

بسمه تعالی

گزارش پروژه درس پردازش داده های حجیم

ابر کلمات برای کلمات همراه کلمات پر تکرار اینستاگرام
(کلماتی که بیشترین تکرار را در کنار کلمات پر تکرار داشته اند)

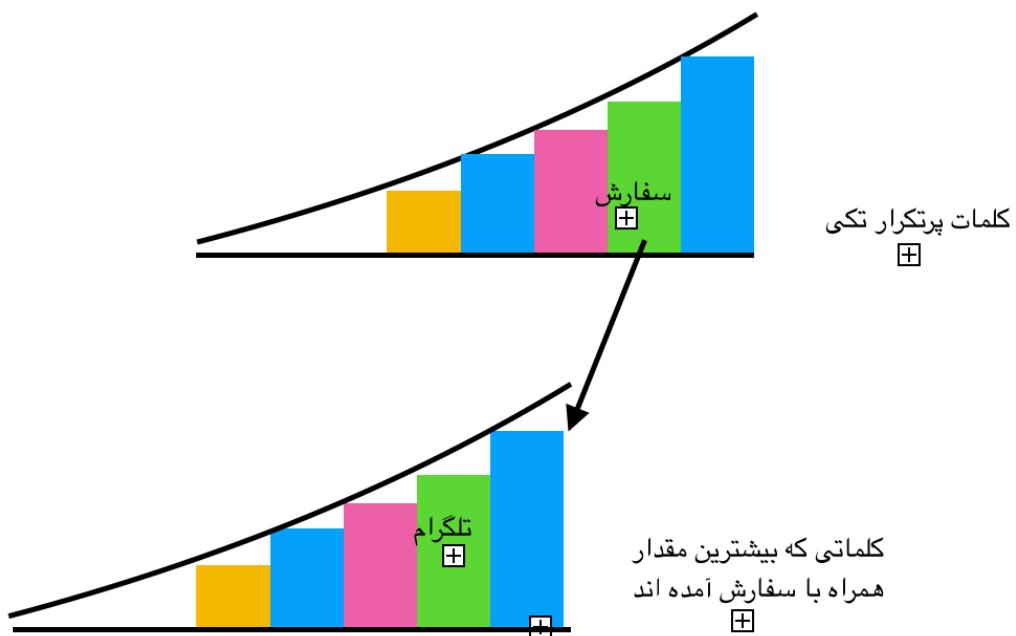
الگوریتم:

این پروژه پیاده سازی دو بعدی از پنجره میرا میباشد. که در سطر اول کلمات پرتکرار در کل کپشن های تشخیص داده میشوند و در لایه بعدی کلماتی که در کنار هر (در شعاع ۱۰ کلمه ای) کلمه تکرار شده اند می آیند و در پنجره میرا داخلی (لایه ۲) قرار میگیرند.

ساخت هر دولایه پنجره همزمان انجام میشود و کل داده ها تنها یکبار استریم میشوند.
در صورتی که یک کلمه از پنجره لایه اول خارج شود، پنجره لایه دوم مربوط به آن کلمه نیز حذف میشود.

شرایط پیاده سازی پنجره :

به دلیل اینکه تعداد کلمات زیاد است و ضرب کردن تکرارهای هر آیتم بر $(C-1)$ در هر مرحله زمانبر است در هر مرحله عددی که جمع میشود $(C-1) / 1$ برابر میشود و پس از مشاهده ۱۰۰۰ پست تمامی تکرارها نرمال میشوند.



بسمه تعالی

انتخاب پارامترها لایه اول:

(پارامترها بر اساس روش های تجربی و صحیح و خطا انتخاب شده اند اما محاسبات زیر هم کمک کننده بوده است.)

از آنجایی که در هر روز حدود ۵۰ تا ۱۰۰ میلیون پست در اینستاگرام منتشر میشود و نیز ۲۴ میلیون از ۸۰۰ میلیون تا ۱ میلیارد کابر ایرانی هستند.

