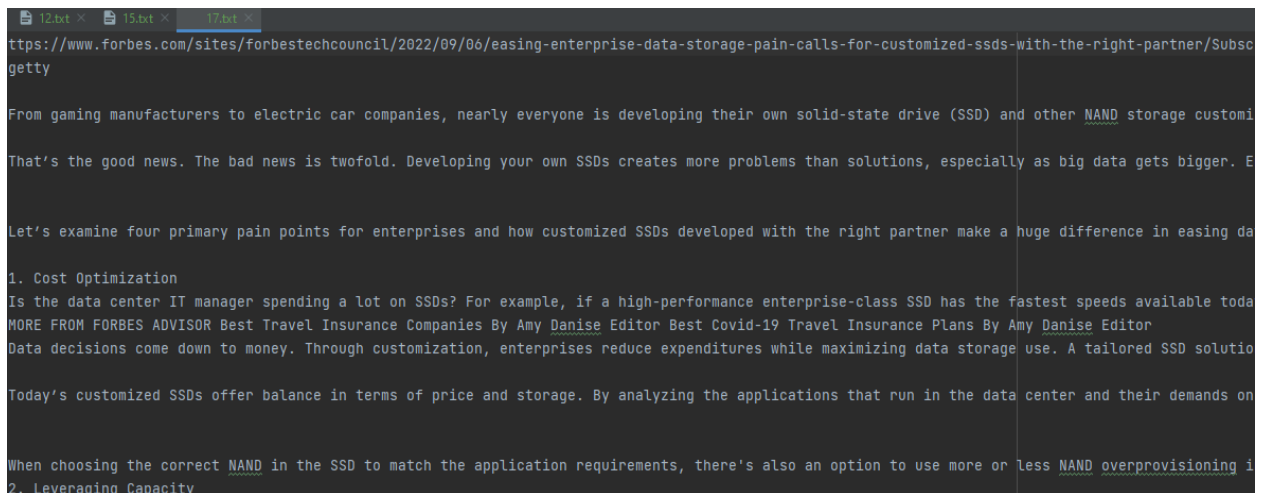


Team Members:

- Blake Oberlander bho180000
- Sharar Siddiqui srs200003

(1) describe how you created your knowledge base, include screen shots of the knowledge base, and indicate your top 10 terms;

- Before creating the knowledge base, we created a web crawler that extracts external urls from a technology subreddit (/r/Newmaxx). The crawler knows how to go to the next reddit page to get more results in case the first page doesn't have enough content.
- The crawler scrapes the text off of external URLs.



The screenshot shows a web browser with three tabs labeled '12.txt', '15.txt', and '17.txt'. The active tab displays a Forbes article titled 'Easing enterprise data storage pain calls for customized SSDs with the right partner'. The article text is as follows:

<https://www.forbes.com/sites/forbestechcouncil/2022/09/06/easing-enterprise-data-storage-pain-calls-for-customized-ssds-with-the-right-partner/Subsc>
getty

From gaming manufacturers to electric car companies, nearly everyone is developing their own solid-state drive (SSD) and other NAND storage customi

That's the good news. The bad news is twofold. Developing your own SSDs creates more problems than solutions, especially as big data gets bigger. E

Let's examine four primary pain points for enterprises and how customized SSDs developed with the right partner make a huge difference in easing da

1. Cost Optimization

Is the data center IT manager spending a lot on SSDs? For example, if a high-performance enterprise-class SSD has the fastest speeds available toda

MORE FROM FORBES ADVISOR Best Travel Insurance Companies By Amy Danise Editor Best Covid-19 Travel Insurance Plans By Amy Danise Editor

