



دانشگاه تهران  
پردیس دانشکده‌های فنی  
دانشکده مهندسی برق و کامپیوتر



توسعه ابزار آدرس‌یابی آدرس‌های پستی  
به کمک مدل پنهان مارکوف

پایان‌نامه برای دریافت درجه کارشناسی  
در رشته مهندسی کامپیوتر گرایش نرم افزار

نام:

محمد رضا بخشایش

شماره دانشجویی:

۸۱۰۱۹۹۳۸۱

استاد راهنما:

دکتر احمد کلهر

بهمن ماه ۱۴۰۳

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

## تعهدنامه اصالت اثر

باسمه تعالی

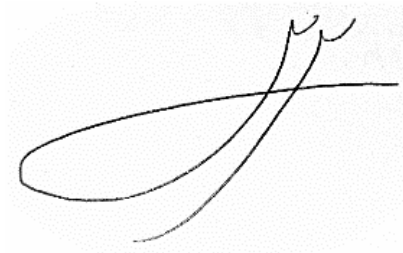
اینجانب محمدرضا بخشایش تأیید می‌کنم که مطالب مندرج در این پایان‌نامه حاصل تلاش اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم سطح یا بالاتر ارائه نشده است.

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

نام و نام خانوادگی دانشجو :

محمدرضا بخشایش

امضای دانشجو :



## چکیده

با گسترش روزافزون خدمات مکان محور در حوزه‌هایی نظیر مدیریت شهری، حمل‌ونقل هوشمند و امداد و نجات، تبدیل آدرس‌های پستی فارسی به مختصات جغرافیایی اهمیت زیادی پیدا کرده است. آدرس‌های پستی در ایران به دلیل ساختار پیچیده و محاوره‌ای خود، چالشی جدی در فرایند آدرسیابی محسوب می‌شوند. هدف این پژوهش، توسعه ابزاری مستقل و کارآمد برای تبدیل این آدرس‌ها به مختصات جغرافیایی است که علاوه بر کاهش وابستگی به داده‌های خارجی، امکان کاربرد در شهرها و مناطق کوچک‌تر را نیز فراهم سازد.

این ابزار مبتنی بر مدل مارکوف پنهان (HMM) طراحی شده و از داده‌های OpenStreetMap به‌عنوان مرجع نقشه استفاده می‌کند. نتایج آزمایش‌ها نشان داد که این ابزار توانسته ۹۴.۷۰٪ از آدرس‌ها را با خطایی کمتر از ۵ کیلومتر مکان‌یابی کند که در مقایسه با ابزارهای مشابه، نرخ قابل قبولی ارزیابی می‌شود.

طراحی ماژولار و سادگی ابزار، آن را به گزینه‌ای مناسب برای محیط‌های کم‌داده و کاربردهایی نظیر خدمات مشتریان، تحلیل‌های جغرافیایی و مدیریت بحران تبدیل کرده است. باوجود برخی محدودیت‌ها در پردازش آدرس‌های پیچیده، ابزار پیشنهادی پتانسیل بالایی برای توسعه و بهبود در پژوهش‌های آینده دارد.

## کلمات کلیدی:

آدرسیابی، آدرس پستی، مدل پنهان مارکوف

## فهرست مطالب

فصل ۱: مقدمه	۱
فصل ۲: مفاهیم اولیه	۶
۱-۲- ساختار آدرس های پستی در ایران	۷
۲-۲- مدل پنهان مارکوف	۸
فصل ۳: معماری ابزار	۱۰
۱-۳- نقشه مرجع	۱۱
۲-۳- داده ارزیابی	۱۲
۳-۳- تابع امتیاز Emission	۱۳
۴-۳- تابع امتیاز Transition	۱۴
۵-۳- امتیاز دهی به کاندید ها	۱۴
۶-۳- مشکل Performance	۱۵
۷-۳- ارزیابی	۱۶
فصل ۴: نتایج	۱۷
فصل ۵: جمع بندی	۲۲
فصل ۶: مراجع	۲۵

## فهرست تصاویر

شکل ۱ شبه کد الگوریتم ویتربی [۹]: این تصویر، مراحل الگوریتم ویتربی را شرح میدهد. به طوری کلی این الگوریتم در هر مرحله احتمال وقوع حالت درونی **qi** را در هر مقطع از زمان، به کمک بالاترین احتمال رسیدن به آن حالت از هر حالت قبلی، احتمال transition بین آن دو حالت و احتمال مشاهده رخ داده در آن مقطع حساب میکند. .... ۹

شکل ۲ توزیع خطای هندسی: در نمودار بالا، محور افقی خطای هندسی پاسخ های سامانه را بر حسب متر و محور عمودی درصد پاسخ ها را مشخص میکند. خط آبی رنگ، به صورت تجمیعی حجم آدرس های آدرس یابی شده با خطای کمتر مساوی هر آستانه خطا را نشان میدهد در حالی که خط قرمز رنگ، توزیع خطا را برای آستانه های مختلف خطا به صورت نقطه ای نشان میدهد. برای رسیدن به درک بهتر از توزیع خطا، در دو نمودار بالاتر محور افقی به خطای ۰ تا ۲ کیلومتر محدود شده تا اثر داده های پرت از نمودار حذف شود. .... ۱۸

## فهرست جدول‌ها

جدول ۱ مقایسه عملکرد ابزار با TehranGeocode: در این جدول، عملکرد ابزار توسعه داده شده در این پژوهش در بازه های مختلف خطای جغرافیایی با ابزار رقیب (TehranGeocode) مقایسه شده..... ۲۱

# فصل ۱

## فصل ۱ : مقدمه

---



با پیشرفت چشمگیر فناوری‌های اطلاعاتی و ارتباطی در دهه‌های اخیر، تقاضا برای خدمات مکان‌محور و تحلیل‌های مکانی به شکل قابل توجهی افزایش یافته است. این خدمات در زمینه‌های گوناگونی مانند حمل‌ونقل هوشمند، مدیریت بحران و امدادسانی، پژوهش‌های علمی، اپلیکیشن‌های مسیریابی، بازاریابی مکان‌محور، مدیریت شهری و خدمات ارسال کالا کاربرد فراوان دارند. بسیاری از داده‌های مورد استفاده در این حوزه‌ها به صورت داده‌های بدون ساختار، نظیر آدرس‌های پستی و توصیفات مکانی محاوره‌ای، ارائه می‌شوند. تبدیل این داده‌ها به فرمت‌های دیجیتال قابل پردازش مانند مختصات جغرافیایی، پیش‌نیاز اصلی برای بهره‌برداری مؤثر از آن‌ها در این خدمات است.

**آدرس‌یابی** یا **Geocoding** به فرایند تبدیل توصیفات متنی مکان به مختصات جغرافیایی اطلاق می‌شود. داده‌های ورودی این فرایند می‌توانند اشکال مختلفی داشته باشند، از جمله استخراج مکان‌های ذکر شده در مقالات خبری و شبکه‌های اجتماعی، تبدیل توصیفات مکانی نظیر "رستوران سنتی نزدیک میدان انقلاب"، یا استخراج اطلاعات مکان‌محور از آدرس‌های پستی، شماره تلفن و کدهای پستی. خروجی آدرس‌یابی بسته به کاربرد می‌تواند شامل مختصات جغرافیایی، کدهای پستی، تقسیمات شهری و کشوری، یا دیگر اطلاعات مکان‌محور باشد.

آدرس‌یابی آدرس‌های پستی در کشورهایی با ساختار استاندارد و یکنواخت، مانند بسیاری از کشورهای اروپایی، به دلیل وجود شبکه راه‌های مشخص، داده‌های مرجع دقیق، و یکتایی نام‌ها روی نقشه، با دقت و سادگی قابل انجام است. در این کشورها، تنها نام خیابان و شماره پلاک برای تعیین مختصات کافی است. اما در کشورهایی نظیر ترکیه، هند، چین و ایران، آدرس‌دهی غالباً مبتنی بر توصیف مسیر از یک نقطه مرجع تا مقصد مورد نظر است. در ایران به‌ویژه، استفاده از نشانه‌های محلی و تنوع در توصیفات، تحلیل آدرس‌ها را پیچیده کرده و نیاز به توسعه روش‌های پیشرفته و انعطاف‌پذیر برای آدرس‌یابی را به‌وضوح نشان می‌دهد.