میتوان گفت روزانه ۱ الی ۲ میلیون پست از ایران منتشر میشود. که(فرض) اگر همه‌ی یک میلیون پست هم بررسی شود و ارزش پست‌های هر روز از لحظه ترند بودن ۲ برابر روز قبلی باشد محاسبات زیر $c = 10^{**5}$ را به ما میدهد.

```
>>> a = 1 / (1 - .000001)
>>> a
1.000001000001
>>> a**1000000
2.71828318761652
```

و نیز چون به حدود ۱۰۰ هزار کلمه فارسی با معنا داریم پس با قرار دادن حد آستانه‌ی میرایی برابر با ۱۰ به عدد حد اکثر ۱۰۰ هزار کلمه در پنجره میرسیم .

Maximum number of words = $1 / th * c = 10^{**5}$

البته ما در اینجا به ازای هر پست یکبار ارزش اعداد را تغییر میدهیم و متمین نیستم که شرایط ریاضی حاکم بر قضیه اصلی که طبق رابطه هندسی در میاید اینجا نیز برقرار باشد. (ولی کار میکند).

```
post index 419999 total words: 19352 len words: 2621544 total counters: 1135382 mean counter: 58.67000826787929 salam len : 313 94.104160745110
28 [(('28.619795633816086', ('آبجه', 'لبن'), ('30.463646706424893', ('خوبی', 'برانو', '31.588111239595356', 'ما', 'u200c3
3.15797645085788), ('33.36630329489955', ('بسیار', 'دوستان', '41.159588023050546', ('سبز', 'پیش', '94.10416074511028
', 142.30488156615343)) c is 1.085544939719912
layer1add: 1.001000500667239 1.5219603532873376
```

طبق عکس بالا : تعداد کلمات در لایه اول ۲۰ هزار کلمه و تعداد کل شمارندها ۱.۱ میلیون است.
و در لایه دوم نیز به ازای هر کلمه به طور میانگیم ۶۰ عدد پر تکرار وجود دارد.

بسمه تعالی

پارامترهای لایه دوم:

تجربی بدست آمده اند

پارامترها C با دو مقدار 0.001 و 0.002. حساب شده است و نتایج در فolder هایی با همین نام آمده است.

نکته مهم در این تغییر . تغییر سایز پنجره لایه دوم است.

و نیز C بیشتر به دادههای جدید حساسیت بیشتری دارد .

به نظر خودم $C = 0.001$ مناسب تر است

The screenshot shows two windows displaying text files. Both files have the title 'result.txt' and show frequency analysis results. The left window (L1 word) has a window size of 437 and 20 samples. The right window (L1 word) has a window size of 3233 and 20 samples. Both windows list words along with their frequencies and other parameters like 'frequency' and 'window size'. The right window's results are generally higher than the left window's.

Word	L1 word (Left)	L1 word (Right)
کار	frequency: 36.7568024737577	frequency: 361.4742828184887
جنس	frequency: 82.8796559722399	frequency: 911.9317277449188
سایز	frequency: 191.6568424852626	frequency: 2331.574038226057
شمان	frequency: 149.1497232	frequency: 3308.4466449610237
رسال	frequency: 74.642409916	frequency: 444.6091666666666
دانان	frequency: 39.316816906401184	frequency: 593.9049325524337
سایزهای	frequency: 27.067779852979776	frequency: 253.36599172969662
کشور	frequency: 62.35644537470187	frequency: 81.77323354580285
قد	frequency: 42.7438585083272	frequency: 233.25484901846667
مراء	frequency: 3.757672448310202	frequency: 45.207291926252694
شماره	frequency: 12.159746545135127	frequency: 268.2361689601753
تماس	frequency: 32.56713984636269	frequency: 2051.76938083308275
رنگ	frequency: 96.5512201729417	frequency: 415.5264546501514
لیست	frequency: 119.576337014195808	frequency: 228.6573307046451
عنوان	frequency: 69.089613451702	frequency: 81.95271800000001
دستورات	frequency: 87.87892692454407	frequency: 300.161690433905
ذکر	frequency: 22.6650958532695	frequency: 95.06654911681605
به	frequency: 12.07643035468715	frequency: 256.0918558877351
کل	frequency: 5.461526791881159	frequency: 200.22171731722605
نطاط		frequency: 258.8941057732453

عکس سمت راست C کوچکتری بزرگتری دارد

خروجی:

ورد کلود کلمات پر تکرار همراه را به ازای ۲۰ کلمه‌ی پر تکرار اصلی (لایه ۱) می‌توانید در فolder result مشاهده کنید.
فایل result.txt نیز شامل کلمات پر تکرار و تعداد تکرار آنها به همراه ۲۰ نمونه (مرتب نشده) از لایه دوم آن کلمه
آمده است.

