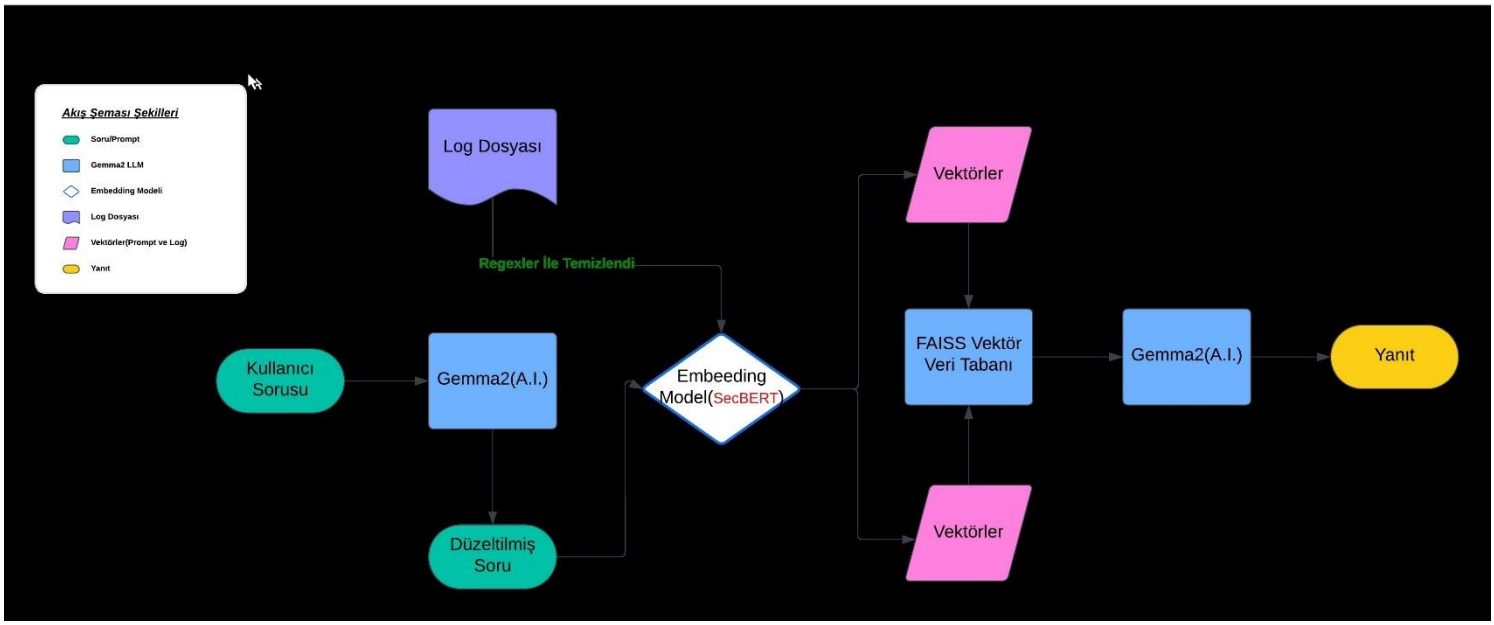
Rag LLM Uygulamaları ile Log Analizi

19.08.2024



Projemiz için akış şeması yukarıda gösterilmiştir. Bu yazıda bu akış şemasındaki elementler adım adım açıklanacaktır. Açıklama oluşturma sırası ile yapılacaktır.

1)Verinin Oluşturulması (Log Dosyası)

Apache24 kurarak local sunucuda bir web sitesi çalıştırıyoruz. Daha sonra bir Python scripti ile bu sunucuya çeşitli etkiler yaparak ..Apache24/logs/Access dosyasından loglarımızı görebiliriz. Bu dosyayı repo üzerinde "generate_apache_logs.py" olarak bulabiliriz.

2)Regexler ile Veri Temizliği

Elde ettiğimiz log dosyası her satırında bir eylemi yazar. Bu eylemleri regexler ile [ip,timestamp,http method,page,status code,byte size] olarak alırız. Aldığımız bu verileri bir Python dictionaryde tutarız.

3)Log Verilerinin SecBERT'e Aktarılması

Python dictionary'deki veriler SecBERT embeeding modele aktarılır ve bir vektör listesi olarak tutulur. Daha sonra vektör veri tabanına kaydedilir

4)Kullanıcı Sorusunun

Kullanıcıdan bir soru alınır daha sonra bu soru hemen embedding modele verilmez. Soru Gemma2 modeline aktarılarak log dosyalarının tutulduğu formata benzetilmeye çalışılır. Bu sayede faisste yapılan aramalarda daha düşük 'distancelar' yakalandığı test edilmiştir. Gemmadan alınan sonuç embedding modele verilir ve daha sonra vektör veri tabanına kaydedilir.