در سال‌های اخیر، پژوهش‌های مختلفی برای افزایش دقت ابزارهای آدرس یاب انجام شده. برخی از این پژوهش‌ها به روش‌های مبتنی بر قواعد (Rule-Based) پرداخته‌اند، درحالی‌که برخی دیگر به استفاده از روش‌های آماری و یادگیری ماشین روی آورده‌اند و در نهایت، تعدادی نیز از روش‌های یادگیری عمیق بهره برده‌اند.

در دسته‌بندی روش‌های مبتنی بر قواعد، مطالعه مازوچی و همکاران سیستمی برای آدرس یابی آدرس‌های فارسی به نام TehranGeocode ارائه داده‌اند که طی سه مرحله، در مرحله اول آدرس را به کمک یک گرامر مستقل از متن به اجزا سازنده تجزیه می‌کند، سپس به کمک یک موتور جستجو متنی، کاندیدهای احتمالی

منطبق هر قسمت آدرس را پیدا کرده و سپس با کمک برنامه‌نویسی پویا مختصات بهترین کاندید را به‌عنوان جواب برمی‌گرداند [۱]. کومارلاس و کروشکه از تکنیک‌هایی مثل CFD برای پیش‌بینی اطلاعات ناقص در یک آدرس و استفاده هم‌زمان از چند منبع داده مرجع برای افزایش دقت آدرس‌یابی استفاده کرده‌اند. [۲] همچنین شیائوجینگ یائو و همکاران با کمک تکنیک‌های تطبیق فازی و جستجوی تمام متن ابزاری برای آدرس‌یابی آدرس‌های چینی ارائه داده‌اند. [۳]

در دسته روش‌های آماری و یادگیری ماشین، جیونگ لی و همکاران سه روش مختلف یادگیری ماشین شامل ماشین بردار پشتیبان (SVM)، جنگل تصادفی (RF) و تقویت گرادیان شدید (XGB) را برای تطبیق اسامی مکان‌ها در آدرس مقایسه کردند و نتایج نشان داد روش تقویت گرادیان شدید بهترین عملکرد را داشت [۴].

در نهایت، برخی پژوهش‌ها به سمت استفاده از روش‌های یادگیری عمیق رفته‌اند. ژنهورگ دو و همکاران یک روش جدید برای تبدیل آدرس به مختصات جغرافیایی ارائه کرده‌اند که شامل سه مرحله اصلی است. ابتدا یک مدل زبان آدرس (ALM) با استفاده از تکنیک BERT و یک مجموعه داده آدرس چینی پیش آموزش داده می‌شود. سپس، ویژگی‌های معنایی و جغرافیایی آدرس‌ها با استفاده از الگوریتم بهبودیافته K-means خوشه‌بندی می‌شوند تا مدل GSAM ایجاد شود. در نهایت، یک وظیفه پیش‌بینی مختصات آدرس برای اعتبارسنجی مدل GSAM طراحی شده است. این روش با ادغام مستقیم ویژگی‌های معنایی و مختصات جغرافیایی، دقت بالاتری در تبدیل آدرس به مختصات جغرافیایی فراهم می‌کند. برای آموزش این مدل از حدود یک میلیون آدرس به‌عنوان داده آموزشی استفاده شده. [۵] در کنار این موارد، برخی مقالات از شبکه‌های عصبی و مدل‌های زبانی برای تجزیه آدرس [۶] و تطبیق دو آدرس پستی [۷] استفاده کرده‌اند.

پژوهش‌های متعددی در زمینه آدرس‌یابی در سال‌های اخیر انجام شده که نتایج چشمگیری به همراه داشته‌اند. این مطالعات با استفاده از رویکردهای مختلف، از روش‌های مبتنی بر قواعد تا تکنیک‌های پیشرفته یادگیری ماشین و یادگیری عمیق، تلاش کرده‌اند دقت و کارایی ابزارهای آدرس‌یابی را بهبود بخشند. با این حال، بسیاری از این پژوهش‌ها به دلیل تفاوت‌های ساختاری و زبانی آدرس‌های فارسی، برای کاربرد در این زبان بهینه‌سازی نشده‌اند.

تنها ابزار شناخته شده‌ای که به‌طور خاص برای آدرس‌یابی آدرس‌های فارسی توسعه یافته **TehranGeocode** است. این ابزار اگرچه رویکرد نوآورانه‌ای در تجزیه و تحلیل آدرس‌های فارسی به کار گرفته و نتایج قابل قبولی در شهر تهران ارائه داده است، اما به دلیل وابستگی به داده‌ها و برنامه‌های خارجی، محدودیت‌هایی برای

گسترش به سایر شهرها، به‌ویژه مناطق کوچک‌تر یا روستایی، دارد. این محدودیت‌ها باعث می‌شود استفاده از آن در مواردی که داده‌های مرجع کامل یا زیرساخت‌های مشابه موجود نیستند، دشوار باشد.

هدف اصلی این پروژه توسعه ابزاری برای تبدیل آدرس‌های پستی محاوره‌ای فارسی به مختصات جغرافیایی است. این ابزار باید قادر باشد آدرس ورودی را به محدوده‌ای مشخص از نقطه مقصد تبدیل کند. با توجه به کاربردهای موردنظر، دقت بالای مختصات تا سطح یک نقطه خاص (Rooftop Accuracy) ضروری نیست و تعیین محدوده‌ای در حدود ۱ تا ۲ کیلومتر کافی است. این ویژگی می‌تواند ابزار را برای استفاده در کاربردهایی مانند سیستم‌های امداد و نجات، مدیریت خدمات مشتریان، و نرم‌افزارهای مسیریابی بسیار مناسب سازد.

یکی از چالش‌های اصلی این پژوهش، کمبود داده‌های آموزشی معتبر و برچسب‌گذاری‌شده در زبان فارسی است. جمع‌آوری و آماده‌سازی این داده‌ها فرآیندی دشوار و زمان‌بر است. به همین دلیل، این پروژه بر استفاده از رویکردی کلاسیک‌تر مبتنی بر مدل پنهان مارکوف (HMM) تمرکز دارد که به داده‌های آموزشی کمتری نیاز دارد و درعین‌حال می‌تواند عملکرد مطلوبی در این زمینه ارائه دهد.

همچنین، ابزار توسعه‌یافته باید به‌گونه‌ای طراحی شود که محدود به شهرهای بزرگ نباشد و در مناطق کوچک‌تر یا روستاها نیز قابل‌استفاده باشد. یکی دیگر از اهداف این پروژه، کاهش وابستگی به سامانه‌های بیرونی و توسعه سیستمی مستقل است که بتواند در شرایط مختلف جغرافیایی و بدون نیاز به منابع خارجی عمل کند.

در فصل بعدی این گزارش، مفاهیم اولیه مرتبط با پژوهش بررسی می‌شود. این بخش شامل معرفی ساختار آدرس‌های پستی در ایران و چالش‌های مرتبط با آن و همچنین تشریح مدل پنهان مارکوف (HMM) به‌عنوان مبنای نظری ابزار توسعه‌یافته است.

در **فصل ۳**، معماری مدل ارائه می‌شود. این بخش شامل توضیحات مربوط به ساختار ابزار، فرایندهای توسعه، و روش‌های آزمایشی است که برای ارزیابی عملکرد مدل استفاده شده‌اند.

**فصل ۴** به نتایج اختصاص دارد و در آن عملکرد ابزار در تست‌های مختلف بیان می‌شود. دقت مدل در شرایط مختلف، محدوده خطای قابل‌قبول، و مقایسه نتایج با ابزارهای مشابه از جمله موضوعات این بخش است.