مثال:

:L1 word
سفرارش
frequency: 13823.08507261614
window size 518
:20samples
سایز 8
frequency: 75.58600802401938
قیمت
frequency: 125.82426407955775
تومان 6
frequency: 105.42532313393886
رنگ 3
frequency: 56.28676037951883
ثبت
frequency: 94.78013283679957
اطلاعات
frequency: 17.426891351382647
تلگرام 3
frequency: 190.58682132058613
کانال 7
frequency: 31.216445224262227
اطلاع
frequency: 20.734727441016958
دایرکت
frequency: 204.16433318199566
پیام 7
frequency: 36.97521735832197
مناسب
frequency: 11.07929109163009
دوست
frequency: 7.633425794414566
سفرارش
frequency: 89.79608603169693
عزیز
frequency: 10.277490469112463
ولنتاین
frequency: 12.009807819729351
مشتری
frequency: 20.41974062882825
شماره
frequency: 49.25384075515014
ارسال
frequency: 126.57289217180063
پرداخت

:L1 word
قیمت
frequency: 15553.672920531886
window size 437
:20samples
کار
frequency: 36.7568024737577

- frequency: 82.8796559722399 جنس
- frequency: 191.6568424852626 سایز
- frequency: 228.0394441497232 تومان
- frequency: 74.64240980549539 ارسال
- frequency: 39.316816906491184 رایگان
- frequency: 27.96777962979776 سایزبندی
- frequency: 62.35644537470187 قیمت
- frequency: 42.7438585083272 قد
- frequency: 3.757672448310202 همراه
- frequency: 12.159746545135127 شماره
- frequency: 34.56713984636269 تماس
- frequency: 96.5512201729417 رنگ
- frequency: 17.782984914195808 ثبت
- frequency: 119.57633177021557 سفارش
- frequency: 69.08961345115912 تکرام
- frequency: 87.87892692454407 دایرکت
- frequency: 22.6650958532695 تخفیف
- frequency: 12.07643035468715 تهران
- frequency: 5.461526791881159 العاده

نمونه تصویر ابر کلمه قیمت:



نمونه تصویر ایر کلمه شب:



مقایسه این روش با روش آیتم ست های پر تکرار (نسخه آیتم سستهای دوتایی) :

تفاوت اول: در نوع خروجی این دو روشن است.

در روش آیتم سست های پر تکرار، شرط لازم برای پر تکرار بودن یک دو تایی پر تکرار بودن تک تک اعضا (نسبت به تمامی اعضا) است.

اما در این روش شرط لازم فقط پر تکرار بودن عضو اصلی (لایه اول است) و پر تکرار بودن عضو دوم فقط نسبت به کلماتی که در کنار عضو اول آماده اند سنجیده میشود و معیار سنجش عضو دوم تعداد تکرار کنار عضو اول است.

