نوشته‌ها

بعد فراکتالی واژگان در یک نوشته

امیرحسین درونه *

پیشگفتار

پیشرفت‌های فناوری در دو دهه گذشته سبب دسترسی ما به انبوهی از اطلاعات شده است که بیشتر آنها اسنادی نوشتاری هستند. یافتن اطلاعات مورد نیاز به صورت دستی در این انبوهه، کاری دشوار است. روش‌های گوناگونی تاکنون معرفی شده‌اند تا بتوان اطلاعات مهم یک نوشته یا شماری از نوشته‌ها را به شکل خودکار بیرون کشید. مجموعه این روش‌ها نوشته کاوی [۱] نامیده می‌شود. نوشته‌کاوی موضوعی میان رشته‌ای است که نخست در علوم رایانه مطرح شد و سپس دانش‌پیشگان دیگر حیطه‌ها مانند زبان‌شناسی محاسباتی، ریاضی، آمار و فیزیک نیز به آن علاقه‌مند شدند. رتبه بندی واژگان بر پایه اهمیت آنها در یک نوشته از موضوعات مهم نوشته کاوی است که می‌توان از آن در یافتن کلید واژگان، چکیده‌سازی نوشته، دسته‌بندی نوشته‌ها و دیگر موضوعات نوشته‌کاوی سود برد. تاکنون روش‌های گوناگونی برای رتبه‌بندی واژگان پیشنهاد شده‌اند.

اگر واژه‌های یک نوشته را بر حسب فراوانی (تعداد تکرار) آنها مرتب کنیم. به فراوان‌ترین واژه رتبه یک و دومین واژه فراوان رتبه دو و به همین ترتیب به باقی واژه‌ها رتبه اختصاص دهیم. در این فهرست مرتب شده از واژگان، واژه‌هایی مانند the، of و a برای زبان انگلیسی در بالای فهرست قرار می‌گیرند. این واژه‌ها در تمام نوشته‌ها یافت می‌شوند و مفهوم خاصی ندارند. اما واژه‌های پر اهمیت که به موضوع نوشته ربط دارند و معنای مورد نظر نویسنده را به خواننده منتقل می‌کنند معمولاً در رتبه‌های میانی قرار می‌گیرند. می‌توان از این واقعیت برای استخراج واژه‌های پر اهمیت سود برد [۱] . وجود واژه‌های کم اهمیت در رتبه‌های میانی دقت این روش را کاهش می‌دهد. افزون بر این محدوده میانی فهرست تعریف دقیقی ندارد. به صورت تجربی دیده شده است که واژه‌های مهم نوشته خوشه تشکیل می‌دهند یعنی در برخی از جاهای نوشته نزدیک یکدیگر دیده می‌شوند و توزیع یکنواختی در طول نوشته ندارند [۲] . اگر فاصله بین ظهور متوالی یک واژه را بر حسب تعداد واژه میان آنها در نوشته اندازه‌گیری کنیم، می‌توان میزان افت و خیز آماری فواصل ظاهر شدن یک واژه را همچون سنجه‌ای برای اندازه‌گیری میزان خوشه شدن آن و در نتیجه رتبه‌بندی واژگان بکار برد [۳] . با استفاده از مکانیک آماری می‌توان نشان داد تابع توزیع فواصل ظهور متوالی یک واژه شکل $(1+(1-q)\beta x)^{\frac{1}{1-q}}$ دارد که در آن x متغیر فاصله، q و β دو پارامتر حقیقی هستند. می‌توان پارامتر q را معیاری برای اندازه‌گیری خوشه شدن دانست و واژگان یک نوشته را بر حسب آن رتبه‌بندی کرد [۴] . ویژگی خوشه‌بندی سبب می‌شود فراوانی واژه در بخش‌های گوناگون نوشته یکسان نباشد. آنتروپی سنجه خوبی برای اندازه‌گیری این ناهمسانی است. آنتروپی به شکل $-\sum_i f_i(w)\ln f_i(w)$ تعریف می‌شود که در آن $f_i(w)$ نسبت فراوانی واژه w در بخش i-ام به فراوانی همان واژه در کل نوشته است. از آنتروپی برای جداسازی نوشته‌های طبیعی و کاتوره‌ای (تصادفی) استفاده شده است [۵، ۶] . گونه‌های دیگر آنتروپی نیز که بر حسب توزیع فواصل واژگان هستند نیز برای رتبه‌بندی پیشنهاد شده است [۷] . هر نوشته را می‌توان با یک گراف نمایش داد. مجموعه واژه‌های مستقل رأس‌های گراف هستند. مجاورت دو واژه در نوشته را با یالی میان رأس‌های آنان در گراف نشان می‌دهیم. با کاربست الگوریتم رتبه پیج [۲] بر روی گراف وابسته به یک نوشته، می‌توان واژگان آن را رتبه‌بندی کرد .[۸] این الگوریتم همانی است که گوگل از آن برای مرتب سازی نتایج جستجوی خود سود می‌برد.