در نهایت، **فصل ۵** به جمع‌بندی مطالب اختصاص دارد. در این بخش، دلایل رسیدن به نتایج به‌دست‌آمده، چالش‌های پیش‌رو در مراحل مختلف توسعه ابزار، و محدودیت‌های موجود مرور شده و پیشنهادهایی برای

ادامه پژوهش و بهبود ابزار در آینده ارائه می‌شود. **فصل ۶** نیز به مراجع استفاده شده در این متن اختصاص دارد.

## فصل ۲

### فصل ۲: مفاهيم اوليه

---

## ۲-۱ ساختار آدرس های پستی در ایران

آدرس پستی، به توصیف یک نقطه جغرافیایی خاص روی سطح زمین گفته میشود. هر آدرس، مجموعه ای از ویژگی های فضایی و ارتباط بین آنها، کد های پستی و قرارداد های مکان یابی است. در نقاط مختلف جهان، از روش های متفاوتی برای نحوه انتخاب این اجزا و توصیف آنها برای مشخص کردن یک نقطه خاص استفاده میشود. برای مثال در اکثر کشور های اروپایی، با توجه به استاندارد بودن ساختار شبکه راه ها، یکتا بودن نام عوارض شهری مانند نام خیابانها و در دسترس بودن داده مرجع برای نقشه شهر ها، آدرس های پستی از نام شهر، خیابان اصلی و شماره پلاک و واحد تشکیل شده. این امر، فرایند تفسیر و درک آدرس پستی و تبدیل آن به مختصات را به امری ساده و بدون ابهام تبدیل میکند.

از سوی دیگر، در کشور هایی مانند ترکیه، برزیل، هند، چین و به خصوص ایران، آدرس های پستی ساختار پیچیده تری دارند. برای مثال، آدرس های پستی در ایران معمولاً یک مسیر را از یک نقطه شناخته شده شهر، مانند میادین اصلی یا بزرگراه های معروف به سمت نقطه کمتر شناخته شده مقصد توصیف می کنند. هر چند استفاده از این قالب، پیدا کردن مسیر را برای عامل انسانی بدون نیاز به دانش قبلی از یک ناحیه به امری ساده تبدیل میکند، درک آن برای یک کامپیوتر به دلیل استفاده از عبارات نسبی، دستورات پیچیده مسیریابی و عدم دسترسی به شبکه کامل راه ها و تغییرات سریع محیط شهری به سادگی انجام نمیشود. کاربرد دیگر اطلاعات اضافی مسیریابی، متمایز کردن دو نقطه در شهر با نام یکسان ولی با مکان متفاوت است. برای مثال، خیابان های زیادی در شهر تهران عناوین یکسانی مانند "امام علی"، "امیر کبیر" و ... دارند.

به طور دقیق تر، آدرس های پستی در ایران را میتوان به صورت "دنباله ای از ویژگی های فضایی (مانند خیابان ها، میدان ها، تابلو ها و غیره) و روابط فضایی بین آنها (نرسیده به، بعد از، روبروی، ۱۰۰ متر قبل از) که یک مسیر را از یک نقطه شناخته شده به یک نقطه منحصر به فرد مشخص میکند" توصیف کرد. برای فهم بهتر، در ادامه این متن به هر بخش از آدرس که یک ویژگی فضایی خاص را به همراه ارتباط آن با بخش قبلی/بعدی مسیر مشخص میکند یک **قسمت** یا **segment** میگوییم. در این سیستم آدرس دهی، با انتخاب نقاط مختلف به عنوان مبدأ، انتخاب مسیر طی شده متفاوت و کم و زیاد کردن جزئیات استفاده شده در توصیف مسیر، میتوان آدرس های متفاوتی برای یک نقطه ارائه داد.

برای آشنایی بیشتر با ساختار این آدرس ها در مثال "تهران، خیابان کارگر شمالی، ۵۰۰ متر بالاتر از بلوار جلال، نرسیده به مرکز پست، دانشکده اقتصاد" از ۵ قسمت تشکیل شده که هر کدام به یک موقعیت جغرافیایی

خاص اشاره میکنند. قسمت اول، context جست و جو که شهر تهران است را مشخص میکند. خیابان کارگر شمالی نقطه شروع مسیر است. ویژگی های فضایی توصیف شده بعدی بلوار جلال و مرکز پست هستند که به کمک عبارات "۵۰۰ متر بالاتر از" و "نرسیده به" رابطه فضایی آنها با قسمت های قبلی و بعدی آدرس مشخص شده، در نهایت، "دانشکده اقتصاد" قسمت نهایی آدرس است که هدف نهایی آدرس را مشخص میکند. [۸]

## ۲-۲- مدل پنهان مارکوف

**مدل پنهان مارکوف** یا HMM یک مدل قدرتمند در تحلیل داده های دارای توالی است. این مدل برای توصیف سیستم هایی استفاده میشود که حالت درونی آنها به طور مستقیم قابل مشاهده نیست، اما میتوان با کمک خروجی های مدل، وضعیت درونی آنها را حدس زد. به عبارت دیگر، HMM به ما اجازه میدهد، صرفاً با مشاهده خروجی های متوالی یک سامانه، یک مدل احتمالی برای حالت درونی سامانه پیدا کنیم.

یکی از ویژگی های کلیدی HMM، توانایی بالای آن در تحلیل داده های دارای توالی، مانند داده های متنی، اصوات و تصاویر ضبط شده است که باعث میشود این مدل کاربرد های قابل توجهی در زمینه های پردازش زبان طبیعی مانند Speech Recognition، Part of Speech Tagging و ... داشته باشد. HMM یک ابزار قدرتمند و انعطاف پذیر برای تحلیل داده های توالی است و در بسیاری از حوزه های علمی و مهندسی کاربرد دارد.

یک مدل پنهان مارکوف از اجزای زیر تشکیل شده است:

- مجموعه ای از **حالات** یا **State** ها که آن را با  $Q = q_1, q_2 \dots q_n$  نمایش میدهیم
- یک **توالی مشاهدات** ایجاد شده توسط مدل که آن را با  $O = o_1 o_2 \dots o_T$  نمایش میدهیم
- **تابع احتمال انتشار** یا **Emission Probability Function** که احتمال رخداد هر مشاهده  $o_i$  را به شرط قرار داشتن در حالت درونی  $q_j$  مشخص میکند.
- **تابع احتمال انتقال** یا **Transition Probability Function** که احتمال انتقال از یک حالت درونی  $q_i$  به حالت درونی دیگر  $q_j$  را مشخص میکند.
- **تابع احتمال اولیه** یا **Initial Probability Function** که احتمال قرار داشتن حالت درونی سیستم در هر کدام از حالت های  $Q_i$  را مشخص میکند.

برای استفاده از مدل مارکوف پنهان درجه یک، دو فرض اساسی وجود دارد که برای استفاده از این مدل، باید از برقرار بودن این شروط اطمینان حاصل کرد، این دو فرض عبارت اند از:

- **فرض مارکوف:** احتمال قرار داشتن در هر حالت درونی صرفاً به حالت درونی قبلی سیستم بستگی دارد.

$$P(q_i | q_1 \dots q_{i-1}) = P(q_i | q_{i-1})$$

- **استقلال خروجی:** احتمال هر خروجی مشاهده شده  $o_t$  در هر زمان صرفاً به حالت درونی مدل در آن زمان و نه به حالت یا مشاهده قبلی و بعدی وابسته است.

$$P(o_i | q_1 \dots q_i, o_1 \dots o_i) = P(o_i | q_i)$$

در مسائل HMM، معمولاً هدف این است که صرفاً با مشاهده توالی خروجی های قابل مشاهده توسط سیستم، حدس هایی درباره حالت درونی سیستم انجام شود. مسئله ای که در این متن مورد بررسی قرار میگیرد، پیدا کردن محتمل ترین حالت درونی سیستم در زمان وقوع آخرین مشاهده است.

