



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر



ایجاد مدل داده برای نمادهای بورس تهران به منظور پیش بینی قیمت سهام
پایان نامه برای دریافت درجه کارشناسی
در رشته مهندسی کامپیوتر

نام

مریم کریمی

شماره دانشجویی

۸۱۰۱۹۵۴۶۱

استاد راهنما

دکتر هشام فیلی

مهر ۱۳۹۹

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

تعهدنامه اصالت اثر

باسمه تعالی

اینجانب مریم کریمی تأیید می‌کنم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان نامه قبلاً برای احراز هیچ مدرک هم سطح یا بالاتر ارائه نشده است.

کلیه حقوق مادی و معنوی این اثر متعلق به دانشکده فنی دانشگاه تهران می‌باشد.

نام و نام خانوادگی دانشجو :

مریم کریمی

امضای دانشجو :



چکیده^۱

امروزه با توجه به پیشرفت روش‌های هوش مصنوعی در تحلیل داده امکان پیش بینی روند برخی از رخدادهای آینده میسر شده است. یکی از حوزه‌هایی که تحلیل داده‌ها و پیش بینی در آن اهمیت بالایی دارد، بازار بورس است. وجود مدل داده‌ای و تحلیل داده‌های موجود می‌تواند به عنوان ابزاری قوی به کمک سرمایه‌گذاران در این حوزه آمده و توانمندی افراد را جهت سرمایه‌گذاری مناسب افزایش دهد. پیش بینی قیمت و ارزش سهام‌ها یکی از مسائل مهم در این بازار است که به وسیله آن می‌توان تصمیم بر خرید، فروش و یا نگهداری یک سهم کرد. همچنین با افزایش آگاهی مردم و الکترونیکی شدن امکان خرید و فروش سهام اقبال برای سرمایه‌گذاری در این بازار افزایش یافته در حالی که همه توانایی و زمان مورد نیاز برای یک سرمایه‌گذاری هوشمندانه را ندارند. نیاز به روش‌های تحلیل داده هوشمندانه و کمک به تصمیم‌گیری انسان بسیار احساس می‌شود. در این پروژه هدف فراهم کردن بستر استفاده از روش‌های هوش مصنوعی با جمع‌آوری داده‌های با ارزش بازار بورس تهران در ساختار مناسب است. در اختیار داشتن این داده‌ها زمینه را برای استفاده از انواع روش‌ها برای پیش بینی آینده بازار میسر می‌سازد. وجود یک مدل داده قوی که تمامی اطلاعات مؤثر بر روند بازار را شامل می‌شود یکی از ارزش‌های اساسی است که کمک می‌کند تا بازار بورس را از جنبه‌های مختلفی که تا قبل از این ممکن نبود بررسی کرد. پس از در اختیار داشتن این مدل داده‌ای می‌توان با روش رگرسیون^۲ قیمت آینده سهام را پیش بینی کرده و یا از طریق طبقه بندی^۳ سهام را در یکی از دسته‌های خرید، فروش و نگهداری قرار داد. برای این کار باید از میان ویژگی^۴ های متعددی که استخراج شده‌اند، چند دسته از ویژگی‌هایی را که تأثیر گذاری بیشتری دارند، در نظر گرفت. از میان روش‌های یادگیری ماشین^۵ روش مناسب را به عنوان پایه^۶ انتخاب کرد و راه‌کاری برای ارزیابی مناسب و اعتبار سنجی نتایج حاصله نمود.

کلمات کلیدی: جمع‌آوری داده بورس، مدل داده بورس، پیاده سازی خزنده

¹ Abstract

² Regression

³ Classification

⁴ Feature

⁵ Machine Learning

⁶ Baseline

فهرست مطالب

فصل ۱: مقدمه و بیان مساله.....	۱
۱-۱- مقدمه.....	۲
۱-۲- شرح مسئله تحقیق.....	۲
۱-۳- تعریف موضوع تحقیق.....	۲
۱-۴- اهداف و آرمان‌های کلی تحقیق.....	۳
۱-۵- روش انجام تحقیق.....	۳
۱-۶- ساختار پایان نامه.....	۴
فصل ۲: مفاهیم اولیه و پیش زمینه ساخت مدل داده ای.....	۵
2-1- بررسی ساختار سایت شرکت مدیریت فناوری بورس تهران.....	۶
۲-۲- ایجاد محیط توسعه خزنده.....	۹
۲-۳- انتخاب پایگاه داده و نگاشت دهنده رابطه به شیء(ORM).....	۱۰
۲-۴- خلاصه و جمع بندی.....	۱۰
فصل ۳: مدل داده ای و ساختار خزنده.....	۱۲
3-1- مدل داده ای.....	۱۳
3-1-2- جدول Instrument.....	۱۳
3-1-3- جدول DayPriceThreshold.....	۱۵
3-1-4- جدول DayTradeSummary.....	۱۵
3-1-5- جدول ClientTradeInfo.....	۱۶
3-1-6- جدول Shareholder.....	۱۸

۱۹.....	Share جدول 3-1-7-.....
۱۹.....	۳-۲- ساختار خزنده.....
۲۲.....	۳-۳- خلاصه و جمع‌بندی.....
۲۳.....	فصل ۴: نحوه استفاده از خزنده و دریافت داده ها.....
۲۴.....	4-1- آماده سازی سیستم.....
۲۴.....	۴-۲- اجرای خزنده.....
۲۶.....	۴-۳- خلاصه و جمع‌بندی.....
۲۷.....	فصل ۵: جمع‌بندی و پیشنهادها.....
۲۸.....	۵-۱- جمع‌بندی.....
۲۸.....	۵-۲- دستاوردها.....
۲۸.....	۵-۳- پیشنهادها.....
۳۰.....	فصل ۶: مراجع.....

فهرست شکل‌ها

شکل (2-1)	اطلاعات نماد فولاد در سایت	۶
شکل (۲-۲)	لیست نماد های بازار بورس تهران	۷
شکل (۳-۲)	اطلاعات یک روز معاملاتی نماد فولاد	۸
شکل (۴-۲)	معماری کتابخانه Scrapy	۱۰
شکل (3-1)	مدل داده ای بازار بورس تهران	۱۳
شکل (3-2)	جدول Instrument	۱۴
شکل (3-3)	جدول DayPriceThreshold	۱۵
شکل (3-4)	جدول DayTradeSummary	۱۶
شکل (3-5)	جدول ClientTradeInfo	۱۸
شکل (۶-۳)	جدول Shareholder	۱۸
شکل (3-7)	جدول Share	۱۹
شکل (۸-۳)	نمونه کد های نماد استخراج شده	۲۰
شکل (۹-۳)	نمونه از log های اجرای خزنده	۲۲
شکل (۱-۴)	فایل settings.py رشته اتصال به پایگاه داده	۲۴
شکل (4-2)	متغیر LOG_FILE در فایل settings.py	۲۵
شکل (4-3)	متغیر JOBDIR در فایل settings.py	۲۵

فصل ۱: مقدمه و بیان مساله

در این فصل نخست به شرح مسئله، تعریف آن و روش کلی تحقیق پرداخته، سپس مساله و موضوع مورد بررسی در این پایان نامه و اهداف و آرمان های کلی تحقیق بیان شده و در نهایت ساختار پایان نامه ی پیش رو ذکر شده است.

