

دانشگاه صنعتي امیرکبیر  
(پلی تکنیک تهران)

دانشكده مهندسی کامپیوتر و فن­­آوری ­اطلاعات

پایان‌نامه کارشناسی

گرایش نرم­افزار

عنوان

طراحی و پیاده‌سازی ابزاری به منظور اِعمال خط مشی امنیتیِ عدم تداخل مبتنی بر روش بازنویسی برنامه

نگارش

سید محمدمهدی احمدپناه

استاد راهنما

جناب آقای دکتر مهران سلیمان فلاح

شهریور 1394

اينجانب سید محمدمهدی احمدپناه متعهد مي‌شوم كه مطالب مندرج در اين پايان نامه حاصل كار پژوهشي اينجانب تحت نظارت و راهنمايي اساتيد دانشگاه صنعتي اميركبير بوده و به دستاوردهاي ديگران كه در اين پژوهش از آنها استفاده شده است، مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذكر گرديده است. اين پایان نامه قبلاً براي احراز هيچ مدرك هم‌سطح يا بالاتر ارائه نگرديده است.

در صورت اثبات تخلف در هر زمان، مدرك تحصيلي صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پيگيري قانوني خواهد داشت.

كليه نتايج و حقوق حاصل از اين پایان نامه متعلق به دانشگاه صنعتي اميركبير مي‌باشد. هرگونه استفاده از نتايج علمي و عملي، واگذاري اطلاعات به ديگران يا چاپ و تكثير، نسخه‌برداري، ترجمه و اقتباس از اين پایان نامه بدون موافقت كتبي دانشگاه صنعتي اميركبير ممنوع است.   
نقل مطالب با ذكر مآخذ بلامانع است.

سید محمدمهدی احمدپناه

امضا

تقدیم به پدرم

کوهی استوار و حامی من در طول تمام زندگی 

تقدیم به مادرم

سنگ صبوری که الفبای زندگی به من آموخت

تقدیم به خواهر و برادرم

همراهان همیشگی و پشتوانه های زندگیم

**تقدیر و تشکر:**

سپاس خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت های او ندانند و کوشندگان، حق او را گزاردن نتوانند. سلام و درود بر محمّد و خاندان پاك او، طاهران معصوم، هم آنان که وجودمان وام‌دار وجودشان است.

بدون شک جایگاه و منزلت معلم، بالاتر از آن است که در مقام قدردانی از زحمات بی شائبه‌ او، با زبان قاصر و دست ناتوان، چیزی بنگارم. اما از آنجا که تجلیل از معلم، سپاس از انسانی است که هدف آفرینش را تامین می‌کند، به رسم ادب دست به قلم برده‌ام، باشد که این خردترین بخشی از زحمات آنان را سپاس گوید.

از پدر و مادر مهربانم، این دو معلم بزرگوار که همواره بر کوتاهی من، قلم عفو کشیده و کریمانه از کنار غفلت‌های گذشته‌اند و در تمام عرصه‌های زندگی یار و یاورم بوده‌اند؛

از استاد فرزانه و دلسوز، جناب آقای دکتر سلیمان فلاح که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ نداشتند؛

از اساتید محترم، جناب آقای دکتر محمدرضا رزازی و جناب آقای دکتر بهمن پوروطن که زحمت داوری این پایان‌نامه را متقبل شدند؛

**کمال تشکر و قدردانی را دارم.**

چكيده

شسیشس

واژه‌های کلیدی:

شسیشسی

|  |  |
| --- | --- |
| فهرست عناوین | صفحه |

[1 ‌ فصل اول مقدمه 1](#_Toc428959128)

[1‌.1‌ نمای کلی پروژه 2](#_Toc428959129)

[1.1.1‌ هسته تشخیص و استخراج موضوع متن 2](#_Toc428959130)

2.[1‌.1‌ جمع آوری و نمایش متن وب‌سایت‌ها 2](#_Toc428959131)

[1‌.1‌.3‌ واحد ارتباط با پایگاه داده 2](#_Toc428959132)

[4.1.1 پایگاه داده متن‌ها به همراه موضوعات استخراج شده آن‌ها 3](#_Toc428959133)

[1‌.2‌ ساختار پایان نامه 4](#_Toc428959134)

[2 فصل دوم تشخیص و استخراج موضوع متن‌های کوتاه 5](#_Toc428959135)

[2‌.1‌ یادگیری با نظارت 6](#_Toc428959136)

[2‌.1‌.1‌ روش DOR 7](#_Toc428959137)

[2‌.1‌.2‌ روش TCOR 8](#_Toc428959138)

