A new method in text mining using fractional entropy

Hossein Mehri-Dehnavi*[,1], Hamzeh Agahi[2], Ali Mehri[1]

[1]Department of Physics, Faculty of Basic Science, Babol Noshirvani University of Technology, Shariati Ave., Babol 47148-71167, Iran

[2]Department of Mathematics, Faculty of Basic Science, Babol Noshirvani University of Technology, , Shariati Ave., Babol 47148-71167, Iran

Received: 07.09.2019 Final revised: 31.12.2019 Accepted: 12.02.2020

Doi link: 10.22055/JRMBS.2020.15568

Abstract

In this paper, we firstly review some definitions related to fractional calculus and fractional entropy, as a generalization of Shannon entropy. Then we introduce the generalized word importance metric based on fractional entropy. Using the proposed definition, we introduce a new text mining method based on fractional entropy. This method for keyword extraction of the Statistical Inference book by Casella and Berger (1990) shows that the F-measure value of the proposed text mining method, is higher than the related value for common text mining method based on Shannon entropy. These results indicate that the proposed text mining method based on fractional entropy is more comprehensive than the traditional text mining based on Shannon entropy.

Keywords: Shannon entropy, Fractional entropy, Text mining, Word ranking

*Corresponding Author: mehri@nit.ac.ir

روشی نوین در متن‌کاوی با آنتروپی کسری

حسین مهری دهنوی[1,*]، حمزه آگاهی[2]، علی مهری[1]

[1]گروه فیزیک، دانشکده علوم پایه، دانشگاه صنعتی نوشیروانی بابل، بابل، ایران

[2] گروه ریاضی، دانشکده علوم پایه، دانشگاه صنعتی نوشیروانی بابل، بابل، ایران

دریافت: 1398/06/16 ویرایش نهائی: 1398/10/10 پذیرش: 1398/11/23

Doi link: 10.22055/JRMBS.2020.15568

چکیده

در این مقاله، ابتدا به مرور برخی تعاریف مربوط به حسابان کسری و آنتروپی کسری، به‌عنوان تعمیمی از آنتروپی شنون، می‌پردازیم. سپس تعمیم معیار اهمیت لغات را بر اساس آنتروپی کسری معرفی می‌کنیم. با استفاده از تعریف پیشنهادی، به ارائه روشی جدیدی در متن‌کاوی بر اساس آنتروپی کسری خواهیم پرداخت. روش ارائه شده برای استخراج نمایهٔ کتاب آماری کسلا و برگر (1990)، نشان می‌دهد که مقدار میانگین هارمونیک بازیابی و صحت برای روش پیشنهادی بیشتر از مقدار به‌دست با روش متن کاوی متداول بر اساس آنتروپی شنون است. این نتایج نشان می‌دهد که روش پیشنهادی برای متن‌کاوی با استفاده آنتروپی کسری، نسبت به روش متداول بر اساس آنتروپی شنون، ابزار جامع‌تری است.

کلیدواژگان: آنتروپی شنون، آنتروپی کسری، متن‌کاوی، رتبه‌بندی کلمات

مقدمه و مفاهیم مقدماتی

یکی از کاربردهای مهم آمار و فیزیک آماری در رتبه بندی کلمات است [1و2]. رتبه‌بندی کلمات اهمیت روزافزونی در زبان شناسی، متن‌کاوی و پیدا کرده است. از آنجایی‌که کلمات مهم متن به شیوه‌ای خاص برای توجیه هدف نویسنده در متن توزیع می‌شوند. لذا کلمات مهم در مناطق خاصی از متن متمرکز می‌شوند و خوشه‌هایی را تشکیل می‌دهند، در حالی که کلمات غیرمهم معمولاً در متن به‌طور تقریباً یکنواخت توزیع می‌شوند. در نتیجه، انتظار داریم توزیع فضایی کلمات کلیدی و غیر کلیدی متن، به‌طور قابل توجهی متفاوت باشد. بر این اساس اخیراً روش‌های متنوعی برای متن کاوی و رتبه‌بندی کلمات پیشنهاد شده است. مراجع [3-5] به ارائهٔ الگوریتم‌هایی برای رتبه‌بندی کلمات براساس انحراف معیارتوزیع کلمات در متن پرداخته‌اند. مراجع [6-8] به متن کاوی با استفاده از توزیع کلمات مهم و غیر مهم در متن و و تفاوت آنتروپی مربوط به این دو دسته از کلمات پرداخته‌اند.

توزیع فضایی کلمات غیر مهم (بر خلاف کلمات مهم) در متن اصلی و متن مخلوط شده، تقریباً یکسان است. با توجه به این امر، مقادیر واگرایی جنسن-شنون[1] برای کلمات مهم باید بیشتر از مقادیر آن برای کلمات غیر مهم باشد [9].

با معرفی یک گراف مربوط به متن و با استفاده از نظریهٔ سیستم‌های پیچیده به تجزیه و تحلیل ارتباط بین کلمات مختلف یک متن پرداخته است [10].

* نویسنده مسئول: mehri@nit.ac.ir

[1] Jensen-Shannon divergence (JSD)

توزیع فرکانس (تعداد تکرار) کلمات از توزیع توانی تبعیت می‌کند [11].

مرجع [12] پیشنهاد داد که کلمات با توزیع میانه لغات کلیدی متن هستند و از این رابطه برای رتبه بندی کلمات استفاده نمود.

یک مفهوم حائز اهمیت در آمار و احتمالات و فیزیک آماری آنتروپی شنون نام دارد. همان‌طور که از نام این مفهوم برداشت می‌شود، این مفهوم نشان‌دهندهٔ میزان بی‌نظمی سیستم مورد مطالعه است. در ادامهٔ این بخش بیان تعریف آنتروپی شنون تعمیم دیگری از آن تحت عنوان آنتروپی کسری خواهیم پرداخت و در بخش‌های بعدی با ارائهٔ روش جدیدی برای متن‌کاوی بر اساس آنتروپی کسری پرداخته و نشان می‌دهیم که آنتروپی کسری، نسبت به آنتروپی شنون، ابزار قدرتمندتری برای متن کاوی است.

