Katedra Systemów Informacyjnych

**Eksploracja zasobów internetowych**

**Kierunek Systemy teleinformatyczne, WE, II st., sem. II**

**Laboratorium nr 2**

**Celem zajęć** jest ocena zastosowania wybranych narzędzi analitycznych do analizy wykorzystania

zasobów wybranego serwera WWW.

Aby zaliczyć laboratorium **przygotuj sprawozdanie**, w którym umieść odpowiedzi do pytań oraz sformułowane wnioski w przeprowadzonej analizy. Sprawozdanie prześlij w systemie ilias.

Dane wejściowe: access.log wybranego serwera www

Jakie dane wykorzystałeś do ćwiczenia:

<https://raw.githubusercontent.com/elastic/examples/master/Common%20Data%20Formats/apache_logs/apache_logs>

**Wykonali:**

**Miłosz Wiczarski Jakub Krajewski**

**1. Dokonaj zgrubnej ceny danych zawartych w pliku access.log Jakiego rodzaju informacje możesz odczytać z zapisu poszczególnych zapytań do serwera?**

Logi z pliku access.log zawierają informacje o żądaniach, jakie otrzymał serwer WWW, wraz z dodatkowymi danymi takimi jak adres IP klienta, datą i czasem żądania, metodą HTTP (np. GET), żądanym zasobem (ścieżką URL), kodem odpowiedzi serwera, rozmiarem odpowiedzi oraz informacjami o refererze (stronie, z której pochodził link do żądanego zasobu) i user-agent (informacje o przeglądarce i systemie operacyjnym klienta).

Na przykład, pierwsza linijka:

83.149.9.216 - - [17/May/2015:10:05:03 +0000] "GET /presentations/logstash-monitorama-2013/images/kibana-search.png HTTP/1.1" 200 203023 "http://semicomplete.com/presentations/logstash-monitorama-2013/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_9\_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.77 Safari/537.36"

oznacza:

Adres IP klienta: 83.149.9.216

Data i czas: 17/May/2015:10:05:03 +0000

Metoda HTTP: GET

Żądany zasób: /presentations/logstash-monitorama-2013/images/kibana-search.png

Kod odpowiedzi serwera: 200 (OK)

Rozmiar odpowiedzi: 203023 bajty

Referer: http://semicomplete.com/presentations/logstash-monitorama-2013/

User-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_9\_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.77 Safari/537.36

**2. Sposób identyfikacji sesji:**

Sesje zostały zidentyfikowane na podstawie interwałów czasowych między żądaniami użytkownika.

**Procedura:**

* Podział logów na adresy IP i znaczniki czasowe.
* Sortowanie czasów dla każdego adresu IP.
* Podział czasów na sesje z uwzględnieniem 15-minutowego interwału.

Parsowanie > grupowanie po adresie IP > sortowanie czasu > identyfikacja sesji > zapis

**3. Czas identyfikacji sesji:**

Czas identyfikacji sesji opiera się na 15-minutowym interwale między kolejnymi żądaniami, co rozdziela sesje.

Wybrany czas powinien być zrównoważony uwzgledniajacy potrzeby uzytkownikow i wymogi bezpieczeństwa.

**4. Długość sesji:**

Sesje zostały ograniczone do maksymalnie 15 minut, co pozwala na lepsze śledzenie aktywności użytkownika. Krótsze sesje są preferowane.

**5. Moment zakończenia sesji:**

Sesje kończą się, gdy czas między żądaniami przekracza 15 minut.

**5. Traktowanie pojedynczych żądań:**

Pojedyncze żądanie bez aktywności przez ponad 15 minut jest traktowane jako osobna sesja.

**6. Uzyskano dwa częste zbiory:**

Zbiór ("Mozilla/5.0) otrzymał wsparcie (support) wynoszące około 65%.

Zbiór ("msnbot/2.0b) otrzymał wsparcie (support) wynoszące około 11%.

Nie zostały znalezione żadne reguły asocjacyjne, co wskazuje na brak istotnych zależności między danymi elementami.

**7. Co możesz powiedzieć o tych regułach w kontekście budowy serwisu oraz użytkowania serwisu?**

Brak znalezionych reguł asocjacyjnych sugeruje brak silnych powiązań między analizowanymi danymi (user-agent string w tym przypadku) w logach. Może to oznaczać, że analizowane dane mogą być losowe lub nieposortowane, co z kolei może wynikać z różnorodności urządzeń, przeglądarek lub sposobów dostępu do serwisu.

W przypadku budowy serwisu może to wskazywać na brak specyficznych trendów lub zachowań użytkowników, które można byłoby zidentyfikować na podstawie analizy danych logów.

**8. Co możesz powiedzieć na temat topologii witryny na podstawie otrzymanej analizy asocjacyjnej?**

Z braku wykrytych reguł asocjacyjnych nie można wyciągnąć konkretnych wniosków na temat topologii witryny.

Brak znalezionych zależności sugeruje, że elementy analizowane (np. user-agent strings) nie mają silnych współzależności ani powiązań między sobą, co uniemożliwia wywnioskowanie o strukturze czy topologii witryny na podstawie tych danych.