۱-۱- مقدمه

امروزه با توجه به پیشرفت روش‌های هوش مصنوعی در تحلیل داده امکان پیش بینی روند برخی از رخدادهای آینده میسر شده است. یکی از حوزه‌هایی که تحلیل داده‌ها و پیش بینی در آن اهمیت بالایی دارد، بازار بورس است. وجود مدل داده‌ای و تحلیل داده‌های موجود می‌تواند به عنوان ابزاری قوی به کمک سرمایه‌گذاران در این حوزه آمده و توانمندی افراد را جهت سرمایه‌گذاری مناسب افزایش دهد.

۱-۲- شرح مسئله تحقیق

پیش بینی قیمت و ارزش سهام‌ها یکی از مسائل مهم بازار بورس است که به وسیله آن می‌توان تصمیم بر خرید، فروش و یا نگهداری یک سهم کرد. اینگونه تصمیمات وقت بسیار زیادی را از سرمایه‌گذاران می‌گیرد در حالی که با استفاده از تکنولوژی‌های تحلیل داده می‌توان بسیار زمان مورد نیاز را کاهش داده و به افراد بیشتری اجازه داد تا سرمایه‌های خود را به طریقه صحیح در این بازار به کار گیرند. همچنین با افزایش آگاهی مردم و الکترونیکی شدن امکان خرید و فروش سهام اقبال برای سرمایه‌گذاری در این بازار افزایش یافته در حالی که همه توانایی و زمان مورد نیاز برای یک سرمایه‌گذاری هوشمندانه را ندارند. نیاز به روش‌های تحلیل داده هوشمندانه و کمک به تصمیم‌گیری انسان بسیار احساس می‌شود.

۱-۳- تعریف موضوع تحقیق

در این پروژه هدف فراهم کردن بستر استفاده از روش‌های هوش مصنوعی با جمع‌آوری داده‌های با ارزش بازار بورس تهران در ساختار مناسب است. پیاده سازی یک خزنده جهت جمع‌آوری داده‌ها مورد نیاز است. این خزنده باید در مقابل چالش‌هایی که برای آن وجود دارد مقاوم باشد. تغییرات مداوم صفحات وب، مسدود کردن IP، سرعت پایین پاسخ‌گویی به درخواست‌ها، محتوای پویای صفحات وب و خطاهای مختلف صورت پذیرفته حین دریافت داده مشکلاتی هستند که باید برای آن‌ها راه حلی ارائه شود [1]. پس از در اختیار داشتن خزنده می‌توان داده‌های مدنظر را جمع‌آوری کرده و به وسیله این داده‌ها زمینه برای استفاده از انواع روش‌ها برای پیش بینی آینده بازار میسر می‌شود. وجود یک مدل داده قوی که تمامی اطلاعات مؤثر

بر روند بازار را شامل می شود یکی از ارزش های اساسی است که کمک می کند تا بازار بورس را از جنبه های مختلفی که تا قبل از این ممکن نبود بررسی کرد. پس از در اختیار داشتن این مدل داده ای می توان یا با روش رگرسیون^۱ قیمت آینده سهام را پیش بینی کرده و یا از طریق طبقه بندی^۲ سهام را در یکی از دسته های خرید، فروش و نگهداری قرار داد. برای این کار باید از میان ویژگی^۳ های متعددی که استخراج شده اند، چند دسته از ویژگی هایی را که تأثیر گذاری بیشتری دارند، در نظر گرفت. از میان روش های یادگیری ماشین^۴ روش مناسب را به عنوان پایه^۵ انتخاب کرد و راه کاری برای ارزیابی مناسب و اعتبار سنجی نتایج حاصله نمود.

۴-۱- اهداف و آرمان های کلی تحقیق

ایجاد یک مدل داده ای قوی از بازار بورس تهران که تمامی داده های موجود مربوط به هر یک از نماد ها را دارا ست. این داده شامل اطلاعات سهامداران مهم و تاریخچه خرید و فروش سهام شان، خرید و فروش مشتری های حقیقی و حقوقی و اطلاعات معاملاتی هر روز یک سهام است. با دستیابی به این مدل می توان ابزار های تحلیل داده را قدرتمند تر اجرا کرده و الگوهای متفاوتی را تشخیص داد. همچنین پیاده سازی یک خزنده قابل گسترش که بتواند در آینده داده های به روز بازار بورس را استخراج نماید.

۵-۱- روش انجام تحقیق

روش انجام تحقیق در گام های زیر خلاصه می شود:

- مطالعه اولیه در مورد بازار بورس تهران و تحقیق در مورد رابط های برنامه نویسی

¹ Regression

² Classification

³ Feature

⁴ Machine Learning

⁵ Baseline

- بررسی ساختار داده‌های سایت tsetmc.ir و طراحی شمای پایگاه داده
- پیاده سازی خزنده برای جمع آوری اطلاعات از سایت tsetmc.ir
- راه اندازی سرور برای اجرای خزنده به صورت منظم
- ارزیابی نتایج و مصور سازی داده‌ها

۶-۱- ساختار پایان نامه

فصل دوم، شامل بررسی تعاریف اساسی مربوط به حوزه‌ی بورس، مفاهیم اولیه و اجزای اساسی یک خزنده و مروری بر پیش‌زمینه‌های مورد نیاز برای درک هرچه بهتر ساخت مدل داده‌ای بورس است.

فصل سوم در برگیرنده‌ی توضیح مربوط به مدل داده ای پیشنهادی و ساختار خزنده مربوطه است.

در فصل چهارم در مورد نحوه اجرای خزنده و استفاده از آن بر روی سایر سیستم ها صحبت خواهیم کرد. ابتدا ابزار های مورد نیاز برای سیستمی که بر روی آن اجرا صورت می گیرد توضیح داده می شود و سپس راه اندازی خزنده و انواع نحوه استفاده از آن ذکر می شود.

در نهایت، در فصل پنجم، نتیجه گیری های کلی حاصل شده در این تحقیق بیان می شود و محدودیت ها مورد بحث قرار می گیرد. همچنین پیشنهادهایی برای ادامه ی مسیر به علاقمندان این حوزه ی ارائه خواهد شد.

فصل ۲: مفاهیم اولیه و پیش زمینه ساخت مدل

داده ای

در فصل پیش رو تعاریف اساسی مربوط به حوزه ی بورس، مفاهیم اولیه و اجزای اساسی یک خزنده و مروری بر پیش زمینه های مورد نیاز برای درک هرچه بهتر ساخت مدل داده ای بورس است.

۱-۲- بررسی ساختار سایت شرکت مدیریت فناوری بورس تهران

تمامی اطلاعات مربوط به بازار بورس تهران در وبسایت به آدرس tsetmc.com توسط شرکت مدیریت فناوری بورس تهران در اختیار سرمایه گذاران قرار می گیرد تا تصمیمات سرمایه گذاران با آگاهی و شفافیت صورت پذیرد. این وبسایت اطلاعات مربوط به تمامی نمادها، معاملات انجام شده، عرضه و تقاضا، سهامداران و خریداران حقیقی و حقوقی را شامل می شود (شکل ۱).