الگوریتم های متفاوتی برای استنتاج این مسئله در مدل مارکوف پنهان توسعه داده شده که از مهم ترین آنها میتوان به الگوریتم ویتربی اشاره کرد. این الگوریتم، به کمک برنامه نویسی پویا توسعه یافته و در هر مرحله، احتمال بودن سیستم در هر کدام از حالات درونی محاسبه میکند. شکل ۱، شبه کد الگوریتم ویتربی را نشان میدهد. [۹]

```

function VITERBI(observations of len  $T$ , state-graph of len  $N$ ) returns best-path, path-prob

create a path probability matrix viterbi[ $N, T$ ]
for each state  $s$  from 1 to  $N$  do                                ; initialization step
    viterbi[ $s, 1$ ]  $\leftarrow \pi_s * b_s(o_1)$ 
    backpointer[ $s, 1$ ]  $\leftarrow 0$ 
for each time step  $t$  from 2 to  $T$  do                            ; recursion step
    for each state  $s$  from 1 to  $N$  do
        viterbi[ $s, t$ ]  $\leftarrow \max_{s'=1}^N \text{viterbi}[s', t-1] * a_{s',s} * b_s(o_t)$ 
        backpointer[ $s, t$ ]  $\leftarrow \operatorname{argmax}_{s'=1}^N \text{viterbi}[s', t-1] * a_{s',s} * b_s(o_t)$ 
    bestpathprob  $\leftarrow \max_{s=1}^N \text{viterbi}[s, T]$                         ; termination step
    bestpathpointer  $\leftarrow \operatorname{argmax}_{s=1}^N \text{viterbi}[s, T]$             ; termination step
    bestpath  $\leftarrow$  the path starting at state bestpathpointer, that follows backpointer[] to states back in time
return bestpath, bestpathprob

```

شکل ۱ شبه کد الگوریتم ویتربی [۹]: این تصویر، مراحل الگوریتم ویتربی را شرح میدهد. به طوری کلی این الگوریتم در هر مرحله احتمال وقوع حالت درونی  $q_i$  را در هر مقطع از زمان، به کمک بالاترین احتمال رسیدن به آن حالت از هر حالت قبلی، احتمال *transition* بین آن دو حالت و احتمال مشاهده رخ داده در آن مقطع حساب میکند.



## فصل ۳

### فصل ۳: معماری ابزار

---

برای تبدیل مسئله آدرس‌یابی به یک مدل پنهان مارکوف (HMM) و تحلیل آن، لازم است ابتدا اجزای اصلی این مدل را برای مسئله موردنظر تعریف کنیم. مدل پنهان مارکوف شامل چند بخش اصلی است که باید با ساختار مسئله آدرس‌یابی تطبیق داده شوند.

در این مدل، می‌توان فرض کرد که حالت‌های پنهان (States) مختصات جغرافیایی مربوط به آدرس هستند. این مختصات اجزای کلیدی سناریوی مسیریابی آدرس را تشکیل می‌دهند، اما در زمان تحلیل، به‌صورت مستقیم قابل مشاهده نیستند. از طرف دیگر، هر قسمت از آدرس به‌عنوان یک مشاهده (Observation) در مدل عمل می‌کند؛ به این معنا که هر قسمت از آدرس به یکی از مختصات جغرافیایی وابسته است. همانند یک مدل پنهان مارکوف، آدرس به‌عنوان زنجیره‌ای از ویژگی‌های جغرافیایی عمل می‌کند و هدف، شناسایی آخرین قسمت از این زنجیره است.

برای استفاده از مدل پنهان مارکوف، باید بررسی کنیم که آیا شرایط لازم برای اعمال این مدل در مسئله وجود دارد یا خیر. در یک آدرس پستی، اگر مشخص باشد که یک بخش از آدرس به چه مختصاتی اشاره دارد، توالی بخش‌های بعدی یا قبلی آدرس به‌صورت مستقل از یکدیگر قابل توصیف هستند. این خاصیت، شرط مارکوف را در مسئله تضمین می‌کند. علاوه بر این، تمرکز اصلی هر قسمت از آدرس بر نام یا توصیف ویژگی جغرافیایی متناظر با آن است. این ویژگی‌ها معمولاً مستقل از سایر بخش‌های آدرس هستند، مگر در مواردی که از روابط پیچیده مکانی برای توصیف مسیر استفاده شود که این روابط در روش ارائه شده در این پروژه ندیده گرفته میشوند. بنابراین، شرط استقلال در این مسئله تا حد قابل قبولی برقرار است.

برای تکمیل مدل‌سازی، نیاز به تعریف توابع احتمال انتقال (Transition) و انتشار (Emission) داریم. این توابع باید احتمال جابه‌جایی میان حالت‌ها و ارتباط میان مشاهدات و حالت‌های پنهان را مدل‌سازی کنند. در ادامه این مقاله، جزئیات بیشتری درباره ویژگی‌های جغرافیایی، اجزای آدرس، و نحوه مدل‌سازی توابع احتمال انتقال و انتشار ارائه خواهیم کرد. همچنین، نحوه ترکیب این توابع و ارتباط آن‌ها با اجزای آدرس و ویژگی‌های جغرافیایی شرح داده می‌شود. در انتها، چالش‌های مربوط به پیاده‌سازی این مدل، راهکارهای پیشنهادی، و نتایج حاصل از ارزیابی آن مورد بررسی قرار خواهند گرفت.

### ۱-۳- نقشه مرجع

در این پروژه، از یک رویکرد کلاسیک برای شناسایی مقصد آدرس استفاده شده است. این رویکرد نیازمند دسترسی به مجموعه‌ای جامع از ویژگی‌های جغرافیایی هر شهر به‌عنوان داده مرجع است. این ویژگی‌ها شامل

اطلاعاتی مانند نام ویژگی، نوع ویژگی (مانند معبر، منطقه شهری یا نقطه توجه)<sup>۱</sup>، هندسه، و مختصات جغرافیایی هستند. مجموعه این اطلاعات به عنوان نقشه مرجع شناخته می شود که مبنای تحلیل آدرس ورودی را فراهم می کند. هدف اصلی الگوریتم در هر مرحله، شناسایی محتمل ترین رکورد از میان این داده ها به عنوان نقطه مورد اشاره در آدرس است. این فرآیند به عنوان تطابق شناخته می شود و رکوردهایی که در این فرآیند بررسی می شوند، کاندیدها نامیده می شوند.

برای تأمین داده مرجع، از OpenStreetMap (OSM) استفاده شده است. OSM یک منبع گسترده از داده های مکانی است که توسط گروهی از داوطلبان در سراسر جهان ایجاد و به روزرسانی می شود. جامعیت بالا، دسترسی آزاد، و کیفیت مناسب این مجموعه داده، آن را به گزینه ای ایده آل برای تأمین داده مرجع در این پروژه تبدیل کرده است.

جهت آماده سازی نقشه مرجع برای استفاده در ابزار، داده های جغرافیایی OSM مربوط به محدوده سیاسی ایران استخراج شدند [۱۰]. این داده ها به صورت شهر به شهر دسته بندی شده و برای بهینه سازی عملکرد ابزار، رکوردهای غیرضروری از مجموعه داده حذف شدند. این رکوردها شامل اطلاعاتی مانند عوارض طبیعی، معابر خصوصی، خطوط راه آهن، رکوردهای بدون نام، و سایر ویژگی هایی بودند که نقشی در فرآیند آدرس یابی ایفا نمی کردند.

در نسخه نهایی نقشه مرجع، متادیتای ذخیره شده به نام و هندسه رکوردها محدود شد و چند تگ متادیتا جدید برای مشخص کردن نوع رکورد (معبر، منطقه شهری یا نقطه توجه) و شهر مربوطه به داده ها اضافه شد. این پیش پردازش باعث شد تا ابزار با دقت و سرعت بیشتری به تحلیل و تطابق آدرس ها پردازد.

## ۲-۳- داده ارزیابی

برای ارزیابی مدل و تعیین مقادیر بهینه برای پارامترهای آن، به مجموعه ای از داده های شامل آدرس های پستی و مختصات جغرافیایی مرتبط با هر آدرس نیاز بود. جمع آوری این داده ها یکی از چالش های اصلی پروژه بود. در نهایت، با استفاده از یک ابزار Web Crawler و اجرای آن بر روی وبسایت های تخفیفان و کتاب اول، موفق به جمع آوری مجموعه ای شامل سه هزار آدرس پستی مرتبط با شهر تهران همراه با مختصات جغرافیایی شدیم.

