دانشکده مهندسی کامپیوتر

بررسی الگوریتم‌های یادگیری ماشین برای جریان‌های متنی

گزارش سمینار کارشناسی ارشد در رشته مهندسی کامپیوتر
گرایش هوش مصنوعی و رباتیک

وحید خرازی

استاد راهنما

دکتر بهروز مینایی بیدگلی

آبان ۹۵

بسم الله الرحمن الرحیم

تقدیم به:

پدر و مادر عزیزم که پیغمبر نگاهشان همیشه مرا چشم و چراغ خواهد بود.

قدردانی

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را زیور عقل آراست.

در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای دکتر مینایی، صمیمانه تشکر و قدردانی کنم که قطعاً بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید.

وحید خرازی

آبان ۹۵

چکیده

این پایان‌نامه، به بحث در مورد نوشتن پروژه، پایان‌نامه و رساله با استفاده از کلاس IUST-Thesis می‌پردازد. حروف‌چینی پروژه کارشناسی، پایان‌نامه یا رساله یکی از موارد پرکاربرد استفاده از زی‌پرشین است. زی‌پرشین بسته‌ای است که به همت آقای وفا خلیقی آماده شده است و امکان حروف‌چینی فارسی در LaTeX 2ε را برای فارسی‌زبانان فراهم کرده است. از جمله مزایای لاتک آن است که در صورت وجود یک کلاس آماده برای حروف‌چینی یک سند خاص مانند یک پایان‌نامه، کاربر بدون درگیری با جزییات حروف‌چینی و صفحه‌آرایی می‌تواند سند خود را آماده نماید.

شاید با قالب‌های لاتکی که برخی از مجلات برای مقالات خود عرضه می‌کنند مواجه شده باشید. اگر نظیر این کار در دانشگاههای مختلف برای اسناد متنوع آنها مانند پایان‌نامه‌ها آماده شود، دانشجویان به جای وقت گذاشتن روی صفحه‌آرایی مطالب خود، روی محتوای متن خود تمرکز خواهند نمود. به علاوه با آشنایی با لاتک خواهند توانست از امکانات بسیار این نرم‌افزار جهت نمایش بهتر دست‌آوردهای خود استفاده کنند. به همین خاطر، یک کلاس با نام IUST-Thesis برای حروف‌چینی پروژه‌ها، پایان‌نامه‌ها و رساله‌های دانشگاه علم و صنعت ایران با استفاده از نرم‌افزار زی‌پرشین، آماده شده است. این فایل به گونه‌ای طراحی شده است که کلیات خواسته‌های مورد نیاز مدیریت تحصیلات تکمیلی دانشگاه علم و صنعت ایران را برآورده می‌کند و نیز، حروف‌چینی بسیاری از قسمت‌های آن، به طور خودکار انجام می‌شود.

واژگان کلیدی: یادگیری ماشین، داده های جریانی، متن کاوی، پردازش متن

فهرست مطالب

فهرست تصاویر چ

فهرست جداول ح

فهرست الگوریتم‌ها خ

فهرست علائم اختصاری د

پیوست آ: مدیریت مراجع در لاتک ذ

آ ـ ۱ مدیریت مراجع با BibTEX . . . . . ذ

آ ـ ۱ ـ ۱ سبک‌های فعلی قابل استفاده در زی‌پرشین . . . . . ر

آ ـ ۱ ـ ۲ نحوه استفاده از سبک‌های فارسی . . . . . ز

پیوست ب: جدول، نمودار و الگوریتم در لاتک ش

ب ـ ۱ مدلهای حرکت دوبعدی . . . . . ش

ب ـ ۲ ماتریس . . . . . ش

ب ـ ۳ الگوریتم با دستورات فارسی . . . . . ص

ب ـ ۴ الگوریتم با دستورات لاتین . . . . . ض

ب ـ ۵ نمودار . . . . . ض

ب ـ ۶ تصویر . . . . . ط

واژه‌نامه فارسی به انگلیسی ظ

واژه‌نامه انگلیسی به فارسی ع

فهرست تصاویر

آـ۱ نمونه خروجی با سبک asa-fa . . . . . س

ب‌ـ۱ دو شیر . . . . . ط

فهرست جداول

ب ـ ۱ مدلهای تبدیل. . . . . ص

فهرست الگوریتم‌ها

ب–۱ الگوریتم DLT برای تخمین ماتریس هوموگرافی. . . . . ض
ب–۲ الگوریتم RANSAC برای تخمین ماتریس هوموگرافی. . . . . ض