شکل (۱-۲) اطلاعات نماد فولاد در سایت

در ابتدا برای به دست آوردن این اطلاعات به بررسی رابط های برنامه نویسی^۱ موجود و مزایا و معایب آنها پرداختیم. به دنبال رابط برنامه نویسی که تمامی اطلاعات را به صورت جامع در اختیار دهد گشتم. چند نمونه از این رابط های ارائه شده که باید مورد بررسی قرار گیرند عبارتند از: وب سرویس شرکت مدیریت فناوری بورس تهران، وب سرویس اطلاعات مالی مبنا و وب سرویس پویا. چنین رابط برنامه

¹ Application Programming Interface (API)

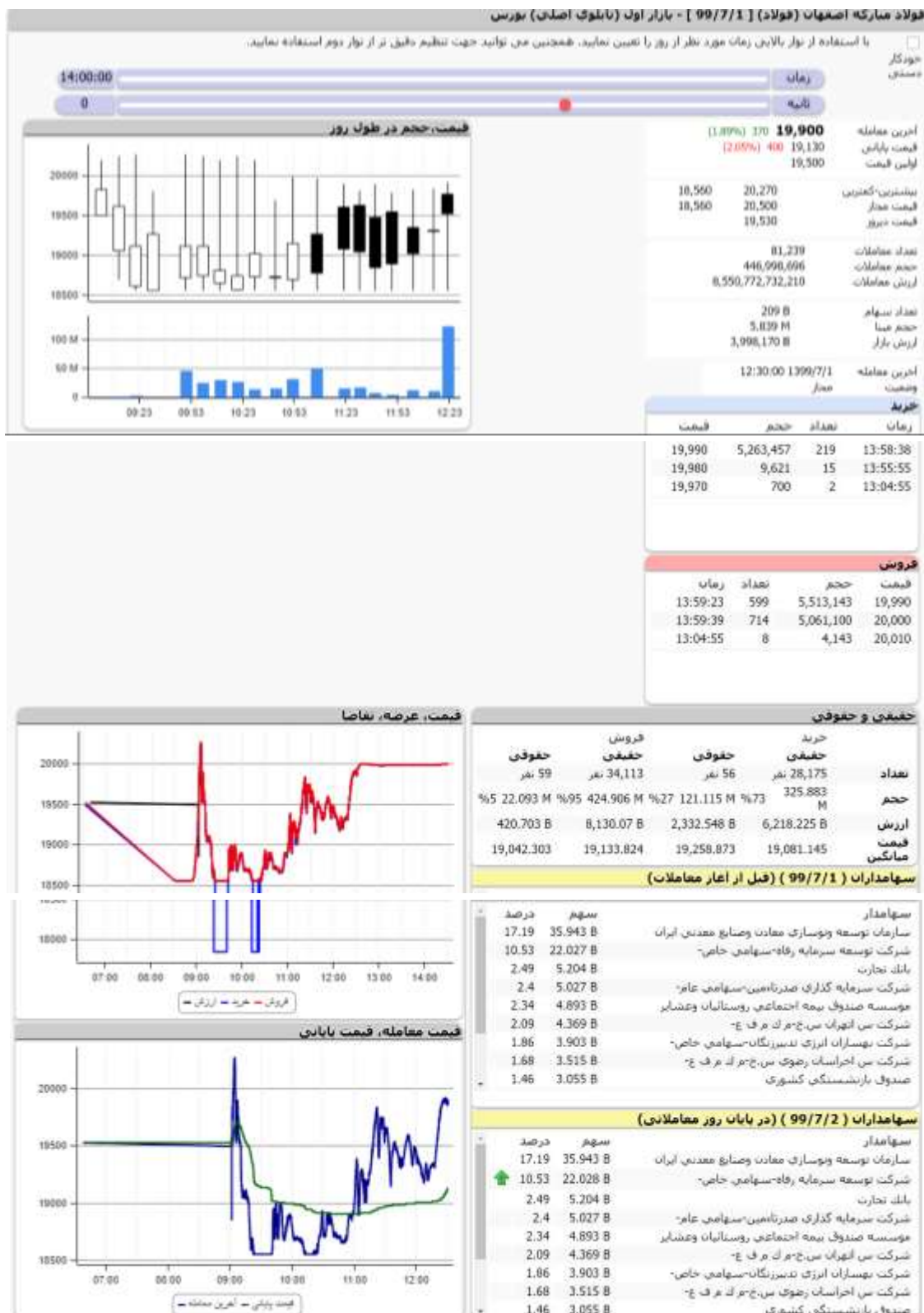
نویسی ویژگی های مدنظر برای این کار را دارا نبودند. در نتیجه به بررسی وبسایت tsetmc.com جهت استخراج اطلاعات پرداختم.

با بررسی ساختار سایت tsetmc متوجه شدم تمامی نماد ها کد منحصر به فرد خود را دارند که در لینک ها از آن کد برای دریافت اطلاعات مربوط به آن نماد استفاده می شود. پس به دنبال صفحه ای گشتم که از طریق آن بتوان کد مربوط به هر نماد را استخراج کرد و برای ایجاد لینک هر نماد استفاده نمود. این [لینک](#) لیستی از تمام نماد ها را دارد که با نوشتن خزنده می توان کد هر نماد را از فایل html آن دریافت نمود.

لیست همه نمادهای بازار عادی						
کد نماد	گروه	گروه های صنعت	نابلو	نماد	نام لاتین	نماد
IRB5IKCO8751	N2	خودرو و ساخت قطعات	فهرست اولیه	IKCQ1	Iran Khodro-D	18719101
IRO1NBAB0001	N2	عرضه برق، گاز، بخار و آب گرم	فهرست اولیه	NBAB1	Abadan PG	آبادا
IRO1APPE0001	N2	رایانه و فعالیت های وابسته به آن	فهرست اولیه	APPE1	Asan Pardakht Pers	آسان پرداخت پرشین
IRO1ASIA0001	N1	بیمه و صندوق بازنشستگی به حزامین اجتماعی	نابلو اصلی	ASIA1	Asia Bime	بیمه آسیا
IRO1ASTC0001	N2	اطلاعات و ارتباطات	فهرست اولیه	ASTC1	Asiatech	آسیاتک
IRO1CONT0001	N1	ابزاربرشگی، اینگی و اندازه گیری	نابلو فرعی	CONT1	Iran Counter	آکتور
IRR1CONT0101	N1	ابزاربرشگی، اینگی و اندازه گیری	نابلو فرعی	CONX1	Iran Counter-R	آکتورج
IRO1OPAL0001	N2	استخراج کانه های فلزی	فهرست اولیه	OPAL1	Opal Kani Pars	آبال

شکل (۲-۲) لیست نماد های بازار بورس تهران

حال برای دریافت اطلاعات نماد نیاز به صفحاتی بود که داده های معاملاتی هر نماد را از گذشته تا الان را شامل شود. صفحات مربوط به هر نماد که تمامی اطلاعات مربوط به یک روز معاملاتی را دارا می باشد پیدا کردم. در سایت tsetmc به ازای هر روز و هر نماد صفحه ای وجود دارد که اطلاعاتی اعم از قیمت ها، معاملات، خرید و فروش اشخاص حقیقی و حقوقی و سهامداران را نمایش می دهد. یک نمونه از این صفحه برای نماد فولاد از طریق این [لینک](#) قابل دسترسی است.