<sup>۱</sup> نقطه توجه یا Point-of-Interest به نقاطی روی نقشه گفته می شود که ممکن است برای افراد مختلف مهم باشند، مواردی مثل بنا های مهم تاریخی، فروشگاه ها، ایستگاه های حمل و نقل عمومی نمونه هایی از این نقاط هستند.

از این مجموعه، ۲۵۰۰ آدرس برای ارزیابی عملکرد ابزار و ۵۰۰ آدرس به عنوان داده‌های **Held-out** برای تنظیم مقادیر **Hyperparameter**ها مورد استفاده قرار گرفت.

برای اطمینان از عملکرد ابزار، آدرس ورودی با استفاده از کتابخانه **هضم** نرمال‌سازی شد. این فرآیند شامل حذف نیم‌فاصله و کاراکترهای اعراب، جایگزینی ارقام و علائم فارسی با معادل انگلیسی، محدودسازی کاراکترها به حروف فارسی و انگلیسی و علائم مجاز، و اصلاح فاصله‌گذاری بود. سپس، آدرس به کمک علائم سجاوندی مانند ویرگول به بخش‌های مجزا تجزیه شد.

ابزار صرفاً شباهت بخش‌های آدرس با نام ویژگی‌های هندسی را بررسی می‌کند؛ بنابراین، روابط مکانی پیچیده مانند "نرسیده به" یا "بعد از" و همچنین جزئیات غیرمرتبط نظیر شماره پلاک و رنگ ساختمان حذف شدند تا تحلیل به اجزای کلیدی محدود شود.

### ۳-۳- تابع امتیاز Emission

تابع امتیاز Emission مشخص می‌کند که احتمال اشاره یک بخش از آدرس به یک کاندیدای خاص چقدر است. با توجه به اینکه اغلب قسمت‌های آدرس ویژگی‌های جغرافیایی را بر اساس نام توصیف می‌کنند، استفاده از فاصله لون اشتاین (**Levenshtein Distance**) میان بخش آدرس و نام کاندیدا، معیار مناسبی به نظر می‌رسد.

مدلی که توسعه داده شده، از فرمول زیر برای محاسبه امتیاز Emission استفاده می‌کند:

$$P(segment_i | candidate_j) = 1 - \frac{EditDistance(segment_i, candidate_j.name)}{len(segment_i) + len(candidate_j.name)}$$

در این فرمول EditDistance فاصله ویرایش بین دو رشته ورودی، len طول کل رشته، **segment<sub>i</sub>** قسمت شماره i از آدرس و **candidate<sub>j</sub>** یک رکورد خاص از داده‌های مرجع را مشخص می‌کند. این مقدار نشان‌دهنده نزدیکی معنایی بین قسمت آدرس و کاندید بوده که مقادیر بالاتر به معنای شباهت بیشتر بین دو رشته ورودی است.

### ۳-۴ تابع امتیاز Transition

در مسئله آدرس‌یابی پستی، هر آدرس مسیری را در سطح شهر مشخص می‌کند که از نقاطی با روابط هندسی معین تشکیل شده است. این روابط می‌توانند نزدیکی جغرافیایی، اتصال مستقیم یا ارتباطات پیچیده‌تر باشند. یکی از چالش‌های اساسی در این فرآیند، وجود کاندیداهایی با نام‌های مشابه است که نیازمند تابعی برای تعیین احتمال وقوع هر کاندید در توالی آدرس است. در مدل HMM، این احتمال توسط تابع امتیاز Transition محاسبه می‌شود.

به منظور حفظ سادگی ابزار، از فاصله اقلیدسی میان مختصات دو ویژگی جغرافیایی به عنوان معیار اصلی استفاده شده است. این روش بیان می‌کند که هرچه فاصله کمتر باشد، احتمال وقوع متوالی این ویژگی‌ها در آدرس بیشتر است. به دلیل محدودیت داده‌ها و عدم امکان آموزش مدل‌های یادگیری ماشین، از تابعی یکنوا و غیرصعودی استفاده شده که فاصله‌ها را به مقادیری بین صفر و یک نگاشت می‌کند. این تابع برای فاصله‌های کوتاه مقدار نزدیک به یک و برای فاصله‌های زیاد مقدار نزدیک به صفر باز می‌گرداند.

به دلیل محدود بودن حجم داده آزمایشی و همچنین هزینه بالای برچسب گذاری هر قسمت از آدرس، امکان رسیدن به تابع امتیاز بهینه برای Transition با کمک شیوه‌های متداول Machine Learning وجود نداشت. بنابراین تلاش شد یک تابع طراحی شده به شکل دستی برای این تابع پیشنهاد شود. بعد از امتحان چند تابع ممکن و ارزیابی نتیجه روی داده آموزش، بهترین تابع پیشنهادی به عنوان تابع امتیاز Transition انتخاب شد.

### ۳-۵ امتیاز دهی به کاندید ها

برای تطابق بخش‌های مختلف آدرس با رکوردهای نقشه مرجع، مسئله به عنوان یک مدل پنهان مارکوف (HMM) تعریف شده و از الگوریتمی مشابه ویتربی استفاده شده است. در این روش، در هر مرحله احتمال تطابق هر رکورد نقشه مرجع با یک بخش از آدرس محاسبه و ذخیره می‌شود. این احتمال بر اساس دو مولفه اصلی تعیین می‌شود: احتمال تطابق کاندیدای جاری با بخش فعلی آدرس و مجموع احتمالات تطابق کاندیداهای مرحله قبلی با کاندیدای فعلی. فرمول زیر نحوه محاسبه این احتمال را نشان می‌دهد:

$$P(segment_i | candidate_j) * \sum_{candidate_k \in ReferenceMap} P_{match_{i-1}}(candidate_k) * P(candidate_j | candidate_k)$$

در این فرمول  $P_{match_i}$  احتمال تطابق یک کاندید با قطعه  $m$  از آدرس را مشخص کند. این مقدار برای  $i=0$  برای همه کاندیدها برابر ۱ فرض شده.

برخلاف الگوریتم کلاسیک ویتربی که تنها از بیشترین احتمال انتقال استفاده می‌کند، این روش مجموع احتمالات را به کار می‌گیرد. این انتخاب اجازه می‌دهد اگر در مرحله‌ای به یک کاندید اشتباهاً امتیاز بالایی داده شود، این خطا در مراحل بعدی اصلاح شود.

برای جلوگیری از تأثیرات منفی امتیازات بسیار بزرگ یا کوچک، مقادیر تطابق هر بخش از آدرس با ترکیبی از نرمال‌سازی گوسی و تابع Softmax نرمال‌سازی می‌شوند. این فرآیند تضمین می‌کند که مقادیر نهایی در بازه‌ای متناسب قرار گرفته و تأثیر مثبتی بر ادامه پردازش داشته باشند.

در پایان، کاندیدایی که برای آخرین بخش آدرس بالاترین امتیاز را دارد به عنوان نتیجه نهایی فرآیند آدرس‌یابی گزارش می‌شود. این رویکرد انعطاف‌پذیر امکان بهبود خطاهای محلی در طی مراحل تطابق را فراهم می‌کند و دقت نهایی سیستم را افزایش می‌دهد.

### ۶-۳- مشکل Performance

راه حل ارائه شده برای مسئله آدرس‌یابی، با وجود قدرت بالای آن در یافتن پاسخ صحیح، دارای پیچیدگی زمانی  $O(nm^2)$  است. این پیچیدگی برای آدرسی با  $n$  بخش و نقشه‌ای با  $m$  رکورد، به‌ویژه برای شهرهای بزرگ با بیش از یک میلیون رکورد در نقشه مرجع، هزینه زمانی بالایی دارد. بنابراین، بهینه‌سازی زمان اجرا بدون کاهش قابل توجه در دقت مدل ضروری است.

