

Subject :

Year :

Month :

Date :

( به نام خدا )

تمرین : تمرین ۶ - چرا Distributed- Learning  
در مدیریت داده ها کاربرد دارد ؟

درس : مباحث ویژه

استاد : احمدزاده

اعضای گروه : همکاران - سمانه بهاری

ترم بهمن : ۱۴۰۲

رشته : کامپیوتر نرم افزار

Subject :

Year :

Month :

Date :

۱- جارا (DLT) Distributed Ledger Technology در مدیریت داده ها کاربرد دارد.

کنترل و توزیع کل دفتر کل (DLT) به دلایل زیر مدیریت داده ها کاربرد دارد.

تمرکززدایی؛ داده ها در شبکه ای از کامپیوترها بخش می شود و نتیجه هیچ نقطه مرکزی

واحد برای کنترل بازاری وجود ندارد این امر امنیت و قابلیت اطمینان افزایش می دهد

شفافیت؛ تمام تراکنش ها و تغییرات داده در دفتر کل ثبت می شود و برای همه

اعضای شبکه قابل مشاهده هستند. این شفافیت اعتماد را افزایش می دهد و امکان

حسابرسی آسان تر را فراهم می کند.

افزایش DLت؛ روش های رمزنگاری برای این سازی داده ها جلوگیری از تغییرات

غیر مجاز استفاده می کند به طور معمول هر تراکنش با امضاء دیجیتال و کلید خصوصی اعضا

می شود و به یک بلوک در زنجیره اضافه می شود.

تغییرات ناپذیری؛ پس از ثبت داده در DLT تغییرات آن بسیار دشوار و پرهزینه خواهد بود

غیر ممکن است؛ این امر یک تاییدیه تغییر ناپذیر از داده ها را تضمین می کند.

کارایی؛ DLT می تواند فرآیندهای سیستم مدیریت داده را تسریع کند زیرا نیاز به واسطه

ها و واسطه ها بین سیستم های مختلف را از بین می برد.

Subject :

Year :

Month :

Date :

قابلیت ردیابی DLT امکان ردیابی کامل مشا در تاریخچه دادن ها و مقایسه فی سئو این  
امر برای کاربردهای فاکتور مدیریت زنجیره تامین و تایید اطلاعات کالا بسیار مفید است  
به طور خلاصه DLT یک روش امن مختلف و کارآمد برای مدیریت دادن ها ارائه می دهد  
که برای کاربردهای مختلف در صنایع گوناگون مناسب است .  
کاهش هزینه ها (Cost Reduction) با حذف واسطه ها و خودکارسازی فرآیندها DLT  
می تواند هزینه های مربوط به مدیریت داده ها را به طور قابل توجهی کاهش دهد  
به طور کلی DLT یک روش نوین و کارآمد برای مدیریت داده ها ارائه می دهد که امنیت، شفافیت  
کارایی و قابلیت اطمینان را در همکاری با سیستم های متمرکز سنتی بهبود می بخشد

۲- برای ذخیره سازی داده های امن استفاده می شود (Blockchain)

هلاک چین به دلیل مختلف برای ذخیره سازی داده های امن استفاده می شود که هم ترین آن

عبارتند از: <sup>۱</sup> در یک سیستم بلاک چین داده ها به بای یک مکان مرکزی در یک شبکه ای کامپیوتری

ها توزیع می شود و در این باعث می شود هک کردن یا دستکاری داده ها بسیار دشوارتر شود

زیر هکرها به منظور هک کردن به تعداد زیادی کامپیوتر دسترسی پیدا کنند.

شفافیت: تمام تراکنش ها در بلاک چین به صورت عمومی قابل مشاهده هستند این

شفافیت امکان تقلب و فساد را کاهش می دهد و به طور مستقل بررسی کنند

تغیرناپذیری: پس از ثبت یک تراکنش در بلاک چین یک سیستم غیر آن غیر ممکن است این ویژگی

بلاک چین را به یک سیستم ذخیره سازی داده بسیار امن تبدیل می کند زیرا داده ها نمی

توانند حذف اصلاح یا دستکاری شوند و هر تلاش برای تغییر یک بلاک نیازمند تغییر تمام بلاک

های بعدی است که از نظر محاسباتی بسیار پرهزینه و غیر عملی است.

