دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات
دانشگاه صنعتی امیرکبیر

گزارش تمرین سوم درس مدل‌های احتمالاتی گرافی

استاد درس:

دکتر نیک‌آبادی

نام دانشجو:

احمد اسدی

۹۴۱۳۱۰۹۱

تیرماه ۱۳۹۵

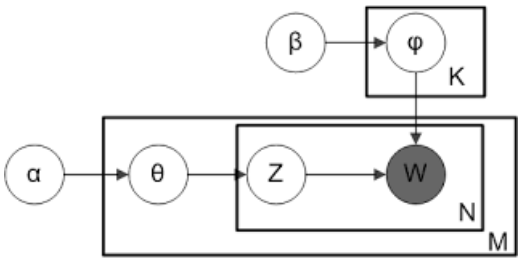
فهرست مطالب

۱ یادگیری در مدل تخصیص پنهان دیریکله[۱]

در این بخش ابتدا، به طور مختصر، فرایند یادگیری در مدل تخصیص پنهان دیریکله را مورد بررسی قرار می‌دهیم و سپس تاثیر پارامترهای مختلف را در روند یادگیری مدل بررسی خواهیم نمود.

۱.۱ یادگیری با نمونه‌برداری گیبس

شکل ۱ مدل تخصیص پنهان دیریکله، متغیرهای تصادفی مورد استفاده در آن و رابطه بین متغیرها را نمایش می‌دهد. همان‌طور که در شکل مشخص است، برای تولید فاکتورهای Θ و Φ که به ترتیب نمایش‌دهنده احتمال شرطی عناوین به شرط اسناد و احتمال شرطی کلمات به شرط عناوین هستند، از دو توزیع دیریکله به ترتیب با پارامترهای α و β استفاده شده است.



شکل ۱: مدل تخصیص پنهان دیریکله. روند تولید یک سند به طور کامل در این شکل مشخص است. هر سند شامل تعدادی کلمه است که در شکل با W نشان داده شده‌اند. کلمات با توجه به توزیع احتمالاتی شرطی کلمات به شرط عناوین (Θ) و بردار عناوین موجود در اسناد (Z) به‌وجود می‌آید. بردار Z، در شکل، با استفاده از توزیع احتمالاتی عناوین به شرط اسناد (Φ) تشکیل می‌شود. پارامترهای α و β، پارامترهای توزیع دیریکله برای تولید متغیرهای تصادفی α و β هستند.

مراحل اجرای الگوریتم نمونه‌برداری گیبس در بخش یادگیری مدل به شرح زیر است: (k تعداد عناوین موجود و l تعداد کلمات موجود هستند.)

۱. کلیه متغیرهای تصادفی z_i با یک مقدار تصادفی بین ۱ تا l مقداردهی می‌شوند. (l تعداد کل کلمات موجود در مجموعه اسناد و z_i عنوان پیشنهادی برای کلمه w_i است.)

۲. تا زمانی که شرایط mixing برقرار نشده مراحل زیر تکرار می‌شوند. [۲]

(آ) به ازای تمام کلمات w_i، توزیع احتمالی تغییر عنوان به هریک از عناوین موجود، مطابق با رابطه ۱ حساب می‌شود.

$$p(z_i = j|.) \propto \frac{n^{w_i}_{-i,j} + \beta}{n^{\cdot}_{-i,j} + W \cdot \beta} \cdot \frac{n^{d_i}_{-i,j} + \alpha}{n^{d_i}_{-i,\cdot} + l \cdot \alpha} \qquad (۱)$$

(ب) یک داده تصادفی از توزیع احتمالاتی $p(z_i = j|.)$ که در مرحله قبل محاسبه شده است، تولید شده و عنوان کلمه w_i مساوی این مقدار جدید قرار می‌گیرد.

(ج) ماتریس شمارش‌های محاسبه شده که در رابطه ۱ مورد استفاده قرار گرفته‌اند، تصحیح می‌شوند.

۳. با توجه به قاعده نمونه‌برداری که در ادامه توضیح داده خواهد شد، یک نمونه از بردار Z تولید می‌شود.

۴. با استفاده از نمونه تولید شده و با در نظر گرفتن روابط ۲ و ۳ فاکتورهای Θ و Φ محاسبه می‌شوند.

---

[۱] Latent Dirichlet Allocation (LDA)

[۲] در این پروژه، شرایط mixing پس از سپری شدن تعداد مشخصی تکرار ، احراز شده فرض می‌شود.

$$\hat{\Theta}_j^d = \frac{n_j^d + \alpha}{n_\cdot^d + k \cdot \alpha} \tag{۲}$$

$$\hat{\Phi}_j^w = \frac{n_j^w + \beta}{n_{\cdot j} + l \cdot \beta} \tag{۳}$$

در ادامه به بررسی تاثیر پارامترهای مختلف بر عملکرد الگوریتم می‌پردازیم.