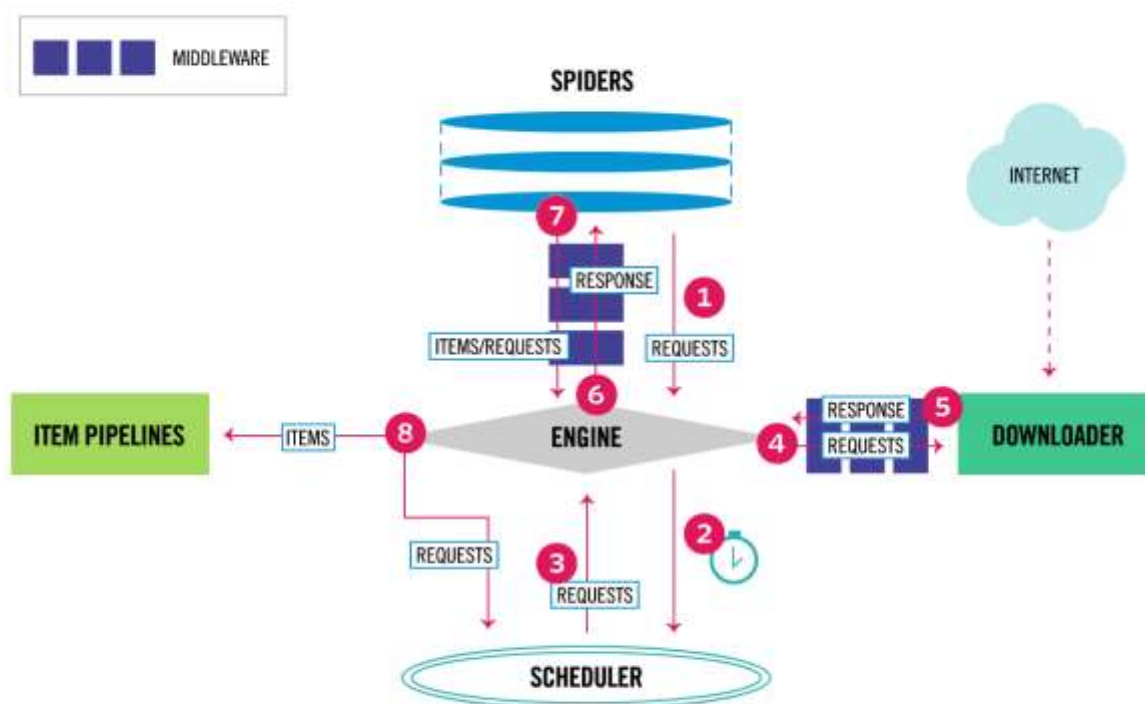
شکل (۲-۳) اطلاعات یک روز معاملاتی نماد فولاد

۲-۲- ایجاد محیط توسعه خزنده

پس از بررسی سایت tsetmc و امکان سنجی استخراج اطلاعات مورد نظر به بررسی انتخاب های موجود برای استخراج اطلاعات از این سایت پرداختیم. یکی از رایج ترین زبان ها برای نوشتن خزنده زبان پایتون است که امکانات و کتابخانه های زیادی را برای توسعه یک خزنده ارائه می دهد. بسته به میزان پیچیدگی خزنده کتابخانه های متفاوتی برای انجام اینکار وجود دارد. BeautifulSoup و Scrapy دو کتابخانه رایج هستند که از BeautifulSoup برای کارهای ساده تر مناسب بوده و استفاده از آن ساده است. در حالی که Scrapy امکانات بسیار بیشتری را ارائه می دهد و استفاده از آن نیاز به یادگیری و مطالعه بیشتر دارد. از آن جایی که میزان اطلاعاتی که در سایت بورس از گذشته تا امروز باید استخراج شود زیاد بوده و اینکار باید مرتباً صورت پذیرد و نیاز به نگهداری^۱ از کد خزنده می باشد، Scrapy را برای نوشتن خزنده انتخاب نمودم. پس از مطالعه [مستندات](#) Scrapy شروع به پیاده سازی خزنده اولیه که تنها لازم بود کد منحصر فرد نماد ها را استخراج کند کردم.

در Scrapy اصطلاحاتی وجود دارد که به توضیح آن ها می پردازیم. کلاس اصلی Spider که با دریافت URL ها درخواست های مربوطه را به برنامه ریز ارسال می کند و پس از برنامه ریزی شدن درخواست ها، این درخواست ها پس از عبور از کلاس های Downloader و انجام فیلتر ها مورد نیاز بر روی درخواست پاسخ را دریافت کرده و برای Spider جهت پردازش آن و استخراج داده از صفحات HTML را فراهم می کند. یک Spider دیکشنری هایی را به نام item پس از پایان کارش بر می گرداند که شامل داده های خام استخراج شده است. سپس این داده ها از کلاس هایی به نام ItemPipeline عبور داده می شود تا پردازش های بیشتری بر روی آن ها انجام شود و آماده ذخیره سازی در پایگاه داده شوند.

¹ Maintanace



شکل (۲-۴) معماری کتابخانه Scrapy

۲-۳- انتخاب پایگاه داده و نگاشت دهنده رابطه به شیء^۱ (ORM)

داده های بورسی ساختار یافته هستند و برخلاف برخی دیگر از انواع خزنده ها که در فایل های csv، json و یا پایگاه داده های NoSql داده های خود را ذخیره می کنند، پایگاه داده SQL برای آن مناسب تر است. پایگاه داده MySQL برای ذخیره داده ها انتخاب شده است. یک ORM بسیار مناسب برای پایتون و این پایگاه داده SQLAlchemy است که امکانات کافی را دارا بود.

۲-۴- خلاصه و جمع بندی

در این فصل با مفاهیم اولیه و پیش زمینه هایی که جهت ایجاد پیش زمینه برای درک مدل داده ای