فهرست علائم اختصاری

شتاب گرانش ..... a (m/s^2)

نیرو ..... F (N)

پیوست آ

مدیریت مراجع در لاتک

در بخش ؟؟ اشاره شد که با دستور \bibitem می‌توان یک مرجع را تعریف نمود و با فرمان \cite به آن ارجاع داد. این روش برای تعداد مراجع زیاد و تغییرات آنها مناسب نیست. در ادامه به صورت مختصر توضیحی در خصوص برنامه BibTeX که همراه با توزیع‌های معروف تِک عرضه می‌شود و نحوه استفاده از آن در زی‌پرشین خواهیم داشت.

آ ـ ۱ مدیریت مراجع با BibTeX

یکی از روش‌های قدرتمند و انعطاف‌پذیر برای نوشتن مراجع مقالات و مدیریت مراجع در لاتک، استفاده از BibTeX است. روش کار با BibTeX به این صورت است که مجموعه‌ی همه‌ی مراجعی را که در پروژه/پایان‌نامه/رساله استفاده کرده یا خواهیم کرد، در پرونده‌ی جداگانه‌ای نوشته و به آن فایل در سند خودمان به صورت مناسب لینک می‌دهیم. کنفرانس‌ها یا مجله‌های گوناگون برای نوشتن مراجع، قالب‌ها یا قراردادهای متفاوتی دارند که به آنها استیلهای مراجع گفته می‌شود. در این حالت به کمک استیل‌های BibTeX خواهید توانست تنها با تغییر یک پارامتر در پرونده‌ی ورودی خود، مراجع را مطابق قالب موردنظر تنظیم کنید. بیشتر مجلات و کنفرانس‌های معتبر یک پرونده‌ی سبک (BibTeX Style) با پسوند bst در وب‌گاه خود می‌گذارند که برای همین منظور طراحی شده است.

به جز نوشتن مقالات این سبک‌ها کمک بسیار خوبی برای تهیه‌ی مستندات علمی همچون پایان‌نامه‌هاست

که فرد می‌تواند هر قسمت از کارش را که نوشت مراجع مربوطه را به بانک مراجع خود اضافه نماید. با داشتن چنین بانکی از مراجع، وی خواهد توانست به راحتی یک یا چند ارجاع به مراجع و یا یک یا چند بخش را حذف یا اضافه نماید؛ مراجع به صورت خودکار مرتب شده و فقط مراجع ارجاع داده شده در قسمت کتاب‌نامه خواهندآمد. قالب مراجع به صورت یکدست مطابق سبک داده شده بوده و نیازی نیست که کاربر درگیر قالب‌دهی به مراجع باشد. در این جا مجموعه سبک‌های بسته Persian-bib که برای زی‌پرشین آماده شده‌اند به صورت مختصر معرفی شده و روش کار با آن‌ها گفته می‌شود. برای اطلاع بیشتر به راهنمای بسته‌ی Persian-bib مراجعه فرمایید.

آـ ۱ ـ ۱ سبک‌های فعلی قابل استفاده در زی‌پرشین

در حال حاضر فایلهای سبک زیر برای استفاده در زی‌پرشین آماده شده‌اند:

unsrt-fa.bst این سبک متناظر با unsrt.bst می‌باشد. مراجع به ترتیب ارجاع در متن ظاهر می‌شوند.

plain-fa.bst این سبک متناظر با plain.bst می‌باشد. مراجع بر اساس نام‌خانوادگی نویسندگان، به ترتیب صعودی مرتب می‌شوند. همچنین ابتدا مراجع فارسی و سپس مراجع انگلیسی خواهند آمد.