در این نوشتار ما با دیدگاهی نو به رتبه‌بندی واژگان می‌پردازیم. بر پایه این دیدگاه فرض می‌کنیم مکان‌هایی که یک واژه در نوشته ظاهر می‌شود الگویی خودهمانند را پدید می‌آورد. هر الگوی خود همانند دارای بعدی فراکتالی است که می‌توان به صورت عملی آن را محاسبه کرد. در یک نوشته که به گونه‌ای تصادفی درهم‌ریخته [۳] شود بعد فراکتالی همه واژه‌ها نزدیک یک است اما در نوشته‌ای که دارای

[1] Text Mining [2] Page Rank [3] Shuffled Text

مفهوم است، واژه‌هایی که بار معنایی نوشته را به دوش می‌کشند دارای بعد فراکتالی کمتر از یک هستند. بعد فراکتالی واژه می‌تواند همچون سنجه‌ای برای رتبه‌بندی واژگان بکار آید.

در بخش بعدی در این باره بیشتر خواهیم گفت و در بخش پایانی از داده‌های تجربی برای پشتیبانی از این فرض استفاده می‌کنیم.

نوشته‌ها و فراکتال‌ها

زبان بشری سامانه‌ای پیچیده است که واژه‌ها اجزای بنیادین آن هستند. زبان سامانه‌ای بسته نیست بلکه ارتباط تنگاتنگی با پیرامون خود دارد.با گذشت زمان و ساخته شدن مفاهیم جدید تعداد واژگان یک زبان تغییر می‌کند و در گذر زمان حتی معنی برخی از واژه‌ها نیز دگرگون می‌شود. نوشته‌ها نیز بخشی مهم از زبان هستند که پیچیدگی آن را به ارث می‌برند. توزیع فراوانی واژه‌ها، رشد تعداد واژه‌های مستقل با افزایش طول نوشته و برخی دیگر از ویژگی‌های آماری، نمودهای این پیچیدگی هستند. در یک نوشته واژه‌ها در حال برهم‌کنش با یکدیگرند. این برهم‌کنش به دو گونه دستوری و معنایی است. برهم‌کنش دستوری در یک نوشته سبب می‌شود که واژه‌ها در یک جمله به ترتیب ویژه‌ای کنار یکدیگر قرار گیرند. این برهم‌کنش کوتاه برد است و برد آن اندازه یک یا دو جمله است. برهم‌کنش بین برخی از واژه‌ها در نوشته برای انتقال مفهوم به خواننده، برهم‌کنش معنایی نام دارد. این برهم‌کنش بلند برد است و چینش واژه‌ها را در سراسر نوشته هماهنگ می‌کند. برهم‌کنش معنایی سبب پیدایش الگویی فضایی برای هر واژه می‌شود. در این نوشتار فرض می‌کنیم این الگو دارای ویژگی خود همانندی [۴] یا فراکتالی است، همان گونه که در بسیاری از سامانه‌های پیچیده دیده می‌شود. بعد این فراکتال‌ها را می‌توان با روش جعبه‌شماری [۵] اندازه گرفت. برای این کار نوشته را به صورت دنباله یک بعدی از واژگان می‌انگاریم. هر s واژه متوالی را درون یک جعبه فرضی قرار می‌دهیم. بدین گونه نوشته به N_s جعبه تقسیم می‌شود. شکل نمایی از این جعبه‌بندی را نشان می‌دهد. شمار جعبه‌هایی را که یک واژه معین در آن یافت می‌شود را با نماد $N_w(s)$ نشان می‌دهیم. اگر توزیع هر واژه الگویی خود همانند داشته باشد، برای اندازه‌های بزرگ s خواهیم داشت:

$$N_w(s) \approx s^{-D} \qquad (۱.۲)$$

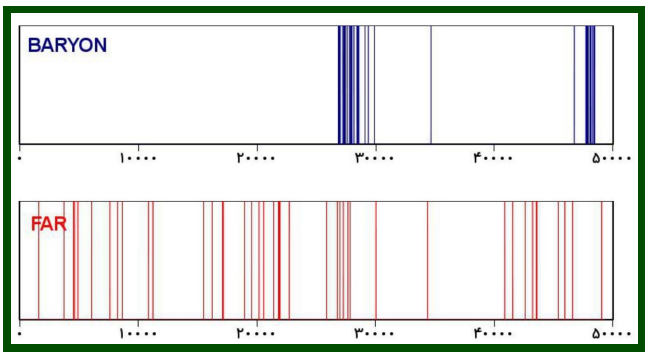
که D بعد فراکتالی واژه نام دارد.

برای محاسبه D کافی است نمودار تمام لگاریتمی $N_w(s)$ را بر حسب s رسم کنیم. شیب نمودار، اندازه D را بدست می‌دهد.

در بخش بعدی درستی این فرض را برای واژگان یک کتاب نمونه می‌آزماییم.

positrons was	almost exactly	equal to	the number	of electrons	in addition	to electrons	and positrons
positrons was almost exactly		equal to the number		of electrons in addition		to electrons and positrons	
positrons was almost exactly equal to the number				of electrons in addition to electrons and positrons			

شکل ۱: سیمایی از جعبه‌بندی برای بخشی از نوشته. در بالا اندازه جعبه ۲ و در لایه میانی و پایینی به ترتیب ۴ و ۸ می‌باشد .



شکل ۲: نمای طیفی دو واژه BARYON و FAR . خوشه شدن و توزیع یکنواخت را برای این دو واژه می‌توان دید. هر دو واژه ۴۴ بار در نوشته تکرار شده‌اند.
. خوشه شدن و توزیع یکنواخت را برای این دو واژه می‌توان دید. هر دو واژه ۴۴ بار در نوشته تکرار شده‌اند.

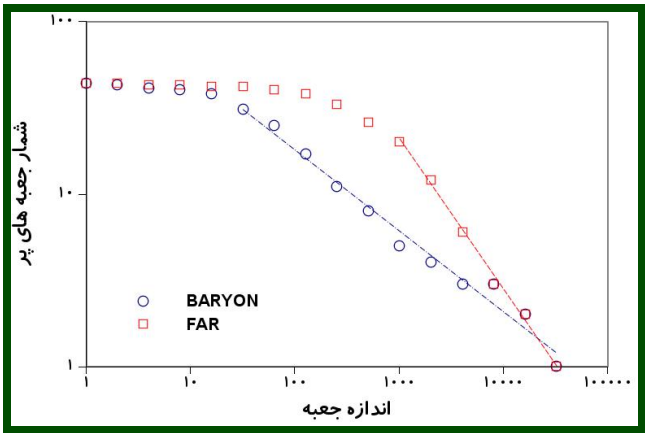
اندازه‌گیری بعد فراکتالی واژگان

در این بخش کتاب ارزشمند؛ سه دقیقه نخست؛ نوشته استیون واینبرگ [۹] را برای بررسی برمی‌گزینیم. این کتاب درباره آغاز کیهان بر پایه نظریه مهبانگ (انفجار بزرگ) [۶] و الگوی متعارف [۷] حاکم بر برهم‌کنش ذرات بنیادی است و چگونگی پیدایش ماده را در سه دقیقه آغاز جهان شرح می‌دهد. طول این کتاب ۴۹۸۶۰ واژه است و واژه‌نامه‌ای با اندازه ۴۰۳۹ واژه دارد. اندازه جعبه ها را توانی از ۲ برمی‌گزینیم. شکل سیمایی از جعبه‌ها را برای بخشی از این نوشته نشان می‌دهد.