رهنماری: بلاک چین از تکنیک های رمزنگاری پیشرفته برای ایمن سازی داده ها

استفاده می کند هر بلاک در زنجیره با استفاده از یک هاش رمزنگاری شده به بلاک در

قبلی متصل است این امر دستکاری داده ها را بسیار دشوار می کند زیرا هر بلاک تابع



Subject :

Year :

Month :

Date :

تغییرات آن و تمام بلاک های جدیدی می شود.

مکانیسم اجماع برای افزودن یک بلاک جدید به بلاک چین باید توسط اکثریت شرکت کنندگان در شبکه تایید شود این مکانیسم اجماع از ثبت تراکنش های جعلی یا نادرست جلوگیری می کند.

به طور ~~مفصله~~ خلاصه بلاک چین با ترکیب تمرکززدایی شفافیت تغییرناپذیری و هزینه های مکانیسم اجماع یک سیستم ذخیره سازی داده بسیار امن را فراهم می کند که در برابر سرقت و تقلب عموماً شور و مقاومت است.

Subject :

Year :

Month :

Date :

۳- چرا Generative Adversarial Networks (GANs) در علم داده پشته کاربرد دارند؟

شبکه‌های مولد رقابتی (GANs) به دلیل جنبه و بازیگری تولیدی در علم داده پشته نقش مهمی را بر دارند. تولید داده‌های مصنوعی با کیفیت بالا (GANs) می‌تواند داده‌های مصنوعی را بیفتی تولید کنند که

شبیه داده‌های دنیای واقعی هستند. این امر به ویژه در مواردی که به دست آوردن داده‌های

در صورت محدودیت است که به دست آوردن داده‌های واقعی به دلیل مسائل مربوط به حریم خصوصی

هزینه یا محدودیت‌های اخلاقی دشوار است افزایش دادن داده‌های GAN می‌تواند با تولید

تفصیلات داده‌های موجود مجموعه داده‌های آموزشی را غنی‌تر کند. این امر به ویژه در زمینه

های مانند تصویربرداری پزشکی که داده‌های برجسته‌گذاری شده می‌باشد به دست

دست

کاربردهای متنوع GANs در طبقه‌بندی و سبب از وظایف از جمله سنتز تصویر، سنتز متن و

به تصویر ترجمه تصویر و غیره قابل استفاده هستند. انعطاف پذیری آن‌ها در کاربردهای

حکمت آن‌ها را به اندازه‌ای قدرتمند در علم داده تبدیل کرده است. یادگیری بدون نظارت GAN

چارچوبی برای یادگیری بدون نظارت ارائه می‌دهد. این یعنی که می‌تواند از داده‌های

بدون برچسب یاد بگیرد. این مزیت بزرگی است زیرا برچسب گذاری مجموعه داده‌ها می‌تواند

زمان به ویژه هزینه باشد.

Subject :

Year :

Month :

Date :

بهبود عملکرد؛ با تولید نه‌های داره اضافی GANS به تقویت قدرت مدل‌های یادگیری ماشین

کمک می‌کنند این امر به ویژه در زمینه‌های مانند مراقبت‌های بهداشتی که داده‌های مصنوعی تولید

شده توسط GANS هستند می‌تواند در تهیه سازی سناریو و آزمایش بدون به خطر

انراختن حریم خصوصی یا امنیت مجوعه داده‌های واقعی کمک کنند

نوع آماری و درام؛ ماهیت رقابتی GANS تجربه بهبود مداوم و نوآوری در تکنیک‌های تولیدی

دادن می‌شود در مجوعه GANS به دلیل توانایی نشان در تولید دادن‌های با کیفیت بالا افزایش

دادن‌ها کاربرد‌های متنوع یادگیری چون نظارت و بهبود عملکرد مدل ابزار قدرتمندی در علم

دادن بیشتر فته هستند.

تولیدی داده‌های مصنوعی با کیفیت بالا GANS می‌توانند داده‌های مصنوعی تولید کنند

که بی‌درستی داده‌های واقعی هستند این ویژگی در صورتی که جمع آماری داده‌های واقعی به دلیل

مسائل مربوط به حریم خصوصی هزینه با محدودیت‌های پهن دشوار است بسیار مفید است

افزایش دادن؛ می‌توانند با تولید دادن‌های بیشتر و متنوع‌تر به بهبود عملکرد مدل‌های

یادگیری ماشین کمک کنند امر به ویژه در زمینه‌های مانند تصویر برداری پزشکی که داده‌ها

های پزشکی گزاری شده محدود است ارزشمند است.