¹ Object-Relational Mapper

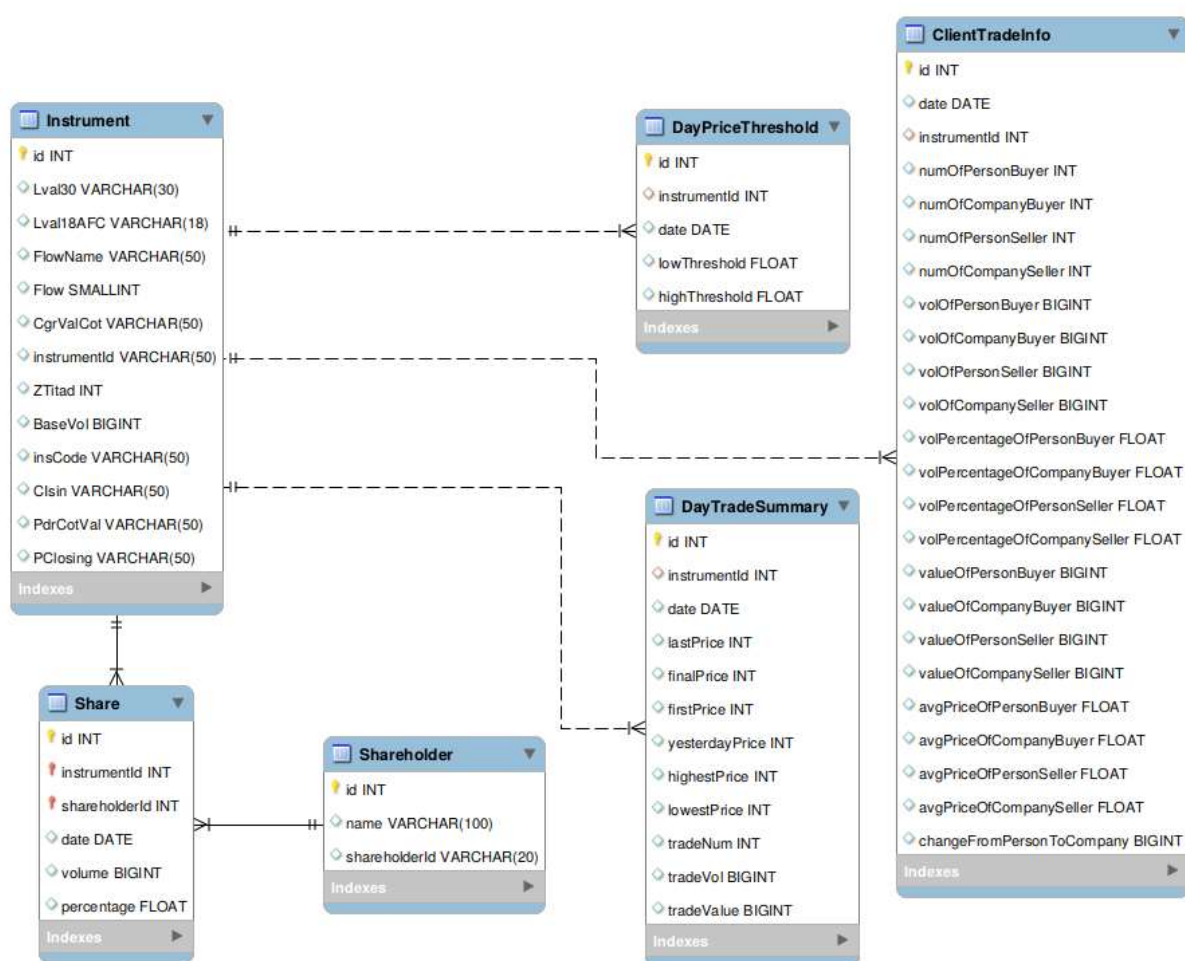
لازم بود آشنا شدیم. ساختار سایت tsetmc و صفحاتی که برای استخراج داده مناسب بودند بررسی شدند. همچنین کتابخانه مورد استفاده برای پیاده سازی خزنده و مفاهیم مرتبط به آن ذکر شدند. پایگاه داده مناسب برای داده های بورسی و نگاشت دهنده رابطه به شیء سازگار با محیط توسعه معرفی شدند.

فصل ۳: مدل داده ای و ساختار خزنده

فصل سوم در برگیرنده‌ی توضیح مربوط به مدل داده ای پیشنهادی و ساختار خزنده مربوطه است.

۳-۱- مدل داده ای

مدل داده ای بازار بورس تهران شامل ۶ جدول است. در زیر به توضیح هر یک از جدول ها می پردازیم.



شکل (۳-۱) مدل داده ای بازار بورس تهران

۳-۱-۲- جدول Instrument

این جدول اطلاعات مربوط به هر نماد را ذخیره می کند.

ستون Lval30: نام فارسی نماد که شامل ۳۰ کاراکتر است.

ستون Lval18AFC: نام فارسی نماد خلاصه شده که شامل ۱۸ کاراکتر است.

ستون FlowName: نام بازاری که نماد در آن قرار دارد. مانند: بازار اول (تابلوی اصلی)، بازار دوم بورس

ستون Flow: کد بازار

- 0: عمومی - مشترک بین بورس و فرابورس
- 1: بورس
- 2: فرابورس
- 3: آتی
- 4: پایه فرابورس
- 5: پایه فرابورس (منتشر نمی شود)
- 6: بورس انرژی
- 7: بورس کالا

ستون CgrValCot: کد گروه نماد

ستون instrumentId: کد ۱۲ کاراکتری لاتین نماد

ستون Ztitad: تعداد سهام - سرمایه

ستون BaseVol: حجم مبنا نماد

ستون insCode: این کد برای هر نماد یکتا بوده و در هنگام اضافه شدن هر نماد جدید به بورس، در بانک اطلاعاتی وب سایت شرکت مدیریت فناوری بورس ساخته می شود.

ستون CIsin: کد ۱۲ رقمی شرکت

ستون PClosing: آخرین قیمت

id	LocDB	Lval18AFC	FlowName	Flow	CgrValCot	instrumentId	Ztitad	BaseVol	insCode	Cisin	PdCotVal	PClosing
1			نامین سرمایه بانک ملت	بازار دوم بورس	N2	IRO1TMLT0001	3481620	7000000000	11129387075131725			
2			فولاد مبارکه اصفهان	بازار اول (تابلوی اصلی)	N1	IRO1FOLD0001	6666667	209000000000	4634858193224090			
3			ایران خودرو	بازار دوم بورس	N2	IRO1IKCO0001	36923077	301656068000	65883838195688438			
4			کارخانجات قدس قزوین	بازار دوم بورس	N2	IRO1GGAZ0001	1537515	330196748	15259343650667588			
5			ج. اسرنگ صنعت و معدن	بازار اول (تابلوی اصلی)	N1	IRR1LSMD0101	1	0	19216136630342675			
6			شرکت سیر استان آذربایجان ...	بازار دوم بورس	N2	IRO1QSO10001	81791148	204477865140	85774725600261203			
7			تولید برق	بازار دوم بورس	N2	IRO1TOP10001	1	250000000	67432066034966650			
8			بانک اقتصاد نوین	بازار اول (تابلوی فرعی)	N1	IRO1NOVN0001	12170284	30425734000	47302318535715632			
9			ج. ایران تابر	بازار اول (تابلوی فرعی)	N1	IRR1TAIRO101	1	0	7503966413749235			
10			ج. پارسین اصفهان	بازار دوم بورس	N2	IRR1PESF0101	1	0	37397450502348400			

شکل (۳-۲) جدول Instrument

۳-۱-۳- جدول DayPriceThreshold

این جدول شامل بازه قیمتی معاملاتی هر روز هر نماد است.

ستون instrumentId: شناسه نمادی که این بازه متعلق به آن است.

ستون date: تاریخ مربوط به روز این بازه قیمتی است.

ستون lowThreshold: کمترین قیمت ممکن معامله برای نماد در آن روز است.

ستون highThreshold: بیشترین قیمت ممکن معامله برای نماد در آن روز است.

id	instrumentId	date	lowThreshold	highThreshold
1	1	2020-09-13	13610	15030
2	2	2020-10-13	16720	18480
3	3	2020-10-14	3690	4070
4	1	2020-10-12	12220	13500
5	3	2020-10-12	3410	3750
6	2	2020-10-12	17320	19140
7	1	2020-10-07	10820	11940
8	1	2020-10-03	9120	10080
9	1	2020-10-11	11640	12860
10	1	2020-10-06	10370	11450

شکل (۳-۳) جدول DayPriceThreshold

۳-۱-۴- جدول DayTradeSummary

این جدول شامل خلاصه ای از معاملات روز نماد نماد است.

ستون instrumentId: شناسه نمادی که این اطلاعات متعلق به آن است.

ستون date: تاریخ مربوط به روز این اطلاعات است.

ستون lastPrice: قیمت آخرین معامله انجام شده است.

ستون finalPrice: قیمت پایانی محاسبه شده برای نماد است و بازه های قیمتی فردا از روی آن محاسبه می شود.

ستون firstPrice: قیمت اولین معامله انجام شده در روز است.

ستون yesterdayPrice: قیمت دیروز نماد است.

ستون highestPrice: قیمت بالاترین معامله انجام شده در روز است.

ستون lowestPrice: قیمت کمترین معامله انجام شده در روز است.