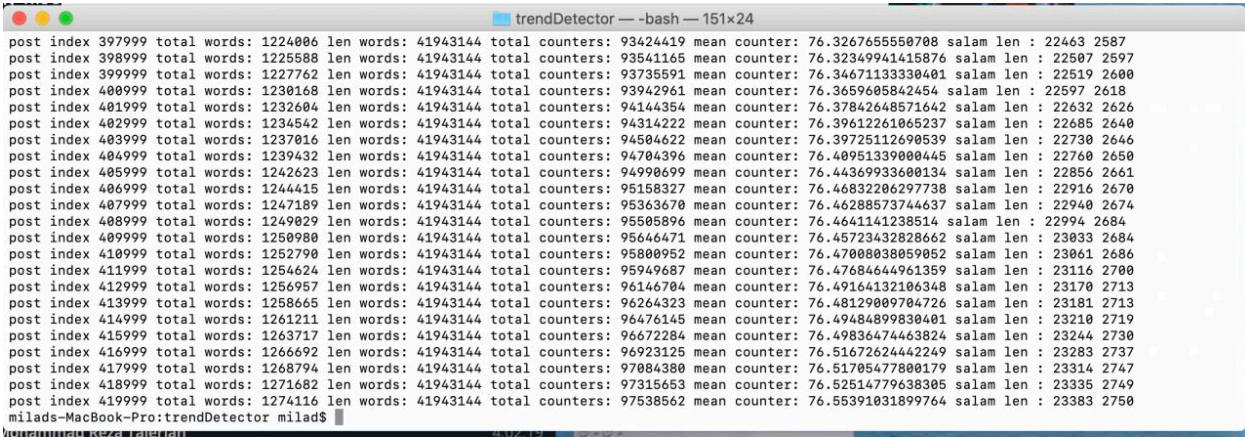
تفاوت دوم: این روش کلیه محاسبات را در یک پس انجام میدهد و نیاز به نگهداری دادهای نیست. و در شرایطی که داده ها استریم هستند و نمیتوان آنها را نگه داری کرد این روش میتواند استفاده شود.

در آینده انشا الله اگر فرصت شه ک مقاسه آماری بین این دو روش انعام ممکن شود.

کاربرد:

یک برنز میتواند مشاهده کند در کنار نام برنزش معمولاً چه کلماتی استفاده شده اند. و نیز مثلاً در کنار کلمه ای مانند بانک بیشتر چه کلماتی(نام چه بانک هایی) آمده است.

مقایسه با روش شمارش بدون پنجره میرا :



```
trendDetector — -bash — 151x24
post index 397999 total words: 1224006 len words: 41943144 total counters: 93424419 mean counter: 76.3267655550708 salam len : 22463 2587
post index 398999 total words: 1225588 len words: 41943144 total counters: 93541165 mean counter: 76.3234941415876 salam len : 22587 2597
post index 399999 total words: 1227762 len words: 41943144 total counters: 93735591 mean counter: 76.346711333380401 salam len : 22519 2600
post index 400999 total words: 1230168 len words: 41943144 total counters: 93942961 mean counter: 76.3659605842454 salam len : 22597 2618
post index 401999 total words: 1232684 len words: 41943144 total counters: 94144354 mean counter: 76.37842648571642 salam len : 22632 2626
post index 402999 total words: 1234542 len words: 41943144 total counters: 94314222 mean counter: 76.39612261065237 salam len : 22685 2640
post index 403999 total words: 1237016 len words: 41943144 total counters: 94504622 mean counter: 76.39725112690539 salam len : 22730 2646
post index 404999 total words: 1239432 len words: 41943144 total counters: 94704396 mean counter: 76.409513390808445 salam len : 22760 2650
post index 405999 total words: 1242623 len words: 41943144 total counters: 94998699 mean counter: 76.44369933680134 salam len : 22856 2661
post index 406999 total words: 1244415 len words: 41943144 total counters: 95158327 mean counter: 76.46832286297738 salam len : 22916 2670
post index 407999 total words: 1247189 len words: 41943144 total counters: 95363670 mean counter: 76.46288573744637 salam len : 22940 2674
post index 408999 total words: 1249029 len words: 41943144 total counters: 95505896 mean counter: 76.4641141238514 salam len : 22994 2684
post index 409999 total words: 1250980 len words: 41943144 total counters: 95646471 mean counter: 76.45723432828662 salam len : 23033 2684
post index 410999 total words: 1252790 len words: 41943144 total counters: 95800952 mean counter: 76.470808038059852 salam len : 23061 2686
post index 411999 total words: 1254624 len words: 41943144 total counters: 95949687 mean counter: 76.47684644961357 salam len : 23116 2700
post index 412999 total words: 1256957 len words: 41943144 total counters: 96146704 mean counter: 76.49164132186348 salam len : 23170 2713
post index 413999 total words: 1258665 len words: 41943144 total counters: 96264323 mean counter: 76.48129089784726 salam len : 23181 2713
post index 414999 total words: 1261211 len words: 41943144 total counters: 96476145 mean counter: 76.49484899830401 salam len : 23210 2719
post index 415999 total words: 1263717 len words: 41943144 total counters: 96672284 mean counter: 76.49836474463824 salam len : 23244 2730
post index 416999 total words: 1266692 len words: 41943144 total counters: 96923125 mean counter: 76.51672624442249 salam len : 23283 2737
post index 417999 total words: 1268794 len words: 41943144 total counters: 97084380 mean counter: 76.51705477800179 salam len : 23314 2747
post index 418999 total words: 1271682 len words: 41943144 total counters: 97315653 mean counter: 76.52514779638305 salam len : 23335 2749
post index 419999 total words: 1274116 len words: 41943144 total counters: 97538562 mean counter: 76.55391031899764 salam len : 23383 2750
milads-MacBook-Pro:trendDetector milad$
```