در مدل پنهان مارکوف، اکثر حالت‌ها برای بسیاری از مشاهدات دارای احتمال نزدیک به صفر هستند. این خاصیت به ما امکان می‌دهد تعداد کاندیداهای بررسی شده را برای هر بخش آدرس به صورت مؤثری محدود کنیم. استراتژی‌های اعمال شده برای کاهش زمان اجرا شامل موارد زیر است:

- حذف رکوردهایی با امتیاز Emission کمتر از یک آستانه مشخص در هر مرحله.
- حذف کاندیداهایی با فاصله اقلیدسی بیشتر از یک آستانه مشخص نسبت به کاندیداهای انتخاب شده قبلی.
- محدود کردن تعداد کاندیداهای بررسی شده در هر مرحله به یک تعداد حداکثر.

مقادیر آستانه‌ها به صورت تجربی و با ارزیابی عملکرد مدل روی داده‌های آزمایشی تعیین شده‌اند. این روش‌ها زمان اجرای الگوریتم را به میزان قابل توجهی کاهش داده‌اند، بدون اینکه تأثیر منفی قابل ملاحظه‌ای بر دقت نهایی داشته باشند.

برای مدیریت حجم بالای داده‌های نقشه مرجع، از پایگاه داده **Elasticsearch** استفاده شده است. این ابزار به عنوان یک موتور جستجوی متن‌باز و توزیع شده، توانایی بالایی در ذخیره‌سازی حجم زیادی از داده‌های بدون ساختار دارد و ابزارهای قدرتمندی برای جستجوی متن و داده‌های جغرافیایی ارائه می‌دهد. استفاده از **Elasticsearch** امکان ارزیابی سریع و مؤثر اطلاعات جغرافیایی و متنی را فراهم کرده و نقش مهمی در بهینه‌سازی کلی سیستم داشته است.

### ۷-۳- ارزیابی

برای ارزیابی عملکرد ابزار توسعه یافته، داده‌های جمع‌آوری شده به عنوان داده تست به مدل وارد شدند و مختصات خروجی مدل با مختصات مرجع مقایسه گردید. فاصله میان نقطه پیش‌بینی شده توسط مدل برای هر آدرس و نقطه مرجع، به عنوان **خطای هندسی** تعریف شد.

با توجه به ماهیت آدرس‌دهی، حتی دقیق‌ترین روش‌های آدرس‌یابی نیز ممکن است به نتایج مختلفی برای یک آدرس مشابه منجر شوند. به عنوان مثال، برای آدرسی که به یک خیابان اشاره دارد، ممکن است نقاط مختلفی روی آن خیابان به عنوان خروجی ارائه شوند یا اختلاف مختصات در حد اعشار وجود داشته باشد. به همین دلیل، تعریف یک **آستانه خطا** برای تعیین مناسب یا نامناسب بودن پاسخ ضروری است.

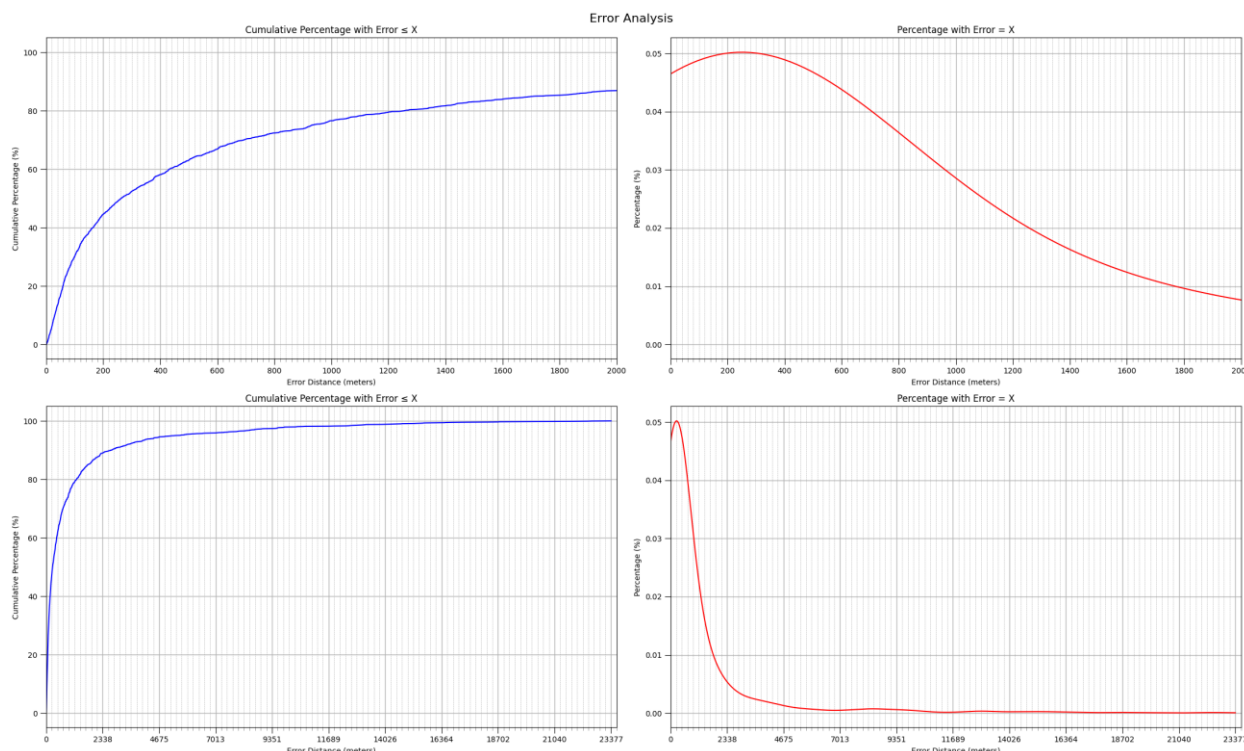
این ارزیابی امکان تحلیل دقیق خطای مدل و بررسی عملکرد آن تحت شرایط مختلف را فراهم می‌کند. فصل بعد، نتایج حاصل از این ارزیابی را ارائه و تحلیل خواهد کرد.

## فصل ٤

### فصل ٤: نتایج

---





**شکل ۲ توزیع خطای هندسی:** در نمودار بالا، محور افقی خطای هندسی پاسخ های سامانه را بر حسب متر و محور عمودی درصد پاسخ ها را مشخص میکند. خط آبی رنگ، به صورت تجمیعی حجم آدرس های آدرس یابی شده با خطای کمتر مساوی هر آستانه خطا را نشان میدهد در حالی که خط قرمز رنگ، توزیع خطا را برای آستانه های مختلف خطا به صورت نقطه ای نشان میدهد. برای رسیدن به درک بهتر از توزیع خطا، در دو نمودار بالاتر محور افقی به خطای ۰ تا ۲ کیلومتر محدود شده تا اثر داده های پرت از نمودار حذف شود

**شکل ۲** توزیع خطای هندسی ابزار توسعه یافته را بر حسب متر و بر روی مجموعه ای از آدرس ها نمایش می دهد. بر اساس این نمودار، مدل پیشنهادی توانسته است ۸۶.۸۵٪ از آدرس های ورودی را با حداکثر خطای هندسی ۲ کیلومتر مکان یابی کند.

این عملکرد نشان می دهد که ابزار توسعه یافته، به ویژه در کاربردهایی که نیازی به دقت مکانی در حد متر ندارند، به خوبی قادر به برآورده سازی نیازها است. نمونه هایی از این کاربردها شامل پژوهش های علمی، تحلیل داده های جغرافیایی، و مراکز خدمات مشتریان می باشد، جایی که تعیین موقعیت تقریبی آدرس ها کافی بوده و نیازی به یافتن نقطه هدف دقیق وجود ندارد.

علاوه بر این، در کاربردهایی که دقت بالا (Rooftop Accuracy) مورد نیاز است، مانند اپلیکیشن های مسیریابی و ناوبری، این ابزار می تواند با ارائه موقعیت حدودی مقصد، فرایند تعیین نقطه دقیق هدف را برای کاربران انسانی تسهیل و تسهیل کند. این ویژگی ابزار را به یک راه حل کارآمد و انعطاف پذیر در شرایط مختلف تبدیل می کند.