[2‌.1‌.3‌ روش‌های دسته بندی متون 9](#_Toc428959139)

[2‌.2‌ یادگیری بدون نظارت 11](#_Toc428959140)

[2‌.2‌.1‌ دسته بندی‌ متن‌های کوتاه بر اساس آنالیز معنایی 12](#_Toc428959141)

[2‌.2‌.2‌ مدل تخصیص پنهان دیریکله 13](#_Toc428959142)

[3 فصل سوم طراحی و پیاده سازی هسته‌ ی اصلی و پایگاه داده فارسی 16](#_Toc428959143)

[3‌.1‌ طراحی و پیاده سازی پایگاه داده 16](#_Toc428959144)

[1.1.3 جدول اطلاعات کاربران 16](#_Toc428959145)

[3‌.1‌.2‌ جدول قابلیت‌های مختلف تشخیص موضوع متن 17](#_Toc428959146)

[3‌.2‌ پایگاه داده آموزش و آزمون 18](#_Toc428959147)

[3‌.3‌ پیش پردازش 20](#_Toc428959148)

[3‌.3‌.1‌ حذف کلمات متداول 20](#_Toc428959149)

[2.3.3 فرآیند ریشه یابی 21](#_Toc428959150)

[3‌.4‌ متن‌های جدید منتشر شده بر روی وب‌سایت‌ها 24](#_Toc428959151)

[3‌.5‌ معرفی روند اصلی کار سیستم تشخیص موضوع متن 25](#_Toc428959152)

[3‌.5‌.1‌ یادگیری با نظارت، روش DOR 25](#_Toc428959153)

[3‌.5‌.2‌ یادگیری بدون نظارت، روش تخصیص پنهان دیریکله 26](#_Toc428959154)

[6.3‌ قابلیت توسعه پذیری سیستم 28](#_Toc428959155)

[4 فصل چهارم پیاده سازی و ایجاد رابط کاربری 30](#_Toc428959156)

[4‌.1‌ تحلیل و طراحی نرم‌افزار 30](#_Toc428959157)

[4‌.2‌ ابزارهای مورد استفاده 37](#_Toc428959159)

[4‌.3‌ ایجاد رابط کاربری گرافیکی 38](#_Toc428959160)

[5 فصل پنجم اندازه گیری و آزمایش و جمع بندی 43](#_Toc428959161)

[5‌.1‌ معیار اندازه گیری 43](#_Toc428959162)

[1.1.5 دقت 43](#_Toc428959163)

[5‌.2‌ نتایج 44](#_Toc428959164)

[5‌.3‌ جمع بندی و کار‌های آینده 45](#_Toc428959165)

[منابع و مراجع 46](#_Toc428959166)

|  |  |
| --- | --- |
| فهرست اشكال | صفحه |

[شکل ‏1‑1: نمای کلی سامانه 3](#_Toc428781304)

[شکل ‏2‑1: نمودار کلی روند کار بخش آموزش و آزمون سیستم‌های یادگیری با نظارت 7](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781305)

[شکل ‏2‑2: الگوریتم ماشین بردار پشتیبان 10](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781306)

[شکل ‏2‑3: روند کلی کار سیستم‌های یادگیری بدون نظارت 11](#_Toc428781307)

[شکل ‏2‑4: مدل تخصیص پنهان دیریکله 13](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781308)

[شکل ‏3‑1:جدول اطلاعات کاربران 16](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781309)

[شکل ‏3‑2: جدول قابلیت‌های مختلف موجود در سامانه 17](#_Toc428781310)

[شکل ‏3‑3: نمونه ‌ای از داده‌های آموزش برای دو موضوع خوشنویسی و سینما 20](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781311)

[شکل ‏3‑4: کاراکترهای زایدی که در زمان پیش پردازش از متن ورودی حذف می‌شوند. 24](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781312)

[شکل ‏4‑1: فرآیند تولید نرم‌افزار به صورت آبشاری 31](#_Toc428781313)

[شکل ‏4‑3: نموار Usecase سیستم 32](#_Toc428781315)

[شکل ‏4‑4: نمودار Sequence ورود به سیستم 32](#_Toc428781316)

[شکل ‏4‑5: نمودار Sequence ثبت نام 33](#_Toc428781317)

[شکل ‏4‑6: نمودار Sequence، خروج از سیستم، انتخاب یک موضوع و جست و جو میان موضوعات 33](#_Toc428781318)

[شکل ‏4‑7: نمودار Sequence تعیین حداکثر بازه‌ی زمانی برای ذخیره شدن اطلاعات 34](#_Toc428781319)

