دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی
مهندسی کامپیوتر

تشخیص اهمیت اخبار فارسی با استفاده از مدل‌های زبانی بزرگ

نگارش

شایان صالحی

استاد راهنما

دکتر مهدی جعفری

بهمن ۱۴۰۴

سپاس

از استاد بزرگوارم، دکتر جعفری به خاطر زحمات و راهنمایی‌هایی که در طول این پروژه داشته‌اند متشکرم و همچنین از دانشجوی دکترا ایشان، آقای معین سلیمی به خاطر زمان و راهنمایی‌هایی که برای پیش‌بردن پروژه انجام داده‌اند قدردانم.

چکیده

این پروژه به بررسی قدرت تشخیص اهمیت یک خبر فارسی توسط مدل‌های زبانی بزرگ پرداخته و قدرت یادگیری از محتوا، قدرت استدلال و قدرت تفکر آن را ارزیابی کرده است. در ابتدا، از دادگان علائم‌گذاری‌شده توسط افراد در حوزه‌های مختلف از جمله ورزشی، سیاسی، اجتماعی، پزشکی و فرهنگی استفاده و محیطی برای ارزیابی مدل‌های زبانی بزرگ توسعه داده شده است. در این محیط مدل‌های مختلف موجود بررسی و ارزیابی شده و در نهایت با تمام حالات مختلف و شرایط مخلتف، قدرت تحلیل آنها در زبان فارسی و انگلیسی بررسی شده است. این پروژه نشان‌داده که دستورهای[۱] شامل زنجیره تفکر[۲] و درخت تفکر[۳] باعث بهبود کارایی مدل‌ها و همچنین روش نتظیم نمادها[۴] باعث حساسیت بسیار زیاد به پرسش داده شده و محتوای آن می‌شود.

کلیدواژه‌ها: مدل‌های زبانی بزرگ، پردازش زبان‌های طبیعی، یادگیری ماشین، تشخیص اهمیت اخبار

---

[۱] Prompt
[۲] Chain-of-Thoughts
[۳] Tree-of-Thoughts
[۴] Symbol Tuning

فهرست مطالب

فهرست جداول

فهرست تصاویر

فصل ۱

مقدمه

نخستین فصل یک پایان‌نامه به معرفی مسئله، بیان اهمیت موضوع، ادبیات موضوع، اهداف پژوهش و معرفی ساختار پایان‌نامه می‌پردازد. در این فصل نمونه‌ی مختصری از مقدمه آورده شده است.

۱-۱ تعریف مسئله

نگارش یک پایان‌نامه علاوه بر بخش‌های پژوهش و آماده‌سازی محتوا، مستلزم رعایت نکات دقیق فنی و نگارشی است که در تهیه‌ی یک پایان‌نامه‌ی موفق بسیار کلیدی و مؤثر است. از آن جایی که بسیاری از نکات فنی مانند قالب کلی صفحات، شکل و اندازه‌ی قلم، صفحات عنوان و غیره در تهیه‌ی پایان‌نامه‌ها یکسان است، می‌توان با ارائه‌ی یک قالب حروف‌چینی استاندارد نگارش پایان‌نامه‌ها را تا حد بسیار زیادی بهبود بخشید.

۱-۲ اهمیت موضوع

وجود قالب استاندارد برای نگارش پایان‌نامه از جهات مختلف حائز اهمیت است، از جمله:

- ایجاد یک‌نواختی در قالب کلی صفحات و شکل و اندازه‌ی قلم‌ها
- تسهیل نگارش پایان‌نامه با در اختیار گذاشتن یک قالب اولیه
- تولید خودکار صفحات دارای بخش‌های تکراری نظیر صفحات ابتدایی و انتهایی پایان‌نامه

- پیش‌گیری از برخی خطاهای مرسوم در نگارش پایان‌نامه

۱-۳ ادبیات موضوع

اکثر دانشگاه‌ها قالب استانداردی برای تهیه‌ی پایان‌نامه‌ها در اختیار دانشجویان خود قرار می‌دهند. این قالب‌ها عموما مبتنی بر نرم‌افزارهای متداول حروف‌چینی نظیر لاتک و مایکروسافت ورد هستند.

لاتک[1] یک نرم‌افزار متن‌باز قوی برای حروف‌چینی متون علمی است [۱، ۲]. در این نوشتار از نرم‌افزار حروف‌چینی زی‌تک[2] و افزونه‌ی زی‌پرشین[3] استفاده شده است.

۱-۴ اهداف پژوهش

کتابخانه‌ی مرکزی دانشگاه صنعتی شریف دستورالعمل جامعی در خصوص نحوه‌ی تهیه‌ی پایان‌نامه‌های کارشناسی ارشد و رساله‌های دکتری ارائه کرده است. در این نوشتار سعی شده است قالب استانداردی برای تهیه‌ی پایان‌نامه‌ها مبتنی بر نرم‌افزار لاتک و بر اساس دستورالعمل مذکور ارائه شده و نحوه‌ی استفاده از قالب به طور مختصر توضیح داده شود. این قالب می‌تواند برای تهیه‌ی پایان‌نامه‌های کارشناسی و کارشناسی ارشد و همچنین رساله‌های دکتری مورد استفاده قرار گیرد.