۲.۱ پارامتر توزیع‌های دیریکله اولیه

در این بخش قصد داریم تاثیر پارامترهای α و β را بر عملکرد الگوریتم مورد بررسی قرار دهیم. در تمام آزمایشات این بخش، برای افزایش سرعت محاسبات، مقدار $k = ۳$ در نظر گرفته شده است. همین‌طور ۷۰ درصد از کل داده‌ها به طور تصادفی به مجموعه آموزشی و ۳۰ درصد باقیمانده به مجموعه تست اختصاص داده شده‌اند.

۱.۲.۱ تاثیر پارامتر آلفا

مطابق با رابطه ۲، که نحوه محاسبه فاکتور Θ را مشخص می‌کند، مقدار پارامتر α در توزیع احتمالی عناوین به شرط اسناد تاثیرگذار است. اگر این پارامتر را برابر با صفر در نظر بگیریم (رابطه ۴)، رابطه تبدیل به یک تخمین MLE ساده (تعداد کلمات موجود از هر عنوان به کل عناوین موجود در بین کلمات یک سند) برای عناوین خواهد شد. با اضافه کردن پارامتر α به رابطه، به شکلی دانش اولیه خود را در تخمین توزیع‌ها دخالت داده‌ایم. رابطه ۲ دقیقا برابر با رابطه ۴ است در صورتی که خودمان به طور دستی به هر جمله α کلمه از همه عناوین وارد کرده باشیم. این کار معادل این است که در ابتدا احتمال رخداد تمام عناوین را در اسناد با یک‌دیگر برابر بدانیم. از طرفی هر چه مقدار α بزرگتر باشد، اطمینان ما از دانش اولیه بیشتر است و برعکس.

$$\hat{\Theta}_j^d = \frac{n_j^d}{n_\cdot^d} \tag{۴}$$

با توجه به توضیحات ارائه شده، انتظار می‌رود، با کاهش میزان α مدل با سرعت بیشتری به داده‌ها برازش شود. اگر مقدار این پارامتر خیلی کوچک باشد، احتمال وقوع بیش‌برازش[۱] به داده‌ها به طور چشم‌گیری افزایش خواهد یافت. از طرفی اگر مقدار این پارامتر، به طور قابل توجهی بزرگ باشد، اطمینان بالای ما از دانش اولیه را نشان می‌دهد. این مساله باعث خواهد شد، قدرت یادگیری مدل از داده‌ها کاهش یافته و مدل نتواند توزیع‌های واقعی موجود در مجموعه اسناد موجود را به درستی نمایش دهد.

آزمایشات انجام شده در این بخش، مؤید انتظارات ما از نحوه رفتار پارامتر است. جدول ۱ نتایج عملکرد الگوریتم را به ازای مقادیر مختلف از پارامتر α نمایش می‌دهد. در تمام موارد $\beta = ۰/۵$ است.

همان‌طور که نتایج جدول ۱ نمایش می‌دهند با کاهش مقدار این پارامتر، میزان سرگشتگی الگوریتم روی مجموعه آموزشی کاهش می‌یابد. بر خلاف نتایج مجموعه آموزشی، سرگشتگی الگوریتم در بین داده‌های تست تا نقطه‌ای کاهش یافته و با کاهش بیش از حد مقدار این پارامتر، دوباره افزایش می‌یابد. این نکته نشان می‌دهد کاهش بیش از حد مقدار این پارامتر، باعث ایجاد بیش‌برازش بر داده‌ها می‌شود. بهترین مقدار به‌دست آمده برای این پارامتر، $\alpha = ۰/۳$ است.

شکل ۲ نمودار تغییرات سرگشتگی مدل را به ازای مقادیر مختلف پارامتر α روی مجموعه‌های آموزشی و تست نمایش می‌دهد.

[۱] Overfit

جدول ۱: تاثیر پارامتر α در عملکرد الگوریتم

α	زمان سپری شده در هر تکرار	سرگشتگی[1] مجموعه آموزشی	سرگشتگی مجموعه تست
۰.۰۱	۲۵	۳۶۱۵.۶۴۸	۴۲۱۲.۸۴۰
۰.۰۳	۲۵	۳۶۴۸.۹۳۸	۴۲۰۵.۳۹۵
۰.۱	۲۵	۳۶۸۳.۴۸۰	۴۱۷۰.۱۲۲

۲.۲.۱ تاثیر پارامتر بتا

مانند پارامتر آلفا می‌توان در مورد پارامتر بتا نیز قضاوت کرد. مطابق با رابطه ۳، که نحوه محاسبه فاکتور Φ را مشخص می‌کند، مقدار پارامتر β در توزیع احتمالی کلمات به شرط عناوین تاثیرگذار است. اگر این پارامتر را برابر با صفر در نظر بگیریم (رابطه ۵)، رابطه تبدیل به یک تخمین MLE ساده (تعداد کلمات موجود از هر عنوان در بین کل کلمات موجود در مجموعه اسناد) برای کلمات خواهد شد. با اضافه کردن پارامتر β به رابطه، به شکلی دانش اولیه خود را در تخمین توزیع‌ها دخالت داده‌ایم. رابطه ۳ دقیقا برابر با رابطه ۵ است در صورتی که خودمان به طور دستی β کلمه از همه عناوین وارد مجموعه اسناد موجود، کرده باشیم. این کار معادل این است که در ابتدا احتمال رخداد تمام کلمات را در عناوین با یک‌دیگر برابر بدانیم. از طرفی هر چه مقدار β بزرگتر باشد، اطمینان ما از دانش اولیه بیشتر است و برعکس.

$$\hat{\Phi}_j^d = \frac{n_j^w}{n_j^{(.)}} \tag{۵}$$

جدول ۲ نتایج عملکرد الگوریتم را به ازای مقادیر مختلف از پارامتر β نمایش می‌دهد.

