کم حجم و خاص‌شده

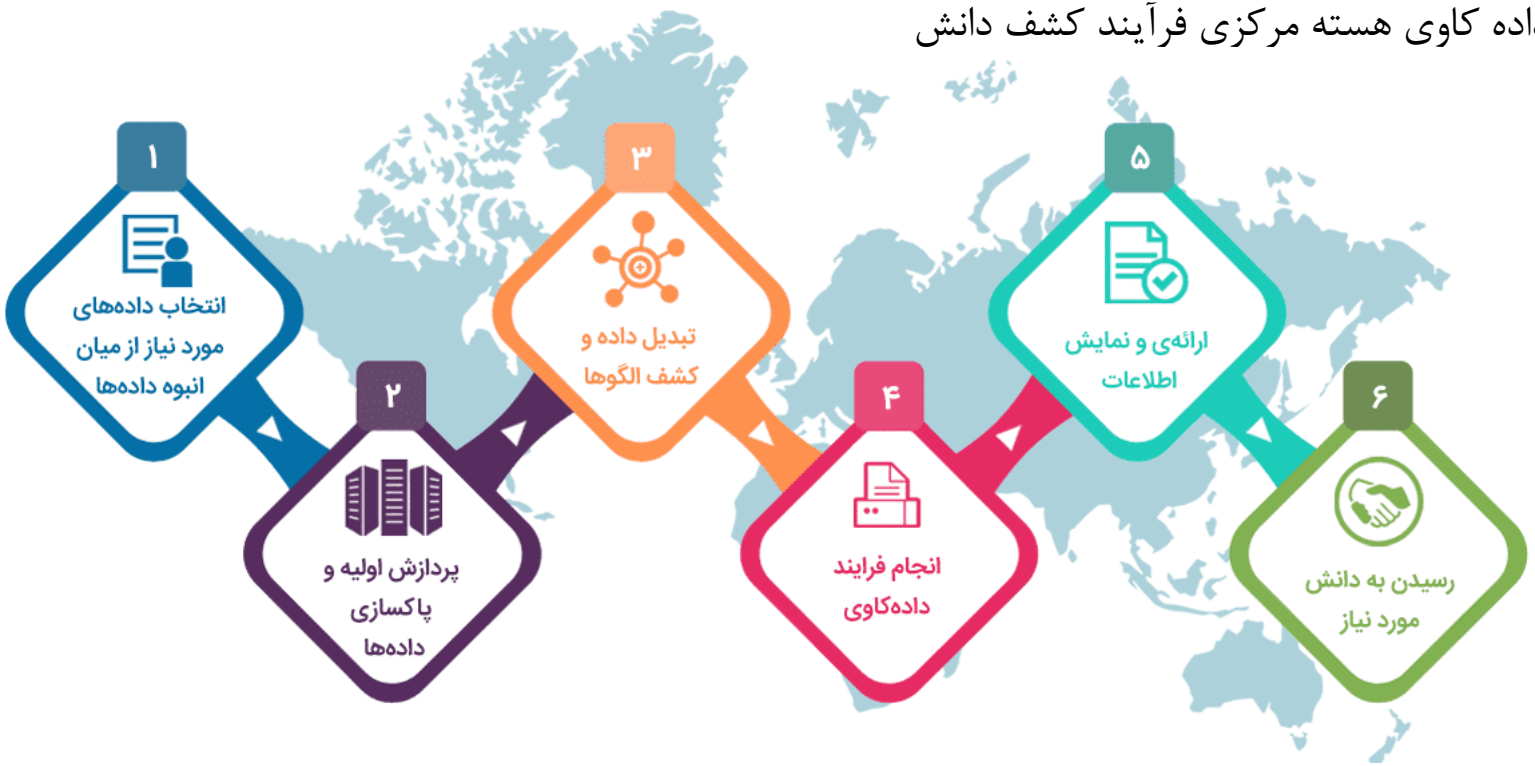



حجم داده
Volume
حجم بالای داده‌های موجود و در حال تولید
سرعت
Velocity
سرعت بالای تولید داده‌ها
تنوع
Variety
تنوع منابع، قالب‌ها و ساختارهای داده
ارزش
Value
ارزش داده برای کسب‌وکار
صحت
Veracity
حجم بالای داده‌های موجود و در حال تولید
اعتبار
Validity
کیفیت داده، حاکمیت، مدیریت داده‌های اصلی
تغییرپذیری
Variability
پویایی، رفتار تکاملی در منبع داده
محل
Venue
داده‌های توزیع شده همگن
از محل‌های داده گوناگون
فرهنگ لغت
Vocabulary
مدل‌ها، معانی که ساختار داده را توصیف می‌کند.
ابهام
Vagueness
سردرگمی درباره معنای کلان‌داده و ابزارهای آن
نوسان
Volatility
مدت زمان اعتبار داده و مدت زمانی که داده باید ذخیره شود.
بصری‌سازی
Visualization
نمایش داده‌ها به صورت گرافیکی
شیوع‌پذیری
Virality
داده‌ها چقدر سریع در شبکه‌های مردم به مردم گردش می‌کند.
وضعیت بیان
Verbality
اشاره دارد به جریان اصلی داده‌ها که به صورت ساختارنیافته هستند.
اطمینان
Verbosity
در میان داده‌های متنوع، تکرار، افزونگی و مقادیر از دست رفته وجود دارد.
تطبیق‌پذیری
Versatility
مفید بودن داده‌ها در سناریوهای گوناگون و کاربردپذیری برای مجموعه ذینفعان گوناگون
ناروانی
Viscocity
قابلیت جریان یافتن داده‌ها به دیگر بررسی‌های موردی که اطمینان را تحت تاثیر می‌دهد.
پدیداری
visibility
پیرامون کنترل دسترسی به داده‌ها است.
کلان‌داده یا مهداده
Big Data

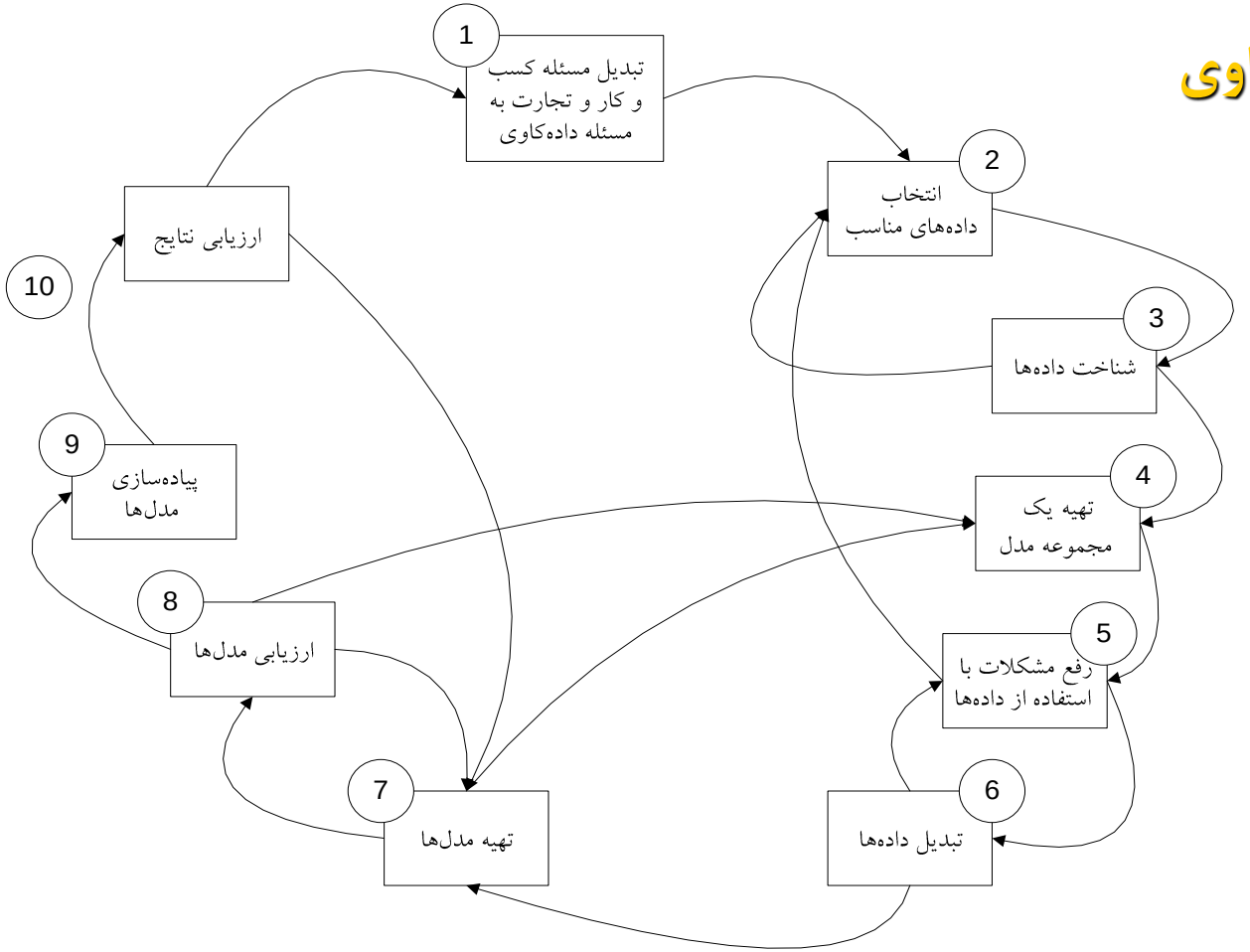
بخش دوم:
معرفی برخی از ابزارها

فرآیند داده کاوی

داده کاوی هسته مرکزی فرآیند کشف دانش



فرآیند داده کاوی



پایه های یک فرآیند داده کاوی

۵ پایه اصلی:

- مجموعه نمونه های آموزشی: باید انتخاب، جمع آوری و پیرایش شوند.
- هدف: نوع دانش مورد انتظار، تکنیک داده کاوی مورد استفاده را مشخص خواهد کرد.
- دانش پایه: انتقال دانش موجود در مورد مسئله به فرآیند داده کاوی، غالبا به صورت سلسله مراتبی از مفاهیم
- معیارهای ارزیابی: ملاک‌های ارزش دانش حاصل از داده کاوی، چه در زمان استخراج دانش و چه در زمان بازنمائی از اهمیت کلیدی برخوردار بوده و راهنمای فرآیند داده کاوی خواهند بود.
- نحوه ارائه: معمولا بر حسب نوع دانش استخراج شده تعیین می شود. در موارد متعددی نیز روش مناسبی برای بازنمائی وجود ندارد

پلتفرم‌های مورد استفاده در فرایند داده‌کاوی

- زبان برنامه‌نویسی R
- زبان برنامه‌نویسی پایتون
- زبان برنامه‌نویسی متلب
- نرم‌افزار SPSS
- نرم‌افزار Weka
- نرم‌افزار RapidMiner
- و ...

هدف داده کاوی

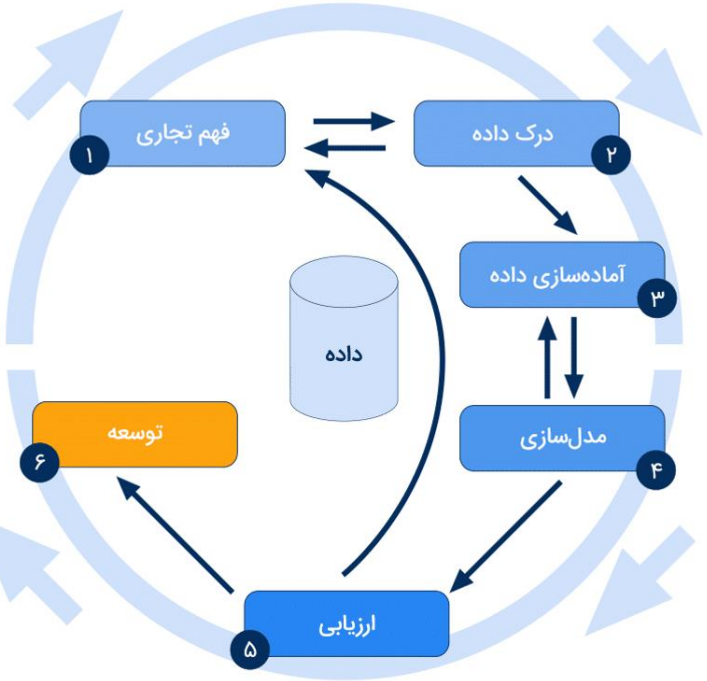
ویژگی الگوهای حاصل از فرآیند داده کاوی

- Valid — معتبر باشد
- Useful — مفید باشد
- Novel — جدید باشد
- Understandable — قابل فهم باشد

مراحل متدلوژی CRISP

- **CRISP: CRoss Industry Standard Process for Data Mining**

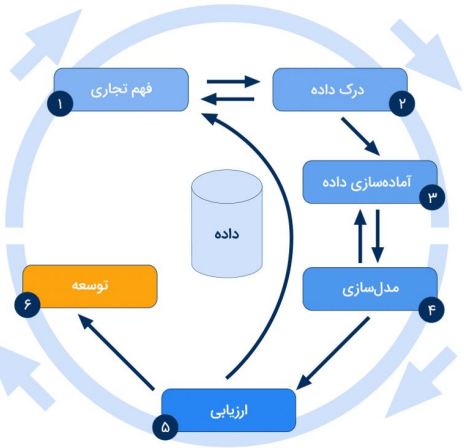
- معنی: فرایندهای استاندارد صنعت متقابل برای داده‌کاوی
- یکی از روش‌های تحلیلی متفاوت برای فرایند داده‌کاوی است



- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Development

- فهم تجاری
- درک داده
- آماده‌سازی داده
- مدل‌سازی
- ارزیابی
- توسعه

مراحل متدلوژی CRISP



- **فهم تجاری:** شامل گردآوری موارد مورد نیاز و گفتگو با مدیران ارشد برای تعیین اهداف
- **درک داده:** نگاه نزدیک و بررسی دسترسی به داده‌ها برای فرایند داده کاوی
  شامل گردآوری، توصیف، کشف و تغییر کیفیت داده‌ها
- **آماده سازی داده:** از مهم ترین و همچنین زمان‌برترین بخش‌های داده‌کاوی
  شامل انتخاب، پاک‌سازی، ساختاربندی، و ادغام داده‌ها
- **مدل سازی:** پس از آن که داده‌ها آماده‌ی فرایند داده‌کاوی
  تکنیک‌های انتخاب مدل‌سازی، ایجاد یک طراحی آزمون، ساخت مدل‌ها، و ارزیابی مدل این مرحله
- **ارزیابی:** در این مرحله نتایج ارزیابی شده، فرایند انجام کار بازبینی و مراحل بعدی انجام می‌شوند.
- **توسعه:** بکار بردن نتایج به‌دست آمده توسعه یافته و برای بهبود عملکرد سازمان

انواع مختلف داده‌ها

- ساختار یافته structured
- نیمه ساختار یافته semi-unstructured
- غیر ساختار یافته unstructured

ساخت یافته

ID	نام رشته	تعداد دانشجو
۱	فنی و مهندسی	۷۰۰
۲	پزشکی	۴۰۰
۳	علوم پایه	۴۰۰

نیمه ساخت یافته

```
<UNIVERSITY>
  <College ID="1">
    <Name>فنی مهندسی</Name>
    <Count>700</Count>
  </College>
  <College ID="2">
    <Name>پزشکی</Name>
    <Count>400</Count>
  </College>
  <College ID="3">
    <Name>علوم پایه</Name>
    <Count>400</Count>
  </College>
</UNIVERSITY>
```

غیر ساخت یافته

این دانشگاه 1500 دانشجو دارد که در گروه های فنی مهندسی، پزشکی و علوم پایه مشغول به فعالیت هستند. 700 دانشجو در رشته ی فنی و مهندسی، 400 دانشجو در رشته ی پزشکی و 400 دانشجو نیز در رشته ی علوم پایه در حال تحصیل هستند

- داده های ساختار یافته دارای فرمتی مشخص هستند و می‌توان آن را درفایل های اکسل در فیلدهای مختلف در سطرها و ستون های جدول قرار داد
- داده ی نیمه ساخت یافته نیز در قابل جدول مشخص نشده دارای ساختاری است که با برچسب هایی از یکدیگر جدا شده اند
- داده های غیرساختار یافته فرمت مشخصی ندارند ،نمی توان آن ها به عنوان یک فیلد اطلاعاتی در سطرها و ستون های جدول قرار داد .