۱-۵ ساختار پایان‌نامه

این پایان‌نامه در پنج فصل به شرح زیر ارائه می‌شود. مفاهیم اولیه مورد استفاده در این پایان‌نامه در فصل دوم به اختصار اشاره شده است. فصل سوم به مطالعه و بررسی کارهای پیشین مرتبط با موضوع این پایان‌نامه می‌پردازد. در فصل چهارم، نتایج جدیدی که در این پایان‌نامه به‌دست آمده است، ارائه می‌شود. فصل پنجم به جمع‌بندی کارهای انجام شده در این پژوهش و ارائه‌ی پیشنهادهایی برای انجام کارهای آتی خواهد پرداخت.

---

[1] LaTeX
[2] XeLaTeX
[3] XePersian

فصل ۲

مفاهیم اولیه

دومین فصل پایان‌نامه به طور معمول به معرفی مفاهیمی می‌پردازد که در پایان‌نامه مورد استفاده قرار می‌گیرند. در این فصل به عنوان یک نمونه، نکات کلی در خصوص نحوه‌ی نگارش پایان‌نامه و نیز برخی نکات نگارشی به اختصار توضیح داده می‌شوند.

۲-۱ نحوه‌ی نگارش

۲-۱-۱ پرونده‌ها

پرونده‌ی اصلی پایان‌نامه در قالب استاندارد[۱] `thesis.tex` نام دارد. به ازای هر فصل از پایان‌نامه، یک پرونده در شاخه‌ی `chapters` ایجاد نموده و نام آن را در `thesis.tex` (در قسمت فصل‌ها) درج نمایید. برای مشاهده‌ی خروجی، پرونده‌ی `thesis.tex` را با زی‌لاتک کامپایل کنید. مشخصات اصلی پایان‌نامه را می‌توانید در پرونده‌ی `front/info.tex` ویرایش کنید.

۲-۱-۲ عبارات ریاضی

برای درج عبارات ریاضی در داخل متن از $...$ و برای درج عبارات ریاضی در یک خط مجزا از $$...$$ یا محیط equation استفاده کنید. برای مثال عبارت $۲x + ۳y$ در داخل متن و عبارت زیر

$$\sum_{k=۰}^{n} \binom{n}{k} = ۲^n \qquad (۱-۲)$$

[۱] قالب استاندارد پایان‌نامه از نشانی github.com/zarrabi/thesis-template قابل دریافت است.

در یک خط مجزا درج شده است. دقت کنید که تمامی عبارات ریاضی، از جمله متغیرهای تک‌حرفی مانند x و y باید در محیط ریاضی یعنی محصور بین دو علامت $ باشند.

۲-۱-۳ علائم ریاضی پرکاربرد

برخی علائم ریاضی پرکاربرد در زیر فهرست شده‌اند. برای مشاهده‌ی دستور معادل پرونده‌ی منبع را ببینید.

- مجموعه‌های اعداد: $\mathbb{N}, \mathbb{Z}, \mathbb{Z}^+, \mathbb{Q}, \mathbb{R}, \mathbb{C}$
- مجموعه: $\{۱, ۲, ۳\}$
- دنباله: $\langle ۱, ۲, ۳ \rangle$
- سقف و کف: $\lceil x \rceil, \lfloor x \rfloor$
- اندازه و متمم: $|A|, \overline{A}$
- همنهشتی: $a \overset{n}{\equiv} ۱$ یا (پیمانه‌ی n) $a \equiv ۱$
- ضرب و تقسیم: $\times, \cdot, \div$
- سه‌نقطه: $۱, ۲, \ldots, n$
- کسر و ترکیب: $\frac{n}{k}, \binom{n}{k}$
- اجتماع و اشتراک: $A \cup (B \cap C)$
- عملگرهای منطقی: $\neg p \vee (q \wedge r)$
- پیکان‌ها: $\to, \Rightarrow, \leftarrow, \Leftarrow, \leftrightarrow, \Leftrightarrow$
- عملگرهای مقایسه‌ای: $\neq, \leqslant, \nleqslant, \geqslant, \ngeqslant$
- عملگرهای مجموعه‌ای: $\in, \notin, \setminus, \subset, \subseteq, \subsetneq, \supset, \supseteq, \supsetneq$
- جمع و ضرب چندتایی: $\sum_{i=۱}^{n} a_i, \prod_{i=۱}^{n} a_i$
- اجتماع و اشتراک چندتایی: $\bigcup_{i=۱}^{n} A_i, \bigcap_{i=۱}^{n} A_i$
- برخی نمادها: $\infty, \emptyset, \forall, \exists, \triangle, \angle, \ell, \equiv, \therefore$

۲-۱-۴ لیست‌ها

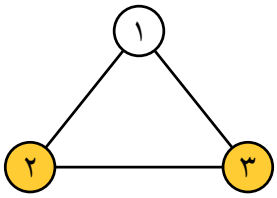
برای ایجاد یک لیست می‌توانید از محیط‌های «فقرات» و «شمارش» همانند زیر استفاده کنید.

- مورد اول
- مورد دوم
- مورد سوم

۱. مورد اول
۲. مورد دوم
۳. مورد سوم

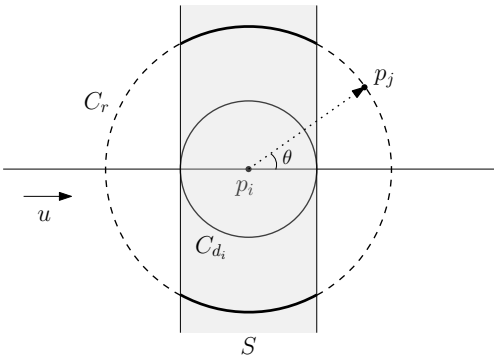
۲-۱-۵ درج شکل

یکی از روش‌های مناسب برای ایجاد شکل استفاده از نرم‌افزار LaTeX Draw و سپس درج خروجی آن به صورت یک فایل tex درون متن با استفاده از دستور fig یا centerfig است. شکل ۲-۱ نمونه‌ای از اشکال ایجادشده با این ابزار را نشان می‌دهد.