جدول ۲: تاثیر پارامتر β در عملکرد الگوریتم

β	زمان سپری شده در هر تکرار	سرگشتگی[2] مجموعه آموزشی	سرگشتگی مجموعه تست
۰.۵	۲۵	۳۶۴۸.۹۳۸	۴۲۰۵.۳۹۵
۱	۲۵	۳۶۷۲.۴۳۶	۴۱۳۳.۳۹۸
۲	۲۵	۳۷۵۸.۳۹۷	۴۲۱۷.۲۷۸

همان‌طور که نتایج جدول ۲ نمایش می‌دهند پارامتر β نیز مانند پارامتر α عمل می‌کند. بهترین مقدار به‌دست آمده برای این پارامتر، $\beta = ۱$ است. شکل ۳ نمودار تغییرات سرگشتگی مدل را به ازای مقادیر مختلف پارامتر β روی مجموعه‌های آموزشی و تست نمایش می‌دهد.

۳.۱ تاثیر تعداد عناوین

مطابق با آزمایشات انجام شده، با افزایش تعداد عناوین، عملکرد الگوریتم بهبود یافت. البته انتظار می‌رود با افزایش بیش از حد تعداد عناوین، مانند حالتی که در خوشه‌بندی داده‌ها تعداد خوشه‌های از پیش تعیین شده بسیار زیاد است، الگوریتم دچار بیش‌برازش شود که به دلیل زمان‌بر بودن این حالت (هر تکرار حدود ۹۸۰ ثانیه با تعداد ۱۰۰ عنوان) امکان تست آن فراهم نشد.

مطابق با نمودارهای به‌دست آمده از این بخش، هرچه تعداد عناوین افزایش می‌یابد، سرگشتگی اولیه الگوریتم نیز بالاتر می‌رود. این امر، کاملا قابل انتظار است زیرا از آنجا که تعداد عناوین زیاد است، با شروع از یک نقطه تصادفی و یک مرحله تکرار، احتمال اشتباه بودن عنوان تخصیص یافته شده به اسناد و کلمات بیشتر است تا زمانی که تعداد عناوین کم باشد.

از طرف دیگر سرعت همگرا شدن الگوریتم با بالا رفتن تعداد عناوین، افزایش می‌یابد. این مورد هم با در نظر گرفتن این که با افزایش تعداد عناوین، نیاز به تعمیم درون‌گروهی هر عنوان کاهش می‌یابد، به راحتی قابل توجیه است. هر چه تعداد عناوین افزایش یابد، اسناد کمتری در یک خوشه قرار می‌گیرند. از آنجا

که یافتن تعداد کم، سندی که شباهت زیادی به یکدیگر دارند، راحت‌تر از یافتن تعداد زیاد سند شبیه به هم است، شباهت اسناد درون یک عنوان در حالتی که تعداد عناوین زیاد است، به مراتب بیشتر از حالات دیگر خواهد بود. با طی تعداد کمی تکرار، بسیاری از اسناد شبیه به هم موجود، تحت عناوین مشترک بیان می‌شوند و باعث بالا رفتن سرعت الگوریتم می‌شود.

جدول ۳ نتایج تاثیر تعداد عناوین بر عملکرد الگوریتم را نمایش می‌دهد.

جدول ۳: تاثیر تعداد عناوین در عملکرد الگوریتم

تعداد عناوین	زمان سپری شده در هر تکرار	سرگشتگی[1] مجموعه آموزشی	سرگشتگی مجموعه تست
۳	۲۵	۳۶۴۸.۹۳۸	۴۲۰۵.۳۹۵
۵	۴۲	۳۵۲۴.۹۴۵	۴۱۵۵.۶۱۰
۱۰	۱۳۵	۳۲۱۸.۸۴۲	۴۴۰۷.۰۰۰
۱۰۰	۹۸۰	محاسبه نشد	محاسبه نشد

شکل ۴ نمودار تغییرات سرگشتگی الگوریتم بر اساس تعداد عناوین را نمایش می‌دهد.

۴.۱ تاثیر روند نمونه‌برداری

پس از رسیدن به شرایط mixing، بردار Z مورد استفاده برای محاسبه فاکتورهای مدل، از طریق فرایند نمونه‌برداری تولید می‌شود. روش‌های مختلف موجود برای این کار، به شرح زیر می‌باشند.