همین طور که در تصویر زیر آمده تعداد کل شمارنده ها از ۹۷ میلیون به ۱.۲ میلیون رسیده .

نکته دیگر افزایش بیرونیه تعداد کلمات بدون استفاده از پنجره میرا است .

در اینجا ۱.۲ میلیون کلمه تشخیص داده شده است !!!

چونکه کلماتی که اشتباہ تایپی بوده اند یا با ایموجی ترکیب شده بودند یا کمترکار بوده اند را یک کلمه فرض کرده و حرص نشده اند ، که پنجره میرا(لایه اول) این مشکل را حل میکند.

بسمه تعالى

این عکس کل کلمات در یک پست به عنوان کلمات مربوط فرض میشند:

Process Name	Memory	Threads	Ports	PID	User
Python	11.32 GB	2	24	2462	milad
Python	21.1 MB	1	14	2331	milad
Python	15.0 MB	1	14	2335	milad
Python	7.8 MB	1	14	2487	milad

```
trendDetector — Python - Python trendDetector.py — 121x24
post index 186999 total words: 658388 len words: 28971616 total counters: 233598936 mean counter: 354.8086758406999
post index 187999 total words: 668969 len words: 28971616 total counters: 234544588 mean counter: 354.8496841417979
post index 188999 total words: 664166 len words: 28971616 total counters: 236115916 mean counter: 355.5073821985969
post index 189999 total words: 667388 len words: 28971616 total counters: 237862311 mean counter: 356.45056186926004
post index 190999 total words: 670428 len words: 28971616 total counters: 238984494 mean counter: 356.46981593627874
post index 191999 total words: 672948 len words: 28971616 total counters: 239941081 mean counter: 356.55642553578897
post index 192999 total words: 675056 len words: 28971616 total counters: 240628464 mean counter: 356.44518973248736
post index 193999 total words: 677425 len words: 28971616 total counters: 241342892 mean counter: 356.263928848212
post index 194999 total words: 680104 len words: 28971616 total counters: 242423880 mean counter: 356.45118981802784
post index 195999 total words: 683826 len words: 28971616 total counters: 244265306 mean counter: 357.203888123587
post index 196999 total words: 686676 len words: 28971616 total counters: 245368659 mean counter: 357.3162728856113
post index 197999 total words: 689482 len words: 28971616 total counters: 246117630 mean counter: 356.96019620526715
post index 198999 total words: 692428 len words: 28971616 total counters: 247658577 mean counter: 357.67103347679154
post index 199999 total words: 695233 len words: 28971616 total counters: 249835513 mean counter: 358.2043904785329
post index 200999 total words: 698318 len words: 28971616 total counters: 250203966 mean counter: 358.2951692495396
post index 201999 total words: 700982 len words: 41943144 total counters: 251144545 mean counter: 358.2753123475353
post index 202999 total words: 702862 len words: 41943144 total counters: 251641626 mean counter: 358.02422950735706
post index 203999 total words: 706234 len words: 41943144 total counters: 253111460 mean counter: 358.3960273790274
post index 204999 total words: 709207 len words: 41943144 total counters: 254107085 mean counter: 358.2973729813722
post index 205999 total words: 711839 len words: 41943144 total counters: 255018279 mean counter: 358.2415110720261
post index 206999 total words: 713983 len words: 41943144 total counters: 255990630 mean counter: 358.42538247717856
post index 207999 total words: 716716 len words: 41943144 total counters: 256455635 mean counter: 357.8204407324519
post index 208999 total words: 719378 len words: 41943144 total counters: 257322062 mean counter: 357.70076649550026
```

Process Name	Memory	Threads	Ports	PID	User
Python	15.23 GB	2	24	2579	milad
Python	7.8 MB	1	14	2604	milad

Heavy memory usage
CleanMyMac has found that you're running out of both physical and virtual memory. Consider quitting some apps.