آنتروپی شنون و آنتروپی کسری

آنتروپی شنون به‌صورت زیر تعریف می‌شود

$$S = \sum_i p_i\, I(p_i) = -\sum_i p_i \ln p_i \ , \qquad 1$$

که $I(p_i) = -\ln p_i$ اطلاعات شنون و p_i احتمال حضور سیستم مورد مطالعه در حالت iام است که در شرط $\sum_i p_i = 1$ صدق می‌کند. با استفاده از مفاهیم مشتق و انتگرال کسری، تعمیمی از آنتروپی شنون تحت عنوان آنتروپی کسری در [13] ارائه شد. که در ادامه به تعاریف و ویژگی‌های آنتروپی کسری می‌پردازیم.

فرض کنید که :

$n \in \mathrm{N}, R(\alpha) \in (n-1, n], \alpha \in C$ و $[a,b]$ یک بازهٔ بسته در R باشد.

مشتق کسری ریمان لیوویل[1] از مرتبهٔ α را که با نماد $D^{\alpha}_{a+} f(x)$ نشان داده می‌شود، به‌صورت زیر تعریف می‌شود:

$$D^{\alpha}_{a+} f(x) = \frac{1}{\Gamma(n-\alpha)} \frac{d^n}{dx^n} \int_a^x \frac{f(\tau)}{(x-\tau)^{1-n+\alpha}} d\tau \qquad 2$$
$$x \geq a,$$

که تابع گاما به‌صورت زیر تعریف می‌شود:

$$\Gamma(\alpha) = \int_0^{\infty} e^{-u} u^{\alpha-1} du.$$

مثال1. فرض کنید $f(x) = \ln(x)$ ، $x \geq 0$. در این صورت طبق رابطهٔ2،

$$D^{\alpha}_{0+} f(x) = D^{\alpha}_{0+} \ln(x) = \frac{x^{-\alpha}}{\Gamma(1-\alpha)} [\ln x + \psi(1) - \psi(1-\alpha)], \qquad 3$$

که $\psi(x)$ نشان دهندهٔ تابع دی گاما است که به‌صورت زیر تعریف می‌شود:

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x).$$

انتگرال کسری ریمان-لیوویل از مرتبهٔ α را که با نماد $D^{-\alpha}_{a+} f(x)$ نشان داده می‌شود، به‌صورت زیر تعریف می‌شود:

$$D^{-\alpha}_{a+} f(x) := \frac{1}{\Gamma(a)} \int_a^x \frac{f(\tau)}{(x-\tau)^{1-\alpha}} d\tau$$
$$x \geq a.$$

برای جزئیات بیشتر در مورد مشتق و انتگرال کسری به [14] مراجعه نمایید.

با توجه به اینکه در نظریهٔ اطلاعات شنون $I(p_i) = -\ln p_i = D^0_{0+} I(p_i)$ است، ماچادو [13] با شروع از رابطهٔ1 و در نظر گرفتن این نکته که

$$D^1_{0+} I(p_i) = \frac{d}{dp_i}(-\ln p_i) = -\frac{1}{p_i}$$

و

$$D^{-1}_{0+} I(p_i) = p_i(1 - \ln p_i)$$

[1] Riemann-Liouville

است، به‌کمک رابطهٔ3 ایدهٔ جدیدی به نام اطلاع کسری شنون ارائه داد که فرم تصحیح شده آن به‌صورت زیر است:

$$I_\alpha(p_i) = D^{\alpha}_{0+} I(p_i) = -D^{\alpha}_{0+} \ln(p_i) = -\frac{p_i^{-\alpha}}{\Gamma(1-\alpha)}[\ln p_i + \psi(1) - \psi(1-\alpha)],$$
$$\alpha \in R$$

که $\Gamma(\cdot)$ و $\psi(\cdot)$ به‌ترتیب نشان دهندهٔ توابع گاما و دی‌گاما می‌باشند. توجه کنید که :

- اگر $\alpha = 0$ انگاه :

$$I_0(p_i) = I(p_i) = -\ln p_i ,$$

- اگر $\alpha = -1$ آنگاه :

$$I_{-1}(p_i) = D^{-1}_{0+} I(p_i) = p_i(1-\ln p_i),$$

- اگر $\alpha \to 1$، آنگاه :

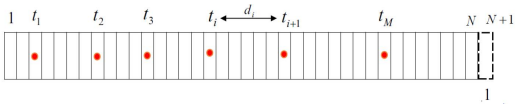
$$\lim_{a\to 1} I_1(p_i) = D^{1}_{0+} I(p_i) = -p_i^{-1},$$

با کمک اطلاع کسری $I_\alpha(p_i)$ تعریف آنتروپی کسری در زیر ارائه شده است.

آنتروپی کسری از مرتبهٔ $\alpha \in R$ به‌صورت زیر تعریف می‌شود

$$S_\alpha = \sum_i p_i I_\alpha(p_i) = \sum_i \{-\frac{p_i^{-\alpha}}{\Gamma(1-\alpha)}[\ln p_i + \psi(1) - \psi(1-\alpha)]\} p_i. \quad (5)$$

آنتروپی کسری تعمیمی از آنتروپی شنون است که در حالت خاص $\alpha = 0$ بر آنتروپی شنون منطبق می‌شود.



شکل1. تصویر نمادین از یک کتاب با طول N کلمه و لغت مشخص M_i با تعداد تکرار M مرتبه که در جایگاه‌های t_i قرار گرفته است. متن به‌طور دوره‌ای فرض شده است (یعنی جایگاه کلمهٔ N+1 بر جایگاه کلمهٔ اول منطبق است).