۱. استفاده از آخرین نمونه تولید شده Z

۲. استفاده از چند نمونه آخر تولید شده Z

۳. استفاده از تعدادی از چند نمونه آخر تولید شده Z

روش اول، ساده‌ترین روش ممکن برای این کار است. در این روش پس از رسیدن به mixing، از نمونه Z تولید شده به طور مستقیم برای محاسبات بعدی استفاده می‌شود. در این روش، پس از اتمام تمام تکرارها و تولید یک نمونه Z، دوباره باید از ابتدا الگوریتم را شروع کرده و انجام دهیم تا به mixing برسیم و سپس یک نمونه Z دیگر تولید کنیم تا به تعداد دلخواه نمونه Z برسیم. همان‌طور که مشخص است این روش در این پروژه قابل انجام نیست. (به دلیل صرف زمان بسیار زیاد)

در روش دوم پس از رسیدن به mixing، به تعداد دلخواه تکرار ها را ادامه می‌دهیم تا نمونه‌های Z دلخواه تولید شوند و سپس با در دست داشتن این نمونه‌ها، می‌توانیم محاسبات فاکتورها را انجام دهیم. ایراد این روش، این است که نمونه‌های تولید شده در آن می‌توانند از هم مستقل نباشند. در این صورت، نمونه‌های تولید شده به یکدیگر وابسته هستند و یادگیری به درستی اتفاق نمی‌افتد. انتظار داریم در صورتی که نمونه‌های تولید شده در واقع از هم مستقل نباشند، عملکرد الگوریتم روی داده‌های تست بسیار ضعیف‌تر از عملکرد الگوریتم روی داده‌های آموزشی باشد.

در روش سوم پس از ادامه دادن تکرارهای الگوریتم به تعداد دلخواه پس از mixing، تعدادی از نمونه‌ها را به طور تصادفی یا با فواصل یکسان انتخاب کرده و از آن‌ها برای محاسبات بعدی استفاده می‌کنیم. این روش مشکل روش‌های قبلی را برطرف می‌سازد.

جدول ۴ تاثیر نحوه نمونه‌برداری را بر عملکرد الگوریتم مشخص می‌کند. همان‌طور که پیداست اگر چه تفاوت‌های جزئی در بین روش‌ها مشخص است اما به نظر می‌رسد از آنجا که تعداد نمونه‌های انتخاب شده (۲۰۰۰ نمونه) بالا بوده و همین‌طور وابستگی موجود بین نمونه‌های پشت سرهم در روند مساله بی‌تاثیر یا کم‌تاثیر به نظر می‌رسد، بهبود قابل توجهی از نتایج قابل استنتاج نیست.

جدول ۴: تاثیر نحوه نمونه‌برداری در عملکرد الگوریتم

نحوه نمونه‌برداری	سرگشتگی[1] مجموعه آموزشی	سرگشتگی مجموعه تست
روش دوم	۳۶۴۸.۹۳۸	۴۲۰۵.۳۹۵
روش سوم با فواصل ۲ تایی	۳۵۲۴.۳۲۴	۴۳۲۸.۵۹۵
روش سوم با فواصل ۴ تایی	۳۵۷۳.۹۲۴	۴۲۹۳.۸۵۶

۵.۱ خوشه‌بندی

در این بخش، فاکتور Θ برای خوشه‌بندی اسناد مورد استفاده قرار گرفته است. خوشه‌بندی اسناد با استفاده از روش KMeans انجام و با استفاده از معیارهای CalinskiHarabasz و silhouette برای ۲۰ خوشه، ارزیابی شده است. نتایج بدست آمده از این ارزیابی نشان می‌دهد استفاده از فاکتور Θ برای خوشه‌بندی داده‌ها (آموزش با ۱۰ عنوان) در ۱۶ خوشه، نتیجه بهینه را می‌دهد.

همین‌طور شکل ۵ و جدول ۵ در حالتی که از ۵ عنوان برای مدل استفاده شده است، ۱۰ مورد از کلمات شاخص هر عنوان را نمایش می‌دهند. همان‌طور که در شکل مشخص است، کلمات به خوبی از یک‌دیگر جدا شده‌اند و عنوان هر یک از ۵ گروه کلمه به طور واضح قابل تشخیص و تمیز از دیگر عناوین است که نشان‌دهنده عملکرد مناسب الگوریتم می‌باشد.