acm-fa.bst این سبک متناظر با acm.bst می‌باشد. شبیه plain-fa.bst است. قالب مراجع کمی متفاوت است. اسامی نویسندگان انگلیسی با حروف بزرگ انگلیسی نمایش داده می‌شوند. (مراجع مرتب می‌شوند)

ieeetr-fa.bst این سبک متناظر با ieeetr.bst می‌باشد. (مراجع مرتب نمی‌شوند)

plainnat-fa.bst این سبک متناظر با plainnat.bst می‌باشد. نیاز به بستهٔ natbib دارد. (مراجع مرتب می‌شوند)

chicago-fa.bst این سبک متناظر با chicago.bst می‌باشد. نیاز به بستهٔ natbib دارد. (مراجع مرتب می‌شوند)

asa-fa.bst این سبک متناظر با asa.bst می‌باشد. نیاز به بستهٔ natbib دارد. (مراجع مرتب می‌شوند)

با استفاده از استیلهای فوق می‌توانید به انواع مختلفی از مراجع فارسی و لاتین ارجاع دهید. به عنوان نمونه مرجع [؟] یک نمونه پروژه دکترا (به فارسی) و مرجع [؟] یک نمونه مقاله مجله فارسی است. مرجع [؟] یک نمونه مقاله کنفرانس فارسی و مرجع [؟] یک نمونه کتاب فارسی با ذکر مترجمان و ویراستاران فارسی است. مرجع [؟] یک نمونه پروژه کارشناسی ارشد انگلیسی و [؟] هم یک نمونه متفرقه می‌باشند.

مراجع [؟، ؟] نمونه کتاب و مقاله انگلیسی هستند. استیل مورد استفاده در این پروژه/پایان‌نامه/رساله acm-fa است که خروجی آنرا در بخش مراجع می‌توانید مشاهده کنید. نمونه خروجی سبک asa-fa در شکل آـ ۱ آمده است.

آـ ۱ ـ ۲ نحوه استفاده از سبک‌های فارسی

برای استفاده از بیب‌تک باید مراجع خود را در یک فایل با پسوند bib ذخیره نمایید. یک فایل bib در واقع یک پایگاه داده از مراجع[۱] شماست که هر مرجع در آن به عنوان یک رکورد از این پایگاه داده با قالبی خاص ذخیره می‌شود. به هر رکورد یک مدخل[۲] گفته می‌شود. یک نمونه مدخل برای معرفی کتاب Digital Image Processing در ادامه آمده است:

```
@BOOK{Gonzalez02image,
  AUTHOR =      {Rafael Gonzalez and Richard Woods},
  TITLE =       {Digital Image Processing},
  PUBLISHER =   {Prentice-Hall, Inc.},
  YEAR =        {2006},
  EDITION =     {3rd},
  ADDRESS =     {Upper Saddle River, NJ, USA}
}
```

در مثال فوق، @BOOK مشخصه‌ی شروع یک مدخل مربوط به یک کتاب و Gonzalez02book برچسبی است که به این مرجع منتسب شده است. این برچسب بایستی یکتا باشد. برای آنکه فرد به راحتی بتواند برچسب مراجع خود را به خاطر بسپارد و حتی‌الامکان برچسب‌ها متفاوت با هم باشند معمولاً از قوانین خاصی به این منظور استفاده می‌شود. یک قانون می‌تواند فامیل نویسنده‌ی اول+دورقم سال نشر+اولین کلمه‌ی عنوان اثر باشد. به AUTHOR و ... و ADDRESS فیلدهای این مدخل گفته می‌شود؛ که هر یک با مقادیر مربوط به مرجع مقدار گرفته‌اند. ترتیب فیلدها مهم نیست.

انواع متنوعی از مدخل‌ها برای اقسام مختلف مراجع همچون کتاب، مقاله‌ی کنفرانس و مقاله‌ی ژورنال وجود دارد که برخی فیلدهای آنها با هم متفاوت است. نام فیلدها بیانگر نوع اطلاعات آن می‌باشد. مثالهای ذکر شده در فایل MyReferences.bib کمک خوبی به شما خواهد بود. با استفاده از سبک‌های فارسی آماده

[۱] Bibliography Database

[۲] Entry

شده، محتویات هر فیلد می‌تواند به فارسی نوشته شود، ترتیب مراجع و نحوه‌ی چینش فیلدهای هر مرجع را سبک مورد استفاده مشخص خواهد کرد.