چیدمان منحصر به فردی از واژگان در نوشته است که معنای دلخواه نویسنده را به خواننده منتقل می‌کند. هرگونه به‌هم ریختگی در این چینش معنای آن را از بین می‌برد. اگر توزیع مکانی واژه‌ای با به‌هم ریختن چیدمان واژگان تغییر چشمگیری نداشته باشد بدین معنی است که آن واژه اهمیت کمی در انتقال معنای نوشته دارد. در نوشته‌ای که به صورت تصادفی به‌هم ریخته شود، واژگان همه

[4]S [5]Box Counting [6]Big Bang [7]Standard Model

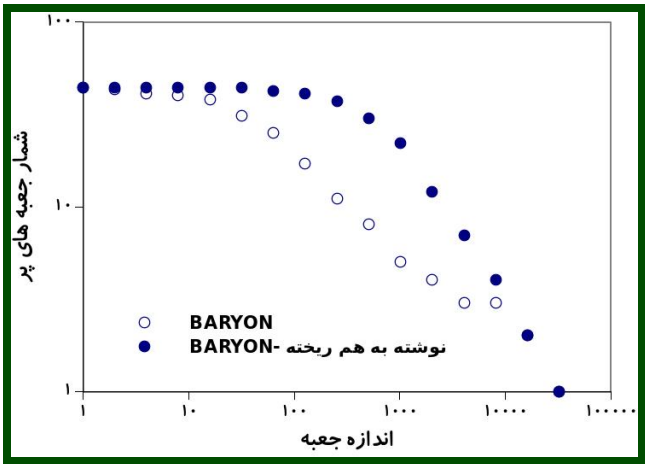
توزیعی یکنواخت دارند بنابر این واژگان کم اهمیت نیز در نوشته اصلی دارای توزیعی یکنواخت هستند. شکل نمایی طیفی از دو واژه BARYON و FAR را نشان می‌دهد. هر دو واژه ۴۴ بار در نوشته تکرار شده‌اند. نخستین واژه در ارتباط با مفاهیم کتاب است و دومی واژه‌ای عمومی است که در هر نوشته‌ای می‌توان آن را یافت. جایگاه‌های ظاهر شدن این واژه‌ها در نوشته با خط‌های پررنگ مشخص شده‌اند. در این شکل، خوشه شدن برای واژه نخست و توزیع یکنواخت را برای دومین واژه می‌توان دید



شکل ۳: جعبه‌شماری برای واژگان BARYON و FAR. خط‌چین و خط‌چین نقطه برازش انجام شده برای بخش خطی نتایج جعبه‌شماری برای این دو واژه را در نوشته اصلی نشان می‌دهند.

فرض کنید در نوشته‌ای به طول N، واژه کم اهمیت w دارای فراوانی M باشد. اگر این نوشته را با جعبه‌هایی به اندازه s واژه افراز کنیم $N_s = N/s$ جعبه خواهیم داشت. برای sهایکوچک ($N_s > M$) می‌دانیم $N_w(s) \simeq M$ است چون توزیع یکنواخت است و اثر خوشه‌بندی وجود ندارد که احتمال قرار گرفتن دو یا چند واژه در یک جعبه را افزایش دهد بنا بر این در هر جعبه یا یک واژه‌است و یا واژه‌ای یافت نمی‌شود. برای s های بزرگ ($N_s < M$) به ناچار همه جعبه‌ها دست کم یک رخداد از واژه را در خود باید جای دهند بدین گونه $N_w(s) = N_s$ می‌شود یعنی $N_w(s) \propto 1/s$ است. با توجه به معادله ؟؟ بعد فراکتالی این دسته از واژگان برابر یک می‌شود.