RAM + Swap Almost Full

Ignore Learn More...

```
trendDetector — Python - Python trendDetector.py — 151x24
post index 247999 total words: 840855 len words: 41943144 total counters: 317851478 mean counter: 378.00985663402133 salam len : 88557 2297
post index 248999 total words: 844008 len words: 41943144 total counters: 319055513 mean counter: 378.024275836248 salam len : 88937 2303
post index 249999 total words: 846613 len words: 41943144 total counters: 320232187 mean counter: 378.258968270836 salam len : 89060 2303
post index 250999 total words: 849403 len words: 41943144 total counters: 321478463 mean counter: 378.4663616681363 salam len : 89248 2307
post index 251999 total words: 851859 len words: 41943144 total counters: 322708153 mean counter: 378.828131181334 salam len : 89415 2309
post index 252999 total words: 854544 len words: 41943144 total counters: 324000000 mean counter: 379.2661220487184 salam len : 89582 2309
post index 253999 total words: 856244 len words: 41943144 total counters: 324492982 mean counter: 379.44400000000005 salam len : 89714 2314
post index 254999 total words: 860574 len words: 41943144 total counters: 326492899 mean counter: 379.38967179844656 salam len : 90184 2317
post index 255999 total words: 862582 len words: 41943144 total counters: 3272233692 mean counter: 379.35361739521574 salam len : 90273 2331
post index 256999 total words: 865188 len words: 41943144 total counters: 328435577 mean counter: 379.61188346461874 salam len : 90525 2341
post index 257999 total words: 867894 len words: 41943144 total counters: 329053219 mean counter: 379.489673553271 salam len : 90672 2342
post index 258999 total words: 871498 len words: 41943144 total counters: 330427148 mean counter: 379.19112692219414 salam len : 90798 2345
post index 259999 total words: 874483 len words: 41943144 total counters: 331440869 mean counter: 379.0135874095288 salam len : 90909 2355
post index 260999 total words: 877721 len words: 41943144 total counters: 332887669 mean counter: 379.26364869930194 salam len : 91261 2363
post index 261999 total words: 881039 len words: 41943144 total counters: 334396543 mean counter: 379.5479462316653 salam len : 91467 2372
post index 262999 total words: 884226 len words: 41943144 total counters: 335880824 mean counter: 379.76809548699804 salam len : 91893 2381
post index 263999 total words: 886980 len words: 41943144 total counters: 336796111 mean counter: 379.7110543624272 salam len : 91951 2381
post index 264999 total words: 890152 len words: 41943144 total counters: 338546899 mean counter: 380.3236862917794 salam len : 92257 2389
post index 265999 total words: 894038 len words: 41943144 total counters: 340385412 mean counter: 380.731532498999 salam len : 92443 2404
post index 266999 total words: 896698 len words: 41943144 total counters: 341397298 mean counter: 380.72717681984346 salam len : 92856 2407
post index 267999 total words: 900816 len words: 41943144 total counters: 343318536 mean counter: 381.4582585198485 salam len : 93246 2416
post index 268999 total words: 902892 len words: 41943144 total counters: 343789396 mean counter: 381.1023664992841 salam len : 93418 2423
post index 269999 total words: 905861 len words: 41943144 total counters: 345575194 mean counter: 381.4881024792987 salam len : 93578 2427
```