نشانه‌گذاری داده‌ها

- **ورودی X**
- X اغلب چندبُعدی است. هر بُعد از X به صورت X_j مشخص شده که به یک ویژگی، یک متغیر اشاره دارد.

- **خروجی Y**
- متغیر پاسخ یا متغیر وابسته نامیده می‌شود. پاسخ تنها هنگامی در دسترس است که یادگیری نظارت شده باشد.

پارامترهای مورد اهمیت در داده‌ها

- اندازه داده ها

BIG DATA
SMAL DATA

- نوع داده ها

  - رکوردها
  - گراف ها
  - مجموعه‌های داده های منظم

پارامترهای مورد اهمیت در داده‌ها

ویژگی‌ها

یک مجموعه داده ازنمونه ها و ویژگی ها (خصیصه‌ها) تشکیل می‌شود.

یک ویژگی، فیلد داده‌ای است که مشخصه‌های یک شی داده را ارائه می‌کند.
نوع یک ویژگی در داده ها توسط مجموعه‌ای از مقادیر ممکن تعیین می‌شود

ابعاد

هر «ویژگی» (Feature | Attribute) در مجموعه داده را که به صورت یک فیلد در فایل مسطح یا ستون در جداول پایگاه داده رابطه‌ای ذخیره شده است را یک «بُعد» (Dimension) گویند.

تعداد کل ویژگی‌ها، ابعاد مجموعه داده را مشخص می‌کند.

انواع مختلف داده‌ها

رکوردها(Records)

- ماتریس‌های داده ای (به طور معمول دو بعدی) (Data matrix)
- اسناد(Documents)
- داده‌های تراکنشی (Transaction data)

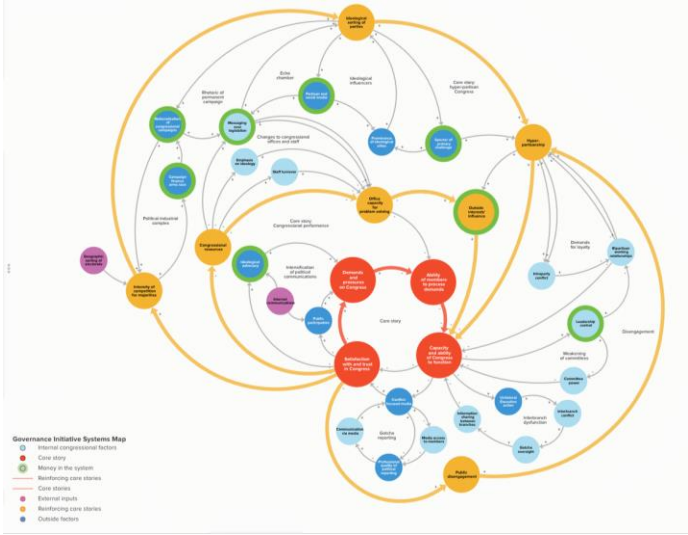
	Sex	Race	Height	Income	Marital Status	Years of Educ.	Liberal-ness
R1001	M	1	70	50	1	12	1.73
R1002	M	2	72	100	2	20	4.53
R1003	F	1	55	250	1	16	2.99
R1004	M	2	65	20	2	16	1.13
R1005	F	1	60	10	3	12	3.81
R1006	M	1	68	30	1	9	4.76
R1007	F	5	66	25	2	21	2.01
R1008	F	4	61	43	1	18	1.27
R1009	M	1	69	67	1	12	3.25

انواع مختلف داده‌ها

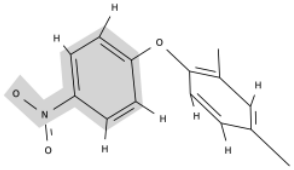
گراف‌ها (Graphs)

- داده های وب(Web data)

شامل ساختار ارتباطی صفحات وب و شبکه های اجتماعی



- ساختارهای مولکولی(Molecular Structures)



انواع مختلف داده‌ها

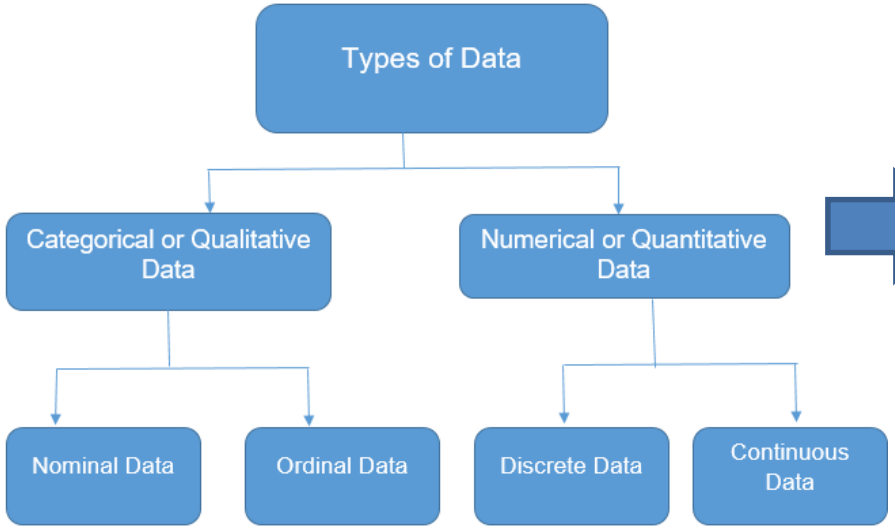
مجموعه‌های داده ای منظم(Ordered Datasets)

- داده‌های زمانی(Time Series)
- داده‌های مکانی(Spatial)
- داده‌های ترتیبی(Sequence Data)

مجموعه داده‌های منظم ترتیبی
(دنباله های ژنوم DNA)

ماهیت مجموعه داده

انواع ویژگی‌های موجود در مجموعه داده



Quantitative	کمی
Qualitative	کیفی
Ordinal	ترتیبی
Nominal/Categorical	اسمی
Numeric	عددی

ماهیت مجموعه داده

انواع ویژگی‌های موجود در مجموعه داده

- **کمی** : اندازه‌گیری‌ها یا شمارش‌هایی که به صورت مقادیر عددی ذخیره شده‌اند
  درجه حرارت و قد افراد

- **کیفی**: گروه یا دسته‌ها، گروه رنگ‌ها (زرد، قرمز و آبی) ، گروه خونی

- **ترتیبی** : دارای یک ترتیب طبیعی هستند. چاقی
  اندازه پیراهن (XL، L، M، S و مدارج تحصیلی (دبستان، راهنمایی، دبیرستان، کارشناسی، کارشناسی ارشد و دکترا)

ماهیت مجموعه داده

انواع ویژگی‌های موجود در مجموعه داده

- **اسمی** :داده‌های دسته‌ای
- به طور کلی، «داده‌های دسته‌ای» (Categorical Data) قرار گیری داده‌ها در تعداد کمی از دسته‌های گسسته است.
- داده‌های دسته‌ای به شیوه مشخصی تعریف می‌شوند.

- اسامی دسته‌ها، مانند وضعیت تاهل، جنسیت و رنگ‌ها

- برخی از این داده‌ها از جمله اسامی شهرها یا جنسیت افراد فاقد ترتیب
- و مواردی مانند دمای هوا (بالا، متوسط و پایین) دارای ترتیب هستند

ماهیت مجموعه داده

انواع ویژگی‌های موجود در مجموعه داده

- **عددی** :داده‌های عددی خود به دو دسته فاصله‌ای و نسبتی تقسیم می‌شوند.

- **داده‌های فاصله‌ای** بر اساس مقیاس واحدهایی با اندازه برابر اندازه‌گیری می‌شوند.
- مقادیر ویژگی‌های عددی دارای ترتیب هستند و می‌توانند مثبت، صفر و یا منفی باشند.

- **داده نسبتی**، خصیصه عددی دارای یک صفر مطلق است.
- اگر اندازه‌ها نسبتی باشند، می‌توان از نسبت مقادیر با یکدیگر سخن گفت.
- به علاوه، مقادیر قابل مرتب‌سازی شدن هستند و می‌توان تفاضل بین آن‌ها، میانگین، میانه و مُد را محاسبه کرد.

ماهیت مجموعه داده

انواع ویژگی‌های موجود در مجموعه داده

- **پیوسته**
- «داده‌های پیوسته» (Continuous) می‌توانند هر مقداری را در یک بازه از اعداد حقیقی بپذیرند.
- این مقدار الزاما نباید صحیح باشد.
- داده‌های پیوسته متفاوت و به نوعی متضاد داده‌های گسسته (Discrete)یا دسته‌ای هستند.

- **گسسته**
- یک قلم داده که دارای مجموعه متناهی از مقادیر است را «گسسته» گویند.

پیش پردازش داده ها Data Preprocessing

پیش پردازش داده ها اولین گام در داده کاوی

- نقص داده ها در پایگاه داده :

1. بعضی داده ها noisyبودن داده ها
2. عدم وجود بعضی از مقادیر داده ها missing
3. وجود ناسازگاری بین داده ها

- این نقایص توی داده های حجم بسیار بیشتر است و به همین خاطر توجه به آنها بسیار مهم است.

پیش پردازش داده ها Data Preprocessing

- بیان داده های نامناسب سبب خروجی‌های داده کاوی نیز غیر مفید

مراحل پیش پردازش

- ۱- شناسایی داده های noisy و missing و ناسازگار
- ۲- رفع نقایص به بهترین شیوه ممکن

هدف: کسب خروجی های مطلوبی از داده کاوی

پیش پردازش داده ها Data Preprocessing

- مهمترین تکنیک های پیش پردازش داده ها عبارتند از:

تکنیک‌های پاکسازی داده Data cleaning

- هدف: از بین برده داده‌های noisyو ناسازگاری‌های بین داده ها

تکنیک‌های یکپارچگی داده Data integration

- هدف: یکپارچگی داده‌ها از منابع مختلفی جمع آوری شده

تکنیک‌های کاهش داده Data reduction

- هدف: حذف داده های غیر مفید از پردازش نهایی درحجم بالای داده

تکنیک‌های Data transformations

- کاربرد در زمانی اعمال نرمال سازی های داده

پاک سازی داده ها

- داده ها در فایل ها و منابع مختلف نگهداری شوند
- نیاز است تا داده ها پیش از اجرای تکنیک های داده کاوی یا آماده سازی برای هوشمندسازی کسب و کار با یکدیگر یکپارچه شوند.

- پاکسازی داده‌ها یا تمیز کردن داده‌ها
- فرآیندی جهت تشخیص، حذف و اصلاح داده‌های نادرست از مجموعه‌ای از رکوردها، جداول یا بانک اطلاعاتی
- شناسایی قسمت‌های ناقص و نادرست داده‌ها
- اصلاح و جایگزینی یا حذف داده‌های فاسد

- **هدف از پاکسازی داده‌ها** : استخراج اطلاعات دقیق
- داده‌های نادرست یا ناسازگار می‌تواند منجر به نتیجه‌گیری غلط و شکست سرمایه‌گذاری بزرگ و کوچک شود

پاک سازی داده ها

- فرآیند پاکسازی محدود به حذف داده‌های نامناسب یا وارد کردن مقادیر از دست رفته نیست
- پاکسازی کشف روابط پنهان شده میان داده ها، شناسایی دقیق ترین منابع داده، تعیین مناسب ترین ستون ها در آنالیز

- **ویژگی داده با کیفیت برای تحلیل:**
- ۱- ارزش یا اعتبار داده
- ۲- دقت و صحت داده
- ۳- دوام یا پایداری داده
- ۴- یکپارچگی ارتباطات و بخش های مختلف داده
- ۵- بردار زمانی داده

پاک سازی داده ها

نکات مهم برای تمیز کردن داده‌های شناسایی و حذف داده‌های تکراری

- استاندارد سازی اعداد
- استانداردسازی زمان‌ها و تاریخ‌ها
- استانداردسازی نحوه نگارش کلماتی که چندین نوع نوشتاری دارند
  مانند مدارک تحصیلی لیسانس و کارشناسی یا نام کشورهایی مانند British-English-UK

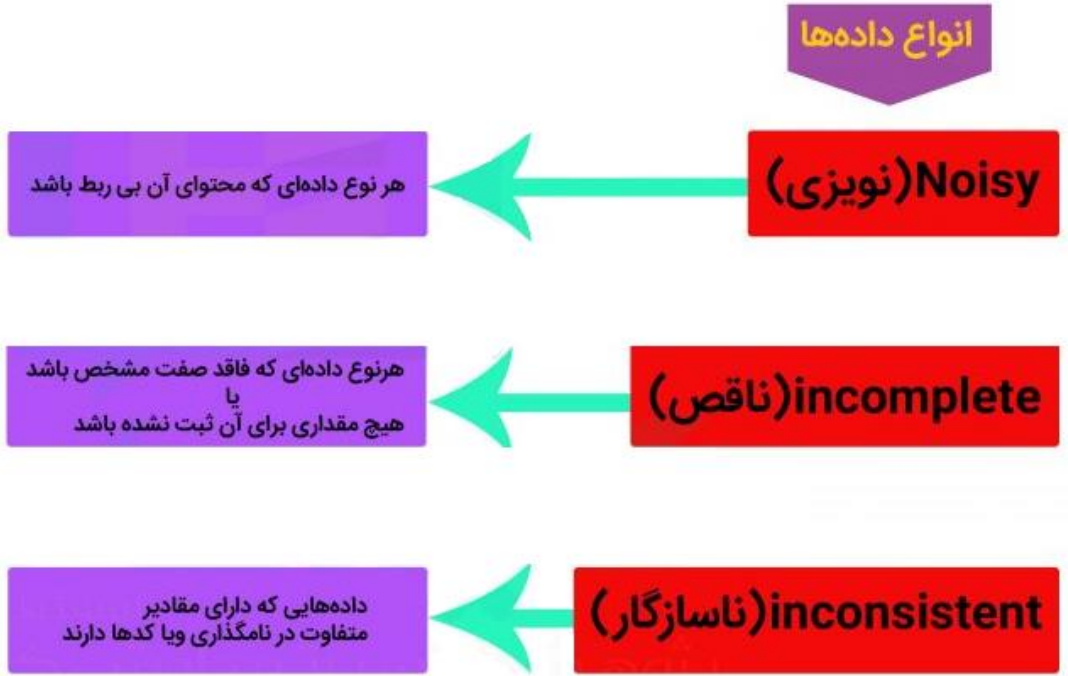
- سراسری کردن استاندارد موارد حساس و تایین کننده مانند نحوه ورود هزینه‌ها

پاک سازی داده ها

- پس از پاکسازی، مجموعه داده‌ها باید با سایر مجموعه داده‌های مشابه در سیستم سازگار باشد.
- داده‌های ناسازگار ناشی از
- اشتباهات ورود داده‌ها از طرف کاربر
- تغییر داده‌ها در حین انتقال پرونده
- یا ذخیره‌سازی با تعاریف غیراستانداردی که بین سازمان‌های مختلف متفاوت می‌باشد