Subject :

Year :

Month :

Date :

۴-  $t-SNE$  و  $PCA$  چه تفاوتی دارند؟ تفاوت های اصلی بین  $t-SNE$  و  $PCA$  به شرح

زیر است: هدف  $PCA$  (تقلیل (حذفه اصلی))؛ هدف اصلی  $PCA$  کاهش ابعاد داده ها با حفظ

بیشترین واریانس ممکن است این روش به دنبال یافتن مولفه های اصلی (Principal Components)

Component است که جهت های داده های هستند که بیشترین تغییرات را نشان می دهند

تقسیم داده های با ابعاد بالا در ابعاد پایین (مثلاً 2 یا 3 بعد) به گونه ای است که ساختار

محلی داده ها حفظ شود  $t-SNE$  -  $t$  به دنبال نگاشت نقاط محلی به در فضای اصلی

به نقاط نزدیک در فضای پایین تر است

حفظ ساختار:  $PCA$  - به خوبی ساختار کلی و جهانی داده ها را حفظ می کند این روش برای

داده های را حفظ می کند این روش برای داده های که دارای ساختار محلی هستند

مناسب است  $t-SNE$  -  $t-SNE$  بیشتر بر حفظ ساختار محلی داده ها تمرکز دارد و ممکن

است ساختار محلی داده ها نیز دارد و ممکن است ساختار جهانی داده ها را به خوبی

حفظ کند به عبارت دیگر تفاوتی که در فضای اصلی به هم نزدیک هستند در فضای پایین تر

نزدیک خواهند بود.

کاربردها:  $PCA$



Subject :

Year :

Month :

Date :

PCA بیشتر برای کاهش ابعاد داده ها حذف نویز و استخراج ویژگی ها استفاده می شود.

همچنین می توان از به عنوان یک مرحله پیش پردازش برای سایر الگوریتم های یادگیری

ماشین مورد استفاده قرار گیرد  $t-SNE$  -  $t-SNE$  چگونه برای تجسم دادن ها استفاده

می شود این روش به خوبی می تواند خوشه ها الگوی ها موجود در داده ها را آشکار

کند.

حقایق پذیری: PCA از نظر محاسباتی کارآمدتر است و برای داده های بزرگ مقیاس

بزرگتر است  $t-SNE$  از نظر محاسباتی گران تر است و برای داده های بسیار بزرگ مناسب

است زمان بر باشد.

پارامترها: پارامترها کمتری برای تنظیم دارد و برای پارامترهای بیشتری است که باید

تنظیم شوند مانند  $\alpha$  می تواند بر نتیجه نهایی تاثیر بگذارد

نتیجه: PCA نتیجه  $R^2$  بسیار قابل تکرار است  $t-SNE$  نتیجه  $t-SNE$  می تواند

متغیر باشد و به مقدار کمی اولیه تصادفی و تنظیم پارامترها حساس است

به طور خلاصه از  $R^2$  زمانی است که نیاز به کاهش ابعاد داده ها با حفظ ساختار

کمی و در اینجا نشانه دار و بازمانده داده ها بزرگ هستند از  $t-SNE$  زمانی است که نیاز به

می توان به داده ها را تجسم کنید و ساختار محلی داده ها برایتان مهم است

Subject :

Year :

Month :

Date :

## ۵- چر UMAP برای Dimensionality Reduction استفاده می شود؟

ما فضا ساختاری کلی را در UMAP تلاش می کند تا ساختار توپولوژیکی داده ها را تا حد امکان حفظ کند این به این معنی است که نقاطی که در فضای اصلی به هم نزدیک هستند در فضای کاهش یافته نیز نزدیک باقی می ماند و روابط کلی بین خوشه ها و گروه ها نیز حفظ می شود. عملکرد خوب در داده های غیر خطی UMAP به خوبی با دادن نهایی که دارای ساختار غیر خطی هستند، کار می کند این الگوریتم قادر است پیچیدگی های موجود در داده ها را شناسایی و در فضای کاهش یافته بازتاب دهد.

۳- سرعت و مقیاس پذیری UMAP معمولاً سریع از برخی روش های دیگر کاهش ابعاد مانند SNE - t است و به خوبی با داده های بزرگ مقیاس می شود.

پارامترهای قابل تنظیم UMAP دارای پارامترهای است که به ما اجازه می دهد تا الگوریتم را برای براساس ویژگی ها خاص داده ها اهداف تحلیل تنظیم کند یا پارامترهای مانند

neighbors - تاثیر زیادی بر نتیجه نهایی دارند ۵- ابعاد نمایشی UMAP اغلب

برای ابعاد نمایشی ۲ یا ۳ اغلب برای نمایش های بصری از داده های راسخ

دیده. انعطاف پذیری UMAP می تواند برای انواع مختلفی از داده ها از جمله

داده های تصویری متنی و ژنومیک استفاده می شود.