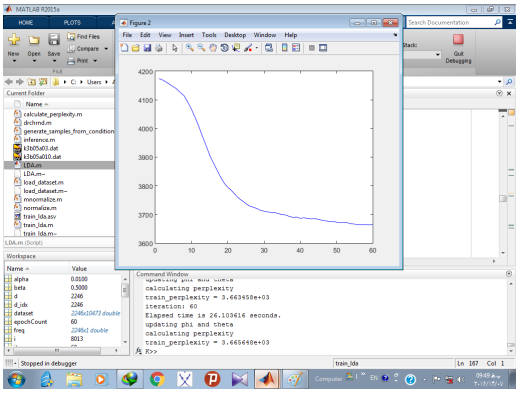
جدول ۵: کلمات نمونه از ۵ عنوان مشخص شده

عنوان اول	عنوان دوم	عنوان سوم	عنوان چهارم	عنوان پنجم
foreign	things	miles	judge	prices
campaign	really	area	convicted	higher
administration	cant	southern	jury	rose
meeting	doesnt	shot	guilty	trading
support	mother	soldiers	alleged	exchange
minister	feel	fighting	sentenced	fell
saying	sure	injured	prosecutors	average
leader	friends	navy	appeals	points
leaders	franks	israeli	enforcement	index
decision	magazine	hundreds	bentsen	analysts

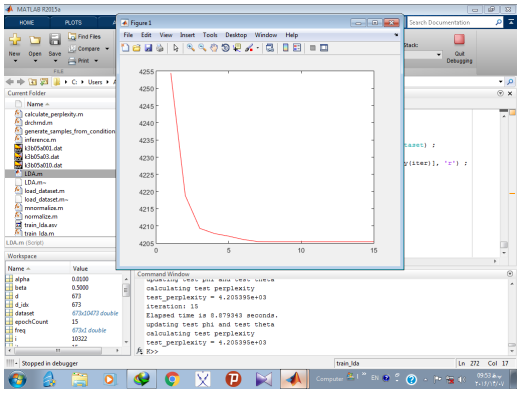
۲ توضیحات

* با توجه به ز مان‌بر بودن اجرای کامل الگوریتم، عموم آزمایشات نتیجه ۶۰ تکرار اول الگوریتم را گزارش داده‌اند. بدیهی است ادامه اجرای الگوریتم بر بهبود پاسخ موثر خواهد بود اما نتایج مقایسات تغییری نخواهند کرد.

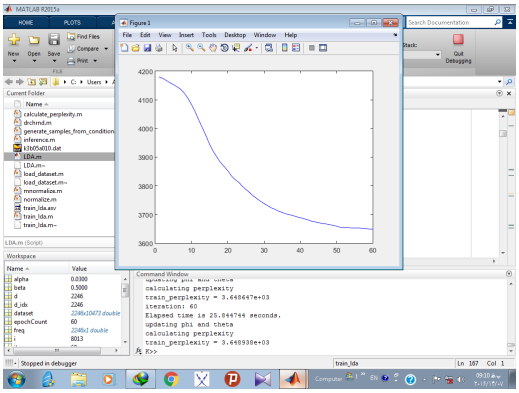
* سورس کد مربوط به پروژه در ضمیمه این گزارش ارسال شده است. همین‌طور این کد از این لینک ، قابل دریافت می‌باشد.



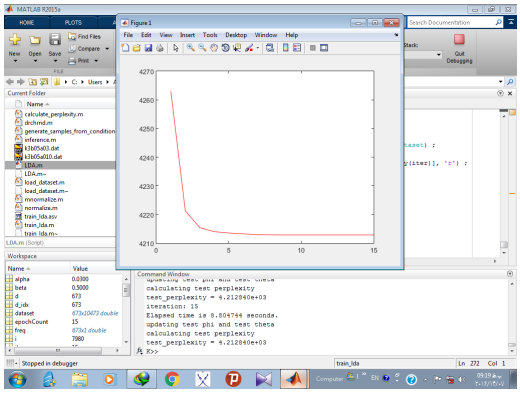
(آ) نمودار تغییرات سرگشتگی مجموعه آموزشی با $\alpha = ۰/۰۱$



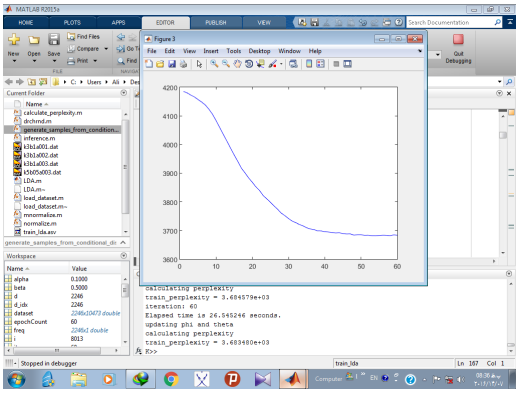
(ب) نمودار تغییرات سرگشتگی مجموعه تست با $\alpha = ۰/۰۱$



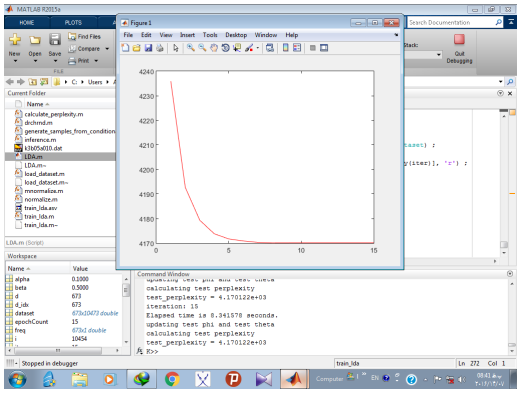
(ج) نمودار تغییرات سرگشتگی مجموعه آموزشی با $\alpha = ۰/۰۳$