**9. Roboty internetowe**

- Tak, w zbiorze danych były wpisy dotyczące robotów internetowych.

Wpisy dotyczące robotów internetowych zawierają informacje o żądaniach HTTP, w tym adresy IP, żądane zasoby (takie jak strony internetowe, pliki), statusy żądań (np. 200 - sukces, 404 - nie znaleziono), a także identyfikatory i nazwy robotów internetowych (np. Googlebot, Baiduspider, Sogou, msnbot, Twitterbot, DuckDuckBot, Exabot, AhrefsBot).

Aby zidentyfikować wpisy dotyczące robotów internetowych, użyto wyrażenia regularnego robot\_pattern, które pasuje do linii logów zawierających nazwy i identyfikatory znanych robotów internetowych. Następnie pasujące linie zostały zapisane do pliku znalezione\_roboty\_v2.txt.

**10. Ile różnych żądań zostało zarejestrowanych w analizowanym pliku log**

1498 roznych żądań

**11**. **Skąd były najczęstsze żądania?**

Najczęstsze adresy IP:

66.249.73.135: 480 żądań

46.105.14.53: 364 żądań

130.237.218.86: 356 żądań

75.97.9.59: 269 żądań

50.16.19.13: 113 żądań

**12. Najczęściej odwiedzane strony:**

- [18/May/2015:04:05:55 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

**: 4 odwiedzeń**

- [18/May/2015:03:05:11 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "http://www.semicomplete.com/blog/tags/puppet?flav=rss20" "Tiny Tiny RSS/1.11 (http://tt-rss.org/)"

**: 2 odwiedzeń**

- [17/May/2015:13:05:11 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

**: 2 odwiedzeń**

- [18/May/2015:03:05:23 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

**: 2 odwiedzeń**

- [18/May/2015:04:05:27 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/"

**: 2 odwiedzeń**

**13. Czy te najczęściej odwiedzane strony były odwiedzane jako pierwsze w sesji?**

Tak, najczęściej odwiedzane strony zawarte w analizowanym pliku logów były odwiedzane jako jedne z pierwszych w sesji. Najczęściej odwiedzaną stroną w kontekście bycia jedną z pierwszych w sesji było żądanie "[18/May/2015:04:05:55 +0000] "GET /blog/tags/puppet?flav=rss20 HTTP/1.1" 200 14872 "-" "UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/", które wystąpiło 4 razy w tej roli.

**14**. **Jak zmieniłaby się odpowiedź na pytanie nr 4, gdyby wydłużyć czas sesji?**

Wydłużenie czasu trwania sesji mogłoby spowodować, że pojedyncza sesja obejmowałaby więcej działań użytkownika na stronie. Jednakże, zwiększenie czasu trwania sesji może również skutkować tym, że użytkownik pozostanie w sesji nawet w przypadku dłuższych okresów nieaktywności, co mogłoby zakłócić analizę danych dotyczących aktywności użytkowników.

**15. Utwórz listę zasobów o największej liczbie żądań.**

/images/jordan-80.png: 533 requests

/images/web/2009/banner.png: 516 requests

/blog/tags/puppet?flav=rss20: 488 requests

/projects/xdotool/: 224 requests

/?flav=rss20: 217 requests

**16. Utwórz listę katalogów o największej liczbie żądań**

presentations: 2304 requests

blog: 1934 requests

images: 1243 requests

projects: 596 requests

files: 547 requests

**17. Ile jest zasobów w poszczególnych katalogach? (przykłady)**

presentations:

/presentations/logstash-scale11x/images/ahhh\_\_\_rage\_face\_by\_samusmmx-d5g5zap.png: 128 requests

/presentations/logstash-puppetconf-2012/: 51 requests

/presentations/puppet-at-loggly/puppet-at-loggly.pdf.html: 37 requests

/presentations/logstash-puppetconf-2012/images/kibana-logstash-downloads.png: 33 requests

/presentations/logstash-scale11x/images/logstash.png: 33 requests

articles:

/articles/dynamic-dns-with-dhcp/: 135 requests

/articles/ssh-security/: 55 requests

/articles/ppp-over-ssh/: 30 requests

/articles/openldap-with-saslauthd/: 18 requests

/articles/arp-security/: 16 requests

images:

/images/jordan-80.png: 533 requests

/images/web/2009/banner.png: 516 requests

/images/googledotcom.png: 101 requests

/images/logstash\_OSCON.pdf: 47 requests

/images/ec2\_m1large\_cost.png: 8 requests

**18. Czy potrafisz coś powiedzieć o strukturze poszczególnych katalogów? Czy potrafiłbyś ocenić tę strukturę oraz zależności zasobów w tych strukturach?**

1. Rozmaite rodzaje zasobów: Każdy katalog zawiera różne typy zasobów. Na przykład:

Katalog "articles" zawiera artykuły z różnymi tematami, np. bezpieczeństwo SSH, LDAP, ARP, itd.

Katalog "blog" wydaje się zawierać wpisy blogowe oraz tagi związane z tematami takimi jak Puppet, Firefox, problemy z Ubuntu, itd.