Subject :

Year :

Month :

Date :

۷- NLTK چار برای پردازش زبان طبیعی (NLP) استفاده می شود؟  
NLTK (Natural Language Toolkit) یکی از مهم ترین و پرمکامبردترین کتابخانه ها برای

پردازش زبان طبیعی (NLP) در زبان برنامه نویسی پایتون است. دلایل استفاده از NLTK برای

NLP به شرح زیر است :

ابزار متنوع : NLTK شامل ابزارهای متعددی است که به کاربران این امکان را می دهد تا

وظایف متن مختلفی از جمله تجزیه و تحلیل متن، شناسایی بخش های گفتاری (POS tagging)

نشانه گذاری (tokenization) و استخراج ویژگی ها را انجام دهند. ۲- داده های آموزشی

این کتابخانه شامل مجموعه ای بزرگ از داده های آموزشی و متن مختلف است که می تواند

برای یادگیری و آزمایش و متن مختلف است که می تواند برای یادگیری و آزمایش الگوریتم های

مختلف NLP مورد استفاده قرار گیرد. ۳- آسانی در استفاده : API ساده و مستندات خوب

NLTK باعث می شود که یادگیری و استفاده از آن برای کاربران آسان باشد و حتی مبتدیان

توانند به راحتی با آن کار کنند. پشتیبانی از الگوریتم های مختلف NLTK : پشتیبانی از انواع

الگوریتم ها و روش های تحلیل زبان طبیعی شامل یادگیری ماشین، نحوی و معنایی را فراهم

می کند. جامعه کاربری بزرگ : به دلیل استفاده گسترده از NLTK، جامعه کاربری فعال و متنوعی

وجود دارد که می تواند در حل مشکلات و پاسخ به سوالات کمک کند.



Subject :

Year :

Month :

Date :

۴ NetworkX چرا برای تحلیل گراف ها کاربرد دارد؟

این کتابخانه پایتون است که برای ایجاد دشواری هلاله ساختار دینامیک و پیچیده شبکه های

پویا استفاده می شود. دلایل متعددی وجود دارد که NetworkX را برای تحلیل گراف ها بسیار

محبوب می کند. سادگی استفاده در NetworkX دارای یک رابط کاربری ساده

و سهمندی است که به کاربران امکان می دهد به راحتی گراف ها را ایجاد و ویرایش و تجزیه و تحلیل

کنند. گراف ها به رهیافت محبوب ساختار دادن به آن گونه ای طراحی شده است

که استفاده از آن برای افراد با سطوح مختلف تجربه برنامه نویسی آسان باشد.

۵- اختلاف پذیری از انواع مختلف گراف ها بستهبانی می کند از جمله گراف ها جهت دار

بدون جهت گراف ها چندگانه گراف های جهت دار بدون جهت گراف های چندگانه و

گراف های با خود حلقه های (K-100) که این اختلاف پذیری به کاربران اجازه می دهد

تا انواع مختلفی از شبکه ها را مدل سازی کنند. الگوریتم های متنوع شامل مجموعه ای

گسترده از الگوریتم های گراف است که برای تحلیل شبکه ها مورد استفاده قرار می گیرند

این الگوریتم ها شامل مورد زیر هستند: مرکزیت معایبه میزان اهمیت گره ها در شبکه

مانند درجه مرکزیت بینایی مرکزیت نزدیکی مرکزیت ویرانگری مرکزیت مسیرهای

کوتاه بافتن کوتاه ترین مسیر بین دو گره در شبکه.



Subject :

Year :

Month :

Date :

۸- چرا scrapy برای web scraping کاربرد دارد؟

scraping یک فریم ورک قدرتمند و متن باز برای web scraping استخراچ دادن

از وب و web crawling بسیارش ربات است دلایل متعددی وجود دارد که scrapy

را به یک انتخاب عالی برای این کارها تبدیل می کند ۱- سرعت و کارایی scrapy برای مدیریت