[شکل ‏4‑8: نمودار Sequence برای مدیریت کاربری و مدیریت پایگاه داده 34](#_Toc428781320)

[شکل ‏4‑9: نمودار Sequence برای افزودن عملیات جدید 35](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781321)

[شکل ‏4‑10: نمودار Activity مدیریت کاربری، مدیریت پایگاه داده و ورود به سیستم 35](file:///F:\Uni\Term%208\payan%20nameh\HamidReza%20Ramezani.docx#_Toc428781322)

|  |  |
| --- | --- |
| فهرست جداول | صفحه |

[جدول ‏3‑1: اطلاعات موجود در پایگاه داده فارسی برای سیستم یادگیری با نظارت 19](#_Toc428781346)

[جدول ‏3‑2: قواعد ریشه یابی 22](#_Toc428781347)

[جدول ‏3‑3: پارامترهای LibSVM 26](#_Toc428781348)

[جدول ‏3‑4: پارامترهای سیستم تخصیص پنهان دیریکله 27](#_Toc428781349)

[جدول ‏3‑5: پایگاه داده اطلاعات فارسی برای آموزش در روش تخصیص پنهان دیریکله 27](#_Toc428781350)

[جدول ‏5‑1: نتایج خروجی برای کم و زیاد کردن تعداد متن‌‌های آموزش سیستم 44](#_Toc428781351)

[جدول ‏5‑2: نتایج خروجی برای طول متن‌های متفاوت ورودی 44](#_Toc428781352)

[جدول ‏5‑3: نتایج خروجی برای طول متن‌های متفاوت ورودی 45](#_Toc428781353)

فهرست علائم اختصاری

|  |  |  |
| --- | --- | --- |
| -- | Document occurrence Representation | DOR |
| -- | Term co-occurrence representation | TCOR |
| -- | Bag of words | BOW |
| ماشین بردار پشتیبان | Support vector machine | SVM |
| تنک | Sparse | -- |
| تنکی | Sparseness | -- |
| k نزدیک‌ترین همسایه | K-nearest neighbors | KNN |
| نایو بیز | naïve Bayes | -- |
| -- | term frequency–inverse document frequency | tf-idf |
| تخصیص دیریکله نهفته | latent Dirichlet allocation | LDA |
|  |  |  |

# ‌ فصل اول مقدمه

**مقدمه**

با گسترش روزافزون سیستم‌های کامپیوتری، امنیت ذخیره‌سازی و انتقال اطلاعات بیش از پیش اهمیت پیدا کرده است. امنیت اطلاعات در جنبه‌های گوناگونی نظیر امنیت شبکه‌های کامپیوتری، امنیت پایگاه داده، امنیت برنامه‌های کاربردی و غیره مورد توجه پژوهشگران این رشته است. در گذشته، مسائل امنیتی بیشتر مورد توجه مراکز نظامی و سیاسی بوده است، اما اکنون برای مردم و کاربران عادی سیستم‌ها نیز حائز اهمیت است.

یکی از زمینه‌های مطرح در امنیت اطلاعات و ارتباطات، امنیتِ برنامه‌های کاربردی و به پیروی آن، امنیتِ زبان‌های برنامه‌نویسی یا امنیتِ زبان‌مبنا[[1]](#footnote-1) می‌باشد. امنیت زبان‌مبنا را می‌توان مجموعه‌ای از تکنیک‌های مبتنی بر نظریه زبان‌های برنامه‌سازی و پیاده‌سازی آن‌ها، شامل معناشناخت[[2]](#footnote-2)، نوع‌ها[[3]](#footnote-3)، بهینه‌سازی و راستی‌آزمایی[[4]](#footnote-4)، برای به کار گیری در مسائل امنیتی تعریف کرد. [1] تلاش این حوزه بر این است که برنامه‌های کاربردی تولید شده توسط برنامه‌نویسان و توسعه‌دهندگان، با توجه به رویکردهای مختلف امنیتی، قابل اعتماد و اطمینان باشند. به همین دلیل، طراحی و توسعه زبان‌های برنامه‌نویسی امن یا ایجاد ابزارهایی بر روی زبان‌های برنامه‌نویسی موجود باعث می‌شود تا توسعه‌دهندگان نرم‌افزار، کمتر درگیر مشکلات امنیتی برنامه‌های خود شده و به کمک این ابزارها، با تلاش کمتری به تولید برنامه‌های امن بپردازند، که این خود هزینه‌های تولید و توسعه نرم‌افزارها را کاهش می‌دهد.