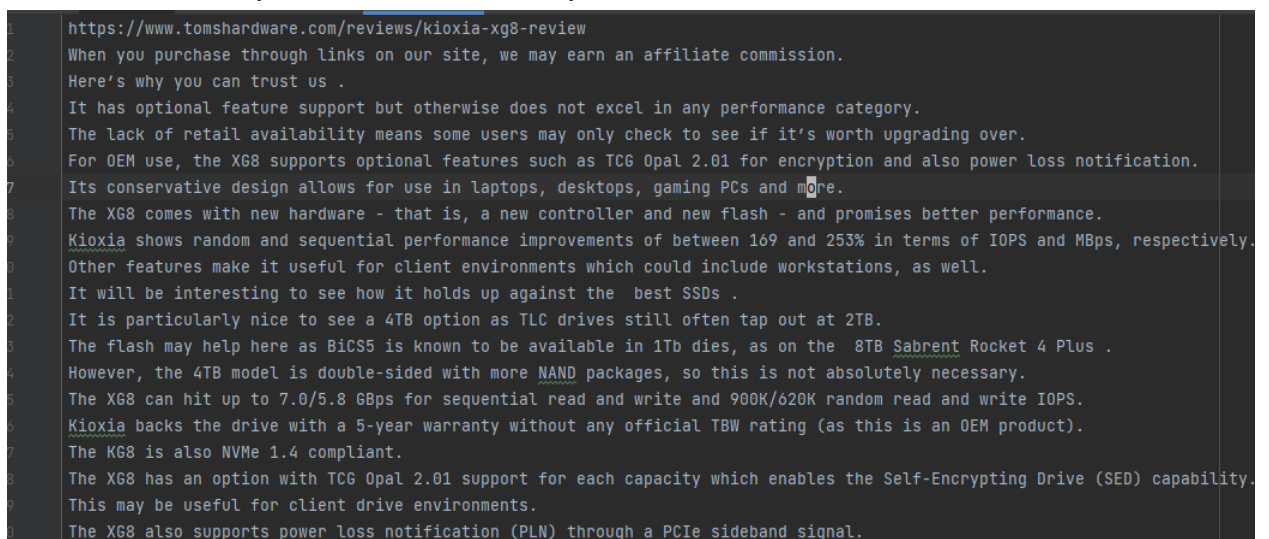
Data decisions come down to money. Through customization, enterprises reduce expenditures while maximizing data storage use. A tailored SSD solutio

Today's customized SSDs offer balance in terms of price and storage. By analyzing the applications that run in the data center and their demands on

When choosing the correct NAND in the SSD to match the application requirements, there's also an option to use more or less NAND overprovisioning i

2. Leveraging Capacity

- After running the crawler, we run 2_top_term_finder.py to parse the sentences from the scraped text and identify the top 25 words. This program uses the tf-idf algorithm. It uses NLTK's Reuters corpus to calculate the 'idf' part.



The screenshot shows a web browser with a single tab displaying a Tom's Hardware article titled 'Kioxia XG8 review'. The article text is as follows:

<https://www.tomshardware.com/reviews/kioxia-xg8-review>

When you purchase through links on our site, we may earn an affiliate commission.

Here's why you can trust us .

It has optional feature support but otherwise does not excel in any performance category.

The lack of retail availability means some users may only check to see if it's worth upgrading over.

For OEM use, the XG8 supports optional features such as TCG Opal 2.01 for encryption and also power loss notification.

Its conservative design allows for use in laptops, desktops, gaming PCs and more.

The XG8 comes with new hardware - that is, a new controller and new flash - and promises better performance.

Kioxia shows random and sequential performance improvements of between 169 and 253% in terms of IOPS and MBps, respectively.

Other features make it useful for client environments which could include workstations, as well.

It will be interesting to see how it holds up against the best SSDs .

It is particularly nice to see a 4TB option as TLC drives still often tap out at 2TB.

The flash may help here as BiCS5 is known to be available in 1Tb dies, as on the 8TB Sabrent Rocket 4 Plus .

However, the 4TB model is double-sided with more NAND packages, so this is not absolutely necessary.

The XG8 can hit up to 7.0/5.8 GBps for sequential read and write and 900K/620K random read and write IOPS.

Kioxia backs the drive with a 5-year warranty without any official TBW rating (as this is an OEM product).