نکته: بدون اعمال تنظیمات موردنیاز BibTEX در TeXWorks، مراجع فارسی در استیل‌هایی که مراجع را به صورت مرتب شده چاپ می‌کنند، ترتیب کاملاً درستی نخواهند داشت. برای توضیحات بیشتر [؟] را ببینید یا به سایت پارسی‌لاتک مراجعه فرمایید. تنظیمات موردنیاز در TeXMaker اصلاح شده اعمال شده‌اند.

برای درج مراجع خود لازم نیست نگران موارد فوق باشید. در فایل MyReferences.bib که همراه با این پروژه/پایان‌نامه/رساله هست، موارد مختلفی درج شده است و کافیست مراجع خود را جایگزین موارد مندرج در آن نمایید.

پس از قرار دادن مراجع خود، یک بار XeLaTeX را روی سند خود اجرا نمایید، سپس bibtex و پس از آن دوبار XeLaTeX را. در TeXMaker کلید F11 و در TeXWorks هم گزینه‌ی BibTeX از منوی Typeset، BibTeX را روی سند شما اجرا می‌کنند.

برای بسیاری از مقالات لاتین حتی لازم نیست که مدخل مربوط به آنرا خودتان بنویسید. با جستجوی نام مقاله + کلمه bibtex در اینترنت سایتهای بسیاری همچون ACM و ScienceDirect را خواهید یافت که مدخل bibtex مربوط به مقاله شما را دارند و کافیست آنرا به انتهای فایل MyReferences اضافه کنید.

از هر یک از سبکهای Persian-bib می‌توانید استفاده کنید، البته اگر از سه استیل آخر استفاده می‌کنید و مایلید که مراجع شما شماره بخورند باید بسته natbib را با گزینه numbers فراخوانی نمایید.

نمونه خروجی با استیل فارسی asa-fa برای BibTeX در زی‌پرشین

محمود امین‌طوسی

مرجع امیدعلی (۱۳۸۲) یک نمونه پروژه دکترا و مرجع واحدی (۱۳۸۷) یک نمونه مقاله مجله فارسی است. مرجع امین‌طوسی و دیگران (۱۳۸۷) یک نمونه مقاله کنفرانس فارسی و مرجع استالینگ (۱۳۸۰) یک نمونه کتاب فارسی با ذکر مترجمان و ویراستاران فارسی است. مرجع خلیقی (۲۰۰۷) یک نمونه پروژه کارشناسی ارشد انگلیسی و خلیقی (۱۳۸۷) هم یک نمونه متفرقه می‌باشند.

مرجع گنزالس و وودس (۲۰۰۶) یک نمونه کتاب لاتین است که از آنجا که دارای فیلد authorfa است، نام نویسندگان آن در استیلهای asa-fa، plainnat-fa و chicago-fa به فارسی دیده می‌شود. مرجع Kanade and Baker (۲۰۰۲) مقاله انگلیسی است که معادل فارسی نام نویسندگان آن ذکر نشده بوده است.

مراجع

استالینگ، ویلیام (۱۳۸۰)، اصول طراحی و ویژگیهای داخلی سیستم‌های عامل. ترجمه‌ی صدیقی مشکنانی، محسن و پدرام، حسین، (ویراستار)برنجکوب، محمود، اصفهان: نشر شیخ بهایی، ویرایش سوم.

امیدعلی، مهدی (۱۳۸۲)، "تابع هیلبرت،" پایان‌نامه دکترا، دانشکده ریاضی، دانشگاه امیرکبیر.

امین‌طوسی، محمود، مزینی، ناصر، و فتحی، محمود (۱۳۸۷)، "افزایش وضوح ناحیه‌ای،" در چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشگاه امیرکبیر، تهران، ایران، صفحات ۱۰۱–۱۰۸.