هم زیادی از داده ها به صورت همزمان طراحی شده است معماری ناچگان آن به این

معنی است که می تواند جنرین درخواست HTTP را به طور همزمان پردازش کند به باعث

افزایش سرعت و کارایی می شود ۲- انعطاف پذیری scrapy به بیت رقابت تنظیم و

انعطاف پذیری است شما می توانید به راحتی spiders عسوت ها سفارشی ایجاد

کنید تا وب سایت های مختلف را ببایش و داده ها را استخراج کنند همچنین می توانید از

middleware ها برای پردازش درخواست ها پاسخ مقابل و بعد از ارسال و دریافت

استفاده کنید - انتخابگرهای قدرتمند scrapy از انتخاب گرهای CSS و XPath

برای استخراج دادن ها از صفحات وب استفاده می کنند این انتخابگرها به شما امکان

می دهند تا به سادگی و با دقت داده های مورد نیاز خود را از HTML استخراج

کنید مدیریت داده ها scrapy دارای مکانیسم های داخل برای مدیریت داده های

استخراچ شده است شما می توانید داده های استخراج شده است شما می توانید داده ها را

به فرمت های مختلفی مانند CSV - JSON - XML ذخیره کنید

Subject :

Year

Month

Date

۹- BeautifulSoup برای تباری استفاده می شود.  
یک کتابخانه پایتون است برای تجزیه (Parsing) اسناد HTML و XML استفاده می شود.  
به بیان ساده تر به شما کمک می کند تا داده ها را از وب سایت ها استخراج کنید حتی HTML.  
آن وب سایت ها استخراج کنید حتی اگر HTML آن وب سایت تیره و مرتب نباشد در اینجا چند دلیل  
اصلی برای استفاده از BeautifulSoup - Soup آورده شده است تجزیه آسان HTML  
ساختار پیچیده HTML و XML را به یک درخت قابل پیمایش تبدیل می کند این درخت  
به شما امکان می دهد تا به راحتی عناصر و ویژگی های خاص متن های مورد نظر خود را پیدا کنید تصحیح  
خط های HTML اغلب HTML وب سایت ها کامل و بدون خطا نیستند Soup  
می تواند این خطاها را تا حدی تصحیح کند و به شما امکان می دهد تا وجود مشکلات در کد HTML داده ها را  
استخراج کنید. انتخابگرهای توی: BeautifulSoup از انتخاب های CSS و XPath  
پشتیبانی می کند این انتخابگرها به شما کمک می کنند تا مشخصه خاصی را در سند HTML یا  
XML پیدا کنید جستجوی آسان BeautifulSoup مقده های برای جستجو در درخت  
HTML یا XML فراهم می کند می توانید به سادگی نام های ویژگی های متن و سایر عناصر  
را جستجو کنید. سادگی سهولت استفاده در BeautifulSoup یک کتابخانه ساده  
و آسان برای یادگیری است API آن بسیار واضح و شهودی است و به شما امکان

Subject :

Year :

Month :

Date :

می دهد نامه سرعت شروع به استخراج دادن ها کنید به طور خلاصه Beautiful برای زیر

کاربرد دارد. استخراج دادن از وب سایت ها این امری تریب کاربرد Beautiful است می توانید

از آن برای استخراج اطلاعاتی مانند قیمت ها نام محصولات فعالیت غیره و غیره از وب سایت

ها استفاده کنید. تمیز کردن دادن های HTML اگر داده های HTML شما نامرتب

بوداری خلاصه Beautiful می تواند به شما کمک کند تا آن ها را تمیز کنید برای پردازش آسان

کنید. تو ماشین و مطالب وب می توانید Beautiful برای خود کار سازی و ظاهر بپی

عانه به کردن فرم ها کلین کردن روی لینک ها و داده های فایل ها از وب سایت ها استفاده کنید

تحلیل کنید این می تواند برای تحقیق بازار با بی حس پید اعراف مفید باشد.

Beautiful اغلب به همراه کتابخانه های دیگری مانند Scrapy برای دریافت

محتوای وب سایت ها استفاده می شود تا یک فرآیند کامل استخراج دادن از وب

را ایج د کند.



۱- چرا (DLT) Distributed Ledger Technology در مدیریت داده‌ها کاربرد دارد؟ (DLT) Distributed Ledger Technology به دلیل ویژگی‌ها و مزایای خاص خود در

مدیریت داده‌ها کاربردهای متنوعی دارد در ادامه به برخی از دلایل اصلی این کاربردها

است:   
 ۱- عدم وابستگی به مرکزیت: DLT به گونه‌ای طراحی شده است که هیچ نهاد

مرکزی برای مدیریت و تأیید داده‌ها وجود ندارد این ویژگی به کاربران این امکان را می‌دهد

که بدون نیاز به واسطه‌ها و نهاد های مرکزی با یکدیگر تعامل کنند و تراکنش‌ها را به سرعت و به صورت

۲- Blockchain چرا برای ذخیره سازی داده های امن استفاده می شود؟

Blockchain به عنوان یک فناوری پیشرفته برای ذخیره سازی داده های امن و بی خطی

ها و مزایای خاصی دارد که آن را به یک گزینه ایده آل برای تأمین امنیت اطلاعات تبدیل می‌کند.