شکل ۲-۱: یک گراف و پوشش رأسی آن

همچنین می‌توانید با استفاده از نرم‌افزار Ipe شکل‌های خود را مستقیما به صورت pdf ایجاد نموده و آن‌ها را با دستورات img یا centerimg درون متن درج کنید. برای نمونه، شکل ۲-۲ را ببینید.



شکل ۲-۲: نمونه شکل ایجادشده توسط نرم‌افزار Ipe

۲-۱-۶ درج جدول

برای درج جدول می‌توانید با استفاده از دستور «جدول» جدول را ایجاد کرده و سپس با دستور «لوح» آن را درون متن درج کنید. برای نمونه جدول ۲-۱ را ببینید.

جدول ۲-۱: عملگرهای مقایسه‌ای

عملگر	عنوان
<	کوچک‌تر
>	بزرگ‌تر
==	مساوی
<>	نامساوی

۲-۱-۷ درج الگوریتم

برای درج الگوریتم می‌توانید از محیط «الگوریتم» استفاده کنید. یک نمونه در الگوریتم ۱ آمده است.

الگوریتم ۱ پوشش رأسی حریصانه

ورودی: گراف $G = (V, E)$

خروجی: یک پوشش رأسی از G

۱: قرار بده $C = \emptyset$

۲: تا وقتی E تهی نیست:

۳: یال دل‌خواه $uv \in E$ را انتخاب کن

۴: رأس‌های u و v را به C اضافه کن

۵: تمام یال‌های واقع بر u یا v را از E حذف کن

۶: C را برگردان

۲-۱-۸ محیط‌های ویژه

برای درج مثال‌ها، قضیه‌ها، لم‌ها و نتیجه‌ها به ترتیب از محیط‌های «مثال»، «قضیه»، «لم» و «نتیجه» استفاده کنید. برای درج اثبات قضیه‌ها و لم‌ها از محیط «اثبات» استفاده کنید.

تعریف‌های داخل متن را با استفاده از دستور «مهم» به صورت **تیره** نشان دهید. تعریف‌های پایه‌ای‌تر را درون محیط «تعریف» قرار دهید.

تعریف ۲-۱ (اصل لانه‌کبوتری) *اگر $n+1$ کبوتر یا بیش‌تر درون n لانه قرار گیرند، آن‌گاه لانه‌ای وجود دارد که شامل حداقل دو کبوتر است.*

۲-۲ برخی نکات نگارشی

این فصل حاوی برخی نکات ابتدایی ولی بسیار مهم در نگارش متون فارسی است. نکات گردآوری‌شده در این فصل به هیچ وجه کامل نیست، ولی دربردارنده‌ی حداقل مواردی است که رعایت آن‌ها در نگارش پایان‌نامه ضروری به نظر می‌رسد.

۲-۲-۱ فاصله‌گذاری

۱. علائم سجاوندی مانند نقطه، ویرگول، دونقطه، نقطه‌ویرگول، علامت سؤال و علامت تعجب بدون فاصله از کلمه‌ی پیشین خود نوشته می‌شوند، ولی بعد از آن‌ها باید یک فاصله قرار گیرد. مانند: من، تو، او.

۲. علامت‌های پرانتز، آکولاد، کروشه، نقل قول و نظایر آن‌ها بدون فاصله با عبارات داخل خود نوشته می‌شوند، ولی با عبارات اطراف خود یک فاصله دارند. مانند: (این عبارت) یا {آن عبارت}.

۳. دو کلمه‌ی متوالی در یک جمله همواره با یک فاصله از هم جدا می‌شوند، ولی اجزای یک کلمه‌ی مرکب باید با نیم‌فاصله[۲] از هم جدا شوند. مانند: کتاب درس، محبت‌آمیز، دوبخشی.

۴. اجزای فعل‌های مرکب با فاصله از یک‌دیگر نوشته می‌شوند، مانند: تحریر کردن، به سر آمدن.

۲-۲-۲ شکل حروف

۱. در متون فارسی به جای حروف «ك» و «ي» عربی باید از حروف «ک» و «ی» فارسی استفاده شود. همچنین به جای اعداد عربی مانند ٥ و ٦ باید از اعداد فارسی مانند ۵ و ۶ استفاده نمود. برای این کار، توصیه می‌شود صفحه‌کلید فارسی استاندارد را روی سیستم خود فعال کنید.

[۲] «نیم‌فاصله» فاصله‌ای مجازی است که در عین جدا کردن اجزای یک کلمه‌ی مرکب از یک‌دیگر، آن‌ها را نزدیک به هم نگه می‌دارد. معمولاً برای تولید این نوع فاصله در صفحه‌کلیدهای استاندارد از ترکیب Shift+Space استفاده می‌شود.

۲. عبارات نقل‌قول‌شده یا مؤکد باید درون علامت نقل قولِ «» قرار گیرند، نه "". مانند: «کشور ایران».