کاربردهایی از آنتروپی کسری در مراجع [16و15] ارائه شده است. اهداف و ساختار کلی مقاله به‌صورت زیر است. در بخش بعدی، ابتدا به مروری بر روش متداول متن‌کاوی بر اساس آنتروپی شنون می‌پردازیم. پس از آن به ارائه مفهوم اهمیت تعمیم‌یافته کسری لغات، WI_α، با استفاده از آنتروپی کسری، پرداخته و بر اساس آن به ارائه روش جدیدی برای متن‌کاوی بر اساس آنتروپی کسری خواهیم پرداخت. در بخش سوم روش ارائه شده خود برای استخراج لغات مهم کتاب آمار استنباطی کسلا و برگر [17] استفاده می‌نماییم. با مقایسه نتایج حاصل از روش متداول متن‌کاوی و روش جدید، نشان خواهیم داد که روش جدید متن‌کاوی بر اساس آنتروپی کسری ابزار قدرتمندتری در متن‌کاوی است. بخش آخر به بحث و نتیجه‌گیری اختصاص دارد.

متن کاوی

مفهوم احتمال در متن کاوی

مطابق با شکل1 متنی به‌طول N را در نظر بگیرید. با این توصیف متن مورد بررسی ما می‌تواند شامل N لغت باشد. فرض بر این است که لغات متن مطابق شکل از ابتدا تا به انتها رتبه‌بندی شده‌اند. همچنین طبق معمول فرض دوره‌ای بودن متن را نیز به کار می‌بریم.

کلمهٔ (مورد نظر) W_1 را در نظر بگیرید که به تعداد (فرکانس) M بار در متن مورد مطالعه تکرار شده باشد. شماره و جایگاه‌های مختلف این کلمه در متن را مطابق با شکل به‌ترتیب با $t_1, t_2, ..., t_M$ نشان می‌دهیم. با توجه به توضیحات فوق فاصلهٔ کلمه (مورد نظر) W_1 شماره i با کلمه مشابه بعدی خود در متن برابر با $d_i = t_{i+1} - t_i$, خواهد بود.

برای محاسبهٔ فاصلهٔ آخرین کلمهٔ متن با کلمهٔ بعدی خود باید فرض چرخه‌ای را در نظر بگیریم.

به‌عبارتی با فرض $t_{M+1} \to t_1 + N$ خواهیم داشت:

$$d_M = t_{M+1} - t_M = t_1 - t_M + N.$$

واضح است که جمع فاصله‌های تمامی کلمات برابر با طول کل متن می‌باشد، به‌عبارتی $\sum_i^M t_i = N$.

حال می‌توانیم برای کلمه مورد نظر W_1 قرار گرفته در جایگاه t_i، یک احتمال به‌صورت $p_i = d_i / N$ ، تعریف نمود. واضح است که جمع احتمال‌های مربوط به یک لغت مورد نظر برابر با یک خواهد بود. روش‌های دیگری هم برای تعریف احتمال مربوط به یک کلمه در متن وجود دارد که تعدادی از آنها در [8،9] مرور شده‌اند.

روش متداول رتبه‌بندی لغات با آنتروپی شنون

در روش متداول رتبه‌بندی لغات یک متن مطابق فرایند فوق برای هر کلمه‌ای یک مجموعه احتمالات $\{p_1, p_2, ... p_M\}$را برای کلمه‌ی دلخواه W_j محاسبه می‌کنند، سپس با استفاده از رابطهٔ1 آنتروپی شنون مربوط به آن کلمه را به‌دست می‌آورند. حال اگر آن کلمه موردنظر یک لغت بی‌اهمیت مانند "است"، "در" و ... باشد، این کلمه تقریباً در تمامی متن به‌طور یکنواخت توزیع شده است. ولی اگر کلمهٔ مربوطه کلمه‌ای مانند "احتمال"، "توزیع"، "همبستگی" و ...باشد، واضح است که این کلمه در بخش‌های خاصی از متن از قبیل بحث روی نتایج، نمودارها و دیگر بخش‌ها در مقایسه با قسمت مقدمات متن بیشتر تکرار می‌شود.

اگر در حالت ایده‌آل یک لغت بی‌اهمیت W_1 با تکرار M بار در متن به‌طور کاملاً یکنواخت توزیع شده باشد. واضح است که

$$p_i = d_i / N = \frac{1}{M}, d_i = N/M,$$

و مقدار آنتروپی شنون مربوطه برابر با

$$S_{uniform} = \sum_{i=1}^{M} \frac{1}{M} \ln M = \ln M,$$

خواهد بود. که بیشترین مقدار برای یک کلمه با تعداد تکرار M عدد در متن است. ولی اگر کلمهٔ موردنظر کلمه‌ای مهم از تعداد تکرار M مرتبه در متن باشد، واضح است که آنتروپی مربوطه از مقدار $\ln M$ کمتر خواهد شد. به‌همین دلیل هر چقدر مقدار آنتروپی یک لغت با مرتبه تکرار M از مقدار آنتروپی لغتی با همان فرکانس و توزیع یکنواخت $(S^M_{uniform} = \ln M)$ متفاوت‌تر باشد این کلمه، کلمهٔ مهم‌تری است. این روش، روش متداول رتبه‌بندی لغات با آنتروپی شنون است.

پس به‌نظر می‌آید که اختلاف آنتروپی یک کلمه با فرکانس مشخص M و آنتروپی کلمه‌ای با توزیع یکنواخت و همان فرکانس معیار خوبی برای تشخیص اهمیت کلمهٔ مربوطه باشد. این امر تا حدی درست می‌باشد! دلیل این امر نرمالیزه نبودن معیار اشاره شده است. برای نرمالیزه نمودن این معیار بهتر است اختلاف به‌دست آمده را بر مقدار آنتروپی لغت با توزیع یکنواخت از همان فرکانس، $S^M_{uniform} = \ln M$ تقسیم نماییم. این معیار را اهمیت لغت می‌نامیم.