خلیقی، وفا (۱۳۸۷)، "زی‌پرشین (XePersian): بسته فارسی برای حروف‌چینی در LaTeX2e،" HTTP://BITBUCKET.ORG/VAFA/XEPERSIAN.

واحدی، مصطفی (۱۳۸۷)، "موضوعی جدید در هندسه محاسباتی،" مجله فارسی نمونه، ۱، ۲۲–۳۰.

Baker, S. and Kanade, T. (2002), "Limits on Super-Resolution and How to Break Them," *IEEE Trans. Pattern Anal. Mach. Intell.*, 24, 1167–1183.

Gonzalez, R. C. and Woods, R. E. (2006), *Digital Image Processing*, Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 3rd ed. .

Khalighi, V. (2007), "Category Theory," Master's thesis, Sydny Univ.

شکل آ ـ ۱: نمونه خروجی با سبک asa-fa

پیوست ب

جدول، نمودار و الگوریتم در لاتک

در این بخش نمونه مثالهایی از جدول، نمودار و الگوریتم در لاتک را خواهیم دید.

ب ـ ۱ مدلهای حرکت دوبعدی

بسیاری از اوقات حرکت بین دو تصویر از یک صحنه با یکی از مدلهای پارامتری ذکر شده در جدول (ب ـ ۱) قابل مدل نمودن می‌باشد.

ب ـ ۲ ماتریس

شناخته‌شده‌ترین روش تخمین ماتریس هوموگرافی الگوریتم تبدیل خطی مستقیم (DLT[۱]) است. فرض کنید چهار زوج نقطهٔ متناظر در دو تصویر در دست هستند، $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ و تبدیل با رابطهٔ $\mathbf{x}'_i = H\mathbf{x}_i$ نشان داده می‌شود که در آن:

$$\mathbf{x}'_i = (x'_i, y'_i, w'_i)^\top$$

[۱] Direct Linear Transform

جدول ب-۱: مدلهای تبدیل.

نام مدل	درجه آزادی	تبدیل مختصات	توضیح
انتقالی	۲	$x' = x + t_x$ $y' = y + t_y$	انتقال دوبعدی
اقلیدسی	۳	$x' = xcos\theta - ysin\theta + t_x$ $y' = xsin\theta + ycos\theta + t_y$	انتقالی+دوران
مشابهت	۴	$x' = sxcos\theta - sysin\theta + t_x$ $y' = sxsin\theta + sycos\theta + t_y$	اقلیدسی+تغییرمقیاس
آفین	۶	$x' = a_{۱۱}x + a_{۱۲}y + t_x$ $y' = a_{۲۱}x + a_{۲۲}y + t_y$	مشابهت+اریب‌شدگی
پروجکتیو	۸	$x' = (m_۱x + m_۲y + m_۳)/D$ $y' = (m_۴x + m_۵y + m_۶)/D$ $D = m_۷x + m_۸y + ۱$	آفین+keystone+chirping
شارنوری	∞	$x' = x + v_x(x, y)$ $y' = y + v_y(x, y)$	حرکت آزاد

و

$$H = \begin{bmatrix} h_۱ & h_۲ & h_۳ \\ h_۴ & h_۵ & h_۶ \\ h_۷ & h_۸ & h_۹ \end{bmatrix}$$

رابطه زیر را برای الگوریتم (ب-۱) لازم دارم.

$$\begin{bmatrix} ۰^\top & -w'_i\mathbf{x}_i^\top & y'_i\mathbf{x}_i^\top \\ w'_i\mathbf{x}_i & ۰^\top & -x'_i\mathbf{x}_i^\top \\ -y'_i\mathbf{x}_i^\top & x'_i\mathbf{x}_i^\top & ۰^\top \end{bmatrix} \begin{pmatrix} \mathbf{h}^۱ \\ \mathbf{h}^۲ \\ \mathbf{h}^۳ \end{pmatrix} = ۰ \qquad (ب-۱)$$

ب-۳ الگوریتم با دستورات فارسی

با مفروضات فوق، الگوریتم DLT به صورت نشان داده شده در الگوریتم (ب-۱) خواهد بود.