(د) نمودار تغییرات سرگشتگی مجموعه تست با $\alpha = ۰/۰۳$

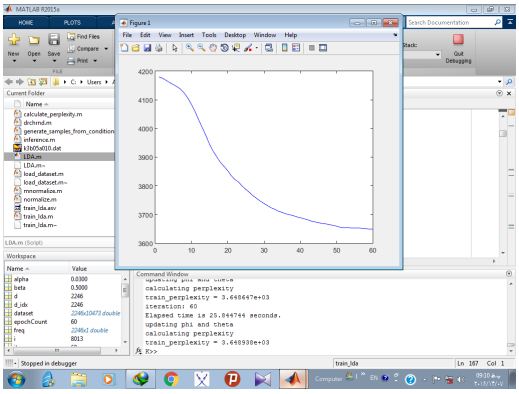


(ه) نمودار تغییرات سرگشتگی مجموعه آموزشی با $\alpha = ۰/۰۱$

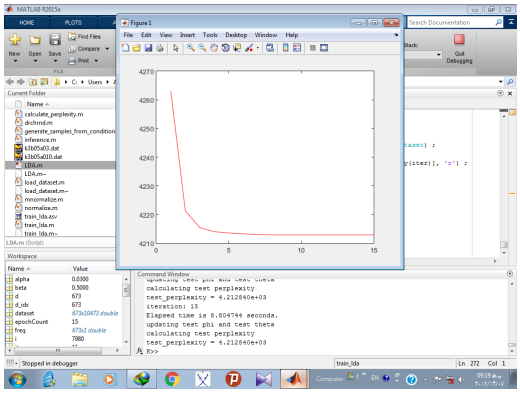


(و) نمودار تغییرات سرگشتگی مجموعه تست با $\alpha = ۰/۰۱$

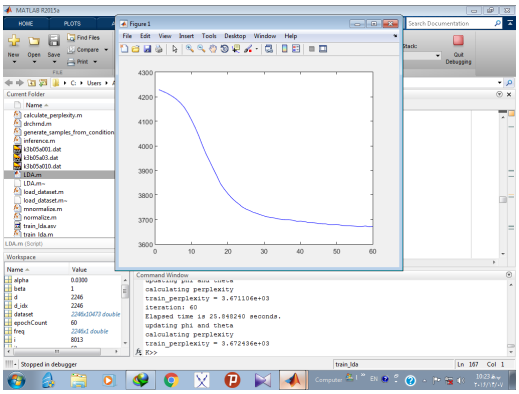
شکل ۲: تغییرات سرگشتگی مدل به ازای مقادیر مختلف پارامتر α برای مجموعه‌های آموزشی و تست



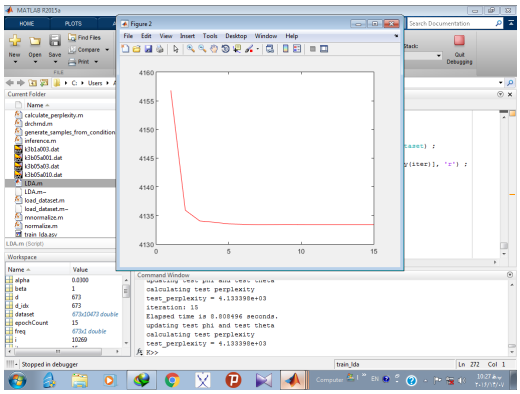
(آ) نمودار تغییرات سرگشتگی مجموعه آموزشی با $\beta = ۰/۵$



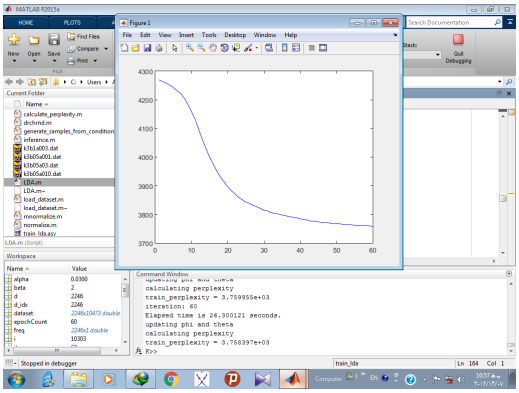
(ب) نمودار تغییرات سرگشتگی مجموعه تست با $\beta = ۰/۵$



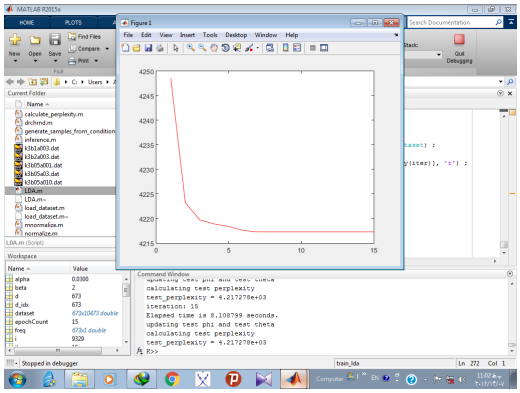
(ج) نمودار تغییرات سرگشتگی مجموعه آموزشی با $\beta = ۱$