اهمیت یک لغت با مرتبهٔ تکرار M در یک متن به‌صورت زیر تعریف می‌شود

$$WI(W_1) = \frac{\left| S^M_{uniform} - S(W_1) \right|}{S^M_{uniform}}, \qquad 6$$

که در آن $S(W_1)$ و $S^M_{uniform} = \ln M$ به‌ترتیب آنتروپی لغت W_1 با تکرار M، مرتبه و آنتروپی یک لغت با توزیع یکنواخت همان مرتبه تکرار می‌باشند.

با تکرار مرتبه و آنتروپی یک لغت با توزیع یکنواخت همان مرتبه تکرار می‌باشند.

اخیراً در مرجع [18] از معیار زیر جهت متن کاوی استفاده شده است:

$$D_\alpha(W_1) = \left| E^M_{\alpha,uniform} - E_\alpha(W_1) \right|,$$

که در آن $E_\alpha(W_1)$ و $E^M_{\alpha,uniform}$ به‌ترتیب آنتروپی دیستورت شده با پارامتر آزاد α برای لغت W_1 با تکرار M مرتبه و آنتروپی دیستورت با پارامتر آزاد α یک لغت با همان مرتبه تکرار و توزیع یکنواخت است.

معیارهای ارزیابی روش‌ها در متن‌کاوی

در این بخش به مرور معیارهای متداول برای متن‌کاوی می‌پردازیم. متنی را در نظر بگیرید که لغات مهم آن از قبل نمایه شده باشند. تعداد لغات حاضر در نمایهٔ این کتاب را N_{rel} فرض می‌کنیم. پس از اجرای برنامه و امتیازبندی لغات طبق رابطهٔ6، به لیستی از لغات که برحسب معیار اهمیت لغت مرتب شده‌اند خواهیم رسید.

تعداد N_{ret} لغت اول به‌دست آمده در این لیست را در نظر بگیرید. از این تعداد، تعداد $N_{rel\cap rel}$ آن لغاتی هستند که در نمایهٔ متن موردنظر نیز وجود دارند. نسبت این دو عبارت را معیار "صحت"[1] الگوریتم می‌نامیم:

$$P = \frac{N_{rel\cap ret}}{N_{ret}}. \qquad 7$$

همچنین نسبت تعداد لغت مشترک به‌دست آمده در هر مرحله با نمایه، $N_{rel\cap ret}$ به تعداد کل لغات لیست نمایه را هم به‌عنوان معیار "بازیابی[2]" تعریف می‌کنند:

$$R = \frac{N_{rel\cap ret}}{N_{rel}} \quad . \qquad 8$$

اگر N_{ret} به‌عنوان متغیر فرض کنیم و معیارهای صحت و بازیابی را برحسب این متغیر رسم کنیم، می‌توانیم رفتار این توابع را برحسب متغیر ذکر شده بررسی کنیم. با توجه به تعریف ارائه شده برای این معیارها، رفتار هر دو آن‌ها برای مقادیر کوچک متغیر N_{ret} رفتار نوسانی دارند. همچنین برای مقادیر بزرگ N_{ret} رفتار معیارهای بازیابی و صحت به‌ترتیب صعودی و نزولی هستند.

معیار دیگری که میانگین هارمونیک دو معیار صحت و بازیابی است را "F-معیار[3]" می‌نامند [19]:

$$F = \frac{2RP}{R+P} \qquad 9$$

رفتار کلی (به‌جز نقاط اولیهٔ نوسانی) این معیار برحسب متغیر N_{ret}، با دو معیار بالا متفاوت است. این معیار ابتدا رفتار صعودی داشته و پس از نقطهٔ بیشینهٔ خود رفتار نزولی خواهد داشت.

برای هر الگوریتم داده‌کاوی و یا متن‌کاوی، مقدار بیشینهٔ F-معیار عددی به‌خصوص برای متن مورد کاوش قرار گرفته است. الگوریتمی که مقدار بیشینهٔ F-معیار آن، عدد بزرگ‌تری را برای متن خاصی به‌خود بگیرد، آن الگوریتم موفق‌تری در بررسی آن متن خواهد بود [19].

روش پیشنهادی برای رتبه‌بندی لغات با استفاده از آنتروپی کسری

در زیر بخش قبلی روشی را بیان نمودیم که برای هر لغت مورد نظر W_1 می‌توان مجموعه احتمال تعریف نمود. آنگاه با مجموعه احتمال به‌دست آمده و با استفاده از رابطهٔ1 آنتروپی شنون مربوط به آن لغت را محاسبه نمود. همچنین نشان دادیم که لغات کم اهمیت آنتروپی

[1] precision
[2] recall
[3] F-measure

بیشتری خواهند داشت. همانند روش اشاره شده در زیر بخش قبلی می‌توان احتمال‌های مربوط به یک لغت در یک متن را به‌دست آورد. پیشنهاد ما این است که برای به‌دست آوردن لغات مهم یک متن به‌جای آنتروپی شنون1 از آنتروپی کسری S^{α} ارائه شده در رابطهٔ5 استفاده کنیم. نتایج حاصل از روش متن‌کاوی با آنتروپی کسری، برای حالت خاص $\alpha = 0$، برابر با نتایج به‌دست آمده برای متن‌کاوی با استفاده از آنتروپی شنون خواهد بود.

اهمیت تعمیم‌یافته کسری یک لغت با مرتبه تکرار M ، در یک متن به‌صورت زیر تعریف می‌شود

$$WI_{\alpha}(W_1) = \frac{\left|S^{M}_{\alpha,uniform} - S_{\alpha}(W_1)\right|}{S^{M}_{\alpha,uniform}}, \qquad 10$$

که در آن $S_{\alpha}(W_1)$ و $S^{M}_{\alpha,uniform}$ به‌ترتیب آنتروپی کسری مرتبهٔ α لغت W_1 با تکرار M مرتبه و آنتروپی کسری با مرتبهٔ α یک لغت با همان مرتبه تکرار و توزیع یکنواخت، می‌باشند.