The XG8 is also NVMe 1.4 compliant.

The XG8 has an option with TCG Opal 2.01 support for each capacity which enables the Self-Encrypting Drive (SED) capability.

This may be useful for client drive environments.

The XG8 also supports power loss notification (PLN) through a PCIe sideband signal.

- The knowledge base is created using top 10 terms manually chosen from 2_top_term_finder.py's output. In this case we print out the top 25 terms.


```
[('pcie', 0.09554880688231629),
 ('ssd', 0.08599392619408465),
 ('ssds', 0.06582251140781789),
 ('drives', 0.0640260164695318),
 ('drive', 0.0494041535628859),
 ('nand', 0.03928117616273003),
 ('nvme', 0.03821952275292651),
 ('raid', 0.0372918076483839),
 ('highpoint', 0.036096215933319485),
 ('gen', 0.03510618252254368),
 ('usb', 0.03397290911371245),
 ('card', 0.032317252571138294),
 ('storage', 0.0300987349062905),
 ('performance', 0.029237963034116757),
 ('flash', 0.024560235495002275),
 ('hardware', 0.02332700031701157),
 ('cards', 0.023313730125519953),
 ('controller', 0.02142026668092856),
 ('write', 0.021101682156929258),
 ('kioxia', 0.020171414786266773),
 ('sequential', 0.019109761376463256),
 ('slot', 0.019109761376463256),
 ('like', 0.018352728098719986),
 ('ie', 0.016986454556856226),
 ('memory', 0.016876519473068775)]
```
- Then, Based on our knowledge of the domain(Flash storage) we chose 10 terms that are relevant.
 1. ssd
 2. usb
 3. controller
 4. drive
 5. pcie
 6. nand
 7. card
 8. raid
 9. nvme
 10. storage

The knowledgebase is generated as a pickle file in 3_kb_creator.py . Executing the program asks the user input to enter the top 10 terms, it combines all the scraped data into one big data input before going through it to make the knowledge base. The resulting output is a pickle file of

the dict of sentences containing the terms given. The pickle file can be viewed using the following command,

```
$ python -mpickle database.pickle
```

Which gives the following output:

```
'nand': ['However, the 4TB model is double-sided with more NAND packages, so '
        'this is not absolutely necessary.',
        'Underneath this we see the controller flanked by two NAND packages '
        'on one side and a DRAM package on the other.',
        'In response to the aforementioned situation, Micron announced last '
        'week that it would cut production of DRAM and NAND Flash, becoming '
        'the first major memory manufacturer to officially reduce its '
        'capacity utilization plan.',
        'In terms of NAND Flash, the market situation is more severe than '
        'that of DRAM.',
        'As the average contract price of mainstream capacity wafers has '
        'fallen to their cash cost and is approaching the periphery of '
        'selling at a loss for various manufacturers, Kioxia also announced '
        'that it will reduce NAND Flash capacity utilization by 30% from '
        '"October on the heels of Micron's announcement.",
        'Therefore, compared with NAND Flash, it remains to be seen whether '
        'there will be a significant reduction in production.',
        'In terms of NAND Flash, Micron originally planned to gradually '
        'increase its proportion of 232-layer products from 4Q22.',
        'More manufactures limiting bit output cannot be ruled out as only '
        'large-scale production reduction can reverse supply/demand '
        'imbalance in 2023 After analyzing 2023 supply and demand in the '
        'memory market, due to a conservative demand outlook, DRAM and NAND '
        'Flash look to be greatly oversupplied in each quarter and inventory '
        'pressure will continue to accelerate in 1H23.',
        'It is imperative to reduce bit supply in the NAND Flash field due ']
```

```
'nvme': ['The K68 is also NVMe 1.4 compliant.',  
        'Many of those folks have largely sequential workloads, and the '  
        'performance of PCIe Gen4 NVMe SSDs is fine.',  
        'What they need is a lot of local storage and so being able to put '  
        'two of these with 8x 8TB NVMe SSDs, as an example, gives 128TB of '  
        'storage in only two PCIe card slots.',  
        'Mere mortals would be much better off with some cheap EPYC board '  
        'and simple passive 4x NVMe card for €100-ish or less.',  
        'I had a lot of problems with overall jankiness with their cards, '  
        'including one of their newer PCIe NVMe RAID cards.',  
        'Count me in the camp that is interested in a hardware NVMe RAID 5 '  
        'solution for Windows.',  
        'Some Control Data mainframes used to have hardware ISAM disks '  
        'drives and these days WD is starting to include pattern/string '  
        'searches into NVMe(CXL) devices: don't know if that has left the '  
        'labs yet.',  
        'If you consider a buffering RAID controller an accelerator, you'll '  
        'be disappointed by NVMe which is all about getting rid of the '  
        'SCSI/SATA abstractions and letting the CPU get closer to the '  
        'hardware.',  
        'So software RAID is all you'll ever get, even with a HBA, which '  
        'will turn into a switch for NVMe and a SAS controller for SAS/SATA '  
        'devices.',  
        'Building a "smart" NVMe RAID controller would be rather hard (and '  
        'somewhat pointless), because designing for >100GB/s bandwidth is '  
        'quite involved: you'd have to resort to GDDR memory or lots of DRAM '  
        'channels, only to face PCIe bottlenecks afterwards.',  
        'However, a quad M.2 bifurcation board using PCIe 3.0 M.2 drives is '
```