(د) نمودار تغییرات سرگشتگی مجموعه تست با $\beta = ۱$

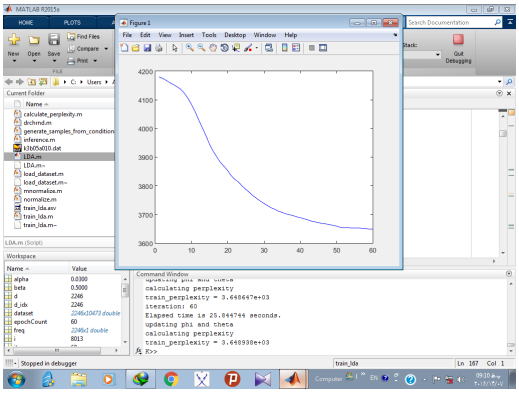


(ه) نمودار تغییرات سرگشتگی مجموعه آموزشی با $\beta = ۲$

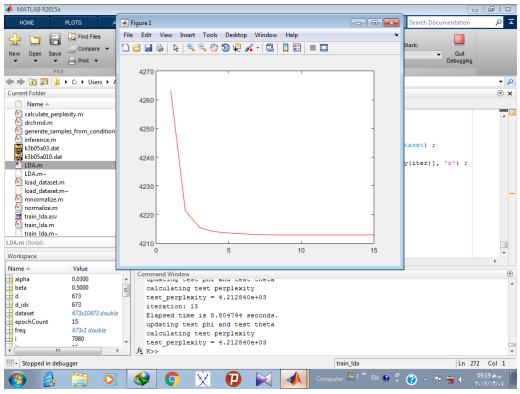


(و) نمودار تغییرات سرگشتگی مجموعه تست با $\beta = ۲$

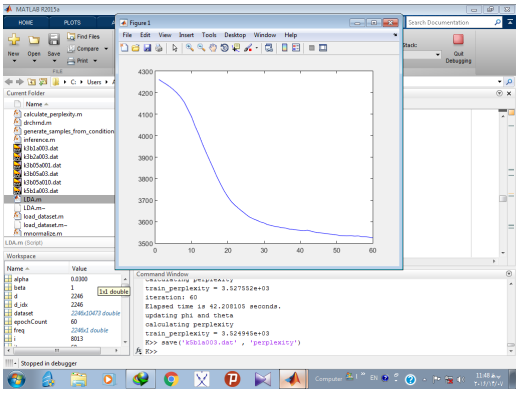
شکل ۳: تغییرات سرگشتگی مدل به ازای مقادیر مختلف پارامتر β برای مجموعه‌های آموزشی و تست



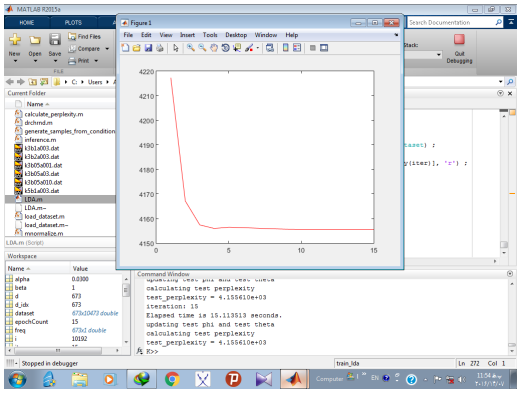
(آ) نمودار تغییرات سرگشتگی مجموعه آموزشی با ۳ عنوان



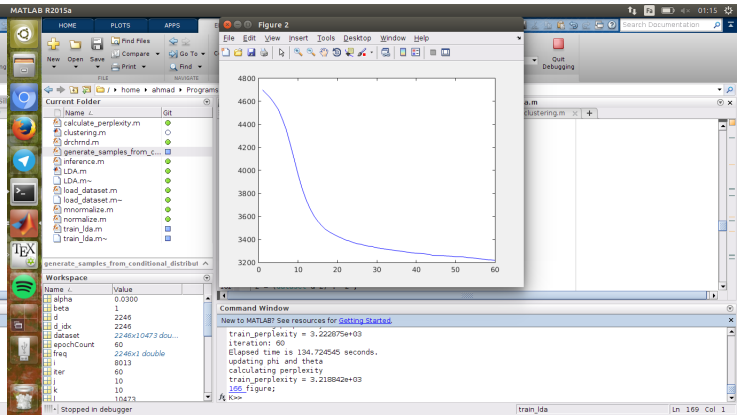
(ب) نمودار تغییرات سرگشتگی مجموعه تست با ۳ عنوان