روش‌های مختلفی برای تولید ابزارهای مرتبط با زبان‌های برنامه‌نویسی با رویکرد برآورده کردن نیازها و خط مشی‌های امنیتی وجود دارد که به طور کلی می‌توان به دو دسته روش‌های تحلیل ایستا[[5]](#footnote-5) یا زمان‌کامپایل[[6]](#footnote-6) و تحلیل پویا[[7]](#footnote-7) یا زمان‌اجرا[[8]](#footnote-8) دسته‌بندی کرد. هر کدام از این روش‌ها نقاط قوت و ضعف مربوط به خود را دارند که بسته به کاربرد، استفاده از هر یک از آن‌ها متفاوت خواهد بود. گرچه شایان ذکر است که تعریف و مشخص کردن دقیق مفهوم امن بودنِ یک سیستم یا برنامه یکی از چالش‌های پیشِ روی متخصصان این حوزه می‌باشد. چنان‌که نحوه و رویکرد اِعمال آن نیازمندی امنیتی، وابستگی زیادی به تعریف ارائه شده خواهد داشت.

هدف از این پروژه، تولید ابزاری برای تشخیص برقراری خط مشی عدم تداخل در کد مبدأ ورودی است که در صورت نقض این خط مشی، با بهره‌گیری از روش بازنویسی برنامه، کد مبدأ به نحوی اصلاح ‌شود تا این نیازمندی برآورده گردد. در این‌جا، خط مشی امنیتیِ عدم تداخل[[9]](#footnote-9) به عنوان نیازمندیِ امنیتی در نظر گرفته می‌شود و برای اِعمال این خط مشی در برنامه‌ها، از یکی از روش‌های تحلیل ایستا؛ یعنی، روش بازنویسی برنامه[[10]](#footnote-10)، استفاده می‌شود که در فصل‌های بعدی، به شرح و توضیح آن‌ها می‌پردازیم.

فصل دوم این پایان‌نامه به توضیح خط مشی امنیتی عدم تداخل و تعریف آن پرداخته خواهد شد و در ادامه، مکانیزم‌های اعمال آن و به ویژه، روش بازنویسی برنامه شرح داده خواهد شد. فصل سوم به توصیف زبان مدل مطرح شده تخصیص یافته است. در فصل چهارم، درباره گراف وابستگی برنامه و کاربرد آن در پروژه بحث خواهد شد. فصل پنجم به توضیح الگوریتم مورد نظر برای بازنویسی کد مبدأ[[11]](#footnote-11) در دو حالت خط مشی عدم تداخل و فصل ششم به فرآیند پیاده‌سازی و تولید ابزار می‌پردازیم. در نهایت، فصل هفتم دربرگیرنده جمع‌بندی و کارهای پیشنهادی آینده پروژه خواهد بود.

# فصل دوم خط مشیِ امنیتیِ عدم تداخل و اِعمال آن

**خط مشی امنیتی عدم تداخل**

به طور کلی، خط‌ مشیِ[[12]](#footnote-12) امنیتی، امن‌بودنِ یک سیستم یا برنامه را تعریف می‌کند. خط مشی امنیتی، قیود روی توابع و جریان‌‌های بین آن‌ها را مشخص می‌کند؛ مثل قیود دسترسی‌ بر روی برنامه‌ها و سطوح دسترسیِ داده‌های بین کاربران که مانع از بروز مشکلات امنیتی از طریق سیستم‌های خارجی و نفوذگران شود.

از دیدگاهی دیگر، یک خط مشی امنیتی را می‌توان به‌ عنوان یک زیرمجموعه از مجموعه‌ توانیِ همه اجراها تعریف کرد که هر اجرا یک دنباله دلخواه از حالت‌[[13]](#footnote-13)ها است. ضمناً می‌‌توان آن را به ‌عنوان مجموعه‌ برنامه‌هایی در نظر گرفت که آن خط مشی را برآورده می‌کنند. بعضی از خط مشی‌های امنیتی، *خاصیت*[[14]](#footnote-14) هستند؛ به‌خاطر این‌که قابل دسته‌بندی و تشخیص توسط مجموعه اجراهای جداگانه می‌باشند. از این نوع خط مشی‌ها می‌توان به خط‌ مشی‌های کنترل دسترسی اشاره کرد. [2] برخی از نیازمندی‌های مهم امنیتی، خاصیت نیستند. یک نمونه مهم از این‌گونه نیازمندی‌ها، خط مشیِ امنیتی عدم تداخل[[15]](#footnote-15) است. عدم تداخل گوگن-مسگر [3]، عدم تداخل تعمیم یافته [4] و قطعیت مبتنی بر مشاهده [5] از مثال‌های خط مشی‌هایی هستند که نمی‌‌توان آن‌ها را در قالب خاصیت بیان کرد. نکته حائز اهمیت این است که روش اِعمال خاصیت‌ها با نحوه اعمال خط مشی‌هایی که خاصیت نیستند، متفاوت است.