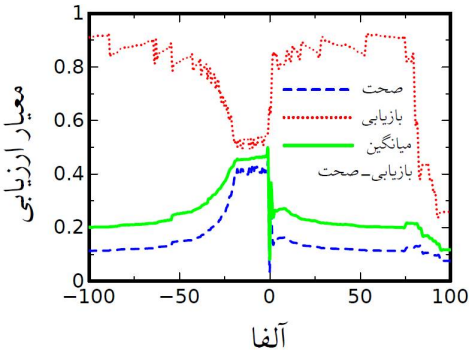
نتایج مدل ارائه شده برای کتاب کسلا و برگر

در این بخش به بررسی روش پیشنهادی برای رتبه‌بندی کلمات در کتاب آماری کسلا و برگر خواهیم پرداخت. این مرجع شامل $N = 218133$ کلمه است. از این تعداد کلمه، تعداد کل $N_v = 11665$ کلمه متفاوت وجود دارد.

در این مقاله ابتدا طبق تعریف 6 لغات کتاب را برحسب اهمیت تعمیم‌یافته کسری لغات برای یک α مشخص رتبه‌بندی کردیم.

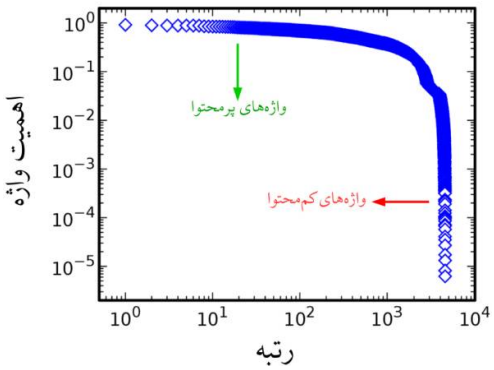
برای یک مقدار مشخص آلفا، با استفاده از نمایهٔ کتاب و روش پیشنهادی در بخش قبلی می‌توانیم نمودار صحت، بازیابی و همچنین میانگین هارمونیک این دو معیار را برحسب تعداد لغت ظاهر شده در لیست لغات مرتب شده بر اساس اهمیت لغات، به‌دست آوریم. سپس بیشینه مقدار F - معیار را برای پارامتر کسری α موردنظر به‌دست می‌آوریم. مقادیر صحت و بازیابی برای مربوط به نقطهٔ بهینهٔ F - معیار را نیز می‌توانیم محاسبه کنیم. شکل2 نتایج به‌دست آمده، را برحسب مقادیر مختلف پارامتر کسری α را رسم نموده است. همان‌طور که در شکل مشاهده می‌نماییم، بیشترین مقدار میانگین هارمونیک بازیابی و صحت در 0/8- $= \alpha$ رخ می‌دهد.

در ادامه سعی نمودیم برای مقدار بهینه 0/8- $= \alpha$ لغات کتاب کسلا و برگر را برحسب امتیاز آنها رتبه‌بندی کنیم. نتایج مربوطه در شکل3 آورده شده‌اند. برای رسم این نمودار کلیدواژه‌های کتاب برحسب مقدار اهمیتشان، WI -0/8 به‌شکل نزولی مرتب شده‌اند. بخش نخست نمودار، با شیب کم، نمایانگر واژه‌های پرمحتوا و مهم است که مقدار اهمیت آن‌ها بزرگ‌تر است. بخش دوم نمودار، با شیب تند واژه‌های کم‌محتوا و دستوری را نشان می‌دهد که مقدار اهمیت آن‌ها پایین است. نکته حائز اهمیت این است که، نتایج این محاسبه از قانون زیف[1] تبعیت می‌کنند [12].

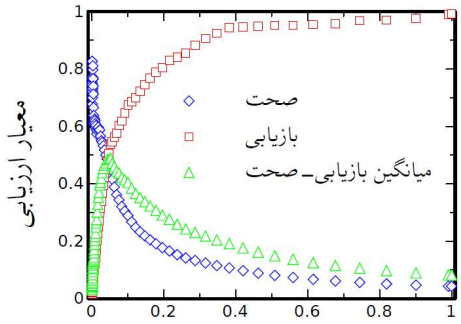


شکل2. بازیابی (خط‌چین آبی)، صحت (نقطه‌چین قرمز) و میانگین هارمونیک آن‌ها (خط سبز) برای نمایه‌های استخراج شده از کتاب آماری کسلا و برگر(1990) به‌کمک آنتروپی کسری با آلفاهای گوناگون. بیشترین مقدار میانگین هارمونیک بازیابی و صحت در 0/8- $= \alpha$ رخ می‌دهد.

[1] Zipf's low



شکل3. میزان اهمیت واژه‌های مرجع [17] برحسب رتبهٔ آنها.



شکل4. صحت (لوزی‌های آبی)، بازیابی (مربع‌های قرمز) و میانگین هارمونیک آنها (مثلث‌های سبز) برای طول‌های نسبی گوناگون ($fr = N_{ret}/N_v$) از نمایه استخراج شده از کتاب آماری کسلا و برگر به‌کمک اهمیت تعمیم یافتهٔ کسری $WI_{\alpha=-0.8}$. بیشینه میانگین هارمونیک بازیابی و صحت در $fr = 0.05$ به‌دست می‌آید.

همچنین شکل4 رفتار نمودارهای صحت، بازیابی و میانگین هارمونیک این دو معیار را برحسب کسر نسبت تعداد لغت اول به‌دست آمده از لیست به تعداد کل لغات کتاب $fr = N_{ret}/N_v$ نشان می‌دهد. این شکل نیز با استفاده از آنتروپی کسری با مقدار بهینهٔ $\alpha = -0.8$، رسم شده است. همان‌طور که در شکل مشاهده می‌شود، بیشینهٔ میانگین هارمونیک بازیابی و صحت در $fr = 0.05$ به‌دست می‌آید. یعنی بهترین انتخاب این است که پنج درصد از واژه‌های ابتدای لیست مرتب شدهٔ واژگان کتاب را به‌عنوان نمایهٔ آن برگزینیم.