برای واژه‌های پر اهمیت اگر چه هنگامی که $s = 1$ است $N_w(s) = M$ می‌شود ولی با بزرگ شدن s به تندی از این مقدار فاصله می‌گیرد. برای sهای بزرگ اثر خوشه‌بندی سبب می‌شود که $N_w(s) \ll N_s \propto 1/s$ باشد و بعد فراکتالی آن کمتر از یک شود. شکل پیامد جعبه‌شماری برای دو واژه BARYON و FAR را نشان می‌دهد. همان‌گونه که در شکل دیده می‌شود بعد فراکتالی واژه FAR از واژه BARYON بیشتر است. برازش انجام شده نشان می‌دهد بعد واژه نخست نزدیک ۱ و دومی برابر ۰/۶ است.



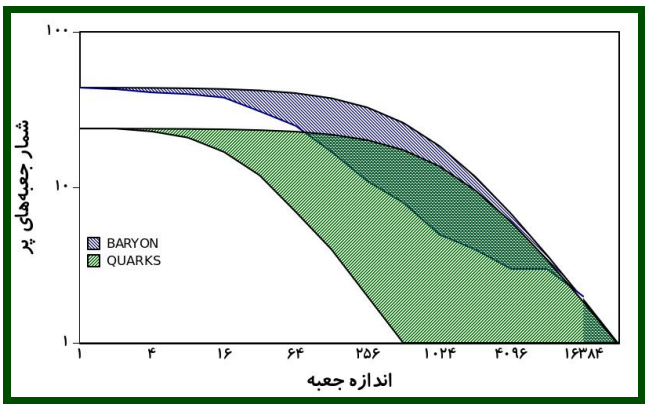
شکل ۴: BARYON در نوشته اصلی و به‌هم ریخته. همان گونه که دیده می‌شود در نوشته به‌هم ریخته بعد فراکتالی این واژه نزدیک ۱ است و با بعد فراکتالی آن در نوشته اصلی تفاوت چشمگیری دارد.

شکل مقایسه‌ای است بین شمار جعبه‌های پر برای واژه پر اهمیت BARYON در متن اصلی و متنی که به صورت تصادفی به‌هم ریخته شده است. در متن به‌هم ریخته بعد فراکتالی این واژه نزدیک ۱ است و در متن اصلی به صورتی چشمگیر کمتر از آن است.

همان گونه که در ابتدای این نوشته اشاره کردیم، تلاش ما برای یافتن راهی است که به صورتی خودکار واژگان را رتبه‌بندی کند. نوشتن الگوریتمی که خودکار بعد فراکتالی را محاسبه کند دشوار است. برای غلبه بر این دشواری به جای استفاده از بعد فراکتالی از کمیتی دیگر بهره می‌بریم که با آن رابطه دارد. این کمیت سطح بین نمودار تمام لگاریتمی جعبه‌شماری یک واژه در متن اصلی و به‌هم ریخته است. اگر واژه‌ای کم اهمیت باشد توزیع آن در نوشته اصلی و به‌هم ریخته تفاوت کمی با یکدیگر دارند که پیامد آن نزدیکی نمودارهای جعبه‌شماری آن در نوشته اصلی و به‌هم ریخته است. بدین گونه برای واژگان کم اهمیت اندازه این سطح ناچیز است. در شکل این دو سطح را برای دو واژه پر اهمیت BARYON و QUARKS نشان داده‌ایم.

جدول ۱.۲: فهرست ۲۰ واژه نخست که بر اساس اندازه سطح‌شان مرتب شده‌اند به همراه فراوانی آن‌ها در سمت راست ونیز فهرست مرتب شده از ۲۰ واژه نخست بر اساس فراوانی آن‌ها به همراه اندازه سطحشان در سمت چپ.