به زبان ساده‌تر‌، خط مشی عدم تداخل بیان می‌کند که یک مشاهده‌گرِ[[16]](#footnote-16) سطحِ پایین که فقط به برنامه و مقادیر عمومیِ زمانِ اجرا دسترسی دارد، نتواند ورودی‌های سطح بالا یا خصوصیِ برنامه را بفهمد. به عبارت دیگر، این خط مشی بیان می‌کند که در هر جفت اجراهای برنامه‌ که ورودی‌های عمومی یکسان دارند، مستقل از ورودی‌های خصوصی متفاوت، باید خروجی‌های عمومی یکی باشند.

نکته مهم این است که خط مشی عدم تداخل، یک خاصیت نیست؛ زیرا توسط اجراهای جداگانه که این خط مشی را برآورده می‌کند، قابل تعریف نیست. [2] این نکته باعث ایجاد محدودیت‌هایی برای اِعمال این خط مشی در برنامه‌ها می‌شود.

خط مشی عدم تداخل را می‌توان به دو دسته *حساس به پیشرفت[[17]](#footnote-17)* و *غیر حساس به پیشرفت[[18]](#footnote-18)* تقسیم کرد. در عدم تداخلِ غیر حساس به پیشرفت، مشاهده‌گرِ سطح پایین، تنها می‌تواند خروجی‌های میانیِ سطح پایین را ببیند؛ در حالی‌که یک مشاهده‌گرِ سطح پایین در عدم تداخلِ حساس به پیشرفت، علاوه بر دسترسی‌های قبلی، به *وضعیت پیشرفتِ[[19]](#footnote-19)* برنامه نیز دسترسی دارد. این باعث می‌شود تا بتواند تفاوت بین واگرایی[[20]](#footnote-20) برنامه با موقعیتی که برنامه خاتمه[[21]](#footnote-21) می‌یابد یا در حال محاسبه مقادیر قابل‌مشاهده‌ بعدی است، را تمیز دهد. [2]

# منابع و مراجع

|  |  |
| --- | --- |
| Aggarwal, Charu C., and Philip S. Yu. *Finding generalized projected clusters in high dimensional spaces*. Vol. 29. No. 2. ACM, 2000. | [1] |
| Cabrera, Juan Manuel, Hugo Jair Escalante, and Manuel Montes-y-Gómez. "Distributional term representations for short-text categorization." Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2013. 335-346. | [2] |
| Song, Ge, et al. "Short text classification: A survey." *Journal of Multimedia* 9.5 (2014): 635-643. | [3] |
| Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections." Proceedings of the 17th international conference on World Wide Web. ACM, 2008. | [4] |
|  |  |
|  |  |
|  |  |
|  |  |



Amirkabir University of Technology  
(Tehran Polytechnic)

Computer and Information Technology Engineering Department

B.Sc. Thesis

Title

Design and Implementation of a Web-App Software with Extendibility Ability on Detection and Extraction of Subject of Short text Systems

By

HamidReza Ramezani

Supervisor

Dr. Ahmad Nickabadi

September 2015

1. Language-based Security [↑](#footnote-ref-1)
2. Semantics [↑](#footnote-ref-2)
3. Types [↑](#footnote-ref-3)
4. Verification [↑](#footnote-ref-4)
5. Static Analysis [↑](#footnote-ref-5)
6. Compile Time [↑](#footnote-ref-6)
7. Dynamic Analysis [↑](#footnote-ref-7)
8. Run-time [↑](#footnote-ref-8)
9. Noninterference [↑](#footnote-ref-9)
10. Program Rewriting [↑](#footnote-ref-10)
11. Source Code [↑](#footnote-ref-11)
12. Policy [↑](#footnote-ref-12)
13. State [↑](#footnote-ref-13)
14. Property [↑](#footnote-ref-14)
15. Noninterference Security Policy [↑](#footnote-ref-15)
16. Observer [↑](#footnote-ref-16)
17. Progress-Sensitive [↑](#footnote-ref-17)
18. Progress-Insensitive [↑](#footnote-ref-18)
19. Progress Status [↑](#footnote-ref-19)
20. Divergence [↑](#footnote-ref-20)
21. Terminate [↑](#footnote-ref-21)