5) Vektör Veri Tabanı Üzerinde Arama İşlemi

Kullanıcının düzeltilmiş promptu ile context yani log dosyalarının içeriği arasında bir benzerlik araması yapılır. Bu aramada en yakın 15 sonuç alınır. Veri tabanında arama için çeşitli arama algoritmaları tercih edilebilir. Burada kısmen daha küçük bir veri tabanına sahip olduğumuz için IndexFlatL2 (L2 distance) kullanılmıştır. Bu algoritmaların seçimi ihtiyaçlara göre değişkenlik gösterebilir.

6) Yapay Zekâ Modeline İçerik ve Sorunun Aktarılması.

Kullanıcının ham yani düzeltilmemiş sorusu ve en yakın bulunan 10 log yapay zeka modeline aktarılarak sorunun cevaplandırılması beklenilir.

Sistem İşleyişine bir göz atalım.

```
query = "What happened on August 29 "
```

Yukarıdaki soru Gemma2'ye aktarılmış ve şu şekilde düzeltilmiştir.

```
Generated Response: IP  accessed  using  on 29/August/  and received status code  with  bytes.
```

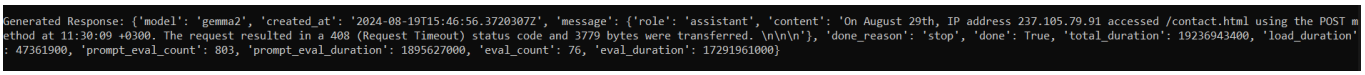
Buradaki amaç embeddinglerle oluşturulan vektörlerin birbiri ile benzerliğini arttırmaktır.

Bir dezavantaj olarak bu formata hiç uymayan soruların sorulabileceği düşünüldüğünde modelin başarımının azalacağı bir gerçektir ancak daha sonradan bahsi geçecek geliştirilmeler yapılmadan loglara dolaylı ya da hiç bağlı olmayan soruların yanıtları zaten iyi bir şekilde verilemeyeceği için böyle bir uygulama yapmaktan geri durulmamıştır.

Aşağıda bu veriler ile yapılan faiss vt. İçin sonuçlar tablo halinde verilmiştir. Tablodaki 'Distance' parametresi soru vektörü ile log vektörleri arasındaki semantik uzaklığı gösterir.

IP Adresi	Erişilen Sayfa	HTTP Yöntemi	Tarih ve Saat	Durum Kodu	Byte Boyutu	Mesafe
237.105.79.91	/contact.html	POST	29/Aug/2024:11:30:09 +0300	408	3779	63.501968383789
::1	/	POST	14/Aug/2024:14:18:29 +0300	200	2866	68.016250610351
::1	/	POST	14/Aug/2024:14:18:09 +0300	200	2866	68.532455444335
216.130.81.77	/contact.html	PUT	17/Aug/2024:23:46:47 +0300	500	-	68.534500122070
247.7.34.98	/login.html	PUT	13/Aug/2024:17:31:26 +0300	408	-	69.031906127929
29.34.137.242	/index.html	DELETE	25/Aug/2024:22:50:15 +0300	200	3633	69.304962158203
::1	/	POST	14/Aug/2024:14:18:19 +0300	200	2866	69.502578735351
::1	/	POST	14/Aug/2024:14:18:38 +0300	200	2866	69.530479431152
::1	/	POST	14/Aug/2024:14:17:54 +0300	200	2866	69.680877685546
::1	/	GET	14/Aug/2024:13:39:57 +0300	304	-	69.885345458984
::1	/	POST	14/Aug/2024:14:18:35 +0300	200	2866	70.036819458007
::1	/	POST	14/Aug/2024:14:17:50 +0300	200	2866	70.086929321289
24.141.187.144	/	PUT	14/Aug/2024:11:53:07 +0300	301	1211	70.239540100097
::1	/	GET	14/Aug/2024:14:18:40 +0300	200	2866	70.255584716796
::1	/	GET	14/Aug/2024:14:17:11 +0300	304	-	70.417816162109

Tabloda görüldüğü üzere model 29 Ağustosta olan bir olayı yakalayabilmiş ancak buna en yakın mesafedeki log kaydını 14 Ağustos olarak görmüş. Bu üzerine düşünülmesi gerek ve iyileştirme yapılması gerek konulardandır. Bu konuya ‘Geliştirme’ kısmında tekrar değineceğiz.



“Generated Response: {'model': 'gemma2', 'created_at': '2024-08-19T15:46:56.3720307Z', 'message': {'role': 'assistant', 'content': **'On August 29th, IP address 237.105.79.91 accessed /contact.html using the POST method at 11:30:09 +0300. The request resulted in a 408 (Request Timeout) status code and 3779 bytes were transferred. \n\n\n'**}, 'done_reason': 'stop', 'done': True, 'total_duration': 19236943400, 'load_duration': 47361900, 'prompt_eval_count': 803, 'prompt_eval_duration': 1895627000, 'eval_count': 76, 'eval_duration': 17291961000}”

Tüm loglar için soru özelinde excel tabloları repo içine aktarılmıştır.