در ادامه به چند دلیل اصلی استفاده از Blockchain به گونه ای طراحی شده است که هیچ

نمی‌تواند داده‌ها را پس از ثبت تغییر دهد یا حذف کند.

۳- چرا GANs (Generative Adversarial Networks) در علم داده پیشرفت دارند؟

دارند؟ به عنوان یکی از تکنیک‌های پیشرفته در یادگیری عمیق و علم داده در کاربردهای مختلفی

صورت استفاده قرار می‌گیرد این شبکه‌ها به دلیل ویژگی‌ها و خصوصیات منحصر به فرد خود قابلیت

های زیادی در ایجاد داده‌های جدید و واقعی دارند در ادامه به برخی از دلایل کاربردهای

اصلی GANs در علم داده پیشرفته است و می‌شود.



۳- PCA و t-SNE چه تفاوتی دارند؟

دو تکنیک متفاوت برای کاهش ابعاد داده هستند که در علم داده و یادگیری ماشین مورد

استفاده قرار می گیرند. با وجود اینکه هر دو برای تحلیل داده های چند بعدی و تجزیه و

تحلیل ویژگی ها استفاده می شوند تفاوت های مهمی در روش ها و کاربردهای آن

موجود دارد. روش های کاهش ابعاد حفظ ساختار داده ها از منظر جامع و پیچیدگی محاسباتی

۴- چرا UMAP برای Dimensionality Reduction استفاده می شود؟

این الگوریتم قدرتمند برای کاهش ابعاد داده ها است که در سال ۲۰۱۸ معرفی شد این روش

به دلایل زیر برای کاهش ابعاد داده ها استفاده می شود. حفظ ساختار همسایگی UMAP

۵- به طور معمول ساختار محلی داده ها را حفظ می کند این به این معناست که نقطه نزدیک به یکدیگر در

فضای اصلی در فضای جدید نیز نزدیک خواهند بود.

۶- NetworkX چرا برای تحلیل گراف کاربرد دارد؟

NetworkX یک کتابخانه پایتون است که برای ایجاد، دستکاری و تحلیل گراف ها و شبکه ها

استفاده می شود دلایل کاربرد آن در تحلیل گراف ها، ساختار داده های گراف و

الگوریتم های متنوع - قابلیت های محاسباتی - سادگی و خوانایی - پشتیبانی

از انواع گراف ها - جامعه کاربری NetworkX یک جامعه کاربری فعال و گسترده

دارد که پشتیبانی و منابع آموزشی زیاد را ارائه می دهد.

۷- NLTK چرا برای پردازش زبان طبیعی (NLP) استفاده می شود؟  
 NLTK (Natural Language Toolkit) یک کتابخانه محبوب پایتون است که برای پردازش زبان طبیعی (NLP) طراحی شده است. دلایلی که NLTK برای پردازش زبان طبیعی کاربرد دارد که کتابخانه جامع NLTK مجموعه ای از ابزارها و منابع را برای تحلیل متن فراهم می آورد این کتابخانه شامل توابع و فاکتورهای مختلفی برای عملیات اساسی NLP مانند تجزیه و تحلیل واژه ها، جملات و متن است.

۸- چرا - scrapy برای web scraping کاربرد دارد؟  
 scrapy یک فریم ورک قدرتمند و انعطاف پذیر برای web scraping است که به دلایل زیر بسیار کاربرد دارد - ۱- ساختار سازمان یافته scrapy با ارائه یک ساختار مشخص و سازمان یافته که نویسی را ساده تر می کند این فریم ورک از الگوی طراحی MVC پیروی می کند که باعث می شود کدهای خوانا و قابل نگهداری باشند.

۹- BeautifulSoup چرا برای کار با استفاده می شود؟  
 BeautifulSoup یک کتابخانه پایتون است که برای تجزیه و تحلیل HTML و XML استفاده می شود به عبارت ساده تر این کتابخانه به شما کمک می کند تا بتوانید محتوای یک صفحه وب (یا صرفاً فایل HTML یا XML) را به شکلی ساختار یافته و قابل

Senobar