از سوی دیگر، در شکل ۲ مشاهده میشود که ۵۳٪ از پاسخ‌های سامانه دارای خطایی بیش از ۵ کیلومتر نسبت به نقطه هدف بوده‌اند. همان‌طور که پیش‌تر بیان شد، این ابزار به نحوی طراحی شده است که حتی در صورت عدم یافتن یک تطابق دقیق برای آدرس ورودی، کاندیدای دارای بالاترین امتیاز را به‌عنوان پاسخ باز می‌گرداند. این مسئله ممکن است منجر به بازگرداندن جوابی شود که حتی در نزدیکی نقطه هدف قرار ندارد. با توجه به اینکه این میزان خطا قابل توجه است و می‌تواند عملکرد مدل را به‌طور جدی تحت تأثیر قرار دهد، به نظر می‌رسد عوامل زیر در ایجاد این خطا نقش داشته باشند:

**وجود نام‌های متعدد برای مکان‌ها:** بسیاری از نقاط جغرافیایی در نقشه شهرهای ایران به دلیل تغییر نام در سال‌های اخیر با عناوین مختلفی شناخته می‌شوند. برای مثال، "بزرگراه شهید رئیسی" و "بزرگراه بعثت" به یک محدوده جغرافیایی اشاره دارند. ابزار توسعه‌یافته صرفاً بخش‌های آدرس را با جدیدترین نام رسمی مقایسه می‌کند و در مواجهه با اسامی قدیمی‌تر مکان‌ها، توانایی شناسایی کاندیدای صحیح را ندارد.

**نقص در داده‌های مرجع:** ابزار موردنظر از داده‌های OpenStreetMap به‌عنوان منبع مرجع نقشه استفاده می‌کند. با وجود اینکه این منبع شامل مجموعه گسترده‌ای از اطلاعات مربوط به عوارض شهری است، در برخی موارد مشاهده شده که اطلاعات مربوط به برخی معابر فرعی در آن موجود نیست. وجود نقص یا اشتباه در داده‌های مرجع می‌تواند منجر به بروز خطا در نتایج خروجی ابزار شود.

**ساختار پیچیده آدرس‌های پستی در ایران:** همان‌طور که پیش‌تر اشاره شد، آدرس‌های پستی فارسی معمولاً شامل مجموعه‌ای از دستورات مسیریابی از یک نقطه شناخته‌شده تا مقصد هستند. این دستورات می‌توانند در قالب‌های متنوعی ارائه شوند و شامل اجزا و روابط مختلفی باشند. ابزار توسعه‌یافته، هر بخش از مسیر را صرفاً به‌صورت یک نام و مختصات جغرافیایی مدل‌سازی می‌کند و روابط مکانی را تنها بر اساس فاصله اقلیدسی نقاط تحلیل می‌نماید. این رویکرد نمی‌تواند تمامی اطلاعات موجود در آدرس‌های فارسی را به‌درستی پردازش کند و ممکن است در برخی موارد، منجر به درک اشتباه از اجزا و روابط آنها با داده‌های نقشه مرجع شود. پیش‌بینی می‌شود با جایگزینی سیستم امتیازدهی به کاندیداها با یک مدل پیچیده‌تر، بتوان از این خطاها جلوگیری کرد.

در جدول ۱، عملکرد مدل در آستانه‌های مختلف خطای هندسی با ابزار TehranGeocode به عنوان state of the art در حوزه آدرس‌یابی آدرس‌های فارسی به مختصات مقایسه شده. مشاهده میشود در بازه‌های مختلف خطا، ابزار توسعه‌داده شده توانایی رقابت با ابزارهای مشابه را داشته و در محدوده خطای هدف موفق به کسب برتری ۲۵ درصدی نسبت به ابزار رقیب شده. این مسئله نشان‌دهنده موفقیت الگوریتم آدرس‌یابی پیشنهادی

بوده و انتظار می‌رود بتوان در پژوهش‌های بعدی، با کمک مدل‌های پیچیده‌تر برای محاسبه امتیازات Emission و Transition به نتایج بهتری رسید.

یکی از تفاوت‌های مشهود بین این دو مدل تفاوت در نرخ آدرس‌های بدون پاسخ است. در حالی که TehranGeocode برای ۷۰٪ از آدرس‌ها هیچ پاسخی ارائه نمی‌دهد، مدل پیشنهادی توانسته برای تمام آدرس‌ها پاسخی برگرداند، حتی اگر برخی از این نتایج دارای خطای مکانی باشند.

بازگرداندن پاسخ، هرچند با خطا، در بسیاری از کاربردها مانند خدمات مشتریان، تحلیل‌های آماری، یا پژوهش‌های علمی می‌تواند مفید باشد. داشتن یک مکان تقریبی بهتر از نداشتن هیچ اطلاعاتی است، چرا که می‌تواند نقطه شروعی برای اصلاح یا تکمیل اطلاعات توسط کاربر انسانی یا سیستم‌های دیگر باشد. این تصمیم همچنین از سردرگمی کاربران جلوگیری کرده و فرآیندهای وابسته به مکان‌یابی را حتی در شرایط عدم دقت کامل تسهیل می‌کند.

TehranGeocode	ابزار توسعه داده شده	حداکثر خطای هندسی
۳۰.۷۰٪	۳۰.۱۰٪	کمتر از ۱۰۰ متر

کمتر از ۲۰۰ متر	۴۴.۵۵٪	۴۳.۲۰٪
کمتر از ۴۰۰ متر	۵۸.۱۰٪	۶۰.۳۹٪
کمتر از ۶۰۰ متر	۶۶.۹۰٪	۶۹.۶۹٪
کمتر از ۸۰۰ متر	۷۲.۳۵٪	۷۵.۶۹٪
کمتر از ۱۰۰۰ متر	۷۶.۵۵٪	۷۹.۳۹٪
کمتر از ۲ کیلومتر	۸۶.۸۵٪	۸۴.۶۹٪
بیشتر از ۲ کیلومتر	۱۳.۱۵٪	۸.۲۰٪
آدرس‌های بدون جواب	۰.۰۰٪	۷.۱۰٪

جدول ۱ مقایسه عملکرد ابزار با *TehranGeocode*: در این جدول، عملکرد ابزار توسعه داده شده در این پژوهش در بازه‌های مختلف خطای

جغرافیایی با ابزار رقیب (*TehranGeocode*) مقایسه شده

## فصل ۵

فصل ۵: جمع بندی

---

هدف اصلی این پژوهش، توسعه ابزاری برای تبدیل آدرس‌های پستی فارسی به مختصات جغرافیایی بود، به شکلی که این مختصات قابل ذخیره‌سازی و پردازش کامپیوتری باشد. این ابزار با رویکردی ساده و بدون نیاز به پیش‌نیازهای پیچیده، طراحی شد تا قابلیت اجرا در محدوده‌های جغرافیایی گوناگون، حتی شهرهای کوچک و روستاها، را داشته باشد. از آنجا که دستیابی به دقت بسیار بالا (Rooftop Accuracy) برای بسیاری از کاربردها الزامی نبود، ابزار پیشنهادی توانست به دقت قابل قبولی دست یابد.

نتایج حاصل از آزمایش‌های انجام‌شده نشان داد که این ابزار قادر است حدود ۹۴.۷۰٪ از آدرس‌های ورودی را با دقتی در حد حداکثر ۵ کیلومتر به مختصات جغرافیایی تطبیق دهد. این دستاورد نشان‌دهنده عملکرد قابل‌توجه این ابزار در مقایسه با مدل رقیب است. علاوه بر این، ویژگی عدم وابستگی به فرایند پیچیده‌ی آموزش و امکان تنظیم مدل تنها با استفاده از داده‌های مرجع محلی، آن را برای استفاده در مناطق فاقد داده‌های گسترده و برچسب‌گذاری‌شده بسیار مناسب ساخته است. این خصوصیت، هزینه و زمان لازم برای پیاده‌سازی ابزار در محیط‌های جدید را به‌طور چشمگیری کاهش می‌دهد. یکی دیگر از دستاوردهای کلیدی این پژوهش، توسعه‌ی مازولار ابزار بود که انعطاف‌پذیری لازم را برای سفارشی‌سازی و بهبود بخش‌های مختلف آن، متناسب با نیازهای کاربران فراهم می‌کند.