ستون tradeNum: تعداد معاملات انجام شده در روز است.

ستون tradeVol: حجم مجموع تمام معاملات انجام شده در روز است.

ستون tradeValue: ارزش مجموع تمام معاملات انجام شده در روز است.

id	instrumentId	date	lastPrice	finalPrice	firstPrice	yesterdayPrice	highestPrice	lowestPrice	tradeNum	tradeVol	tradeValue
1	1	2020-09-13	13610	13730	13610	14320	13610	13610	481	2892669	39369225090
2	2	2020-10-13	17630	17580	17600	17600	18000	17020	46400	177118806	3112948836350
3	3	2020-10-14	3690	3690	3690	3880	3690	3690	14809	182383213	672994055970
4	1	2020-10-12	12360	12570	12870	12860	13000	12220	6850	24241360	304665531410
5	3	2020-10-12	3750	3730	3720	3580	3750	3580	127895	3315229973	12380778347070
6	2	2020-10-12	18050	17600	18000	18230	18260	17320	64970	322364710	5672976346350
7	1	2020-10-07	11940	11940	11940	11380	11940	11940	2492	7415246	88538037240
8	1	2020-10-03	10080	10060	9980	9600	10080	9800	2834	7579841	76243452840
9	1	2020-10-11	12860	12860	12860	12250	12860	12860	1594	6623943	85183906980
10	1	2020-10-06	11450	11380	11060	10910	11450	11060	1850	5443129	61940906720

شکل (۳-۴) جدول DayTradeSummary

۳-۱-۵- ClientTradeInfo جدول

این جدول شامل اطلاعات مربوط به خریداران و فروشندگان حقیقی و حقوقی به تفکیک است.

ستون instrumentId: شناسه نمادی که این اطلاعات متعلق به آن است.

ستون date: تاریخ روزی که این اطلاعات مربوط به آن روز است.

ستون numOfPersonBuyer: تعداد خریداران حقیقی

ستون numOfCompanyBuyer: تعداد خریداران حقوقی

ستون numOfPersonSeller: تعداد فروشندگان حقیقی

ستون numOfCompanySeller: تعداد فروشندگان حقوقی

ستون volOfPersonBuyer: حجم خریداران حقیقی

ستون volOfCompanyBuyer: حجم خریداران حقوقی

ستون volOfPersonSeller: حجم فروشندگان حقیقی

ستون volOfCompanySeller: حجم فروشندگان حقوقی

ستون volPercentageOfPersonBuyer: درصد حجم معامله خریداران حقیقی

ستون volPercentageOfCompanyBuyer: درصد حجم معامله خریداران حقوقی

ستون volPercentageOfPersonSeller: درصد حجم معامله فروشندگان حقیقی

ستون volPercentageOfCompanySeller: درصد حجم معامله فروشندگان حقوقی

ستون valueOfPersonBuyer: ارزش معاملات خریداران حقیقی

ستون valueOfCompanyBuyer: ارزش معاملات خریداران حقوقی

ستون valueOfPersonSeller: ارزش معاملات فروشندگان حقیقی

ستون valueOfCompanySeller: ارزش معاملات فروشندگان حقوقی

ستون averagePriceOfPersonBuyer: میانگین قیمت معاملات خریداران حقیقی

ستون averagePriceOfCompanyBuyer: میانگین قیمت معاملات خریداران حقوقی

ستون averagePriceOfPersonSeller: میانگین قیمت معاملات فروشندگان حقیقی

ستون averagePriceOfCompanySeller: میانگین قیمت معاملات فروشندگان حقوقی

ستون changeFromPersonToCompany: میزان تغییر از حقیقی به حقوقی

id	date	instrumentId	numOfPersonBuyer	numOfCompanyBuyer	numOfPersonSeller	numOfCompanySeller	volOfPersonBuyer	volOfCompanyBuyer	volOfPersonSeller	volOfCompanySeller
1	2020-09-13	1	548	8	54	1	982569	1900000	2803669	90000
2	2020-10-13	3	22447	18	14649	44	147304921	29813887	162980938	14438270
3	2020-10-14	3	12713	7	330	2	180870125	1513088	181398213	985000
4	2020-10-12	1	2226	6	3325	7	16061804	8179756	22820599	620761
5	2020-10-12	3	58031	34	36061	87	3158990640	146238333	2986576989	328652985
6	2020-10-12	2	33345	40	16306	46	229381778	92982932	240569112	81795598
7	2020-10-07	1	878	1	1883	35	7117349	297887	2984480	4428768
8	2020-10-03	1	478	4	2116	4	6775841	804000	3958421	4013420
9	2020-10-11	1	346	0	1136	3	6823943	0	3120533	3500420
10	2020-10-06	1	351	3	1386	4	3270140	2172989	5441786	1363

شکل (۳-۵) جدول ClientTradeInfo

۳-۱-۶- Shareholder جدول

این جدول شامل اطلاعات سهامداران است.

ستون name: نام سهامدار

ستون shareholderId: شناسه یکتای مختص به سهامدار

id	name	shareholderId
1	شرکت گروه مالی ملت-سهام عام-	8113
2	شرکت مدیریت سرمایه آتیه خواهان-سهامی خاص-	19316
3	شرکت تدبیرگران بهسازملت-سهامی خاص-	7514
4	شرکت واسپاری ملت-سهامی خاص-	60521
5	شرکت صرافی ملت-سهامی خاص-	60799
6	شرکت بازرگانی پویاگستر دنیا-سهامی خاص-	781
7	صندوق سرمایه گذاری گسترش فردای ایرانیان	20904
8	صندوق سرمایه گذاری تجربه ایرانیان	25092
9	شرکت قندهکمتان-سهامی عام-	8091
10	شخص حقیقی	50281

شکل (۳-۶) جدول Shareholder

۷-۱-۳- جدول Share

این جدول شامل تاریخچه سهم هایی است که یک سهام دار از یک نماد داشته است. هر بار که میزان سهم یک سهام دار از نماد مربوطه تغییر پیدا کند یک نمونه جدید با مقادیر به روز شده ذخیره می شود.

ستون instrumentId: نمادی که این سهم مربوط به آن است.

ستون shareholderId: سهام داری که این سهم متعلق به آن است.

ستون date: تاریخی که تغییری در سهام به وجود آمده و در نتیجه این اطلاعات جدید ثبت شده است.

ستون volume: میزان حجم سهام سهام دار از نماد در تاریخ مربوطه را نشان می دهد.

ستون percentage: میزان درصد سهام دار از کل نماد را نشان می دهد.