الگوریتم ب-۱ الگوریتم DLT برای تخمین ماتریس هوموگرافی.

ورودی: $n \geq ۴$ زوج نقطهٔ متناظر در دو تصویر $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$،

خروجی: ماتریس هوموگرافی H به نحوی‌که: $\mathbf{x}'_i = H\mathbf{x}_i$.

۱: برای هر زوج نقطهٔ متناظر $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ ماتریس $\mathbf{A}_i$ را با استفاده از رابطهٔ ب-۱ محاسبه کنید.

۲: ماتریس‌های ۹ ستونی $\mathbf{A}_i$ را در قالب یک ماتریس $\mathbf{A}$ ۹ ستونی ترکیب کنید.

۳: تجزیهٔ مقادیر منفرد (SVD) ماتریس $\mathbf{A}$ را بدست آورید. بردار واحد متناظر با کمترین مقدار منفرد جواب $\mathbf{h}$ خواهد بود.

۴: ماتریس هوموگرافی H با تغییر شکل $\mathbf{h}$ حاصل خواهد شد.

الگوریتم ب-۲ الگوریتم RANSAC برای تخمین ماتریس هوموگرافی.

Require: $n \geq 4$ putative correspondences, number of estimations, N, distance threshold T_{dist}.

Ensure: Set of inliers and Homography matrix H.

```
for k = 1 to N do
    Randomly choose 4 correspondence,
    Check whether these points are colinear, if so, redo the above step
    Compute the homography H_curr by DLT algorithm from the 4 points pairs,
    ...
end for
Refinement: re-estimate H from all the inliers using the DLT algorithm.
```

ب-۴ الگوریتم با دستورات لاتین

الگوریتم ب-۲ یک الگوریتم با دستورات لاتین است.

ب-۵ نمودار

لاتک بسته‌هایی با قابلیت‌های زیاد برای رسم انواع مختلف نمودارها دارد. مانند بسته‌های Tikz و PSTricks. توضیح اینها فراتر از این پیوست کوچک است. مثالهایی از رسم نمودار را در مجموعه پارسی‌لاتک خواهید یافت. توصیه می‌کنم که حتماً مثالهایی از برخی از آنها را ببینید. راهنمای همه آنها در تک‌لایو هست. نمونه مثالهایی از بسته Tikz را می‌توانید در http://www.texample.net/tikz/examples/ ببینید.

(آ) شیر ۱ (ب) شیر ۲

شکل ب_۱: دو شیر

ب_۶ تصویر

نمونه تصاویری در بخش قبل دیدیم. دو تصویر شیر کنار هم را هم در شکل ب_۱ مشاهده می‌کنید.

واژه‌نامه فارسی به انگلیسی

احتمالی ..... Probabilistic
ارزیابی ..... Valuation
اندازه ..... Measure
پایدار ..... Stably
توپولوژی ضعیف ..... Weak Topology
دامنه‌توانی ..... Powerdomain
فضای تابع ..... Function Space
دامنه معنایی ..... Semantic Domain
قطعه‌برنامه ..... Program Fragment
مجموعه جزئاً مرتب کامل جهت‌دار ..... Dcpo
مرتب ..... Ordered

واژه‌نامه انگلیسی به فارسی

Dcpo ..... مجموعه جزئاً مرتب کامل جهت‌دار
Function Space ..... فضای تابع
Measure ..... اندازه
Ordered ..... مرتب
Powerdomain ..... دامنه‌توانی
Probabilistic ..... احتمالی
Program Fragment ..... قطعه‌برنامه
Semantic Domain ..... دامنه معنایی
Stably ..... پایدار
Valuation ..... ارزیابی
Weak Topology ..... توپولوژی ضعیف

Abstract:

-abstract

Keywords: Machine Learning, Data Stream, Text Mining, Text Proccessing

Iran University of Science and Technology
Computer Engineering Department

A Survey on Learning Algorithms for Text Streams

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Engineering

By:

Vahid Kharazi

Supervisor:

Dr. Behroz Minaei Bidgoli

Octobr 2016