- **مهمترین فعالیت های این بخش عبارت است:**
- **تخمین مقادیر ناموجود در پایگاه داده ها**
- **از بین بردن اختلال noise در داده ها**
- **حذف کردن داده های پرت و نامربوط**
- **از بین بردن ناسازگاری در داده ها**

پاک سازی داده ها



پیش پردازش داده ها Data Preprocessing

قابل جستجو کردن داده

- **تکنیک های جستجو**
- محاسبه حداقل و حداکثر مقادیر،
- محاسبه میانگین و انحراف معیار
- توجه به توزیع داده

- **هدف:** برآورد مناسب بودن داده ها برای نمایش فرایند های مرتبط با مشتریان ، کسب و کار
- یا بازنگری فرضیات و کسب داده های مناسب

پیش پردازش داده هاData Preprocessing

نرمال سازی داده data Normalization

- مثال: مسابقات ورزشی وزن ها و گروه های مختلف

- نمی‌توان یک مجموعه‌ی داده که در بازه‌ی بین ۰ تا ۲۰ متغیر هستند را با مجموعه‌ی که در بازه‌ی بین ۰ تا ۱۰۰۰ ۰ قرار دارد، مقایسه کرد

- داده‌ها باید در یک بازه‌ی range مساوی نسبت به یکدیگر قرار بگیرند
- مثلاً همه در یک بازه‌ای مانند ۰ تا ۱ قرار داشته باشند

نرمال سازی داده data Nnormalization

- **MinMaxNormalization:**
- $$\frac{x-MIN(X)}{MAX(X)-MIN(X)}$$

داده اولیه	فرمول	نرمال
15	15-15/30-15	0
25	25-15/30-15	0.66
30	30-15/30-15	1

پیش پردازش داده ها Data Preprocessing

نرمال سازی داده data Nnormalization

Z-SCORE Normalization:

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

$$\mu = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sigma = S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$$

پیش پردازش داده ها Data Preprocessing

تبدیل داده به فرم قابل فهم برای الگوریتم های داده کاوی Data Transformation

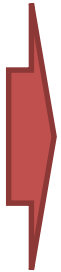
ساختار الگوریتم‌های داده‌کاوی، یادگیری از ماتریس‌های عددی

همیشه داده‌ها به صورت عددی آماده نیستند
تبدیل داده به فرمت استاندارد

پیش پردازش داده ها Data Preprocessing

تبدیل داده به فرم قابل فهم برای الگوریتم های داده کاوی Data Transformation

جنسیت	قد	سن
مرد	160	12
زن	167	37
مرد	159	68



جنسیت	قد	سن
0	160	12
1	167	37
0	159	68

پیش پردازش داده ها Data Preprocessing

داده گم‌شده Missing Values و راهکارهای مقابله با آن‌ها

راه‌حل‌ها:

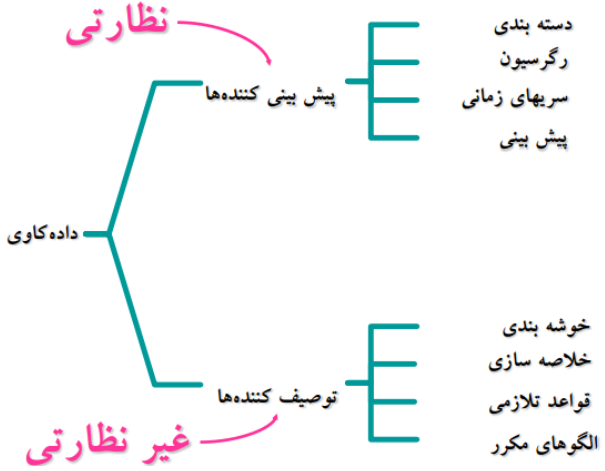
- حذف آن سطر
- حذف آن ستون
- پر کردن آن توسط داده‌ی واقعی.
- جایگزینی میانگین mean یا میانه‌ی median اعداد موجود رکوردهای دیگر را برای یک ستون خاص
- همچنین می‌توانیم یک مقدار ثابت را در نظر گرفته و به جای مقادیر مفقود شده قرار دهیم

- استفاده از KNN

یادگیری نظارت شده در مقایسه با یادگیری نظارت نشده

- اگر Yدر داده‌های آموزش وجود داشته باشد، روش یادگیری «نظارت شده» Supervised است.

- اگر Yوجود نداشته باشد (یا در صورت وجود از آن چشم‌پوشی شود)، یادگیری «نظارت نشده» Unsupervised است.



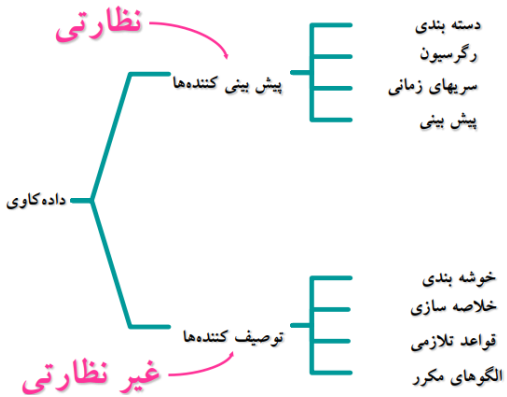
عناصر داده کاوی

توصیف و **کمک به پیش بینی** دو کارکرد اصلی داده کاوی هستند.

- تحلیل داده مربوط به مشخصه های انتخابی متغیرها؛ از گذشته و حال، و درک الگو مثالی از تحلیل توصیفی است

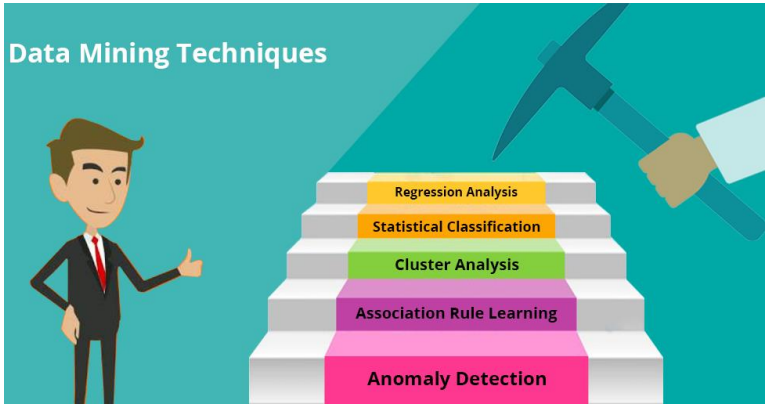
.

- برآورد ارزش آینده یک متغیر و طرح ریزی کردن روند مثالی از توانایی پیشگویانه داده کاوی است.

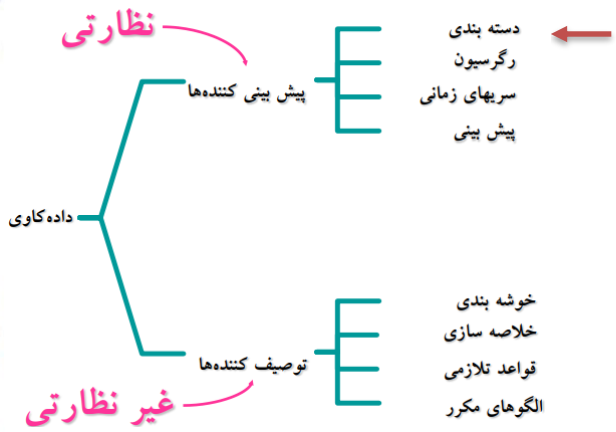


تکنیک های داده کاوی

- داده‌کاوی وابسته به کاربرد
- کاربردهای مختلف نیازمند روش‌ها و تکنیک‌های داده‌کاوی مختلف



تکنیک های داده کاوی



- دسته‌بندی/طبقه بندی (Classification)

  - تعداد دسته ها یا کلاس های دسته بندی از قبل مشخص
  - پس از دریافت تعدادی نمونه آموزشی، یادگیرنده باید دسته نمونه های جدید را مشخص نماید.

  - نتیجه: تولید یک درخت تصمیم یا مجموعه‌ای از قوانین دسته‌بندی،

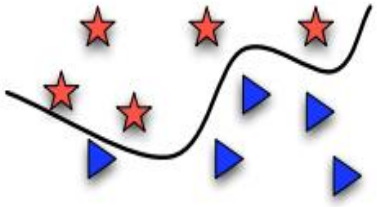
مثال:

ابتلا یا عدم ابتلا به یک بیماری مشخص

تکنیک های داده کاوی

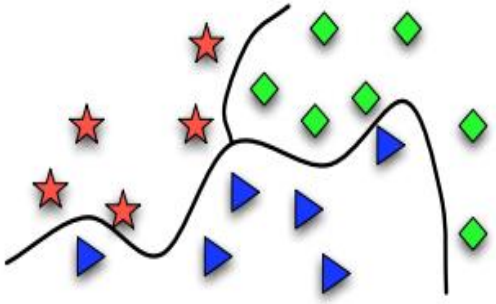
دسته‌بندی/طبقه بندی (Classification)

1



Binary Classification (1)

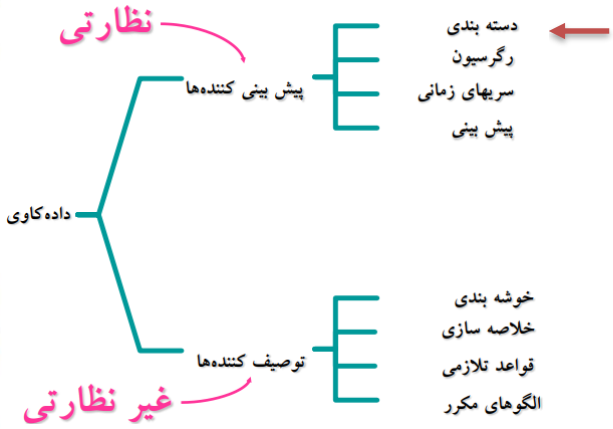
2



Multiclass Classification (2)

تکنیک های داده کاوی

دسته‌بندی/طبقه بندی (Classification)



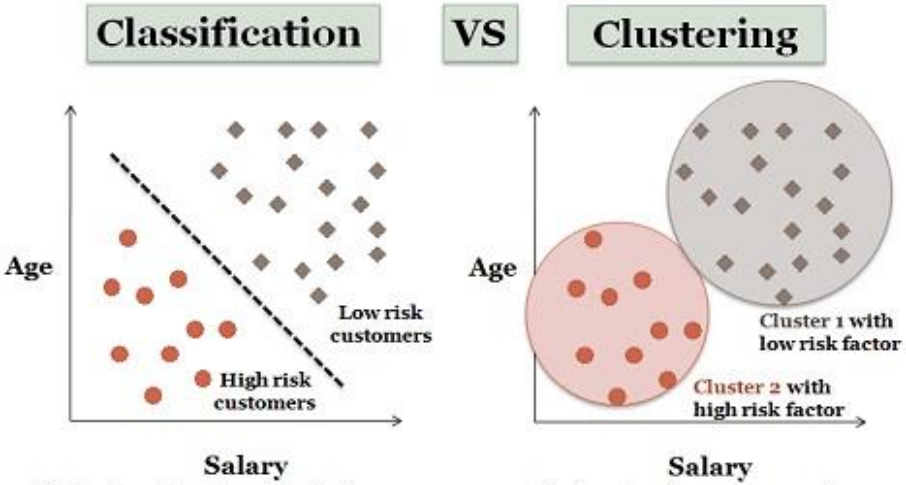
- مثال:
  - یک شرکت با بیش از ۱۰۰۰۰۰ مشتری
  - یک کاتالوگ با هزینه سنگین چاپ و توزیع
  - لزوم ارسال انتخابی کاتالوگ (نه برای همه)
  - دسته بندی: با توجه به سوابق ارسال کاتالوگها و پاسخ مشتریان، چه افرادی احتمالا در گروه "علاقه مند به محصول معرفی شده" قرار می گیرند؟
  - کاهش هزینه ها

تکنیک های داده کاوی

تفاوت خوشه بندی با طبقه بندی :

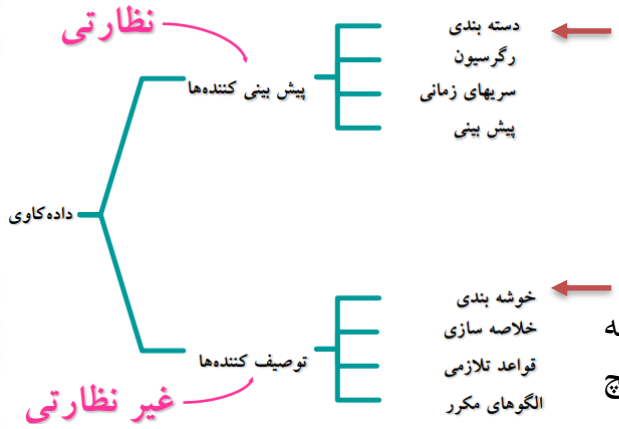
طبقه بندی در تکنیک یادگیری نظارت شده استفاده می شود که در آن برچسب های از پیش تعریف شده به خصوصیات به نمونه ها اختصاص می یابد،

خوشه بندی در یادگیری بدون نظارت در جایی که نمونه های مشابه در آن گروه بندی می شوند ، استفاده می شود



Risk classification for the loan payees on the basis of customer salary

تکنیک های داده کاوی



تفاوت خوشه بندی با طبقه بندی :

- در طبقه بندی هر طبقه از قبل تعریف شده است اما در خوشه بندی هیچ بینشی از قبل در مورد خوشه ها وجود ندارد
- در طبقه بندی هر رکورد بر اساس یک مدل که از مثال های از قبل طبقه بندی شده به دست آمده است، طبقه بندی می‌شود. در خوشه بندی هیچ کلاس یا طبقه‌ای از قبل تعریف نشده است.
- خوشه بندی معمولا قبل از دیگر تکنیک‌ها؛ به منظور بیشتر همگون شدن داده‌ها انجام می شود.

تکنیک‌های داده‌کاوی

تفاوت خوشه بندی با طبقه بندی :

Classification vs Clustering

Criteria	Classification	Clustering
Prior Knowledge of classes	Yes	No
Use case	Classify new sample into known classes	Suggest groups based on patterns in data
Algorithms	Decision Trees, Bayesian classifiers	K-means, Expectation Maximization
Data Needs	Labeled samples from a set of classes	Unlabeled samples

Profiling

یک توصیف خوب می تواند راهنمای خوبی برای شروع فعالیت ها

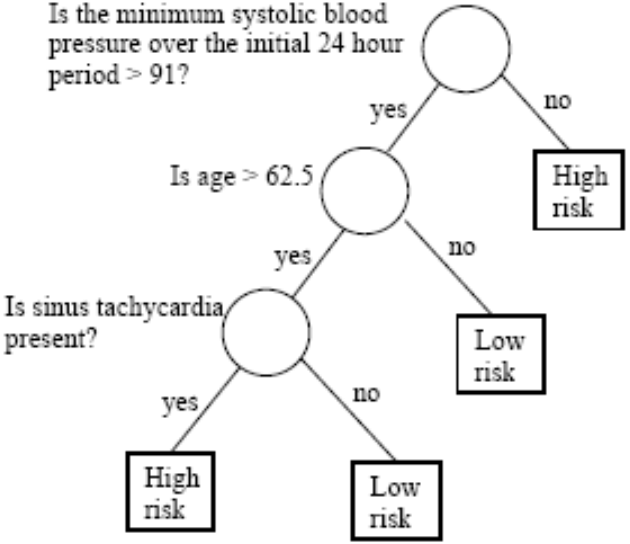
مثال: تمایل زنان به جراحی‌های زیبایی بیشتر از مردان است.
لذا برای بررسی اثرات جانبی جراحی‌های زیبایی، ابتدا مطالعه روی زنان انجام شود.