۳. کسره‌ی اضافه‌ی بعد از «ه» غیرملفوظ به صورت «ه‌ی» یا «هٔ» نوشته می‌شود. مانند: خانه‌ی علی، دنباله‌ی فیبوناچی.

تبصره: اگر «ه» ملفوظ باشد، نیاز به «ی» ندارد. مانند: فرمانده دلیر، پادشه خوبان.

۴. پایه‌های همزه در کلمات، همیشه «ئ» است، مانند: مسئله و مسئول، مگر در مواردی که همزه ساکن است که در این صورت باید متناسب با اعراب حرف پیش از خود نوشته شود. مانند: رأس، مؤمن.

۲-۲-۳ جدانویسی

۱. علامت استمرار، «می»، توسط نیم‌فاصله از جزء بعدی فعل جدا می‌شود. مانند: می‌رود، می‌توانیم.

۲. شناسه‌های «ام»، «ای»، «ایم»، «اید» و «اند» توسط نیم‌فاصله، و شناسه‌ی «است» توسط فاصله از کلمه‌ی پیش از خود جدا می‌شوند. مانند: گفته‌ام، گفته‌ای، گفته است.

۳. علامت جمع «ها» توسط نیم‌فاصله از کلمه‌ی پیش از خود جدا می‌شود. مانند: این‌ها، کتاب‌ها.

۴. «به» همیشه جدا از کلمه‌ی بعد از خود نوشته می‌شود، مانند: به نام و به آن‌ها، مگر در مواردی که «بـ» صفت یا فعل ساخته است. مانند: بسزا، ببینم.

۵. «به» همواره با فاصله از کلمه‌ی بعد از خود نوشته می‌شود، مگر در مواردی که «به» جزئی از یک اسم یا صفت مرکب است. مانند: تناظر یک‌به‌یک، سفر به تاریخ.

۶. علامت صفت برتری، «تر»، و علامت صفت برترین، «ترین»، توسط نیم‌فاصله از کلمه‌ی پیش از خود جدا می‌شوند. مانند: سنگین‌تر، مهم‌ترین.

تبصره: کلمات «بهتر» و «بهترین» را می‌توان از این قاعده مستثنی نمود.

۷. پیشوندها و پسوندهای جامد، چسبیده به کلمه‌ی پیش یا پس از خود نوشته می‌شوند. مانند: همسر، دانشکده، دانشگاه.

تبصره: در مواردی که خواندن کلمه دچار اشکال می‌شود، می‌توان پسوند یا پیشوند را جدا کرد. مانند: هم‌میهن، هم‌ارزی.

۸. ضمیرهای متصل چسبیده به کلمه‌ی پیش از خود نوشته می‌شوند. مانند: کتابم، نامت، کلامشان.

فصل ۳

کارهای پیشین

در فصل سوم پایان‌نامه، کارهای پیشین انجام‌شده روی مسئله به تفصیل توضیح داده می‌شود. نمونه‌ای از فصل کارهای پیشین در زیر آمده است.[۱]

۳-۱ مسائل خوشه‌بندی

مسئله‌ی **خوشه‌بندی**[۲] یکی از مهم‌ترین مسائل در زمینه‌ی داده‌کاوی به حساب می‌آید. در این مسئله، هدف دسته‌بندی تعدادی شیء به‌گونه‌ای است که اشیاء درون یک دسته (خوشه)، نسبت به یکدیگر در برابر دسته‌های دیگر شبیه‌تر باشند (معیارهای متفاوتی برای تشابه تعریف می‌گردد). این مسئله در حوزه‌های مختلفی از علوم کامپیوتر از جمله داده‌کاوی، جست‌وجوی الگو[۳]، پردازش تصویر[۴]، بازیابی اطلاعات[۵] و رایانش زیستی[۶] مورد استفاده قرار می‌گیرد [۳].

تا کنون راه‌حل‌های زیادی برای این مسئله ارائه شده است که از لحاظ معیار تشخیص خوشه‌ها و نحوه‌ی انتخاب یک خوشه، با یک‌دیگر تفاوت بسیاری دارند. به همین خاطر مسئله‌ی خوشه‌بندی یک مسئله‌ی بهینه‌سازی چندهدفه[۷] محسوب می‌شود.

همان طور که در مرجع [۴] ذکر شده است، خوشه در خوشه‌بندی تعریف واحدی ندارد و یکی از

---

[۱] مطالب این فصل نمونه از پایان‌نامه‌ی آقای بهنام حاتمی گرفته شده است.
[۲] Clustering
[۳] Pattern recognition
[۴] Image analysis
[۵] Information retrieval
[۶] Bioinformatics
[۷] Multi-objective

دلایل وجود الگوریتم‌های متفاوت، همین تفاوت تعریف‌ها از خوشه است. بنابراین با توجه به مدلی که برای خوشه‌ها ارائه می‌شود، الگوریتم متفاوتی نیز ارائه می‌گردد. در ادامه به بررسی تعدادی از معروف‌ترین مدل‌های مطرح می‌پردازیم:

- **مدل‌های مرکزگرا:** در این مدل‌ها، هر دسته با یک مرکز نشان داده می‌شود. از جمله معروف‌ترین روش‌های خوشه‌بندی بر اساس این مدل، خوشه‌بندی kـ مرکز، خوشه‌بندی kـ میانگین[۸] و خوشه‌بندی kـ میانه[۹] است.