Şimdi soruyu daha özel hale getirelim ve sonuçları inceleyelim.

```
query = "What happened on August 29th over the ip adress 108.63.206.102"
```

```
Generated Response: IP 108.63.206.102 accessed  using  on 29/August/  and received status code  with  bytes.
```

IP Adresi	Erişilen Sayfa	HTTP Yöntemi	Tarih ve Saat	Durum Kodu	Byte Boyutu	Mesafe
108.63.206.102	/login.html	GET	29/Aug/2024:06:36:23 +0300	408	1764	27.1681861877
237.105.79.91	/contact.html	POST	29/Aug/2024:11:30:09 +0300	408	3779	43.0872726440
244.244.158.241	/login.html	GET	20/Aug/2024:23:02:36 +0300	200	897	43.1893310546
107.234.168.63	/products.html	GET	08/Aug/2024:00:45:55 +0300	200	-	45.0846023559
187.111.135.214	/login.html	GET	17/Aug/2024:21:35:50 +0300	408	800	47.3297348022
107.138.163.223	/	GET	11/Aug/2024:10:07:48 +0300	404	-	47.3651733398
179.149.181.242	/login.html	PUT	11/Aug/2024:18:31:36 +0300	408	-	48.0912551879
145.238.179.103	/favicon.ico	POST	04/Aug/2024:22:58:30 +0300	200	3651	48.4378356933
87.95.2.4	/products.html	GET	16/Aug/2024:22:01:37 +0300	200	1376	48.8327789306
244.133.16.42	/contact.html	PUT	28/Aug/2024:16:04:40 +0300	500	-	48.9561996459
163.144.121.114	/	POST	27/Aug/2024:15:06:16 +0300	408	-	49.7485198974
19.70.49.123	/products.html	POST	10/Aug/2024:22:52:07 +0300	404	-	49.9998588562
152.182.167.135	/favicon.ico	POST	05/Aug/2024:00:23:24 +0300	200	786	50.1221160888
44.230.179.10	/index.html	GET	03/Aug/2024:04:51:31 +0300	500	2081	50.5326271057
175.108.164.111	/about.html	GET	1[illegible]g/2024:04:24:45 +[illegible]0	301	-	50.7031402587

Kolaylıkla fark edileceği üzere log formatına sorular yaklaştırabildiğimiz ölçüde ‘Mesafe-Distance’ azalmaktadır.

Modelin cevabı aşağıdaki gibi olmuştur.

Generated Response: {'model': 'gemma2', 'created_at': '2024-08-19T15:59:03.7342689Z', 'message': {'role': 'assistant', **'content': 'On August 29th, 2024 at 06:36:23 +0300, IP address 108.63.206.102 accessed the /login.html page using the GET method. \n\nThe request received a 408 status code (Request Timeout), indicating that the server did not receive a complete request within the allotted time. The request also transferred 1764 bytes of data. \n\n\n'**}, 'done_reason': 'stop', 'done': True, 'total_duration': 28412158100, 'load_duration': 29910600, 'prompt_eval_count': 968, 'prompt_eval_duration': 2142719000, 'eval_count': 104, 'eval_duration': 26238986000}

Geliştirmeler

Yukarıdaki bilgileri incelediğimizde Soru ve log embeddinglerinin birbirlerine olan uzaklığının çeşitli faktörlerle artıp azaldığı izlenmiştir. Bu nedenle daha küçük mesafeler(distances) oluşturmak için çeşitli iyileştirmeler yapılabilir. Bunlar şu şekilde başlıklandırılabilir:

- **A)** Query kısmında kullanıcıdan soru alan model daha kapsamlı seçilebilir ya da daha iyi bir prompt yazarak daha iyi bir formatta düzeltilmiş soru oluşturulabilir.
- **B)** Log kısa olduğu için daha iyi bir embedding model kullanılabilir ya da bu iş için özel bir model oluşturulabilir yani fine-tune yöntemi uygulanabilir.
- **C)** Farklı “Query translation” işlemleri denenebilir.
- **D)** Crag gibi doğrulayıcı modellerle modele sorulacak loglardan bağımsız sorular içinde cevaplar oluşturması sağlanabilir. Bu tamamen gerekli mi emin değilim ancak kullanılacak senaryolar olabileceğini düşünüyorum.
- **E)** Bu proje rag llm kapsamında olmasaydı ‘mongodb’ gibi bir veri tabanı ile çok daha iyi sonuçlar alınabilirdi.

Deniz Yaldız