یک ابزارهای توصیف و دسته بندی، **درخت تصمیم** است.
خوشه بندی و قوانین وابستگی نیز می توانند به منظور توصیف استفاده شوند

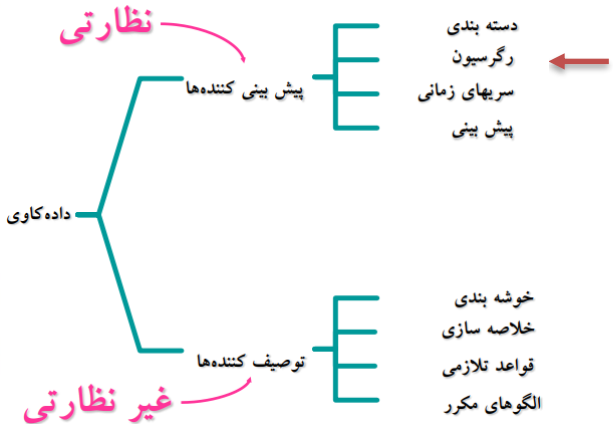
تکنیک های داده کاوی

درخت تصمیم – ابزار دسته بندی

- درخت‌های تصمیم ازساده‌ترین تکنیک‌های داده‌کاوی است
- نمایش یک سری از قوانین که به یک کلاس یا مقدار منجر می شود



تکنیک های داده کاوی



رگرسیون یا Regression

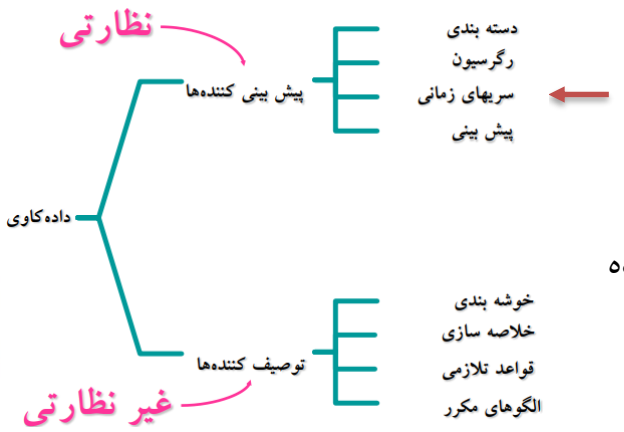
هدف : رسیدن به یک رابطه ریاضی، برای توصیف یک پدیده

مثال۱: رابطه میان ساعت مراجعه به یک سایت، محل سکونت، سن، سرویس ایمیل مورد استفاده و مقدار سفارش انجام شده توسط یک کاربر.

مثال۲: پیش بینی سری های زمانی، به صورت حالت خاصی از مسأله رگرسیون قابل طرح و حل

تکنیک های داده کاوی

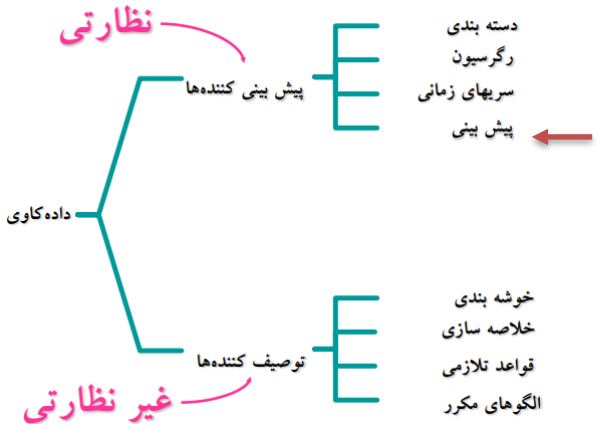
تحلیل سری‌های زمانی (Time Series Analysis)



هدف: یافتن خصوصیات جالب توجه و نظم‌های مشخص در حجم بالای داده
روندها و انحراف‌ها در رخداد وقایع متوالی

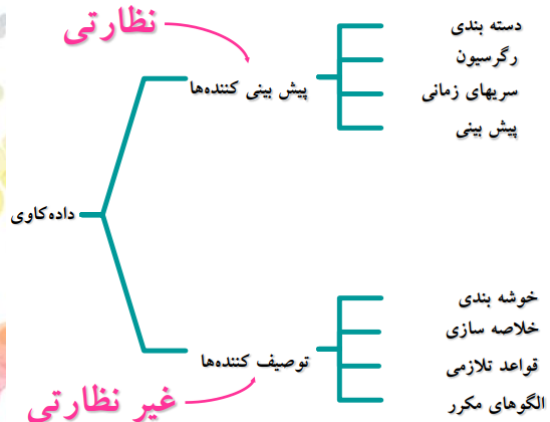
تکنیک های داده کاوی

- پیش‌بینی (Prediction)



- مقادیر ممکن برای متغیرهای نامعلوم پیش‌بینی می‌شوند
- استفاده از شبکه‌های عصبی و الگوریتم ژنتیک برای پیش‌بینی

تکنیک های داده کاوی



کاوش قواعد وابستگی (Association Rules Mining)

کشف وابستگی‌ها و ارتباطات بین داده‌های موجود در یک پایگاه داده
کشف پدیده هایی که با هم رخ می دهند

نتیجه: دسته‌ای از قواعد است که به آن‌ها قواعد وابستگی گفته می‌شود

مثال: امکان وقوع بیماری در افراد با ویژگی‌های مشخص.

تکنیک های داده کاوی

تشخیص قوانین تداعی

- ارتباط محصولات مختلف هنگام خرید مشتریان در فروشگاه های زنجیره ای
- چه محصولاتی با یکدیگر به فروش می روند .
- چینش محصولات در فروشگاه
- تخفیفها و جایزه های هدف دار

- مثال: طی یک عملیات داده کاوی گسترده در یک فروشگاه زنجیره ای مشخص گردید مشتریانی که تلویزیون خریداری می کنند، غالبا گلدان کریستالی نیز می خرند.

تکنیک های داده کاوی

قوانین تداعی

- قوانین وابستگی به شکل اگر و آنگاه به همراه معیارهای پشتیبان و اطمینان مربوط به قوانین

مجموعه ای از کلیه ایتم‌های خریداری شده $I = \{ i1 , i2 , \dots , im \}$

T: زیر مجموعه ای از I به عنوان تراکنش

D: مجموعه تراکنش‌های موجود در T

TID: شناسه یکتایی اختصاص یافته به هر یک از تراکنش ها

تکنیک های داده کاوی

قوانین تداعی

- نمایی از قانون وابستگی

$$X \Rightarrow Y \ [support\ ,\ Confidence]$$

- بطوریکه داریم:

$$X \subset I,\ Y \subset I\ ,X \cap Y = \varnothing$$

- **پشتیبان یا support**
- نشان دهنده تراکنش هایی در D که شامل هر دوی X, Y $(X \cup Y)$ باشد

- **اطمینان یا Confidence**
- میزان وابستگی یک قلم کالای خاص به دیگری رابیان می‌کند

$$confidence = support(X \cup Y)/support(X)$$

تکنیک های داده کاوی

قوانین تداعی

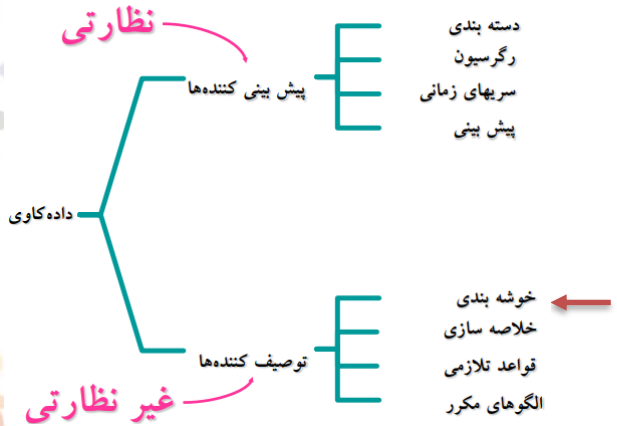
Transaction 1	
Transaction 2	
Transaction 3	
Transaction 4	
Transaction 5	
Transaction 6	
Transaction 7	
Transaction 8	

$$\text{Support}\{🍎\} = \frac{4}{8}$$

$$\text{Confidence}\{🍎 \rightarrow 🍺\} = \frac{\text{Support}\{🍎, 🍺\}}{\text{Support}\{🍎\}}$$

تکنیک های داده کاوی

خوشه‌بندی (Clustering)



گروه بندی داده ها

برخلاف دسته‌بندی تعداد کلاس‌ها در ابتدا مشخص نیست

معیار ارزیابی:

- خوشه‌بندی داده‌ها براساس اصل مفهومی حداکثرسازی شباهت‌های بین اعضای هر کلاس

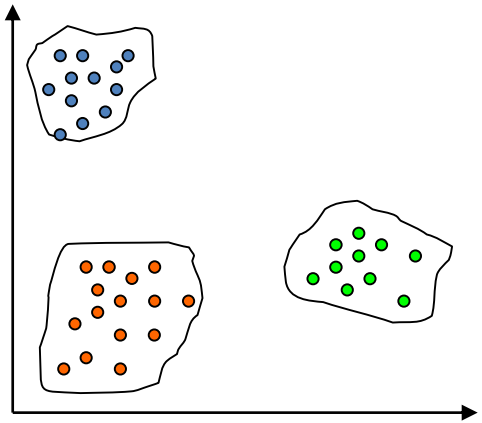
و حداقل‌سازی شباهت‌ها بین اعضای مربوط به کلاس‌های مختلف صورت می گیرد

خوشه‌بندی، بخش بندی یک جمعیت ناهماهنگ به زیر بخش‌های هماهنگ‌تر است.

تکنیک های داده کاوی

خوشه بندی

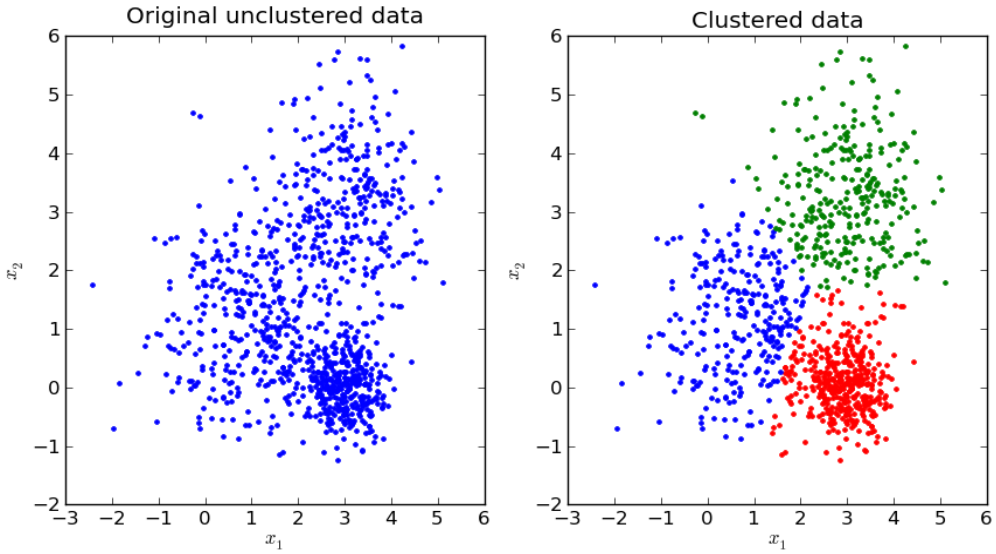
- از شاخه های داده کاوی و یادگیری بدون نظارت
- برچسب‌های اولیه نامشخص
- کشف خودکار خوشه های موجود در نمونه ها
- خوشه: نمونه های آموزشی نزدیک به هم / گروه‌هائی از اشیاء مشابه
- عملکرد بر روی نمونه های دارای ابعاد نسبتا زیاد



- سه خوشه بدیهی از نمونه ها
- تشخیص چنین خوشه هائی در ابعاد زیاد، ساده نیست!

تکنیک های داده کاوی

خوشه‌بندی (Clustering)



تکنیک های داده کاوی

خوشه بندی

هر خوشه می‌تواند خود به چند زیر خوشه تبدیل شود.
برای درک بهتر مشکلات تصمیم گیری برای تشکیل یک خوشه بندی به شکل زیر توجه کنید.
در این شکل مشخص که می‌توانند بصورت سه روش در خوشه بندی تقسیم شوند، نمایش داده شده است.

الف) خوشه بندی با دو خوشه

ج) خوشه بندی با شش خوشه

ب) خوشه بندی با چهار خوشه

با توجه به شکل ممکن است که گرفتن چهار خوشه بهینه نباشد (به علت شباهت نزدیک دو گروه)،

تکنیک های داده کاوی

کاربرد خوشه بندی

- متن کاوی و خوشه بندی اسناد
  - تشکیل سلسله مراتبی از عناوین، با بررسی متن
  - استخراج دانش از نمونه های فاقد ساختار مشخص
  - تشخیص اسناد مرتبط

- بازیابی اطلاعات
  - بازیابی مجموعه ای از نمونه های مشابه

- فشرده سازی
- بدست آوردن داده های پرت

- تجسم و درک نمونه ها
  - تشکیل سلسله مراتبی از نمونه ها
  - کاهش ابعاد

تکنیک های داده کاوی

- **روش های خوشه بندی:**
  - روش های مبتنی بر بخش بندی و تخصیص مجدد
    - ارائه ساختار مسطح از خوشه ها
    - یادگیری مستقیم خوشه ها
    - انتخاب تصادفی خوشه های اولیه
    - بهبود پاسخ از طریق جا به جائی نقاط بین بخش (خوشه) ها
  - روش مبتنی بر بخش بندی K-Means

تکنیک های داده کاوی

روش های خوشه بندی:

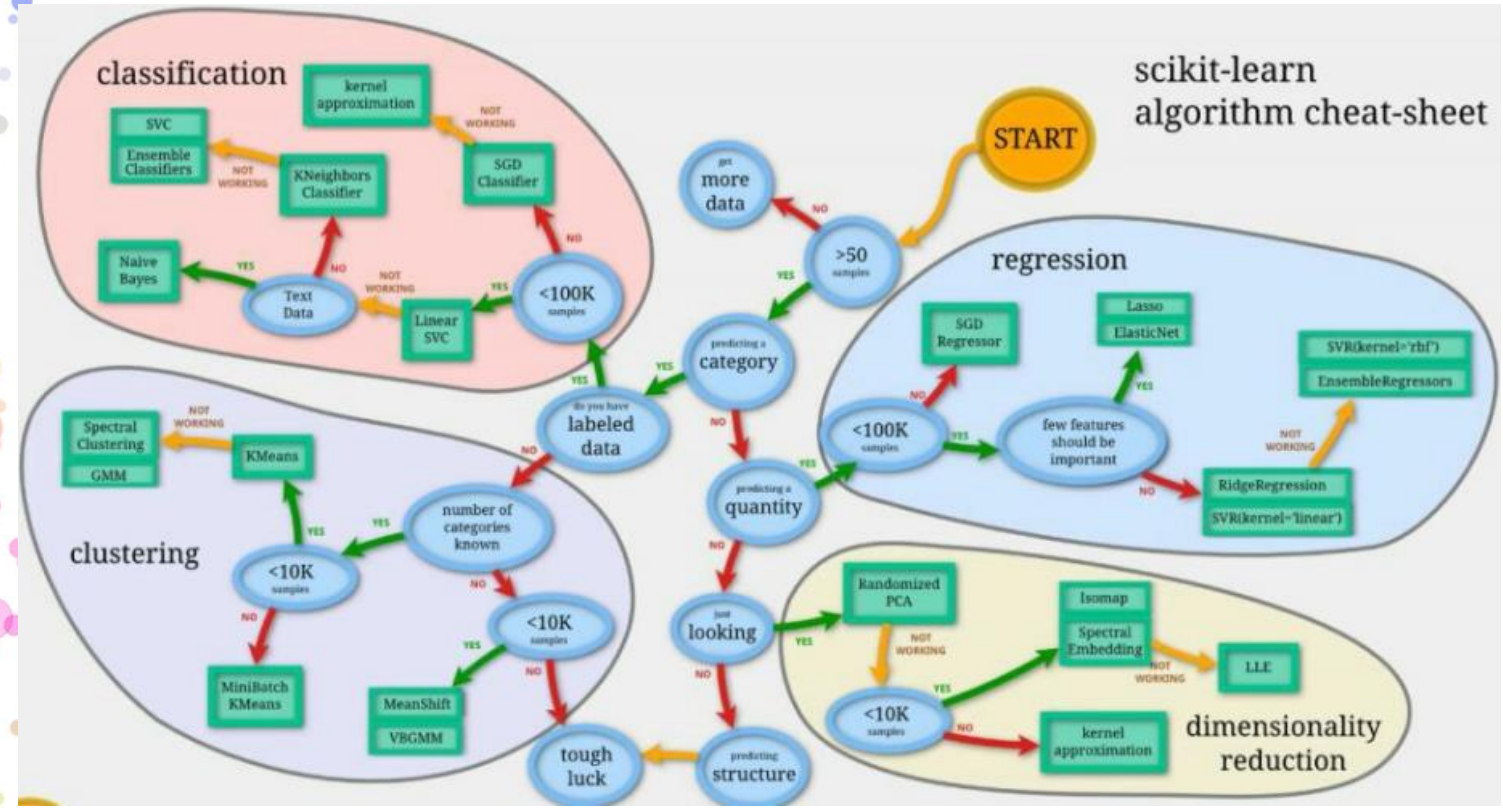
- روش های سلسله مراتبی
  - تشکیل سلسله مراتبی از خوشه ها
  - یادگیری تدریجی خوشه ها
  - دو رویکرد عمده
    - بالا به پائین (تجزیه ای): ساخت یک خوشه بزرگ و تجزیه آن
    - پائین به بالا (ترکیبی): ساخت خوشه های کوچک و ادغام آنها

تکنیک های داده کاوی

خوشه بندی: معیارهای ارزیابی

- امکان اعمال بر روی تعداد نمونه های زیاد
- امکان اعمال بر روی نمونه های دارای ابعاد زیاد
- امکان پردازش مجموعه های حاوی نویز
- کشف خوشه های دارای شکل هندسی نامنظم
- میزان وابستگی به پارامترهای ورودی

روش مناسب داده کاوی:



بخش سوم:
معرفی معیار ارزیابی

ماتریس درهم ریختگی (confusion matrix)

- به ماتریسی گفته می‌شود که در آن عملکرد الگوریتم‌های مربوطه را نشان می‌دهند.
- معمولاً برای الگوریتم‌های یادگیری با ناظر استفاده می‌شود،
- در یادگیری بدون ناظر نیز کاربرد دارد.
- معمولاً به کاربرد این ماتریس در الگوریتم‌های بدون ناظر ماتریس تطابق می گویند.

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

ماتریس درهم ریختگی (confusion matrix)

- هدف دسته‌بندی: دستیابی به بالاترین دقت و صحت ممکن در دسته‌بندی و تشخیص دسته‌ها

- در برخی از مسائل، تشخیص صحیح نمونه‌های مربوط به یکی از دسته‌ها برای ما اهمیت بیشتری دارد.
- مثال: تحقیقی که آن هدف شناسایی افراد مبتلا به یک نوع خاص از یک بیماری خطرناک است.

فرض کنید برای افرادی که مبتلا به این بیماری هستند، خطر مرگ وجود دارد و جهت رفع این خطر، نیاز به دریافت نوعی داروی خاص دارند. در این شرایط، تشخیص درست بیماران دارای اهمیت بسیار زیادی است.

خطا در تشخیص افراد سالم قابل چشم پوشی است اما برای شناسایی افراد بیمار نمی‌توان این احتمال را به جان خرید.
هدف ما تشخیص تمام افراد بیمار است، حتی اگر فرد سالمی به اشتباه جز افراد بیمار دسته‌بندی شود.
در چنین مواقعی، که دقت و صحت تشخیص یک دسته در مقایسه با دقت و صحت تشخیص کلی، اهمیت بیشتری دارد، مفهوم «ماتریس درهم‌ریختگی»Confusion Matrix ، مطرح می‌شود

ماتریس درهم ریختگی (confusion matrix)

- تعلق به دسته افراد بیمار را مثبت بودن Positive
- عدم تعلق به این دسته را منفی بودن Negative

- هر نمونه یا فردی در واقعیت، متعلق به یکی از کلاسهای مثبت یا منفی.
- از هر الگوریتمی که برای دسته‌بندی داده‌ها استفاده شود،
- در نهایت هر نمونه عضو یکی از این دو «دسته» Class دسته‌بندی خواهد شد.

ماتریس در هم ریختگی (confusion matrix)

برای هر نمونه داده، یکی از چهار حالت زیر ممکن است اتفاق بیفتد

True Positive(مثبت صحیح):
نمونه عضو دسته مثبت باشد و عضو همین کلاس تشخیص داده شود

False Negative (منفی کاذب)
نمونه عضو کلاس مثبت باشد و عضو کلاس منفی تشخیص داده شود

True Negative (منفی صحیح)
نمونه عضو کلاس منفی باشد و عضو همین کلاس تشخیص داده شود

False Positive (مثبت کاذب)
و در نهایت، نمونه عضو کلاس منفی باشد و عضو کلاس مثبت تشخیص داده شود

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

معیار اندازه‌گیری کیفیت یک دسته‌بند

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

- معیار صحت (**Accuracy**)

- متداول‌ترین، اساسی‌ترین و ساده‌ترین معیار اندازه‌گیری کیفیت

- مفهوم:میزان تشخیص صحیح دسته‌بند در مجموع دو دسته.
- نشان‌گر میزان الگوهایی است که درست تشخیص داده
- پارامتر صحت معمولا به صورت درصد بیان می‌شود.

$$Accuracy = (TP+TN) / (TP+FN+FP+TN)$$

معیار اندازه‌گیری کیفیت یک دسته‌بند

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

- معیار حساسیت **Sensitivity**
- «نرخ پاسخ‌های مثبت درست» True Positive Rate
- نسبتی از موارد مثبت است که آزمایش آن‌ها را به درستی به عنوان نمونه مثبت تشخیص داده

$$\text{Sensitivity (TPR)} = TP / (TP+FN)$$

معیار اندازه‌گیری کیفیت یک دسته‌بند

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

معیار حساسیت Sensitivity

- دسته‌بند، به چه اندازه در تشخیص تمام افراد مبتلا به بیماری موفق بوده‌است.

- تعداد افراد سالمی که توسط دسته‌بند به اشتباه به عنوان فرد بیمار تشخیص داده شده‌اند(FP)،
- تاثیری در محاسبه این پارامتر ندارد
- معمولا به صورت درصد بیان می‌شود

- هدف: دستیابی به نهایت صحت در تشخیص نمونه‌های کلاس مثبت

معیار اندازه‌گیری کیفیت یک دسته‌بند

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

- **معیار خاصیت Specificity**
- «نرخ پاسخ‌های منفی درست» True Negative Rate

- صحت تشخیص کلاس منفی حائز اهمیت باشد.
- خاصیت به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به درستی به عنوان نمونه منفی تشخیص داده است.
- معمولا به صورت درصد بیان می‌شود

$$\text{Specificity (TNR)} = TN / (TN+FP)$$

معیار اندازه‌گیری کیفیت یک دسته‌بند

«منحنی مشخصه عملکرد سیستم» Receiver Operating Characteristic | ROC
منحنی بیانگر ارتباط بین دو پارامتر حساسیت و خاصیت

- پیش‌بینی عالی: مقادیر Sensitivityو Specificityهر دو صد درصد

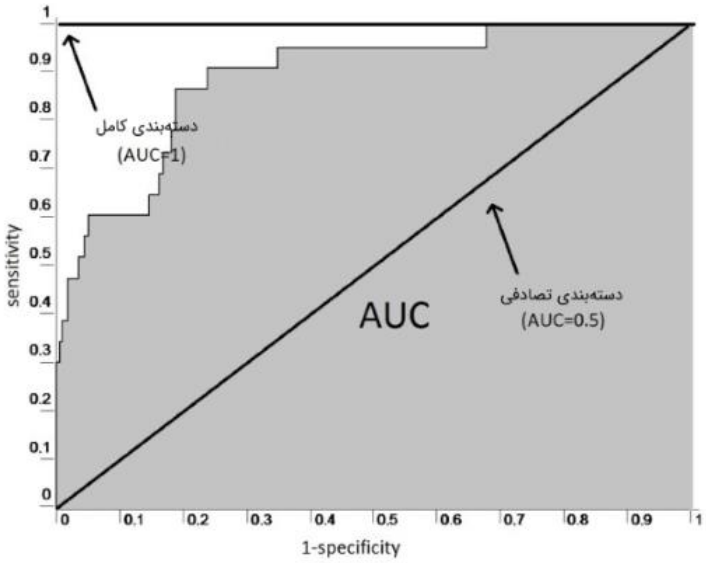
احتمال وقوع این اتفاق در واقعیت بسیار کم است و همیشه یک حداقل خطایی وجود دارد.

- پارامترهای حساسیت و خاصیت، بنابر ماهیتی که دارند همواره در رقابت با یکدیگر
- افزایش یکی با کاهش دیگری همراه است و برعکس.

- ضرورت ابزاری دیگر برای ارزیابی کیفیت دسته‌بندها شد
- . «منحنی مشخصه عملکرد سیستم» ((Receiver Operating Characteristic | ROC،
- منحنی بیانگر ارتباط بین دو پارامتر حساسیت و خاصیت

معیار اندازه‌گیری کیفیت یک دسته‌بند

«منحنی مشخصه عملکرد سیستم» Receiver Operating Characteristic | ROC
منحنی بیانگر ارتباط بین دو پارامتر حساسیت و خاصیت



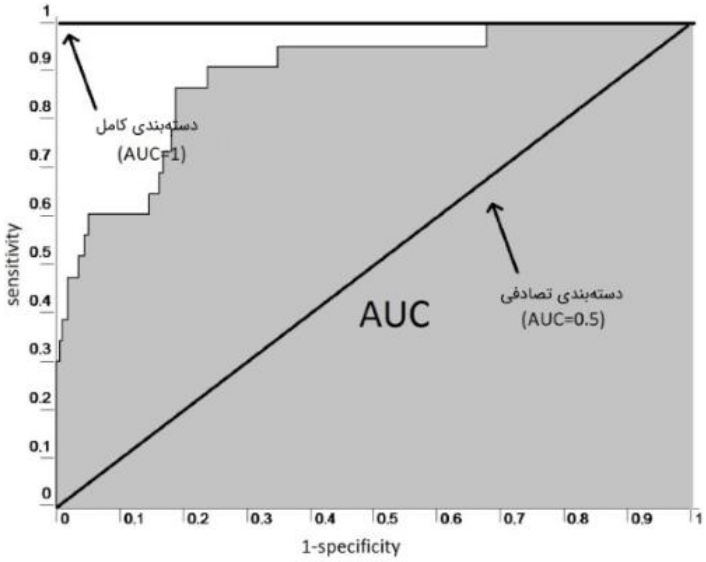
- محور عمودی این نرخ مثبت صحیح Sensitivity،
- محور افقی نشان‌دهنده مقدار نرخ مثبت غلط 1-Specificity

- نتایج مختلف دسته‌بندی نشانگر نقاط مختلف بر روی این نمودار
- در نهایت یک منحنی را تشکیل می‌دهند.

معیار اندازه‌گیری کیفیت یک دسته‌بند

«منحنی مشخصه عملکرد سیستم» Receiver Operating Characteristic | ROC

منحنی بیانگر ارتباط بین دو پارامتر حساسیت و خاصیت

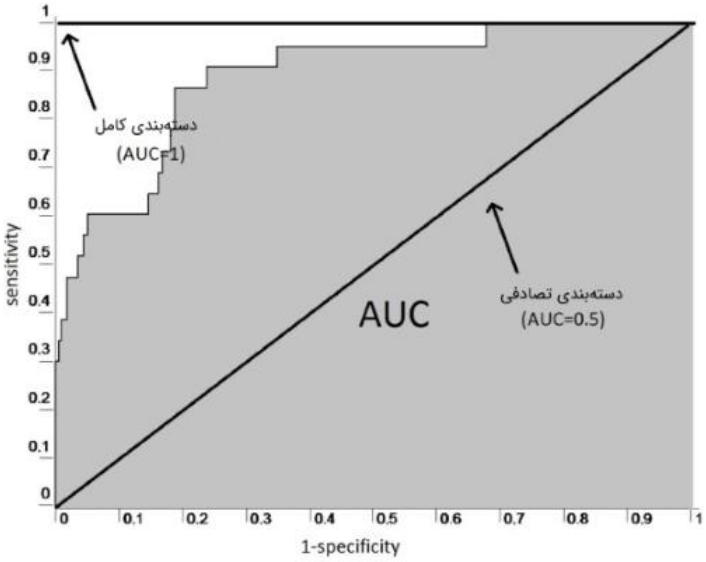


بهترین حالت و با فرض طبقه‌بندی صد درصد صحیح در هر دو دسته، نقطه مربوطه نقطه (۰,۱)
با فرض دسته‌بندی به صورت تصادفی، نقطه متناظر در منحنی، یکی از نقاط موجود روی خط واصل نقطه (۰,۰) و نقطه (۱,۱) خواهد بود.

در واقعیت، منحنی حاصل از یک دسته‌بندی،منحنی بین این دو حالت است.

معیار اندازه‌گیری کیفیت یک دسته‌بند

«منحنی مشخصه عملکرد سیستم» Receiver Operating Characteristic | ROC
منحنی بیانگر ارتباط بین دو پارامتر حساسیت و خاصیت



- مساحت زیر این نمودار Area Under Curve، به عنوان یک معیار برای ارزیابی عملکرد دسته‌بند

- در حالت ایده‌آل، مساحت زیر منحنی برابر با بیشترین مقدار خود، یعنی یک است.

- بنابراین، هر چه مساحت زیر نمودار به عدد یک نزدیک تر باشد، به معنای بهتر بودن عملکرد دسته‌بند است.

معیار اندازه‌گیری کیفیت یک دسته‌بند

- پارامتر مقدار پیش‌بینی شده مثبت Positive Prediction Value
- مقدار پیش‌بینی شده منفی Negative Predictive Values

- بیان «نسبت پاسخ‌های درست در هر دسته»
- **ارزش اخباری مثبت:** چند درصد از الگوهایی که دسته‌بند آن‌ها را مثبت تشخیص داده، در واقعیت هم مثبت هستند
- **ارزش اخباری منفی:** چند درصد از نمونه‌هایی که عضو دسته منفی تشخیص داده شده‌اند، در واقعیت هم عضو همین دسته هستند.

$$PPV=TP / (TP+FP)$$

$$NPV=TN / (TN+FN)$$

معیار اندازه‌گیری کیفیت یک دسته‌بند

- پارامتر «معیار اف» **F-Measure**

- ترکیب دو پارامتر حساسیت و ارزش اخباری مثبت

- پارامتر ارزش اخباری مثبت را اصطلاحا دقت (Precision)
- حساسیت را اصطلاحا یادآوری (Recall) می‌نامند
-

F-measure= 2 * (Recall * Precision) / (Recall + Precision)

بخش چهارم:
سایر کاربردهای داده کاوی

علوم انسانی دیجیتال

- علوم انسانی دیجیتال شکل‌های جدیدی از فعالیت‌های تحقیقاتی و دانشگاهی
- **هدف:** پژوهش، تدریس و انتشار فعالیت‌های مشترک بین‌رشته‌ای و مبتنی بر رایانه است.
- علوم انسانی دیجیتال چیزی فراتر از رواج ابزار رایانه‌ای و اینترنت است
- دیجیتالی شدن مقالات، کتاب‌ها و متون علوم انسانی به معنای علوم انسانی دیجیتال **نیست.**
- علوم انسانی دیجیتال شامل استفادهٔ سیستماتیک منابع دیجیتال در علوم انسانی و همچنین بازتاب در کاربرد آن‌هاست.
- علوم انسانی دیجیتال، ابزار و متدهای دیجیتالی را ارائه می‌دهدکه برای تولید و نشر دانش مورد استفاده قرار می‌گیرند
- منابع علوم انسانی صرفا دیگر فقط کلمات چاپ شده نیستند.
- با تولید و استفاده از برنامه‌های کاربردی و تکنیک‌های جدید، در علوم انسانی نوع جدیدی از آموزش و پژوهش ممکن شده است.