- **مدل‌های مبتی بر توزیع نقاط:** در این مدل، دسته‌ها با فرض پیروی از یک توزیع احتمالی مشخص می‌شوند. از جمله الگوریتم‌های معروف ارائه شده در این مدل، الگوریتم بیشینه‌سازی امید ریاضی[۱۰] است.

- **مدل‌های مبتنی بر تراکم نقاط:** در این مدل، خوشه‌ها متناسب با ناحیه‌های متراکم نقاط در مجموعه داده مورد استفاده قرار می‌گیرد.

- **مدل‌های مبتنی بر گراف:** در این مدل، هر خوشه به مجموعه از رئوس گفته می‌شود که تمام رئوس آن با یک‌دیگر همسایه باشند. از جمله الگوریتم‌های معروف این مدل، الگوریتم خوشه‌بندی HCS[۱۱] است.

الگوریتم‌های ارائه شده تنها از نظر نوع مدل با یک‌دیگر متفاوت نیستند. بلکه، می‌توان آن‌ها را از لحاظ نحوه‌ی تخصیص نقاط بین خوشه‌ها نیز تقسیم‌بندی کرد:

- **تخصیص قطعی داده‌ها:** در این نوع خوشه‌بندی هر داده دقیقاً به یک خوشه اختصاص داده می‌شود.

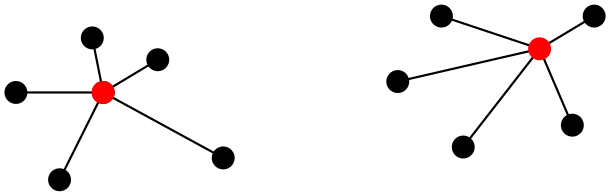
- **تخصیص قطعی داده‌ها با داده‌ی پرت:** در این نوع خوشه‌بندی ممکن است بعضی از داده‌ها به هیچ خوشه‌ای اختصاص نیابد، اما بقیه داده‌ها هر کدام دقیقاً به یک خوشه اختصاص می‌یابد.

- **تخصیص قطعی داده:** در این نوع خوشه‌بندی هر داده دقیقاً به یک خوشه اختصاص داده می‌شود.

- **خوشه‌بندی هم‌پوشان:** در این نوع خوشه‌بندی هر داده می‌تواند به چند خوشه اختصاص داده شود. در گونه‌ای از این مدل، می‌توان هر نقطه را با احتمالی به هر خوشه اختصاص می‌یابد. به این گونه از خوشه‌بندی، خوشه‌بندی نرم[۱۲] گفته می‌شود.

---

[۸] k-Means
[۹] k-Median
[۱۰] Expectation-maximization
[۱۱] Highly Connected Subgraphs
[۱۲] Soft clustering



شکل ۳-۱: نمونه‌ای از مسئله‌ی ۲ـ مرکز

- **خوشه‌بندی سلسه‌مراتبی**: در این نوع خوشه‌ها، داده‌ها به گونه‌ای به خوشه‌ها تخصیص داده می‌شود که دو خوشه یا اشتراک ندارند یا یکی به طور کامل دیگری را می‌پوشاند. در واقع در بین خوشه‌ها، رابطه‌ی پدر فرزندی برقرار است.

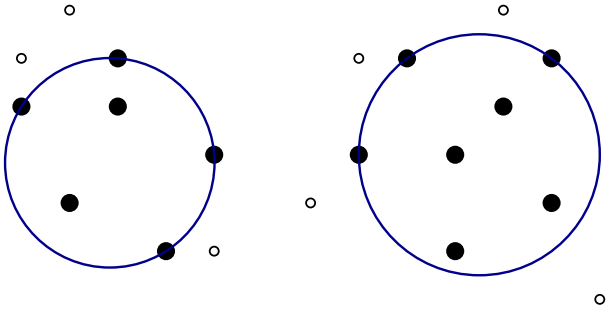
در بین دسته‌بندی‌های ذکر شده، تمرکز اصلی این پایان‌نامه بر روی مدل مرکزگرا و خوشه‌بندی قطعی با داده‌های پرت با مدل kـ مرکز است. همان‌طور که ذکر شد علاوه بر مسئله‌ی kـ مرکز که به تفصیل مورد بررسی قرار می‌گیرد، kـ میانه و kـ میانگین از جمله معروف‌ترین خوشه‌بندی‌های مدل مرکزگرا هستند. در خوشه‌بندی kـ میانه، هدف افراز نقاط به k خوشه است به گونه‌ای که مجموع مربع فاصله‌ی هر نقطه از میانه‌ی نقاط آن خوشه، کمینه گردد. در خوشه‌بندی kـ میانگین، هدف افراز نقاط به k خوشه است به گونه‌ای که مجموع فاصله‌ی هر نقطه از میانگین نقاط داخل خوشه (یا مرکز آن خوشه) کمینه گردد.