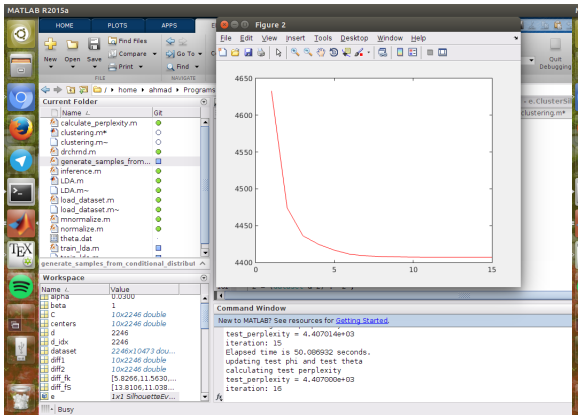
(ج) نمودار تغییرات سرگشتگی مجموعه آموزشی با ۵ عنوان



(د) نمودار تغییرات سرگشتگی مجموعه تست با ۵ عنوان

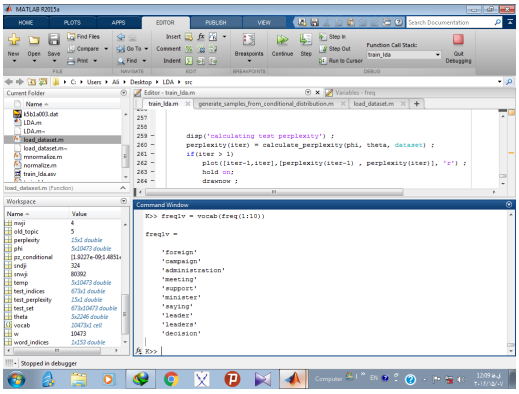


(ه) نمودار تغییرات سرگشتگی مجموعه آموزشی با ۱۰ عنوان

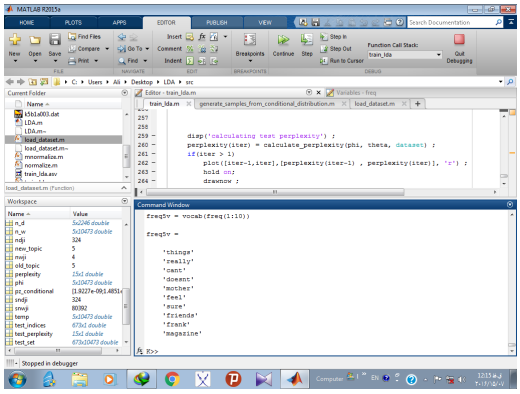


(و) نمودار تغییرات سرگشتگی مجموعه تست با ۱۰ عنوان

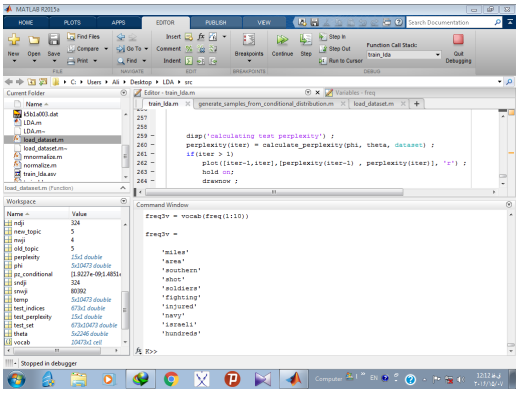
شکل ۴: تغییرات سرگشتگی مدل به ازای تعداد عناوین مختلف برای مجموعه‌های آموزشی و تست



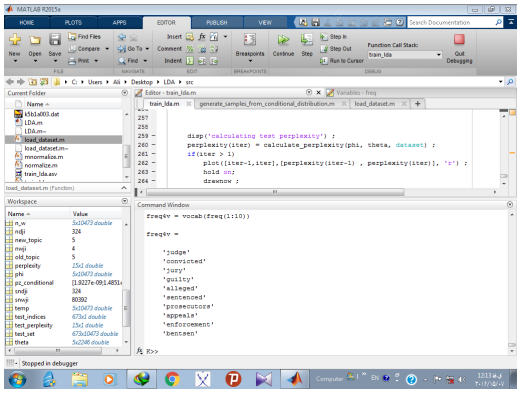
(آ) کلمات شاخص عنوان اول



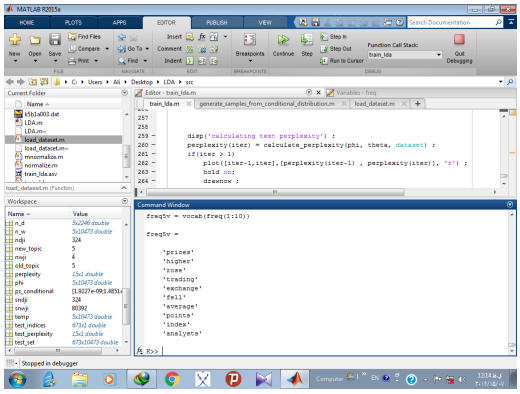
(ب) کلمات شاخص عنوان دوم



(ج) کلمات شاخص عنوان سوم



(د) کلمات شاخص عنوان چهارم



(ه) کلمات شاخص عنوان پنجم

شکل ۵: کلمات شاخص ۵ عنوان