کاربردهای داده کاوی در علوم انسانی

علوم انسانی دربردارنده طیف وسیعی از رشته ها بوده و درهریک از آن‌ها می‌توان کاربردهای متنوعی از داده‌کاوی را برشمرد، مانند:

- متن کاوی
- داده کاوی در اقتصاد و علوم اقتصادی
- داده کاوی در علوم اجتماعی
- داده کاوی در حسابداری و حسابرسی
- داده کاوی در زمینه های مختلف رشته حقوق
- داده کاوی در گردشگری (مدیریت جهانگردی و هتلداری)
- داده کاوی در جغرافیا
- داده کاوی در کتابداری
- داده کاوی در گرایش های مختلف مدیریت مانند مدیریت بازرگانی، مدیریت دولتی، مدیریت صنعتی، مدیریت بیمه
- داده کاوی در مطالعات خانواده
- داده کاوی در علوم قضایی
- داده کاوی در علوم قرآن و حدیث
- داده کاوی در مدیریت مالی
- داده کاوی در مدیریت امور بانکی
- داده کاوی در مشاوره
- داده کاوی در روان شناسی
- داده کاوی در علوم سیاسی
- داده کاوی در تربیت بدنی و علوم ورزشی
- داده کاوی در زبان و ادبیات

متن کاوی چیست؟

متن کاوی :

فرایندذخیره سازی، پردازش تحلیل و کاوش متن

داده ها در انواع مختلف:اعداد، تصویر، صوت و متن

طبق گزارشی ۸۰ درصد داده های موجود در سراسر دنیا به صورت متن

متن ها از نوع داده های بدون ساختار یافته اند. البته متن می تواند حالت نیمه ساخت یافته هم داشته باشد.

.

ساخت یافته

ID	نام رشته	تعداد دانشجو
۱	فنی و مهندسی	۷۰۰
۲	پزشکی	۴۰۰
۳	علوم پایه	۴۰۰

نیمه ساخت یافته

```
<UNIVERSITY>
  <College ID="1">
    <Name>فنی مهندسی</Name>
    <Count>700</Count>
  </College>
  <College ID="2">
    <Name>پزشکی</Name>
    <Count>400</Count>
  </College>
  <College ID="3">
    <Name>علوم پایه</Name>
    <Count>400</Count>
  </College>
</UNIVERSITY>
```

غیر ساخت یافته

این دانشگاه 1500 دانشجو دارد که در گروه های فنی مهندسی، پزشکی و علوم پایه مشغول به فعالیت هستند. 700 دانشجو در رشته ی فنی و مهندسی، 400 دانشجو در رشته ی پزشکی و 400 دانشجو نیز در رشته ی علوم پایه در حال تحصیل هستند

مراحل متن کاوی

مراحل متن کاوی

پنج گام اساسی در متن کاوی:

- در گام اول اطلاعات مورد نیاز جمع آوری می شود.
  برای مثال اگر شما بخواهید نظر مردم درباره ی برند و یا محصولات خود را بدانید، نظرات و بحث های مردم را در سایت‌های مختلف جمع آوری می کنید.

- در مرحله ی دوم داده ها باید پیش پردازش و پاک سازی شوند.
  برای مثال ریشه یابی، حذف کلمات توقف، تبدیل حروف بزرگ به حروف کوچک، حذف اعداد و عملیات دیگر می توانند در مرحله ی پیش پردازش استفاده شوند.

مراحل متن کاوی

پنج گام اساسی در متن کاوی:

- در گام سوم داده های متنی به داده های دارای ساختار تبدیل می شوند و در واقع دارای ساختاری می شوند که برای متن کاوی مناسب باشند.
- انتخاب ویژگی هایی که به تحلیل بهتری منجر می شوند در این گام انجام می شود و برداری از کلمات ایجاد می شود.
- Bag of Words، TFIDF، Word2Vec از جمله روش های مورد استفاده برای تبدیل متن به بردار هستند.

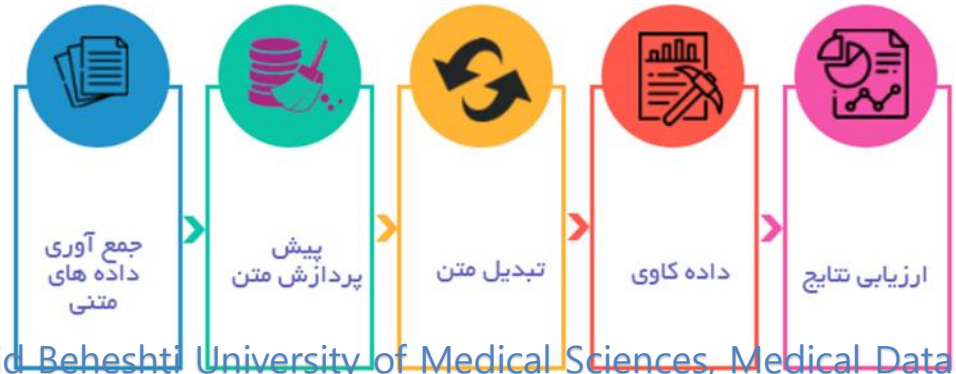
مراحل متن کاوی

پنج گام اساسی در متن کاوی:

- در گام چهارم، کار تحلیل و داده کاوی بر روی داده هایی که دارای ساختار شدند، انجام می شود.
  دسته بندی، خوشه بندی و **استخراج اطلاعات** روش‌های مورد استفاده برای متن کاوی هستند.

در دسته بندی(Classification) باید تعدادی متن برچسب دار داشته باشیم و از این داده‌ها برای برچسب گذاری متن‌های جدید که برچسب ندارند استفاده کنیم.

در خوشه بندی (Clustering) نیازی به داده‌ی برچسب‌دار نداریم و متن‌های مختلف بر اساس محتویاتی که دارند به خوشه های مختلفی دسته بندی می شوند.



مراحل متن کاوی

- **پنج گام اساسی در متن کاوی:**
- در گام آخر نتایج به دست آمده از مرحله ی قبل بررسی می شود.

تکنیک های متن کاوی

- دسته بندی متن
- استخراج اطلاعات
- بازیابی اطلاعات
- خوشه بندی
- خلاصه سازی
- فرکانس کلمات
- نظرکاوی

تکنیک های متن کاوی

دسته بندی متون

دسته بندی متن یکی از روش های آگاهانه ی یادگیری ماشین برای برچسب زنی متن ها در یکی از دسته های مشخص
متن های داده شده به دسته های از پیش مشخص شده اختصاص داده می شوند
دسته بندی عمل جمع آوری اسناد متنی و پردازش آن ها برای کشف دسته ی مناسب شان
نظرکاوی، شناسایی زبان، تعیین عنوان از کاربردهای دسته بندی متن هستند

فرض کنید تعدادی متن دارید که موضوع هر یک مشخص است. حال متن جدیدی به این متن ها اضافه می شود.
با استفاده از الگوریتم های دسته بندی و همچنین داشتن متن های با موضوع مشخص می توان، موضوع متن جدید را پیدا کرد.

تکنیک های متن کاوی

استخراج اطلاعات

فرایند استخراج اطلاعات معنی دار از مقادیر زیاد داده های متنی
تمرکز بر روی استخراج اسامی، ویژگی ها و ارتباط آن ها

ذخیره اطلاعات استخراج شده برای دسترسی و بازیابی در آینده در یک پایگاه داده

اثر بخشی و کارایی نتایج بر اساس دقت و صحت آن ها مورد ارزیابی قرار می گیرد

مثال : استخراج اسامی انسان ها، مکان ها، صفات و دیگر ویژگی ها

تکنیک های متن کاوی

بازیابی اطلاعات

فرایند استخراج الگوهای مرتبط بر اساس مجموعه از خاصی از کلمات یا عبارات

استفاده سیستم های استخراج اطلاعات از الگوریتم های متفاوتی برای ردیابی و نظارت بر رفتار کاربران و کشف داده های مرتبط با آن

موتور جستجوی گوگل و از مشهورترین سیستم های استخراج اطلاعات

تکنیک های متن کاوی

خوشه بندی

یکی از مهم ترین تکنیک های متن کاوی است

هدف: شناسایی ساختارهای درونی در اطلاعات متنی و سازماندهی آن ها در گروه ها یا همان خوشه هاست تا بتوان آن ها را تجزیه و تحلیل کرد.

یکی از چالش های مهم در خوشه بندی :
تشکیل خوشه های معنی دار از داده های متنی بدون برچسب و داشتن اطلاعات قبلی در مورد آن هاست.

تکنیک های متن کاوی

خلاصه سازی

پردازش خودکار داده ها برای تولید یک متن خلاصه که شامل اطلاعات ارزشمند برای کاربر است.

هدف :

دریافت اطلاعات متنی از چند منبع و خلاصه سازی آن بگونه ای است که مفهوم کلی و منظور متن حفظ شود.

تکنیک های متن کاوی

فرکانس کلمات

برای یافتن کلمات پر تکرار در یک متن

این تکنیک می تواند برای موارد متعددی مفید باشد.

یک مثال از این کاربرد:

وقتی که درخواست کاربر تحلیل می شود و بیشترین تعداد تکرار کلمات در متن درخواستی او مثلا در "سرویس ارسال کالا" باشد، در این صورت ممکن است درخواست او هم در این مورد باشد.

تکنیک های متن کاوی

نظرکاوی(sentiment analysis)

مطالعه ی نظرها، احساسات، ارزیابی ها، رفتار و عواطف افراد نسبت به موجودیت هایی مانند محصولات، افراد، سازمانها، موضوعات، حوادث

کاربرد:

مشخص کردن بازخوردهای مشتریان نسبت به سازمان یا شرکت و همچنین اطلاعاتی در مورد رقبا و روند کنونی بازار

افراد عادی هم می توانند از مزیت نظرکاوی بهره ببرند به این صورت که قبل از تصمیم گیری در مورد خرید یا اقدام به انجام کاری از نظرات افراد دیگر مطلع شوند.

کاربردهای های متن کاوی

کاربرد متن کاوی در بسیاری از صنایع از جمله:

- مراکز آموزشی
- مراکز بهداشتی
- شبکه های اجتماعی
- صنایع مرتبط با داروسازی
- پیش بینی آب و هوا
- حمل و نقل
- بیمه

کاربردهای های متن کاوی

مدیریت ریسک

یکی از دلایل شکست در کسب و کارها تحلیل نامناسب و ناکافی ریسک است

.

استفاده از نرم افزارهای مدیریت ریسک مانندSAS Text Miner که از متن کاوی استفاده می‌کند

سبب کمک به کسب کارها می‌شود تا همراه روندهای کنونی حرکت کرده و توانایی خود را برای کاهش ریسک ارتقا دهند.

به دلیل توانایی متن کاوی برای جمع آوری اطلاعات از منابع گوناگون و ارتباط دادن آنها به یکدیگر، سازمان ها می توانند به اطلاعات درست در زمان مناسب دسترسی پیدا کرده

در نتیجه فرایند مدیریت ریسک سازمان خود را بهبود دهند.

کاربردهای های متن کاوی

سرویس مراقبت از مشتریان

تکنیک های متن کاوی از جمله پردازش زبان طبیعی در فرایند مراقبت از مشتریان از اهمیت زیادی برخوردار است.

شرکت ها در حال سرمایه گذاری بر روی نرم افزارهای پردازش متن برای ارتقای تجربه ی کاربری مشتریان خود هستند که این کار با دریافت اطلاعات متنی از منابع متفاوت مانند نظرسنجی ها، بازخوردها، تماس های مشتریان انجام می شود.

هدف تحلیل متن در این کاربرد، کاهش زمان پاسخ گویی است.
استفاده از نرم افزارهای مدیریت دانش که از متن کاوی استفاده می کنند می تواند راه حل مناسبی برای مدیریت این داده ها باشد.

کاربردهای های متن کاوی

کشف کلاهبرداری

متن کاوی فرصت های بی شماری را برای صنایعی که اطلاعات متنی زیادی دارند، فراهم می کند.

از جمله ی این شرکت ها می توان به شرکت های مالی و بیمه اشاره کرد.

با ترکیب نتایج تحلیل متن با دیگر داده های ساختار یافته، زمان پاسخگویی به درخواست ها کمتر شده از طرفی کلاهبرداری شناسایی می شوند.

کاربردهای های متن کاوی

هوش تجاری

سازمان ها و کسب کارها از متن کاوی به عنوان بخشی از هوش تجاری خود استفاده می کنند.

در کنار کمک به ایجاد دیدی عمیق نسبت به رفتارهای مشتریان، متن کاوی به تحلیل نقاط قوت و ضعف آن ها نیز کمک کرده در نتیجه نوعی مزیت رقابتی برای آن ها خواهد بود.

ابزارهای متن کاوی مانند Cogito Intelligence Platform و IBM text analytics دیدی از عملکرد بازاریابی شرکت، مشتریان اخیر و روند بازار به آن ها می دهد.

کاربردهای های متن کاوی

تحلیل شبکه های اجتماعی

ابزارهای متن کاوی متنوعی برای تحلیل عملکرد شبکه های اجتماعی طراحی شده است.

این ابزارها به ردیابی و تفسیر متن های تولید شده در سایت های خبری، بلاگها، ایمیل و دیگر موارد می پردازد.

علاوه بر این، می توانند تعداد پستها، لایکها و فالوئرهای برند شما را در شبکه های اجتماعی مشخص کنند.

در نتیجه با استفاده از این ابزارها می توانید عکس العمل مردم نسبت به برند خود را ارزیابی کنید.

کاربردهای های متن کاوی

مدیریت دانش

در هنگام مدیریت مقدار زیادی از داده های متنی، یافتن اطلاعات مهم به صورت سریع، دشوار است.

سازمانها مخصوصا مراکز بهداشتی و درمانی با این چالش رو به رو هستند.

استفاده از نرم افزارهای مدیریت دانش که از متن کاوی استفاده می کنند می تواند راه حل مناسبی برای مدیریت این داده ها باشد.

کاربردهای های متن کاوی

غنی سازی محتوا

داده های متنی اطلاعات با ارزشی دارند که به صورت خام و بدون تحلیل نمی‌توان از آن ها بهره مند شد.

بهره برداری و استخراج اطلاعات مفید از این داده ها نیازمند صرف زمان زیادی از جانب یک انسان است تا تمام متن ها را خوانده و اطلاعات مفید آن را بصورت دستی استخراج نماید.

با توجه به اینکه تکنیک های تحلیل متن می توانند مقدار زیادی اطلاعات را مدیریت کنند، استفاده از این تکنیک ها هنگام کار با داده های متنی بسیار مفید خواهد بود.