۳-۲ خوشه‌بندی kـ مرکز

یکی از رویکردهای شناخته‌شده برای مسئله‌ی خوشه‌بندی، مسئله‌ی ***k*ـ مرکز** است. در این مسئله هدف، پیدا کردن k نقطه به عنوان مرکز دسته‌ها است به‌طوری‌که شعاع دسته‌ها تا حد ممکن کمینه شود. مثالی از مسئله‌ی ۲ـ مرکز در شکل ۳-۱ نشان داده شده است. در این پژوهش، مسئله‌ی kـ مرکز با متریک‌های خاص و برای kهای کوچک مورد بررسی قرار گرفته است و هر کدام از تعریف رسمی مسئله‌ی kـ مرکز در زیر آمده است:

مسئله‌ی ۳-۱ **(kـ مرکز)** *گراف کامل بدون جهت $G = (V, E)$ با تابع فاصله‌ی d، که از نامساوی مثلثی پیروی می‌کند داده شده است. زیرمجموعه‌ی $S \subseteq V$ با اندازه‌ی k را به‌گونه‌ای انتخاب کنید که عبارت زیر را کمینه کند:*

$$\max_{v \in V} \{\min_{s \in S} d(v, s)\} \qquad (۱-۳)$$



شکل ۳-۲: نمونه‌ای ازمسئله‌ی ۲ـ مرکز با داده‌های پرت

گونه‌های مختلفی از مسئله‌ی kـ مرکز با محدودیت‌های متفاوت توسط پژوهشگران مورد مطالعه قرار گرفته است. از جمله‌ی این گونه‌ها، می‌توان به حالتی که در بین داده‌های ورودی، داده‌های پرت وجود دارد، اشاره کرد. در واقع در این مسئله، قبل از خوشه‌بندی می‌توانیم تعدادی از نقاط ورودی را حذف نموده و سپس به خوشه‌بندی نقاط بپردازیم. سختی این مسئله از آنجاست که نه تنها باید مسئله‌ی خوشه‌بندی را حل نمود، بلکه در ابتدا باید تصمیم گرفت که کدام یک از داده‌ها را به‌عنوان داده‌ی پرت در نظر گرفت که بهترین جواب در زمان خوشه‌بندی به دست آید. در واقع اگر تعداد نقاط پرتی که مجاز به حذف است، برابر صفر باشد، مسئله به مسئله‌ی kـ مرکز تبدیل می‌شود. نمونه‌ای از مسئله‌ی ۲ـ مرکز با ۷ داده‌ی پرت را در شکل ۳-۲ می‌توانید ببینید. تعریف دقیق‌تر این مسئله در زیر آمده است:

مسئله‌ی ۳-۲ **(kـ مرکز با داده‌های پرت)** *یک گراف کامل بدون جهت $G = (V, E)$ با تابع فاصله‌ی d، که از نامساوی مثلثی پیروی می‌کند داده‌شده است. زیرمجموعه‌ی $Z \subseteq V$ با اندازه‌ی z و مجموعه‌ی $S \subseteq V - Z$ با اندازه‌ی k را انتخاب کنید به طوری‌که عبارت زیر را کمینه کند:*

$$\max_{v \in V-Z} \{\min_{s \in S} d(v, s)\} \qquad (۳-۲)$$

گونه‌ی دیگری از مسئله‌ی kـ مرکز که در سال‌های اخیر مورد توجه قرار گرفته است، حالت جویبار داده‌ی آن است. در این‌گونه از مسئله‌ی kـ مرکز، در ابتدا تمام نقاط در دسترس نیستند، بلکه به‌مرور زمان نقاط در دسترس قرار می‌گیرند. محدودیت دومی که وجود دارد، محدودیت حافظه است، به‌طوری‌که نمی‌توان تمام نقاط را در حافظه نگه داشت و بعضاً حتی امکان نگه‌داری در حافظه‌ی جانبی نیز وجود ندارد و به‌طور معمول باید مرتبه‌ی حافظه‌ای کم‌تر از مرتبه حافظه‌ی **خطی**[۱۳] متناسب با تعداد نقاط استفاده نمود. از این به بعد به چنین مرتبه‌ای، مرتبه‌ی **زیرخطی**[۱۴] می‌گوییم. مدلی که ما در این پژوهش بر روی آن تمرکز داریم مدل جویبار داده تک‌گذره[۱۵] [۵] است. یعنی تنها یک بار می‌توان از ابتدا تا انتهای داده‌ها را بررسی کرد و پس

[۱۳] Linear
[۱۴] sublinear
[۱۵] Single pass

از عبور از یک داده، اگر آن داده در حافظه ذخیره نشده باشد، دیگر به آن دسترسی وجود ندارد. علاوه بر این، در هر لحظه باید بتوان به پرسمان (برای تمام نقاطی از جویبار داده که تاکنون به آن دسترسی داشته‌ایم) پاسخ داد.

مسئله‌ی ۳-۳ (kـ **مرکز در حالت جویبار داده**) *مجموعه‌ای از نقاط در فضای dـ بعدی به مرور زمان داده می‌شود. در هر لحظه از زمان، به ازای مجموعه‌ی U از نقاطی که تا کنون وارد شده‌اند، زیرمجموعه‌ی $S \subseteq U$ با اندازه‌ی k را انتخاب کنید به طوری که عبارت زیر کمینه شود:*

$$\max_{u \in U} \{ \min_{s \in S} d(u, s) \} \tag{۳-۳}$$

از آنجایی که گونه‌ی جویبار داده و داده پرت مسئله‌ی kـ مرکز به علت به‌روز بودن مبحث داده‌های حجیم[۱۶]، به تازگی مورد توجه قرار گرفته است. در این تحقیق سعی شده است که تمرکز بر روی این‌گونه‌ی خاص از مسئله باشد. همچنین در این پژوهش سعی می‌شود گونه‌های مسئله را برای انواع متریک‌ها و برای kهای کوچک نیز مورد بررسی قرار داد.