علاوه بر این، توانایی ابزار در تحلیل و تطبیق اجزای مختلف آدرس‌ها، امکان کاربرد آن در حوزه‌های مختلفی را فراهم ساخته است. به‌عنوان مثال، این ابزار در **مراکز خدمات مشتریان** می‌تواند با افزایش سرعت یافتن مختصات، بهره‌وری تکنسین‌ها را ارتقا دهد و تجربه‌ای رضایت‌بخش‌تر برای مشتریان ایجاد کند. در **اپلیکیشن‌های مبتنی بر مکان**، این ابزار نقش مهمی در تسهیل مسیریابی و مدیریت آدرس‌های محلی ناشناخته ایفا می‌کند. همچنین، در **مدیریت بحران و خدمات اورژانسی**، سرعت عمل ابزار در یافتن موقعیت جغرافیایی افراد می‌تواند در کاهش زمان واکنش به حوادث نقش حیاتی داشته باشد. علاوه بر این، در **پژوهش‌های علمی**، قابلیت تبدیل آدرس‌های متنی به مختصات جغرافیایی، امکان استفاده از داده‌های بدون ساختار را برای تحلیل‌های آماری و مدلسازی فراهم کرده و به تحقیقات مرتبط با تحلیل‌های مکانی، الگوهای جغرافیایی و داده‌های شبکه‌های اجتماعی کمک می‌کند.

با وجود موفقیت‌های ابزار پیشنهادی، محدودیت‌هایی نیز وجود دارد. به‌عنوان مثال، این ابزار قادر به پردازش دقیق آدرس‌های پیچیده و محاوره‌ای نیست و برای کاربردهایی که نیاز به موقعیت‌یابی بسیار دقیق دارند، به نظارت عامل انسانی نیاز خواهد بود. این محدودیت‌ها عمدتاً به دلیل ساختار ساده‌ی ابزار و عدم استفاده از تکنیک‌های مدرن‌تر مانند شبکه‌های عصبی است.

چالش‌های متعددی در مسیر توسعه این ابزار وجود داشت. از جمله این چالش‌ها، کمبود پژوهش‌های قبلی در زمینه تبدیل آدرس‌های پستی فارسی به مختصات بود که باعث سردرگمی در انتخاب تکنیک‌ها و منابع اولیه شد. علاوه بر این، محدودیت منابع داده‌ای و هزینه بالای برچسب‌گذاری داده‌ها، استفاده از تکنیک‌های پیشرفته‌تر یادگیری ماشین را غیرممکن ساخت. با این حال، انتخاب یک روش کلاسیک و قابل اجرا بدون نیاز به داده‌های گسترده، به ساده‌سازی توسعه ابزار کمک کرد و کاربرد آن را در محیط‌های کم‌داده تضمین نمود.

برای پژوهش‌های آینده، پیشنهاد می‌شود از مدل‌های داده‌محور مانند شبکه‌های عصبی برای بهبود عملکرد ابزار استفاده شود. به کارگیری مدل‌های پیشرفته برای تعیین امتیازات Emission و Transition در شبکه HMM و ترکیب این امتیازات می‌تواند به دقت بالاتری منجر شود. همچنین، بهره‌گیری از مدل‌های زبانی برای تجزیه و نرمال‌سازی آدرس‌های متنی می‌تواند عملکرد ابزار را در مواجهه با آدرس‌های محاوره‌ای و پیچیده بهبود بخشد. ابزار توسعه‌یافته در این پژوهش می‌تواند به عنوان پایه‌ای مناسب برای توسعه روش‌های پیشرفته‌تر و بهبود عملکرد ابزارهای آدرس‌یابی در آینده مورد استفاده قرار گیرد.

## فصل ٦

### فصل ٦: مراجع

---



## مراجع

- [١] R. Mazochi, s. Bourbour, M. R. Ghofrani and S. Momtazi, "Persian Address Geocoding: an LALR Parsing and Dynamic Programming Approach," *Journal of AI and Data Mining*, vol. ١١, pp. ٢٩١-٣٠٢, ٢٠٢٣.
- [٢] I. Koumarelas, A. Kroschk, C. Mosley and F. Naumann, "Experience: Enhancing Address Matching with Geocoding and Similarity Measure Selection," *J. Data and Information Quality*, vol. ١٠, no. ٢, pp. ١-١٦, ٢٠١٨.
- [٣] X. Yao, X. Li, L. Peng and T. Chi, "A novel fuzzy Chinese address matching engine based on full-text search technology," in *The 5th International Conference on Computer Engineering and Networks*, ٢٠١٥.
- [٤] K. Lee, A. R. C. Claridades and J. Lee, "Improving a Street-Based Geocoding Algorithm Using Machine Learning Techniques," *Applied Sciences*, vol. ١٠, no. ١٦, ٢٠٢٠.
- [٥] L. Xu, Z. Du, R. Mao, F. Zhang and R. Liu, "GSAM: A deep neural network model for extracting computational representations of Chinese addresses fused with geospatial feature," *Computers, Environment and Urban Systems*, vol. ٨١, p. ١٠١٤٧٣, ٢٠٢٠.
- [٦] Z. Yin, D. Li and D. W. Goldberg, "Is ChatGPT a game changer for geocoding: a benchmark for geocoding address parsing techniques," in *GeoSearch '23: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data*, Hamburg, Germany, ٢٠٢٣.
- [٧] Y. Lin, M. Kang, Y. Wu, Q. Du and T. Liu, "A deep learning architecture for semantic address matching," *International Journal of Geographical Information Science*, vol. ٣٤, no. ٣, p. ٥٥٩-٥٧٦, ٢٠١٩.
- [٨] A. Javidaneh, F. Karimipour and N. Alinaghi, "How Much Do We Learn from Addresses? On the Syntax, Semantics and Pragmatics of Addressing Systems," *ISPRS International Journal of Geo-Information*, vol. ٩, no. ٥, p. ٣١٧, ٢٠٢٠.
- [٩] D. Jurafsky and J. H. Martin, "Hidden Markov Models," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, ٣rd ed., ٢٠٢٥, pp. ١-١٥.
- [١٠] O. contributors, "Planet dump retrieved from <https://planet.osm.org>," ٢٠٢٤. [Online]. Available: <https://www.openstreetmap.org>.

....

**Abstract:**

With the rapid expansion of location-based services in fields such as urban management, smart transportation, and emergency response, converting Persian postal addresses to geographic coordinates has become critically important. In Iran, postal addresses present a significant challenge for geocoding due to their complex and colloquial structure. This study aims to develop an independent and efficient tool for converting these addresses into geographic coordinates, thereby reducing reliance on external data sources and enabling application in smaller cities and regions. The proposed tool is based on a Hidden Markov Model (HMM) and utilizes OpenStreetMap data as its mapping reference. Experimental results indicate that the tool successfully geolocates 94.7% of addresses with an error of less than 0.5 kilometers, which is considered an acceptable performance compared to similar tools. Its modular design and simplicity make it a suitable option for low-data environments and applications such as customer service, geographic analysis, and crisis management. Despite some limitations in processing complex addresses, the proposed tool shows high potential for further development and improvement in future research.

**Keywords:**

Geocoding, Postal Address, Hidden Markov Model



University of Tehran



College of Engineering

School of Electrical and Computer Engineering

## **Development of a Geocoding Tool for Postal Addresses Using a Hidden Markov Model**

A thesis submitted to the Undergraduate Studies Office

In partial fulfillment of the requirements for

The degree of Bachelor of science in

Computer Engineering

**By:**

**MohammadReza Bakhshayesh**

**Supervisor:**

**Dr. Ahmad Kalhor**