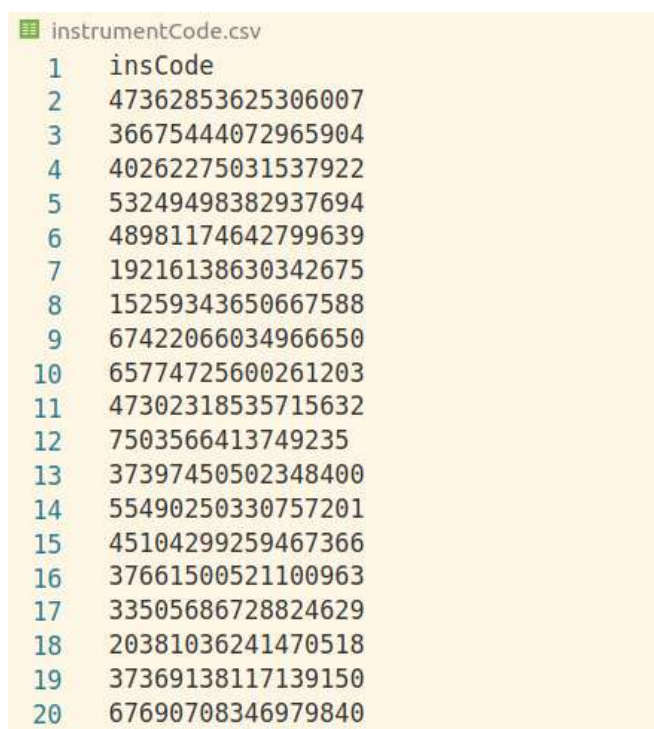
id	instrumen	shareholderId	date	volume	percentage
962	343	53	2020-10-03	400582341	57.22
963	343	145	2020-10-03	56654421	8.09
964	343	228	2020-10-03	20070224	2.86
965	343	603	2020-10-03	9750912	1.39
966	343	275	2020-10-03	9413724	1.34
967	343	183	2020-10-03	9099662	1.29
968	343	430	2020-10-03	7838247	1.11
961	340	310	2020-10-03	306904027	57.9
978	338	606	2020-10-03	5653594	11.08
979	338	607	2020-10-03	4624462	9.06

شکل (۷-۳) جدول Share

۲-۳- ساختار خزنده

برای استخراج اطلاعات از سایت tsetmc ابتدا یک خزنده بسیار ساده پیاده سازی شده است که شناسه یکتای هر نماد را که از طریق آن لینک های مرتبط با هر نماد ساخته می شود به دست می آورد و در یک فایل csv

ذخیره می کند تا توسط خزنده دیگر استفاده شود. از آن جایی که لیست نماد ها معمولاً ثابت است اجرای این خزنده تنها زمان هایی صورت می گیرد که می دانیم نماد جدیدی اضافه شده است.



	insCode
1	47362853625306007
2	36675444072965904
3	40262275031537922
4	53249498382937694
5	48981174642799639
6	19216138630342675
7	15259343650667588
8	67422066034966650
9	65774725600261203
10	47302318535715632
11	7503566413749235
12	37397450502348400
13	55490250330757201
14	45104299259467366
15	37661500521100963
16	33505686728824629
17	20381036241470518
18	37369138117139150
19	67690708346979840
20	

شکل (۸-۳) نمونه کد های نماد استخراج شده

این فایل csv باید در کنار خزنده اصلی قرار گیرد تا بتواند لینک های مربوطه را بسازد. خزنده اصلی با دریافت دو آرگومان تاریخ شروع و تاریخ پایان شروع و خواندن کد نماد ها از فایل لینک درخواست ها را می سازد. در صورت عدم دریافت آرگومان تاریخ پایان خزنده تا تاریخ آخرین روز معاملاتی داده ها را دریافت می کند. در صورت عدم دریافت تاریخ شروع به صورت پیش فرض مقدار آن برابر تاریخ آخرین روز موجود (دیروز) قرار داده می شود. در ابتدا به اندازه تعداد کل نماد ها در تاریخ شروع درخواست ها اجرا می شوند.

برای جلوگیری از ارسال درخواست هایی که داده های مربوط به آن در پایگاه داده موجود است یک middleware به نام IgnoreDuplicateDownloaderMiddleware پیاده سازی شده است که با بررسی وجود نماد و داده بازه قیمتی آن تاریخ تشخیص می دهد که آیا درخواست را ارسال کند یا از آن جلوگیری کند.

پس از دریافت پاسخ داده های مورد نیاز توسط Selector ها استخراج می شود و برای ذخیره به ItemPipeline ها ارسال می گردد. پس از برگرداندن هر Item مربوط به یک نماد در آن روز لینک مربوط

به روز بعدی آن محاسبه شده و در صورت بودن در بازه تاریخ مورد نظر آن درخواست ارسال شده و این فرآیند دوباره بر روی آن طی می شود.

برای این خزنده سه ItemPipeline پیاده سازی شده است. اولین آن ها که Item از آن عبور می کند DropEmptyPipeline است که در صورتی که Item دریافت شده خالی باشد آن را باز می گرداند و برای ذخیره در پایگاه داده ارسال نمی کند. دومین آن ها DuplicatesPipeline است که در صورتی که Item دریافت شده تکراری باشد آن را باز می گرداند. سومین آن ها TehranStockExchangeScraperPipeline است که عملیات ذخیره سازی Item را در پایگاه داده بر عهده دارد.

فایل models.py شامل شمای جداول است که باید برای پایگاه داده ایجاد شود که توسط SQLAlchemy پیاده سازی شده است.

فایل settings.py شامل اطلاعات اعم از ترتیب Middleware ها، ترتیب ItemPipeline ها، ConnectionString برای اتصال به پایگاه داده، نام فایل Log و مسیر پوشه JOBDIR است.

برای هر بار اجرای کامل خزنده باید مسیر JOBDIR یا در فایل settings.py و یا از طریق آرگومان هنگام اجرای خزنده عوض شود. کاربرد این تنظیم برای زمانی است که در میان اجرای خزنده بخواهیم آن را موقتاً متوقف کنیم و دوباره در زمانی دیگر بخواهیم از ادامه آن را اجرا کنیم بدون آن که درخواست های تکراری ارسال شوند.

فصل ۴: نحوه استفاده از خزنده و دریافت داده

ها

پس از توضیحات انجام شده در فصل قبل، در این فصل در مورد نحوه اجرای خزنده و استفاده از آن بر روی سایر سیستم ها صحبت خواهیم کرد. ابتدا ابزار های مورد نیاز برای سیستمی که بر روی آن اجرا صورت می گیرد توضیح داده می شود و سپس راه اندازی خزنده و انواع نحوه استفاده از آن ذکر می شود.

۴-۱- آماده سازی سیستم

بر روی سیستم عامل اوبونتو نیاز به نصب پایتون، MySQL است. سپس پروژه ها را از طریق GitLab دانلود کرده و با فعال سازی محیط مجازی برای پوشه خزنده از طریق فایل requirements.txt پکیج های مورد نیاز برای هر دو خزنده را نصب می کنید.

۴-۲- اجرای خزنده

ابتدا خزنده instrument_code_scraper را به شکل زیر در پوشه پروژه اجرا کنید.

```
maryam@maryam-TP301UJ:~$ scrapy crawl InstrumentCode
```

فایل instrumentCode.csv تولید شده را در پوشه خزنده tehran_stock_exchange_scraper کپی کنید. اگر برای تست و بار اول قصد اجرای خزنده را دارید و یا به هر دلیلی قصد دریافت تنها داده یک یا چند نماد محدود را دارید می توانید با تغییر فایل csv و حذف سایر کدها تنها نمادهای مدنظر خود را استخراج نمایید.