۳-۳ مدل جویبار داده

همان‌طور که ذکر شد مسئله‌ی kـ مرکز در حالت داده‌های پرت و جویبار داده، گونه‌های تعمیم‌یافته از مسئله‌ی kـ مرکز هستند و در حالت‌های خاص به مسئله‌ی kـ مرکز کاهش پیدا می‌کنند. مسئله‌ی kـ مرکز در حوزه‌ی مسائل ان‌پی‌ـ سخت[۱۷] قرار می‌گیرد و با فرض $P \neq NP$ الگوریتم دقیق با زمان چندجمله‌ای برای آن وجود ندارد [۶]. بنابراین برای حل کارای[۱۸] این مسائل از الگوریتم‌های تقریبی[۱۹] استفاده می‌شود.

برای مسئله‌ی kـ مرکز، دو الگوریتم تقریبی معروف وجود دارد. در الگوریتم اول، که به روش حریصانه[۲۰] عمل می‌کند، در هر مرحله بهترین مرکز ممکن را انتخاب می‌کند به طوری تا حد ممکن از مراکز قبلی دور باشد [۷]. این الگوریتم، الگوریتم تقریبی با ضریب تقریب ۲ ارائه می‌دهد. در الگوریتم دوم، با استفاده از مسئله‌ی مجموعه‌ی غالب کمینه[۲۱]، الگوریتمی با ضریب تقریب ۲ ارائه می‌گردد [۸]. همچنین ثابت شده است، که بهتر از این ضریب تقریب، الگوریتمی نمی‌توان ارائه داد مگر آن‌که $P = NP$ باشد.

---
[۱۶] Big data
[۱۷] NP-hard
[۱۸] Efficient
[۱۹] Approximation algorithm
[۲۰] Greedy
[۲۱] Dominating set

جدول ۳-۱: نمونه‌هایی از کران پایین تقریب‌پذیری مسائل خوشه‌بندی

مسئله	کران پایین تقریب‌پذیری
kـ مرکز	۲ [۸]
kـ مرکز در فضای اقلیدسی	۱/۸۲۲ [۱۷]
۱ـ مرکز در حالت جویبار داده	$\frac{۱+\sqrt{۲}}{۲}$ [۱۳]
kـ مرکز با نقاط پرت و نقاط اجباری	۳ [۱۲]

برای مسئله‌ی kـ مرکز در حالت جویبار داده برای ابعاد بالا، بهترین الگوریتم موجود ضریب تقریب $۲ + \varepsilon$ دارد [۹، ۱۰، ۱۱] و ثابت می‌شود الگوریتمی با ضریب تقریب بهتر از ۲ نمی‌توان ارائه داد. برای مسئله‌ی kـ مرکز با داده‌ی پرت در حالت جویبار داده نیز، بهترین الگوریتم ارائه شده، الگوریتمی با ضریب تقریب $۴ + \varepsilon$ است که با کران پایین ۳ هنوز اختلاف قابل توجهی دارد [۱۲].

برای kهای کوچک به خصوص، $k = ۱, ۲$، الگوریتم‌های بهتری ارائه شده است. بهترین الگوریتم ارائه شده برای مسئله‌ی ۱ـ مرکز در حالت جویبار داده برای ابعاد بالا، دارای ضریب تقریب ۱/۲۲ است و کران پایین $\frac{۱+\sqrt{۲}}{۲}$ نیز برای این مسئله اثبات شده است [۱۳، ۱۴]. برای مسئله ۲ـ مرکز در حالت جویبار داده برای ابعاد بالا، اخیرا راه‌حلی با ضریب تقریب $۱/۸ + \varepsilon$ ارائه شده است [۱۵]. برای مسئله‌ی ۱ـ مرکز با داده‌ی پرت، تنها الگوریتم موجود، الگوریتمی با ضریب تقریب ۱/۷۳ است [۱۶].

۳-۴ تقریب‌پذیری

یکی از راه‌کارهایی که برای کارآمد کردن راه‌حل ارائه شده برای یک مسئله وجود دارد، استفاده از الگوریتم‌های تقریبی برای حل آن مسئله است. یکی از عمده‌ترین دغدغه‌های مطرح در الگوریتم‌های تقریبی کاهش ضریب تقریب است. در بعضی از موارد حتی امکان ارائه‌ی الگوریتم تقریبی با ضریبی ثابت نیز وجود ندارد. به طور مثال، الگوریتم تقریبی با ضریب تقریب کم‌تر از ۲، برای مسئله‌ی kـ مرکز وجود ندارد مگر این‌که $P = NP$ باشد. برای مسائل مختلف، معمولاً می‌توان کران پایینی برای میزان تقریب‌پذیری آن‌ها ارائه داد. در واقع برای برخی مسائل انپی‌ـ سخت، علاوه بر این که الگوریتم کارآمدی وجود ندارد، بعضاً الگوریتم تقریبی با ضریبی تقریب کم و نزدیک به یک نیز وجود ندارد. در جدول ۳-۱ میزان تقریب‌پذیری مسائل مختلفی که در این پایان‌نامه مورد استفاده قرار می‌گیرد را می‌بینید.

فصل ۴

نتایج جدید

در این فصل نتایج جدید به‌دست‌آمده در پایان‌نامه توضیح داده می‌شود. در صورت نیاز می‌توان نتایج جدید را در قالب چند فصل ارائه نمود. همچنین در صورت وجود پیاده‌سازی، بهتر است نتایج پیاده‌سازی را در فصل مستقلی پس از این فصل قرار داد.