در جدول1 با اهمیت‌ترین و کم‌اهمیت‌ترین لغات کتاب کسلا و برگر لیست شده‌اند. در سمت چپ و راست این جدول به‌ترتیب پنجاه لغت با مقدار بیشینه و کمینهٔ اهمیت کسری تعمیم‌یافته با مقدار بهینه آلفا، $WI_{\alpha=-0.8}$ مرتب شده‌اند. همان‌طور که در این جدول مشاهده می‌شود، پنجاه لغت کم اهمیت، دارای اهمیت کسری تعمیم‌یافته‌ای با سه مرتبهٔ بزرگی کوچک‌تر از پنجاه لغت پراهمیت، می‌باشند

در ادامه سعی کردیم مرتب‌سازی لغات توسط اهمیت کسری تعمیم‌یافته لغات با آلفای بهینه، $WI_{\alpha=-0.8}$، را با امتیازدهی لغات توسط آنتروپی متداول شنون، $WI = WI_{\alpha=0}$ را در جدول2 مقایسه نماییم. برای این کار، سی‌وپنج جفت واژه از کتاب انتخاب نموده‌ایم. در هر جفت یک واژهٔ کلیدی (فونت سیاه) و یک واژهٔ دستوری (فونت کج) با بسامدهای نزدیک به‌هم قرار دارند. در ستون نخست جفت واژه‌ها به‌همراه بسامدشان نوشته شده‌اند. ستون دوم و سوم هم نتایج مرتب‌سازی آن‌ها به‌ترتیب با آنتروپی کسری و آنتروپی شنون نمایش می‌دهند.

برای مقایسه کمی توانایی روش‌های استخراج نمایه کتاب، با استفاده آنتروپی کسری و آنتروپی شنون، مقادیر بازیابی، صحت و میانگین هارمونیک آن‌ها را برای کتاب آمار توصیفی کسلا-برگر به‌دست آوردیم. مقادیر مربوطه در جدول3 ارائه شده‌اند. ردیف اول نشان‌دهندهٔ نتایج به‌دست آمده برای اهمیت لغات تعمیم‌یافته با آلفای بهینه، $WI_{\alpha=-0.8}$، است. همچنین ردیف دوم نتایج به‌دست آمده برای اهمیت لغات به‌دست آمده از آنتروپی شنون، WI، به‌دست آمده‌اند. کوچک‌تر بودن جمعیت نسبی نمایهٔ استخراج شده، fr ، توسط آنتروپی کسری و همچنین بزرگ‌تر بودن بازیابی و میانگین بازیابی و صحت بیان‌گر این است که این نوع آنتروپی ابزار جامع‌تری از آنتروپی شنون برای

تشخیص واژگان کلیدی از واژه‌های دستوری و کم‌محتوا است.

بحث و نتیجه‌گیری

آنتروپی کسری نتیجه‌ای از حسابان کسری می‌باشد. با استفاده از این آنتروپی به معرفی اهمیت تعمیم‌یافته کسری لغات از مرتبهٔ α، WI_α، پرداختیم، که در حالت خاص $\alpha = 0$، با اهمیت لغات به‌دست آمده از آنتروپی متداول شنون، WI برابر است. سپس با استفاده از این معیار به استخراج لغات با اهمیت کتاب آماری کسلا و برگر پرداختیم. نتایج حاصل نشان دادند که مقدار بهینه پارامتر برای آنتروپی کسری برای متن‌کاوی کتاب مورد مطالعه برابر با $\alpha = -0/8$، است.

همان‌طور که به‌صورت کیفی در جدول2 مشاهده می‌شود، روش رتبه‌بندی کلمات پیشنهادی از روش رتبه‌بندی با استفاده از آنتروپی شنون، در جداسازی لغات مهم یک متن بهتر عمل می‌کند. همچنین جدول 3 به‌صورت کمی نیز نشان که آنتروپی کسری، نسبت به آنتروپی متداول شنون، ابزار جامع‌تری در متن‌کاوی است.

لازم به‌ذکر است که مقدار بهینهٔ آلفا برای سامانه‌های گوناگون می‌تواند متفاوت باشد و یافتن این مقدار بهینه نخستین گام برای استفاده از الگوریتم ارائه شده است. برای این کار لازم است که یک نمایهٔ از پیش تعیین شده در اختیار داشته باشیم یا یک نمایهٔ فرضی برای متن در نظر بگیریم تا به کمک آن مقدار بهینهٔ آلفا را تخمین بزنیم. یافتن مقدار بهینهٔ پارامتر آلفا و مفهوم آن برای سامانه های مختلف یک مسئلهٔ باز می‌باشد و نیازمند پژوهش‌های بیشتر است.

تشکر و قدردانی

از داوران محترم که نظرات سازندهٔ آنها باعث بهبود علمی مقاله شد کمال تشکر را داریم. همچنین نویسندگان مقاله مراتب قدردانی خود را از حمایت دانشگاه صنعتی نوشیروانی بابل از طریق اعتبارهای پژوهشی با شماره‌های BNUT/390012/98، BNUT/392100/98 و BNUT/391023/98 اعلام می‌دارند.

مرجع‌ها

[1] C.D. Manning, H. Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, (1999).

[2] M.W. Berry, J. Kogan, *Text Mining Applications and Theory*, Wiley, New York, (2010).