این تکنیک ها، با ایجاد تگ هایی، محتوای در دسترس را مدیریت و خلاصه می کنند در نتیجه این داده ها می توانند برای اهداف مختلفی مفید واقع شوند.در واقع می توان با تکنیک های متن کاوی، محتوای موجود را غنی کرد.

کاربردهای های متن کاوی

جلوگیری از جرایم اینترنتی

طبیعت ناشناخته ی اینترنت و ارتباطات آن منجر به افزایش جرایم اینترنتی می شود.

امروزه برنامه هایی برای جلوگیری از این جرایم با بکارگیری متن کاوی توسعه یافته است که هر سازمانی می تواند از آن استفاده کند.

تشخیص متن های غیر اخلاقی و فیلتر کردن آن می تواند با تکنیک های مختلف متن کافی انجام شود.

کاربردهای های متن کاوی

فیلتر ایمیل های اسپم

ایمیل یکی از راه های ارتباطی ارزان، سریع و موثر است که با مشکل ایمیل های اسپم مواجه است.

متن کاوی می تواند به بهبود فرایندهای فیلتر این اسپم ها کمک کند.

با تحلیل متن های ایمیل و بکار بردن الگوریتم های متن کاوی می توان الگوهای نشان دهنده ی اسپم را در ایمیل ها شناسایی نمود.

کاربردهای های متن کاوی

تبلیغات شخصی سازی شده

با در اختیار داشتن اطلاعات کاربران می توان تبلیغاتی را به آن ها نشان داد که مطابق میل و سلیقه ی آن ها باشد.

در این صورت دیگر همه ی کاربران یک نوع تبلیغ را مشاهده نمی کنند.

با تحلیل محتوای متنی که یک نفر در اینترنت منتشر می کند، می توان ترجیحات او را شناسایی کرد. نظرات، بحث ها و گفت و گو های افراد منبع ارزشمندی برای پی بردن به گرایشات آن هاست.

داده کاوی در روان شناسی

- روان‌شناسی علمی است که رفتار و فرآیندهای ذهنی موجودات زنده بخوصوص انسان را مورد تجزیه و تحلیل قرار می‌دهد ودارای شاخه‌های متعدد و کاربردهای گسترده‌ای است.
- گستردگی علم روان‌شناسی باعث تولید داده‌های فراوانی در این حوزه شده‌است.
- استفاده از ابزارهای داده‌کاوی در این داده ها سبب تولید دانش می‌گردد که روان‌کاوان با استفاده از آن می‌توانند به نتایج کامل تر ی برای درمان بیماران بپردازند.
- مثال:
- پیش بینی شخصیت یک فرد با استفاده از روند محتوا منتشر شده در شبکه های اجتماعی
- پیش‌بینی عوامل موثر در وقوع یک اختلال با استفاده از داده های ثبت شده مراجعه کنندگان و..

- نمونه داده روان شناسی در سایت های معتبری از جمله kaggle.com
- وب سایت ایرانی http://dataheart.ir/

تاریخ دیجیتالی

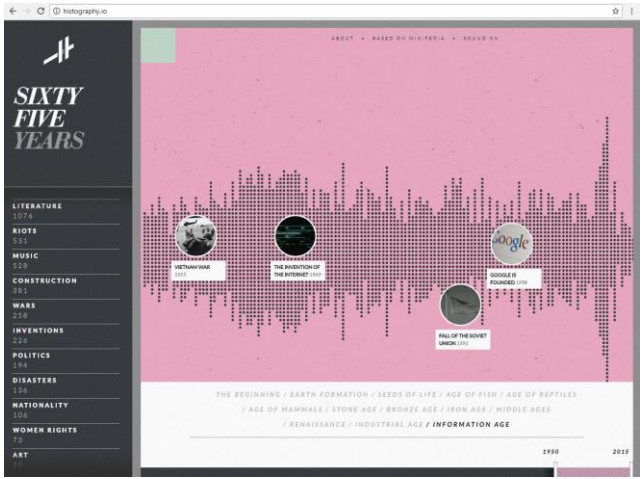
- این رشته در برگیرندۀ گسترۀ وسیعی از اطلاعات و امکانات و ابزار است که برخی از کاربردهای آن می‌توان به موارد زیر اشاره نمود:
- امکان پژوهش موضوعی در حجم عظیمی از منابع و آرشیوها
- داده کاوی و ساخت ماشین‌های یادگیرنده از طریق داده‌های تاریخی
- استفاده از ابزارها و تکنولوژی های جغرافیایی و فضایی برای درک تعامل محیطی و مردمی
- مدل سازی سه بعدی فضاها و ساختمان های تاریخی و...

تاریخ دیجیتالی

سایت Histography، درواقع یک جدول زمانی است که حوادث ۱۴ میلیون سال را نشان می‌دهد (از ابتدا تا سال ۲۰۱۵)
این سایت وقایع مختلف را از سایت Wikipedia به صورت اتوماتیک دریافت کرده و خود را به روزرسانی می‌کند.

کاربران می‌توانند تاریخ را بر اساس دسته‌بندی‌های مختلف از جمله: ادبیات، جنگ‌ها، موسیقی، شورش‌ها و ... در این جدول زمانی مشاهده کنند.

http://histography.io/

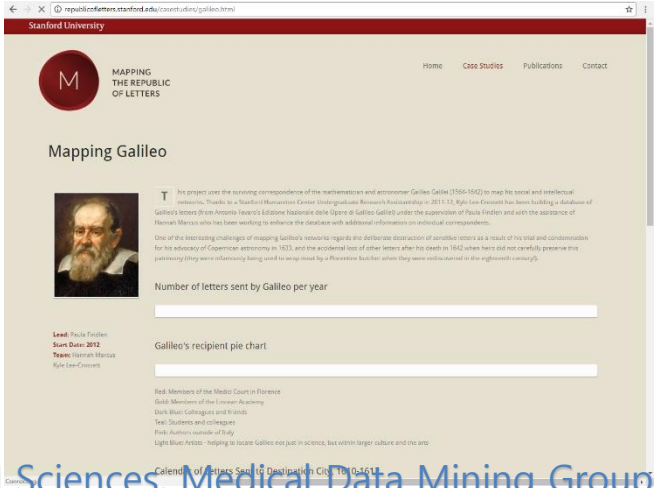


تاریخ دیجیتالی

پروژه Repulic of letters توسط دانشگاه استنفورد و با همکاری دانشگاه آکسفورد و دیگر موسسات، راه‌اندازی شده است.

این سایت، یک سایت بصری سازی پویای تاریخی است که کاربران می‌توانند در آن مکاشفه و جست‌وجو کنند.
به این منظور ابتدا مراحل آماده‌سازی اولیه داده‌های مورد مطالعه صورت گرفته است و بصری سازی شده‌اند،
سپس امکان بررسی فرضیه‌ها و پاسخ به سوالات پژوهشگران تا حدی فراهم آماده است.

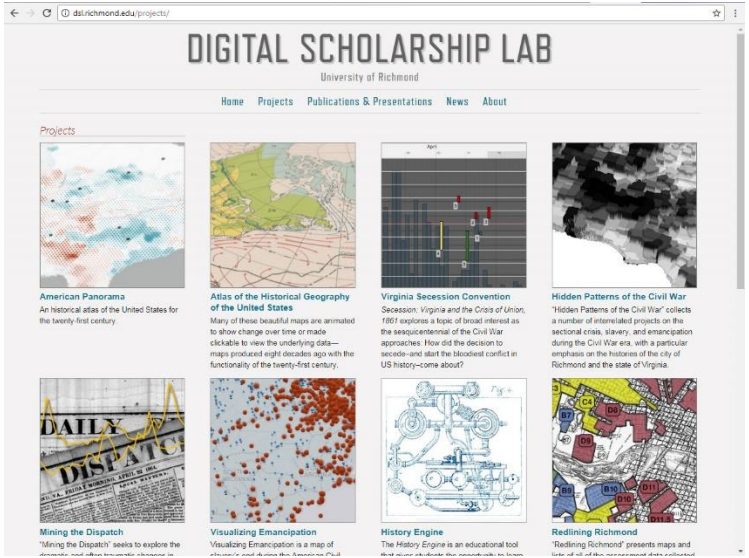
http://republicofletters.stanford.edu/



تاریخ دیجیتالی

- Digital Scholarship Lab، ارائه‌دهندهٔ پروژه‌های خلاقانهٔ علوم انسانی دیجیتال است که برای پژوهش و تدریس در دانشگاه ریچموند و دیگر جاها مورد استفاده قرار می‌گیرد. از جمله پروژه‌های انجام شده می‌توان به موارد زیر اشاره کرد :

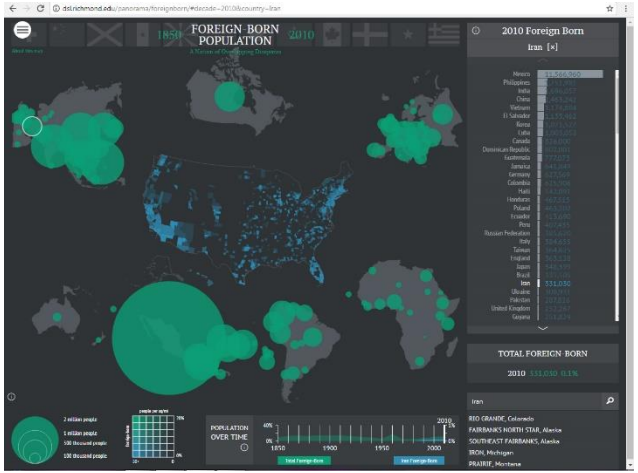
- http://dsl.richmond.edu



تاریخ دیجیتالی

- American Panorama: اطلسی است برای نشان دادن تاریخ آمریکا در قرن ۲۱ که شامل نقشه‌های مختلف مانند جمعیت متولد شدگان خارجی (به تفکیک کشورهای مختلف و امکان گزارش‌گیری و نمایش گزارشات روی نقشه)، نقشه کانال‌های مختلف در سال‌های ۱۸۶۰-۱۸۲۰ (با امکان گزارش‌گیری‌های مختلف، از جمله مواد سازنده کانال، نوع و میزان کالاهای جا به جا شده از طریق این کانال و ...)و دیگر نقشه‌ها و داده‌هاست.

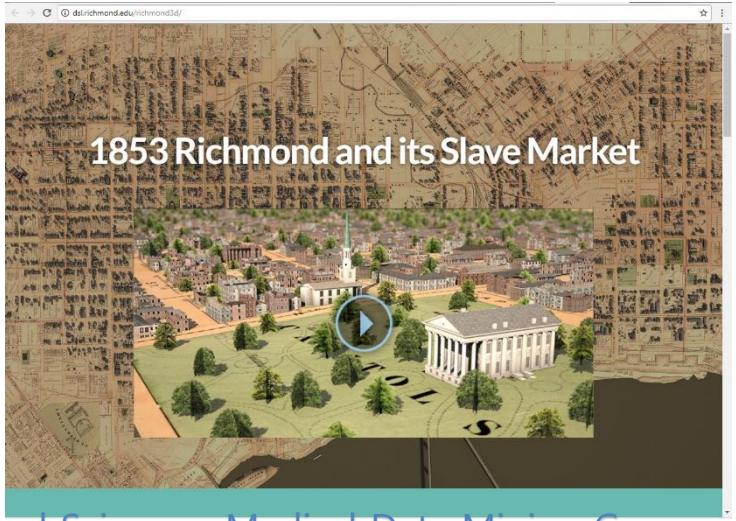
http://dsl.richmond.edu/panorama



تاریخ دیجیتالی

Hidden Pattern of Civil War این پروژه، اطلاعات و داده‌های مربوط به بحران‌ها، برده‌داری و موارد مرتبط دیگر در دوران جنگ داخلی با تاکید بر ریچموند و ویرجینیارا جمع‌آوری کرده و به صورت نقشه و متن اراءه داده است و از ابزار دیجیتال برای ارائه الگوهایی بصری برای سهولت تحقیق و بررسی استفاده کرده است که باعث ظهور جلوه‌های مختلف از این دورهٔ دراماتیک در زمینه‌های اجتماعی، سیاسی و نظامی گشته است. به عنوان مثال بازارهای برده‌فروشی را با نقشه نشان می‌دهند و یا نقشه‌هایی برای نشان دادن ازدواج، برده‌داری و یا رهایی در زمان جنگ داخلی طراحی نموده‌اند.

http://dsl.richmond.edu/richmond3d/



زبان شناسی رایانشی و ادبیات الکترونیکی(دیجیتالی)

- زبان شناسی رایانشی حوزه‌ای میان رشته‌ای است
- با بهره‌گیری از روش‌های آماری و قاعده بنیاد، به مدل سازی زبان طبیعی بپردازد.
- به شکل سنتی امر زبان‌شناسی رایانشی توسط دانشمندان کامپیوتری صورت می‌گرفت که در حوزه پردازش یک زبان خاص توسط کامپیوتر تخصص لازم را کسب کرده بودند .

- امروزه زبان شناسان رایانشی به عنوان اعضای گروه های میان رشته ای به فعالیت می‌پردازند که اعضای این تیم ها می توانند شامل زبان شناسانی که به شکل خاص در زمینه زبان‌شناسی همگانی تخصص دارند،
- کارشناسان زبان، افرادی با پیش زمینه و تا حدی دارای مهارت‌های عملی مرتبط با پروژه مورد نظر، و دانشمندان علم کامپیوتر باشند.

زبان شناسی رایانشی و ادبیات الکترونیکی(دیجیتالی)

ادبیات دیجیتالی، ادبیات کاغذی نیست که به صورت دیجیتال بازنشر می‌شود بلکه به صورت دیجیتال متولد شده است و هدف آن استفاده از ظرفیت‌های کامپیوتری

- **انواع مختلفی از ادبیات دیجیتالی :**

از جمله انیمیشن‌های کامپیوتری، ادبیات تصویری دیجیتالی، ادبیات آزمایشی ویدئویی و...

انواع ادبیاتی که از طبیعت قابل برنامه‌ریزی کامپیوتر بهره برده و سبب تولید آثاری پویا و ایجاد انواع متن‌ها و ادبیات با رویکردی ترکیبی

زبان شناسی رایانشی و ادبیات الکترونیکی(دیجیتالی)

- **ابزارهای علوم انسانی دیجیتال**
- امکان سنجش مسائل مختلف و تغییرات فرهنگی را با استفاده از Big Data(داده های حجیم)
- با دورخوانی به جای خوانش دقیق یک متن، می‌توان سنجید که نوشته‌ها بر چه واژگانی تأکید دارند
- و این واژگان چه جریان‌هایی را نشان می‌دهند.