بسمه تعالی

پردازش کل کلمات و نیز کل کلمات در یک پست به عنوان کلمات مربوط فرض میشند:

```
trendDetector — bash — 151x24
post index 983999 total words: 3256066 len words: 167772264 total counters: 206300200 mean counter: 63.35872798647202 salam len : 13610 1372
post index 984999 total words: 3258697 len words: 167772264 total counters: 206482309 mean counter: 63.363457541465195 salam len : 13635 1375
post index 985999 total words: 3261126 len words: 167772264 total counters: 206663607 mean counter: 63.37185591725067 salam len : 13634 1375
post index 985999 total words: 3261126 len words: 167772264 total counters: 206663607 mean counter: 63.37185591725067 salam len : 13634 1375
post index 986999 total words: 3264052 len words: 167772264 total counters: 206870301 mean counter: 63.378371729371956 salam len : 13643 1375
post index 987999 total words: 3266670 len words: 167772264 total counters: 207077001 mean counter: 63.3908539889245 salam len : 13658 1376
post index 988999 total words: 3269677 len words: 167772264 total counters: 207277348 mean counter: 63.39383003275247 salam len : 13672 1376
post index 989999 total words: 3273573 len words: 167772264 total counters: 207528966 mean counter: 63.39524611181727 salam len : 13692 1380
post index 990999 total words: 3276532 len words: 167772264 total counters: 207784007 mean counter: 63.415833265171834 salam len : 13699 1380
post index 991999 total words: 3279397 len words: 167772264 total counters: 208802037 mean counter: 63.43280756160355 salam len : 13726 1381
post index 992999 total words: 3281298 len words: 167772264 total counters: 2088193179 mean counter: 63.44842163870834 salam len : 13734 1382
post index 993999 total words: 3283821 len words: 167772264 total counters: 2088409238 mean counter: 63.46546842839485 salam len : 13769 1389
post index 994999 total words: 3285724 len words: 167772264 total counters: 2088558436 mean counter: 63.47411894608312 salam len : 13772 1390
post index 995999 total words: 3287888 len words: 167772264 total counters: 2088696276 mean counter: 63.47426554675828 salam len : 13782 1392
post index 996999 total words: 3289738 len words: 167772264 total counters: 2088807469 mean counter: 63.472370444090875 salam len : 13804 1395
post index 997999 total words: 3292376 len words: 167772264 total counters: 209031684 mean counter: 63.48961479490799 salam len : 13833 1398
post index 998999 total words: 3295755 len words: 167772264 total counters: 209279874 mean counter: 63.49982750538192 salam len : 13868 1404
post index 999999 total words: 3298753 len words: 167772264 total counters: 209505845 mean counter: 63.51061901269965 salam len : 13900 1404
post index 1000999 total words: 3302077 len words: 167772264 total counters: 209735686 mean counter: 63.51629171578979 salam len : 13909 1404
post index 1001999 total words: 3304446 len words: 167772264 total counters: 209908426 mean counter: 63.523031092049926 salam len : 13941 1405
post index 1002999 total words: 3308610 len words: 167772264 total counters: 210273434 mean counter: 63.55340581089944 salam len : 14000 1416
^CTraceback (most recent call last):
  File "trendDetector.py", line 22, in <module>
    test_file = open("testfile.txt", 'a')
```

همانطور که در تصویر آمده تعداد شمارنده ها به ۲۰۰ میلیون برای پردازش ۱ میلیون پست میرسد.

نواص:

۱: چون دیتاست محدود بوده است نمیتوانیم بگوییم حافظه‌ی مورد استفاده در این روش برای استریم بینهایت محدود است یا خیر . یا شاید بتوان با تغییر پارامتر ها به این مهم دست یافت.

۲: حد آستانه لایه اول ۱۰ در نظر گرفته شده است و اگر در ۱۰۰۰ پست متوالی کلمه ای که از قبل در پنجره وجود نداشته به تعداد کمتر از حد آستانه (۱۰ بار) تکرار شود آن کلمه از پنجره جذف مشود. ولی اگر در ۱۰۰۰ پست ۱۱ بار بباید با همان اندازه ۱۱ باقی میماند. این ناپیوستگی به نظرم ممکن است مشکل زا باشد و نمیتوان گفت همهی کلمات در پنجره پر تکرار ترین ها بوده اند.

ممnon میشویم اگر مشکلی در سازوکار این روش میبینید به ما متنظر شوید .

با تشکر از توجه شما .
یا حق.