[3] M. Ortuno, P. Carpena, P. Bernaola-Galvan, E. Munoz, A.M. Somoza, Keyword detection in natural languages and DNA, *Europhysics Letter* **57** (2002) 759-764. https://doi.org/10.1209/epl/i2002-00528-3

[4] H. Zhou, G.W. Slater, A metric to search for relevant words, *Physica A* **329** (2003) 309-327. https://doi.org/10.1016/S0378-4371(03)00625-3

[5] P. Carpena, P. Bernaola-Galvan, M. Hackenberg, A.V. Coronado, J.L. Oliver, Level statistics of words: Finding keywords in literary texts and symbolic sequences, *Physical Review E* **79** (2009) 035102. https://doi.org/10.1103/PhysRevE.79.035102

[6] J.P. Herrera, P.A. Pury, Statistical keyword detection in literary corpora, *European Physical Journal B* **63** (2008) 135-146. https://doi.org/10.1140/epjb/e2008-00206-x

[7] Z. Yang, J. Lei, K. Fan, Y. Lai, Keyword extraction by entropy difference between the intrinsic and extrinsic mode, *Physica A* **392** (2013) 4523-4531. https://doi.org/10.1016/j.physa.2013.05.052

[8] A. Mehri, A.H. Darooneh, The role of entropy in word ranking, *Physica A* **390** (2011) 3157-3163. https://doi.org/10.1016/j.physa.2011.04.013

[9] A. Mehri, M. Jamaati, H. Mehri, Word ranking in a single document by Jensen-Shannon divergence, *Physics Letters A* **379** (2015) 1627-

[15] D. Baeanu, K. Diethelm, E. Scalas, J.J. Trujillo, *Fractional Calculus*, world Scientific, Singapore, (2012).

[16] G.B. Bagci, The third law of thermodynamics and the fractional entropies, *Physics Letters A* **380** (2016) 2615-2618. https://doi.org/10.3390/e16042350

[17] G. Casella, R.L. Berger, *Statistical Inference*, Wadsworth, California, (1990).

[18] A. Mehri, H. Agahi, H. Mehri-Dehnavi, A novel word ranking method based on distorted entropy, *Physica A: Statistical Mechanics and its Applications*, **521** (2019) 484-492. DOI: https://doi.org/10.1016/j.physa.2019.01.080

[19] D.L. Olson, D. Delen, *Advanced Data Mining Techniques*, Springer-Verlag, Berlin, (2008).

1632. https://doi.org/10.1016/j.physleta.2015.04.030

[10] R. Mihalcea, Random walks on text structures. CICLing 2006, LNCS, **3878** (2006) 249-262, *Springer Heidelberg*. https://doi.org/10.1007/11671299_27

[11] G. Zipf, *Human Behavior and the Principle of Least Effort: An introduction to Human Ecology*, Addison-Wesley Press, Cambridge, (1949).

[12] H.P. Luhn, The automatic creation of literature abstracts, *IBM Journal of Research and Development* **2** (1958) 159-165. https://doi.org/10.1147/rd.22.0159

[13] M. Mezard, A. Montanari, *Information, Physics and Computation*, Oxford University Press, Oxford, (2009).

[14] J.T. Machado, Fractional order generalized information, *Entropy*, **16** (2014) 2350-2361. https://doi.org/10.3390/e16042350

جدول1. پنجاه واژهٔ معنادار پرمحتوا/کم محتوا با بیشترین/کم‌ترین امتیاز از آغاز/پایان لیست مرتب شدهٔ واژگان کتاب آماری کسلا و برگر [17].

ردیف	واژه پر محتوا	امتیاز	واژه کم محتوا	100× امتیاز
1	confidence	0/9143	programming	0/1934
2	interval	0/8838	opinion	0/1922
3	estimator	0/8837	extend	0/1858
4	regression	0/8826	weaker	0/1852
5	test	0/8797	explanations	0/1833
6	statistic	0/8535	problems	0/1832
7	anova	0/8456	heavy	0/1821
8	mle	0/8415	radius	0/1799
9	versus	0/8392	arising	0/1741
10	principle	0/8379	field	0/1712
11	likelihood	0/8358	areas	0/1651
12	unbiased	0/8289	across	0/1614
13	interval	0/8280	accounted	0/1415
14	squares	0/8235	fine	0/1346
15	estimators	0/8164	chore	0/1278
16	asymptotic	0/8077	surmise	0/1155
17	linear	0/8069	appeared	0/1048
18	testing	0/8018	characterizes	0/1047
19	tests	0/8013	departure	0/0932
20	irt	0/7986	constitute	0/0871
21	minimal	0/7978	differently	0/0729
22	ump	0/7939	iterations	0/0705
23	coverage	0/7927	initial	0/0662
24	posterior	0/7921	normalized	0/0656
25	sufficient	0/7892	wisdom	0/0635
26	risk	0/7868	reality	0/0604
27	joint	0/7855	brief	0/0589
28	hypothesis	0/7832	unchanged	0/0508
29	prior	0/7812	comfortable	0/0504
30	loss	0/7767	rewritten	0/0494
31	error	0/7750	nor	0/0457
32	level	0/7696	individually	0/0457
33	acceptance	0/7661	presence	0/0435
34	mse	0/7647	suspect	0/0426
35	estimate	0/7645	methodologies	0/0405

ادامه جدول1

ردیف	واژه پر محتوا	امتیاز	واژه کم محتوا	100× امتیاز
36	family	0٫7627	eliminate	0٫0386
37	bayes	0٫7621	planning	0٫0361
38	ancillary	0٫7598	crucial	0٫0323
39	sided	0٫7574	simplification	0٫0215
40	credible	0٫7499	passes	0٫0141
41	data	0٫7445	disappears	0٫0107
42	families	0٫7426	unaffected	0٫0104
43	show	0٫7420	formidable	0٫0099
44	reduction	0٫7395	claimed	0٫0072
45	estimation	0٫7389	graphed	0٫0068
46	bootstrap	0٫7370	plethora	0٫0057
47	marginal	0٫7308	presents	0٫0035
48	algorithm	0٫7307	three	0٫0008
49	generating	0٫7276	powers	0٫0006
50	hypotheses	0٫7270	posed	0٫0001

جدول2. مقایسه نتایج رتبه‌بندی سی و پنج جفت واژه برگزیده با فرکانس تقریباً برابر کتاب آماری کسلا-برگر، با استفاده از آنتروپی کسری، WI -0/8 و آنتروپی شنون، WI. در ستون دوم جفت واژه‌ها به‌همراه بسامدشان نوشته شده‌اند. ستون چهارم و ششم، به‌ترتیب، نتایج مرتب‌سازی لغات ستون دوم با استفاده از آنتروپی کسری و آنتروپی شنون، را نمایش می‌دهند.