- تحلیل پیکره‌های دیجیتالی
- امکان بررسی فرهنگ و روحیات غالب یک پدیده یا دوران را از نظر آماری و علمی

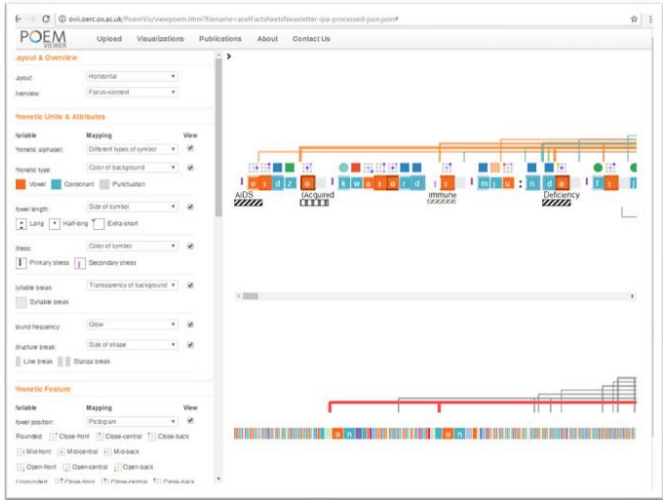
زبان شناسی رایانشی و ادبیات الکترونیکی(دیجیتالی)

Poem Viewer، یک ابزار تحت وب آزمایشگاهی تحت نظارت دانشگاه اکسفورد تهیه و در حال تکمیل

شعرهای موجود در پایگاه داده این سایت توسط نمودارها، رنگ‌ها و علائم مختلف بصری سازی شده‌اند

امکانات برای افزودن شعر به پایگاه داده این سایت

http://ovii.oerc.ox.ac.uk/PoemVis/index.html



زبان شناسی رایانشی و ادبیات الکترونیکی(دیجیتالی)

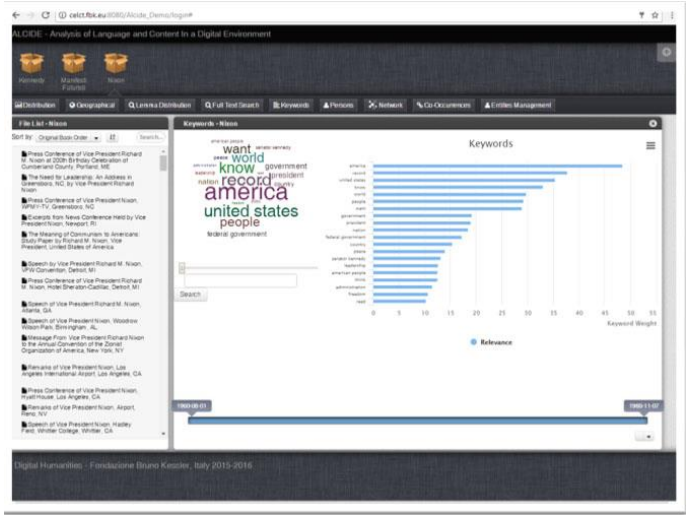
طراحی پایگاه داده های زبان فارسی به منظور ایجاد مجموعه ای بزرگ از پیکره های گوناگون زبان فارسی امروز
شامل متن های برگزیده ادبی، علمی، هنری، سیاسی و مانند اینها از گونه های نوشتاری و گفتاری فارسی
امکان جستجوی واژه ها،ترکیب‌ها، باهمایی‌ها و بررسی بسامد آن‌ها را به همراه گزارش‌های آماری متنوعی از متون
افزون بر اصل متن‌ها و واژه‌های به کاررفته در آن‌ها، معنی، مقوله دستوری، آوانگاشت و ریشه یا (بنواژه) بسیاری از واژه ها نیز در پایگاه وجود دارد.

http://pldb.ihcs.ac.ir/

زبان شناسی رایانشی و ادبیات الکترونیکی(دیجیتالی)

ALCIDE ابزاری است تحت وب، برای تحلیل زبان و مطالب در محیط دیجیتالی
جهت کمک به پژوهش‌های علوم انسانی در تحلیل داده‌های ادبیاتی و یا تاریخی در حجم بالا.
گزارش‌هایی که می‌توان از این ابزار دریافت کرد:
بسامد کلمات کلیدی، جغرافیای مطلب، بسامد اشخاص نام برده در متن، جست‌وجو در کل متن، جدول زمانی متون و ...

http://celct.fbk.eu:8080/Alcide_Demo/



زبان شناسی رایانشی و ادبیات الکترونیکی(دیجیتالی)

- مجموعه‌ای بسیار کامل از اختصارات و در واقع پلتفرمی برای آموزش الکترونیکی در دانشگاه زوریخ می‌باشد،
- طراحی منحصر به فرد این برنامه، جست‌وجوی اختصارات را هم از طریق متنی و هم از طریق عکی برای کارکترهایی که قابل خواندن و تشخیص نیستند مهیا می کند.

http://www.adfontes.uzh.ch/5231.php

علوم قرآنی و اسلامی دیجیتال

اهداف این رشته

- اشاعه فرهنگ غنی اسلام، تسهیل دست‌یابی به منابع و متون اصیل دینی
- سرعت بخشیدن به امر کاوش و پژوهش در حوزه علوم اسلامی با بهره‌گیری از فناوری اطلاعات
- به کارگیری امکانات رایانه‌ای

علوم قرآنی و اسلامی دیجیتال

این سایت ارائه دهندۀ نفل قول های کتاب مقدس با استفاده از ابزارهای مختلف و امکانات گوناگون است.

امکان جستو جو در کتاب مقدس بر اساس شاخص‌های مختلف

کتابخانه دیجیتال

این سایت کتابخانهٔ دیجیتالی است که متن‌های لاتین قدیمی را شامل می‌شود که برای مقایسهٔ زبانی در دوران قدیم بسیار کارآمد می‌باشد. امکان جست‌وجوی تمام و یا قسمتی از متن، اطلاعات و یا نویسنده را داراست.

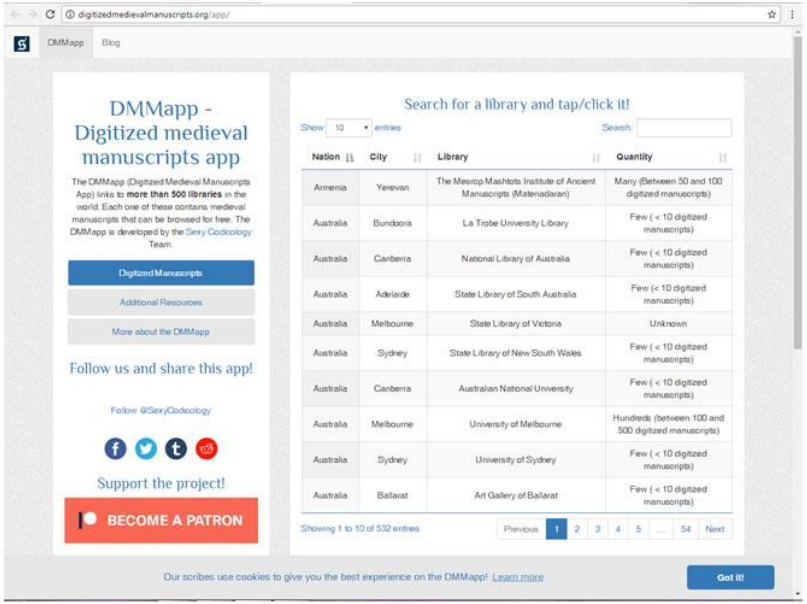
http://www.digiliblt.uniupo.it/index.php

کتابخانه دیجیتال

این اپلیکشن به بیش از ۵۰۰ کتابخانه در جهان وصل است که هر یک از آنان شامل نسخه‌های خطی قرون وسطی دیجیتال است می‌توان به صورت رایگان از آن‌ها استفاده کرد.

این اپلیکشن دسترسی به این منابع را تسهیل کرده و آن‌ها را ترویج می دهد.

http://digitizedmedievalmanuscripts.org/app/



برخی از ابزار دیجیتال مناسب پژوهش‌های مختلف

- **نرم افزارهای کیفی:** Nvivo, Atlas.ti, Maxqda, Qualtrics
- **نرم‌افزارهای کمی:** Spss, Stata, GNU PSPP, R and R Studio
- **ابزار کتابشناختی:** Zotero, EndNote, Mendeley, ReadCube

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

- پایگاه‌های داده در خارج از کشور

- این کتابخانه و پایگاه اطلاعاتی به پژوهشگران این امکان را می‌دهد که در راستای موضوعات مورد تحقیق خود، از هر موضوع و هر فرمتی را جست‌وجو کنند.
- وجود چکیده‌های حرفه‌ای و شاخص‌های مشخص و تکنولوژی‌های خلاقانه،
- جست‌وجو با دقت بالا و خروجی با کیفیت برتر را برای هر نوع پژوهش
- این پایگاه داده مناسب برای کاربران از دانش‌آموزان تا اساتید در رشته‌های متعدد.

- http://www.proquest.com

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

- پایگاه‌های داده در خارج از کشور

- پایگاه داده تحقیقاتی، ژورنال الکترونیکی، اشتراک مجلات، کتاب‌های الکترونیکی و ارائه‌دهندهٔ خدمات کتابخانه‌ای است
- پیشرو در بالا بردن سطح کیفی مطالب با استفاده از تکنولوژی و پلتفرم‌های بصری

- این پایگاه مجهز به ابزار قدرتمند همه‌کاره برای جست‌وجو و تحقیق در تمام منابع کتابخانه‌ای می‌باشد.
- مطالب و ابزار تکنولوژی قدرتمند این پایگاه برای تحقیق در هر زمینه و هر سطحی مناسب است.
- https://www.ebsco.com

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

- پایگاه‌های داده در خارج از کشور

- EMERALDپایگاه داده، بانک مقالات، ناشر و بانک مجلاتی است که امکان جست‌وجو را با ابزاری سریع و قدرتمند فراهم می‌نماید.

- http://www.emeraldinsight.com

- علاوه بر این سازمان‌های دیگری را جز زیرمجموعه خود دارد مانند crossref،
- که خدمات تحقیق و توسعه و به اشتراک‌گذاری زیرساخت‌های مناسب برای جوامع علمی را ارائه می‌دهند.

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

- پایگاه‌های داده در خارج از کشور

- بر اساس ابرداده‌های metadataجمع‌آوری شده و استفاده از تکنولوژی‌های استاندارد وب، ابزار متن‌باز و سرویس‌هایی را برای کمک به ناشران عضو و حل مشکلات آنان و رسیدن به بهترین راه‌حل‌های ممکن تولید نموده‌اند.
- از جمله خدمات آنان می‌توان به موارد زیر اشاره نمود:

- الف. Cited-by این سرویس مثل برعکس لینک دادن به منبع است، به این معنی که پژوهشگر می‌تواند افرادی که در مقاله و نشریات خود به مقاله و یا اثر آن‌ها لینک داده است را پیدا کند.

- https://www.crossref.org/services/cited-by

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

- پایگاه‌های داده در خارج از کشور

- از جمله خدمات آنان می‌توان به موارد زیر اشاره نمود:

- ب. Crossmarkاین سرویس به خوانندگان امکان سریع و راحت دسترسی به محتویات به روز را می‌دهد به این معنی که آنان با یک کلیک می‌توانند آیا محتویات مطلب مورد نظر، به‌روز، تصحیح و یا حتی رد شده است یا خیر.

- https://www.crossref.org/services/crossmark

- ج. Metadata Deliveryابرداده‌ها را به روشی معین برای این سرویس فرستاده، همه را با هم تجمیع کرده و برای استفاده موتورها ی جست‌وجو، بانک‌های اطلاعاتی، پایگاه‌های داده‌ای و ... آماده می‌نماید.

- https://www.crossref.org/services/metadata-delivery

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

پایگاه‌های داده در خارج از کشور

jstor

این پایگاه امکان دسترسی برای مقالات آکادمیک،کتاب و دیگر منابع را در ۷۵ رشته مختلف فراهم می‌کند.
این پایگاه به کشف طیف وسیعی از مطالب علمی توسط پلتفرم قدرتمند تحقیق و آموزش می‌کند.

https://www.jstor.org

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

پایگاه‌های داده در خارج از کشور

از ویژگی‌های این پلتفروم Jstor می‌توان به موارد زیر اشاره نمود:

- امکان دسترسی در هر زمان
- امکان دانلود متن کامل مطالب بدون هیچ‌گونه محدودیت
- استفاده از ابزار جدید برای جستو جوی مطالب مرتبط
- دارای ویژگی‌های شخصی سازی و امکان ذخیره‌سازی مرحله به مرحله کار پژوهشگر
- ارائه دهندۀ طرح کلی برای شروع یک مقاله و نحوۀ نوشتن و تهیه آن
- امکان جستوجوی مطالب فایل آپلود شده توسط کاربر
- *Text Analyzerجست و جوی مقالات و کتاب‌ها برای پیداکردن مطلب مشابه از طریق آپلود فایل
- https://www.jstor.org/analyze

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

پایگاه‌های داده در خارج از کشور

Elsevier

یک شرکت جهانی تجزیه و تحلیل اطلاعات است که به موسساتت، حرفه‌ای‌ها در علوم پیشبردی و مراقبت‌های بهداشتی کمک می‌کند. در تما می زمینه‌ها دارای مقالات و منابع مختلفاست ولی در زمینهٔ علوم پزشکی و بهداشتی قوی‌تر می‌باشد.

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

پایگاه‌های داده داخل از کشور

پایگاه‌های خوب و قدرتمندی هم در داخل کشور، ارائه‌دهندهٔ مقالات و منابع علمی هستند که در زیر به بعضی از آن‌ها اشاره شده است:

پرتال جامع علوم انسانی: وجه تمایز پرتال جامع علوم انسانی با سایر پایگاه داده های مشابه، دسته بندی موضوعی داده های علمی در آن می‌باشد که در قالب بیش از ۳۲۰۰ موضوع، حوزه های مختلف علوم انسانی را شامل می شود. مزیت دیگر پرتال جامع علوم انسانی دسترسی آسان و رایگان مخاطبان به منابع پرتال جامع علوم انسانی می باشد.

http://www.ensani.ir

پایگاه مجلات تخصصی نور: پایگاه مجلات تخصصی نور (نورمگز)در مهرماه سال ۱۳۸۴ به همت مرکز تحقیقات کامپیوتری علوم اسلامی، در راستای تسهیل و ترویج امر پژوهش و با هدف ایجاد بزرگ ترین بانک مجلات تخصصی علوم اسلامی و انسانی پایه ریزی گردید.

http://www.noormags.ir

برخی از پایگاه‌های داده و بانک مقالات مربوط به علوم انسانی

پایگاه‌های داده داخل از کشور

پایگاه‌های خوب و قدرتمندی هم در داخل کشور، ارائه‌دهندهٔ مقالات و منابع علمی هستند که در زیر به بعضی از آن‌ها اشاره شده است:

پایگاه اطلاعات علمی جهاد دانشگاهی:

همواره این پایگاه سعی نموده است که منابع علمی را با دو ویژگی "جامعیت و روزآمدی" به همراه سرویس‌ها و خدمات ویژه و کارآمد در ج هت اشاعه فرهنگ تحقیق و پژوهش به صورت دسترسی آزاد Open Accessدر اختیار محققان و دانش‌پژوهان قرار دهد.

http://fa.projects.sid.ir

بانک اطلاعات نشریات کشور:

هدف این پایگاه ایجاد مرجعی کامل و کارآمد از نشریات کشور در اینترنت، به منظور رفع نیاز محققین و علاقمندان و معرفی عناوین متنوع و بعضا مهجور نشریات و بسترسازی برای حضور موثر این رسانه دیرپا در صنعت نو پای اطلاع رسانی کشور است.

http://www.magiran.com

کاربردهای داده کاوی: جمع بندی

- ✓ حوزه های اصلی شامل کاربردهای علمی، تجاری و امنیتی
- ✓ حجم بسیار زیاد اطلاعات و خصایص متعد در تمام حوزه ها

- ✓ کاهش شدید هزینه ها، افزایش درآمدها و نجات زندگی انسان‌ها از دستاوردهای داده کاوی در هریک از حوزه های کاربردی آن است.

- ✓ کاربردهای تجاری:

  تشخیص صحت ادعای خسارت در بیمه، تشخیص سوء استفاده از کارت‌های اعتباری،

  تحلیل اطلاعات مشتریان یک سازمان،...
- ✓ کاربردهای علمی:

  حوزه های پزشکی، جغرافیائی و اقلیمی، فضا و سفرهای فضائی
- ✓ کاربردهای امنیتی:

  مبارزه با تروریسم، مقابله با نفوذگران به شبکه های کامپیوتری

ACECR

با تشکر از توجه شما