(2) write up a sample dialog you would like to create with a chatbot based on your knowledge base

StorageTips(Chat bot): Hello, my name is StorageTips. My knowledge base is full of facts regarding storage technology and news. What do you want to know about?

- ssd
- usb
- controller
- drive
- pcie
- nand
- card
- raid
- nvme
- storage

User: Tell me about NAND

StorageTips(Chat bot): TLC NAND is used for mainstream SSD applications today, but for many applications, the more affordable QLC NAND is a perfect choice.

Anything else you'd like to know?

User: What about NVME?

StorageTips(Chat bot): The black 2280 PCB is a single sided design with two packages of SK hynix 176-layer NAND flash memory, a SK hynix 2GB DDR4 DRAM buffer chip followed by the SK hynix NVMe 8-channel 'Aries' controller on the far right.

Anything else you'd like to know?

User: Interesting! Tell me more

StorageTips(Chat bot): Mere mortals would be much better off with some cheap EPYC board and simple passive 4x NVMe card for €100-ish or less.

What else can I help you with?

Anything else you'd like to know?

User: Tell me about cards, please.

StorageTips(Chat bot): Using its innovative BiCS FLASH 3D flash memory and an in-house designed controller, basic functions of the KIOXIA 2 TB

microSDXC UHS-I memory card working prototypes were confirmed in the microSDXC standard's maximum density.
Anything else you'd like to know?

User: Info about controller.

StorageTips(Chat bot): On the bridge side, the JHL7440 Titan Ridge Thunderbolt 3 controller is augmented by the ASMedia ASM2362 USB 3.2 Gen 2 - NVMe bridge chip.
Anything else you'd like to know?

User: Give me a random fact

StorageTips(Chat bot): KIOXIA will present a video showing its PCIe 5.0 drives demonstrating high performance, low latency and high bandwidth, in a 12th Gen Intel Core processor-based workstation.
Anything else you'd like to know?

User: Tell me about Apple memory technology

StorageTips(Chat bot): That information is not within my database, but I am always learning more.
Anything else you'd like to know?

User: No thanks.

StorageTips(Chat bot): Thank you for using StorageTips. Come back to know more about storage technology and new. BYE!!!

We used the output of our 4th program to generate this sample:

```
Terms:
- raid
- storage
- nvme
- usb
- ssd
- controller
- pcie
- card
- drive
- nand

Enter a term to explore, or type ! to quit
Term: nand
TLC NAND is used for mainstream SSD applications today, but for many applications, the more affordable QLC NAND is a perfect choice.
```