اندازه سطح	فراوانی	واژه	اندازه سطح	فراوانی	واژه
۳۹۷۴.۲	۴۱۵۹	the	۸۰۸.۱۷	۲۴	quarks
۴۹۶۱.۲	۲۴۴۷	of	۸۴۸۸.۱۶	۱۴	ions
۶۱۰۳۷.۲	۱۱۵۴	to	۲۹۵.۱۵	۷	diagram
۵۴۶۲.۲	۱۰۷۸	a	۷۲۷۲.۱۳	۷	rotating
۵۵۷۷.۲	۱۰۶۰	in	۴۰۲۲.۱۳	۱۰	quark
۵۶۳۵.۲	۹۴۵	and	۸۷۹۹.۱۲	۷	hydroxyl
۴۹۲۶.۲	۸۹۸	is	۴۴۶۱.۱۲	۱۴	nebulae
۶۷۵۷.۲	۷۳۱	that	۲۳۹.۱۲	۱۰	kmsec
۲۱۵۷.۲	۵۰۲	universe	۲۲۷.۱۲	۳۵	antenna
۵۲۷۳.۲	۴۳۴	this	۱۴۹۳.۱۲	۶	stone
۵۵۶۳.۲	۴۲۲	it	۰۷۱۲.۱۲	۴	slab
۹۵۲۴.۲	۴۱۸	as	۵۰۸.۱۱	۲۴	luminosity
۶۰۹۴.۲	۳۹۵	be	۳۸۷۳.۱۱	۱۰	cyanogen
۵۳۲۳.۲	۳۹۱	at	۳۷۸.۱۱	۴	angstroms
۸۱۷۳.۲	۳۶۲	by	۲۶۰۳.۱۱	۴	circumference
۹۰۴۹.۲	۳۴۳	we	۹۸۹۵.۱۰	۲۳	conservation
۱۷۳۴.۲	۳۲۵	for	۴۷۴۵.۱۰	۸	mc
۸۶۲۳.۲	۲۹۹	are	۴۴۳۶.۱۰	۱۸	planck
۱۲۷۵.۲	۲۹۳	was	۳۹۷۲.۱۰	۲	candidate
۷۵۶۴.۲	۲۸۹	with	۳۷.۱۰	۶	disc



شکل ۵: سطح بین نمودارهای تمام لگاریتمی جعبه‌شماری برای دو واژه پر اهمیت BARYON و QUARKS. واژه دوم دارای فراوانی ۲۴ در نوشته است ولی اندازه سطح بزرگتری دارد.

با آن که واژه QUARKS تنها ۲۴ بار تکرار شده اما اندازه سطح بزرگتری دارد.

در جدول ۱.۲ فهرستی از واژگان این کتاب دیده می‌شود که بر اساس اندازه سطحشان و نیز فهرستی از واژگانی که بر اساس فراوانی مرتب شده‌اند دیده می‌شوند. به سادگی می‌توان درستی این روش در رتبه‌بندی واژگان بر اساس اهمیت آن‌ها را تشخیص داد. خوانندگان علاقه‌مند را برای ارزیابی دقیق‌تر به مقاله نجفی و درونه [۱۰] ارجاع می‌دهیم.

مراجع

[1] Luhn HP. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2. 1958: 159–165.
doi: 10.1147/rd.22.0159

[2] Ortuño M, Carpena P, Bernaola-Galvan P, Munoz E, Somoza AM. Keyword detection in natural languages and DNA. Europhysics Letters 57. 2002: 759–764.
doi:10.1209/epl/i2002-00528-3

[3] Zhou H, Slater GW. A metric to search for relevant words. Physica A 329. 2003: 309–327.
doi: 10.1016/S0378-4371(03)00625-3

[4] Mehri A, Darooneh AH. Keyword extraction by nonextensivity measure. Physical Review E 83. 2011:056106.

doi: 10.1103/PhysRevE.83.056106

[5] Cohen A, Mantegna RN and Havlin S. Numerical analysis of word frequencies in artificial and natural language texts, Fractals 05. 1997: 95-104. doi:10.1142/S0218348X97000103

[6] Herrera JP, Pury PA. Statistical keyword detection in literary corpora. Eur. Phys. J. B 63. 2008: 135. doi: 10.1140/epjb/e2008-00206-x

[7] Mehri A, Darooneh AH. The role of entropy in word ranking. Physica A 390. 2011: 3157–3163. doi: 10.1016/j.physa.2011.04.013

[8] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts. Proceedings of conference on Empirical Methods in Natural Language Processing (EMNLP). 2004: 404–411.

[9] Weinberg S. The First Three Minutes. Cambridge: Pegasus Press; 1993.

[10] Najafi E, Darooneh AH. The Fractal Patterns of Words in a Text: A way for Automatic Keyword Extraction, PLoS ONE, 10(6). 2015:e0130617. doi:10.1371/journal.pone.0130617

* دانشگاه زنجان، گروه فیزیک