در MySQL یک پایگاه داده به نام tehran_stock_exchange می سازید. اگر نام متفاوتی مدنظر دارید از طریق فایل settings.py رشته connection را تغییر دهید. نام کاربری و رمز عبور MySQL خود را نیز در رشته قرار دهید.

```
CONNECTION_STRING = "{drivename}://{user}:{passwd}@{host}:{port}/{db_name}?charset=utf8".format(
    drivename="mysql",
    user="root",
    passwd="*****",
    host="localhost",
    port="3306",
    db_name="tehran_stock_exchange",
)
```

شکل (۴-۱) فایل settings.py رشته اتصال به پایگاه داده

در صورتی که تمایل دارید می توانید نام فایل log را تغییر دهید. اگر این متغیر را حذف کنید log های مربوطه در ترمینال نمایش داده می شود.

```
#Logging
LOG_FILE = "tehran_exchange_scraper.log"
```

شکل (۲-۴) متغیر LOG_FILE در فایل settings.py

برای هر بار اجرای کامل خزنده باید مسیر JOBDIR یا در فایل settings.py و یا از طریق آرگومان هنگام اجرای خزنده عوض شود. کاربرد این تنظیم برای زمانی است که در میان اجرای خزنده بخواهیم آن را موقتاً متوقف کنیم و دوباره در زمانی دیگر بخواهیم از ادامه آن را اجرا کنیم بدون آن که درخواست های تکراری ارسال شوند. اگر یکبار خزنده را به طور کامل اجرا کرده باشید و این فایل را برای اجرای جدید تغییر ندهید، داده جدیدی دانلود نمی‌شود چرا که با نگاه به این دایرکتوری فرض می‌کند که اجرای خزنده قبلاً کامل شده است.

```
#JOBS
JOBDIR = "crawls/tsetmc-0002"
```

شکل (۳-۴) متغیر JOBDIR در فایل settings.py

برای اجرای خزنده با استفاده از آرگومان های تاریخ شروع و پایان باید در پوشه خزنده در ترمینال دستور زیر را اجرا نمود. از a- برای قبل از نوشتن هر آرگومان استفاده می‌شود. فرمت تاریخ حتماً باید به شکل زیر و به تاریخ میلادی باشد. برای نمونه در دستور زیر اطلاعات مربوط به روز ۳ اکتبر سال ۲۰۲۰ تمامی نمادها موجود در فایل csv دانلود می‌شود.

```
maryam@maryam-TP301UJ:~$ scrapy crawl tsetmc -a start-date=20201003 -a end-date=20201004
```

در صورتی که تنها تاریخ شروع داده شود تاریخ انتهایی تا آخرین روز ممکن پیش می‌رود. در صورتی که هیچ یک از تاریخ ها داده نشود تنها داده مربوط به آخرین روز دریافت می‌شود.

در صورتی که قصد تغییر JOBDIR از طریق دادن آرگومان دارید باید دستور را به شکل زیر وارد کنید.

```
maryam@maryam-TP301UJ:~$ scrapy crawl tsetmc -a start-date=20201003 -a end-date=20201004 -s JOBDIR=crawls/tsetmc-0001
```

دقت شود که اطلاعات این صفحه ها برای روز های تعطیل بازار خالی است و تنها جدولی که به ازای تمام روز ها و حتی روز های تعطیل داده ها را ذخیره می‌کند جدول DayPriceThreshold است.

از طریق log های خزنده می توان متوجه پیشرفت روند شد و همچنین در صورت بروز خطا آن را بررسی کرد. Log ها شامل اطلاعات کاملی از تعداد درخواست های ارسال شده، تعداد درخواست هایی که تکراری تشخیص داده شده اند، تعداد item های به دست آمده و خطاهای رخ داده ارائه می کند.

۳-۴- خلاصه و جمع بندی

در این فصل به توضیح آماده سازی محیط اجرای خزنده ها و دستورات لازم برای اجرای هر دو خزنده پرداخته شد.

فصل ۵: جمع‌بندی و پیشنهادها

در این فصل نتیجه‌گیری‌های کلی حاصل شده در این تحقیق بیان می‌شود و محدودیت‌ها مورد بحث قرار می‌گیرد. همچنین پیشنهادهایی برای ادامه‌ی مسیر به علاقمندان این حوزه‌ی ارائه خواهد شد.

۱-۵- جمع‌بندی

در این تحقیق در گام نخست ما به مطالعه روش های مختلف برای دریافت داده بازار بورس تهران پرداختیم. سپس ساختار سایت tsetmc را مورد بررسی قرار دادیم تا صفحات مناسب برای استخراج داده را بیابیم. محیط توسعه و کتابخانه ها و ابزار مناسب برای اینکار را انتخاب کردیم. سپس مدل داده ای برای اطلاعات موجود در سایت طراحی کردیم. به پیاده سازی دو خزنده جهت استخراج اطلاعات پرداختیم و در نهایت داده ها را دانلود کردیم و از صحت عملکرد آن اطمینان حاصل کردیم.

۲-۵- دستاوردها

ایجاد یک مدل داده ای کامل که شامل اطلاعات نماد، معاملات روز، خریداران و فروشندگان حقیقی و حقوقی و سهامداران اصلی است. این داده ها زمینه را برای بسیاری از روش های تحلیل داده فراهم می کنند. همچنین توسعه یک خزنده با قابلیت گسترش پذیری و اجرای آسان که به راحتی قابل اجرا است یکی دیگر از دستاوردهای این پروژه است.

۳-۵- پیشنهادها

در ادامه این پروژه می توان اطلاعات مربوط به معاملات در طول یک روز برای یک نماد را نیز به دست آورد. همچنین می توان اطلاعات صف خرید و فروش را نیز به داده ها اضافه نمود. یکی دیگر از بهبود ها می تواند ایجاد سرور برای اجرای دوره ای خزنده به وسیله [Scrapy](#) باشد که امکانات بسیاری را ارائه می دهد. در آینده و با در اختیار داشتن داده های بازار بورس تهران می توان از روش های هوش مصنوعی جهت تحلیل داده و پیش بینی روند بازار استفاده کرد. دو نوع مدل متفاوت رگرسیون یا طبقه بندی برای این کار وجود دارد. در رگرسیون خود مقدار قیمت در آینده پیش بینی می شود اما در طبقه بندی با استفاده از تمامی ویژگی های

موجود می‌توان به یک طبقه‌بندی کلی خرید، فروش و یا نگهداری برای نماد رسید. ماشین بردار پشتیبان^۱ یک روش برای این طبقه‌بندی است که در بازارهای بررسی شده عملکرد خوبی برای بازارهای با تغییر زیاد داشته است [2]. یکی از مسائل مؤثر در تعیین کیفیت تشخیص مدل بسته به ویژگی‌های انتخاب شده است. مدل طراحی شده باید با دسته‌های متفاوتی از ویژگی‌ها مورد بررسی قرار گیرد و ویژگی‌های با تأثیر بیشتر انتخاب شوند [3].

¹ Support Vector Machine (SVM)

فصل ٦: مراجع

مراجع

- [۱] Ph. Meschenmoser, N. Meuschke, M. Hotz, B. Gipp, "Scraping scientific web *D-Lib* ",repositories: Challenges and solutions for automated content extraction Corporation for National Research Initiatives, p. 15, 2016, شماره ۲۲, جلد ۲۲, *Magazine*
- [۲] *Journal* ",R. Rosillo, J. Giner, "Stock market simulation using support vector machines *of Forecasting* جلد ۳۳, 2014, pp. 488-500,
- [3] S. Madge, S. Bhatt, "Predicting stock price direction using support vector machines," *Independent work report spring*, 2015.



University of Tehran



College of Engineering

School of Electrical and Computer Engineering

Creation of Data Model for Tehran Stock Exchange to Predict Stock Price

A thesis submitted to the Undergraduate Studies Office

In partial fulfillment of the requirements for

The degree of Master in

Computer Engineering