فصل ۵

نتیجه‌گیری

در این فصل، ضمن جمع‌بندی نتایج جدید ارائه‌شده در پایان‌نامه یا رساله، مسائل باز باقی‌مانده و همچنین پیشنهادهایی برای ادامه‌ی کار ارائه می‌شوند.

Bibliography

[1] D. E. Knuth. *The TEXbook*. Addison-Wesley, 1984.

[2] L. Lamport. *LATEX—A Document Preparation System*. Addison-Wesley, 1985.

[3] J. Han and M. Kamber. *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.

[4] V. Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75, 2002.

[5] C. C. Aggarwal. *Data streams: models and algorithms*. Springer Science & Business Media, 2007.

[6] M. R. Garey and D. S. Johnson. Computers and intractability: a guide to the theory of NP-completeness. *Freeman & Co.*, 1979.

[7] N. Megiddo and K. J. Supowit. On the complexity of some common geometric location problems. *SIAM Journal on Computing*, 13(1):182–196, 1984.

[8] V. V. Vazirani. *Approximation Algorithms*. Springer-Verlag New York, Inc., 2001.

[9] R. M. McCutchen and S. Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. In *Proceedings of the 11th International Workshop on Approximation Algorithms*, pages 165–178, 2008.

[10] S. Guha. Tight results for clustering and summarizing data streams. In *Proceedings of the 12th International Conference on Database Theory*, pages 268–275, 2009.

[11] H.-K. Ahn, H.-S. Kim, S.-S. Kim, and W. Son. Computing k centers over streaming data for small k. *International Journal of Computational Geometry and Applications*, 24(02):107–123, 2014.

[12] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms*, pages 642–651, 2001.

[13] P. K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. In *Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms*, pages 1481–1489, 2010.

[14] T. M. Chan and V. Pathak. Streaming and dynamic algorithms for minimum enclosing balls in high dimensions. *Computational Geometry: Theory and Applications*, 47(2):240–247, 2014.

[15] S.-S. Kim and H.-K. Ahn. An improved data stream algorithm for clustering. In *Proceedings of the 11th Latin American Symposium on Theoretical Informatics*, pages 273–284, 2014.

[16] H. Zarrabi-Zadeh and A. Mukhopadhyay. Streaming 1-center with outliers in high dimensions. In *Proceedings of the 21st Canadian Conference on Computational Geometry*, pages 83–86, 2009.

[17] M. Bern and D. Eppstein. Approximation algorithms for geometric problems. In D. S. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, pages 296–345. PWS Publishing Co., 1997.

واژه‌نامه

الف

ابتکاری ..... heuristic
ابعاد بالا ..... high dimensions
اریب ..... bias
آستانه ..... threshold
اصل لانه‌ی کبوتری ..... pigeonhole principle
ان‌پی‌ـسخت ..... NP-Hard
انتقال ..... transition

ب

برخط ..... online
برنامه‌ریزی خطی ..... linear programming
بهینه ..... optimum
بیشینه ..... maximum

پ

پرت ..... outlier
پرسمان ..... query
پوشش ..... cover
پیچیدگی ..... complexity

ت

تجربی ..... experimental
تراکم ..... density
تقریب ..... approximation
تقسیم‌بندی ..... partition
توری ..... mesh
توزیع‌شده ..... distributed

ج

جداپذیر ..... separable
جعبه سیاه ..... black box
جویبار داده ..... data stream

ح

حدی ..... extreme
حریصانه ..... greedy

خ

خوشه ..... cluster
خطی ..... linear

د

دستور Prompt

درخت تفکر Tree-of-Thoughts

ر

رأس vertex

رسمی formal

ز

زنجیره تفکر Chain-of-Thoughts

س

سرشکن amortized

سلسه‌مراتبی hierarchichal

ش

شبه کد pseudocode

شیء object

ص

صدق‌پذیری satisfiability

غ

غلبه dominate

ف

فاصله distance

فضا space

ق

قطعی deterministic

ک

کارا efficient

کاندیدا candidate

کمینه minimum

م

مجموعه set

مجموعه هسته coreset

مسطح planar

موازی‌سازی parallelization

میان‌گیر buffer

ن

نابه‌جایی inversion

ناوردا invariant

نقطه‌ی مرکزی center point

نیم‌فضا half space

هـ

هزینه‌ی آشوب price of anarchy (POA)

ی

یال edge

پیوست آ

مطالب تکمیلی

پیوست‌های خود را در صورت وجود می‌توانید در این قسمت قرار دهید.

Abstract

This work examines the capability of large language models (LLMs) to messure the importance of Persian news, evaluating their learning ability from content, reasoning skills, and overall cognitive capacities. Initially, annotated datasets were collected from various domains, including sports, politics, social issues, medicine, and culture, to develop an evaluation framework for LLMs. Within this framework, various existing models were analyzed and assessed under different scenarios and conditions to evaluate their analytical performance in both Persian and English. The findings indicate that prompts incorporating Chain-of-Thoughts and Tree-of-Thoughts significantly improve the models' performance. Additionally, the Symbol Tuning method enhances sensitivity to the input queries and their content.

Keywords: Large Language Models, Natural Language Processing, Machine Learning, News Importance Detection

Sharif University of Technology

Department of Computer Engineering

B.Sc. Thesis

News Importance Detection With the Use of Large Langauge Models

By:

Shayan Salehi

Supervisor:

Dr. Mahdi Jafari

January 2025