ردیف	واژگان	فرکانس	واژگان	WI ۰.۸-	واژگان	WI
1	An	1060	interval	0٫8838	Anova	0٫8617
2	random	1042	estimator	0٫8837	Pearson	0٫6509
3	have[1]	772	regression	0٫8826	regression	0٫6355
4	sample	761	statistic	0٫8535	coverage	0٫5869
5	not	702	anova	0٫8456	bootstrap	0٫5690
6	probability	696	likelihood	0٫8358	Linear	0٫5058
7	pdf	586	linear	0٫8069	bayesian	0٫4955
8	see	573	coverage	0٫7927	interval	0٫4880
9	estimator	551	hypothesis	0٫7832	Error	0٫4775
10	Or	550	error	0٫7750	hypothesis	0٫4536
11	will	545	data	0٫7445	estimator	0٫4368
12	set	491	bootstrap	0٫7370	likelihood	0٫4354
13	has	464	marginal	0٫7308	approximate	0٫4169

ادامهٔ جدول2

ردیف	واژگان	فرکانس	واژگان	WI ۰.۱۸	واژگان	WI
14	variance	458	approximate	0,7024	quadratic	0,4137
15	interval	454	pdf	0,6949	statistic	0,4101
16	given	432	sample	0,6926	marginal	0,3890
17	mean	410	Variance	0,6848	my	0,3674
18	at	408	poisson	0,6557	independence	0,3609
19	statistic	403	probability	0,6536	Maximum	0,3259
20	thus	398	bayesian	0,6528	lehman	0,3127
21	data	301	maximum	0,6506	cauchy	0,3050
22	these	296	exponential	0,6462	poisson	0,2991
23	than	283	set	0,6423	data	0,2981
24	binomial	276	binomial	0,6370	samples	0,2752
25	now	269	random	0,6260	you	0,2743
26	likelihood	266	statistics	0,5996	exponential	0,2599
27	regression	237	independence	0,5880	usual	0,2529
۲۸	consider	۲۳۴	cauchy	0,5861	statistics	۰,۲۴۱۴
29	however	204	mean	0,5771	variances	0,2371
30	exponential	204	pearson	0,5453	binomial	0,2370
31	would	167	samples	0,5116	variance	0,2190
32	statistics	167	parameters	0,5053	pdf	0,2168
33	poisson	160	you	0,4956	set	0,2070
34	was	159	lehman	0,4922	sample	0,1907
35	hence	157	quadratic	0,4842	probability	0,1834
36	linear	156	variances	0,4829	parameters	0,1824
37	hypothesis	150	will	0,4549	should	0,1698
38	its	149	assume	0,4540	mean	0,1658
39	what	142	my	0,4523	without	0,1634
40	parameters	141	usual	0,4497	assume	0,1620
41	marginal	133	not	0,4469	far	0,1611
42	how	131	an	0,4462	random	0,1477
43	error	128	should	0,4450	found	0,1449
44	above	128	its	0,4409	its	0,1422
45	assume	116	thus	0,4385	was	0,1415
46	anova	116	or	0,4374	would	0,1394
47	maximum	107	at	0,4373	what	0,1361
48	here	107	was	0,4332	next	0,1318
49	coverage	103	would	0,4329	how	0,1300
50	way	102	given	0,4165	above	0,1187
51	approximate	98	what	0,4132	were	0,1187
52	should	97	have	0,4110	somewhat	0,1145
53	were	83	has	0,4025	At	0,1119

ادامهٔ جدول2

ردیف	واژگان	فرکانس	واژگان	WI -0/8	واژگان	WI
54	cauchy	83	How	0,4014	way	0,1093
55	found	69	however	0,3892	hence	0,1082
56	bootstrap	69	These	0,3890	however	0,1073
57	variances	61	consider	0,3816	will	0,1073
58	Next	61	without	0,3733	here	0,1061
59	without	58	Than	0,3726	or	0,1046
60	samples	58	hence	0,3696	consider	0,1042
61	you	51	Now	0,3646	thus	0,1009
62	independence	51	found	0,3630	given	0,0998
63	somewhat	47	See	0,3612	not	0,0979
64	bayesian	47	above	0,3578	than	0,0971
65	usual	35	Here	0,3353	these	0,0968
66	lehman	35	Were	0,3342	has	0,0954
67	quadratic	22	Way	0,3313	now	0,0922
68	my	22	Next	0,3279	an	0,0914
69	pearson	20	somewhat	0,2745	have	0,0841
70	far	20	Far	0,2583	see	0,0754

جدول3. مقایسه کمی نتایج حاصل از آنتروپی کسری و آنتروپی شنون در استخراج نمایه برای کتاب آمار توصیفی کسلا-برگر. ردیف اول نشان‌دهندهٔ نتایج به‌دست آمده با استفاده از اهمیت لغات تعمیم‌یافته WI -0/8، ردیف دوم نتایج به‌دست آمده با استفاده از اهمیت لغات WI به‌دست آمده‌اند.

	جمعیت نسبی نمایه	بازیابی	صحت	میانگین هارمونیک بازیابی و
آنتروپی کسری	0,05	0,45	0,54	0,49
آنتروپی شنون	0,56	0,05	0,64	0,09