Katalogi takie jak "files", "scripts", "icons", "misc" mają zróżnicowane typy zasobów, od plików do skryptów czy ikon.

1. Różnorodność struktury:

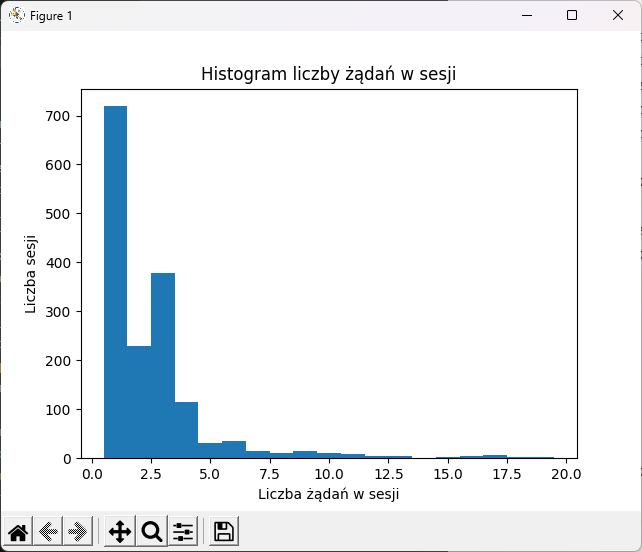
Struktura katalogów może być różnorodna w zależności od typu zasobów. Niektóre katalogi zawierają podkatalogi, inne posiadają pliki bezpośrednio w katalogu głównym.

1. Struktura na poziomie głównym:

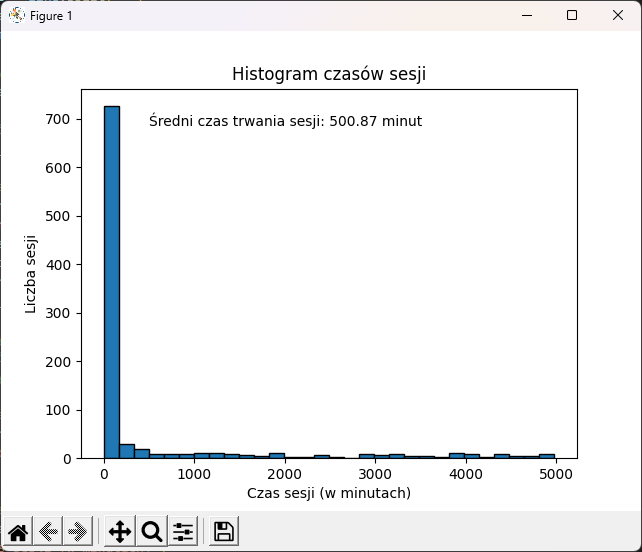
Katalog główny "/", jak również niektóre inne katalogi, zawiera zasoby bezpośrednio na tym poziomie, bez specyficznych podkatalogów.

1. Niezorganizowane zasoby:

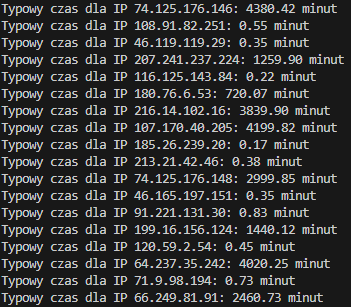
W przypadku niektórych katalogów można zaobserwować specyficzne adresy URL, np. "/administrator", "/test.xml", "/favicon.ico", które mogą wskazywać na strony z funkcjami administracyjnymi, testowymi, czy też plikami konfiguracyjnymi.

19. Histogram rozkładu – liczba żądań w sesji. 

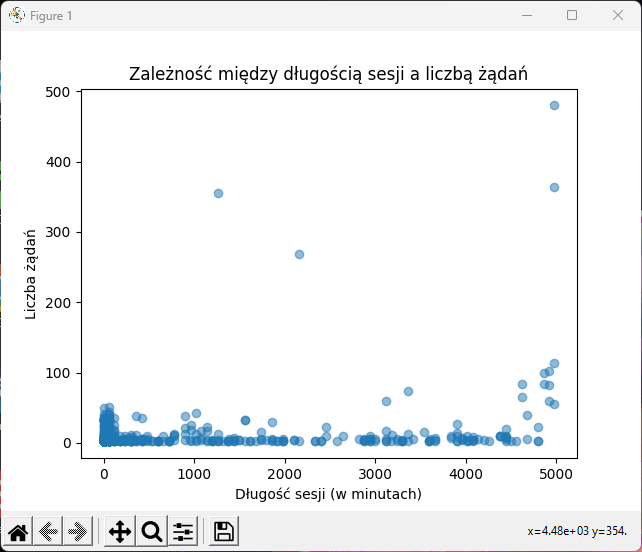
**20-23. Sredni czas trwania sesji - histogram**



Pzrykladowo:



24. **Korelacja** **miedzy dlugoscia sesji a liczba zadan**



**25**. **O ile średnio wydłuży się sesja, jeśli użytkownik zażądałby dodatkowej strony?**

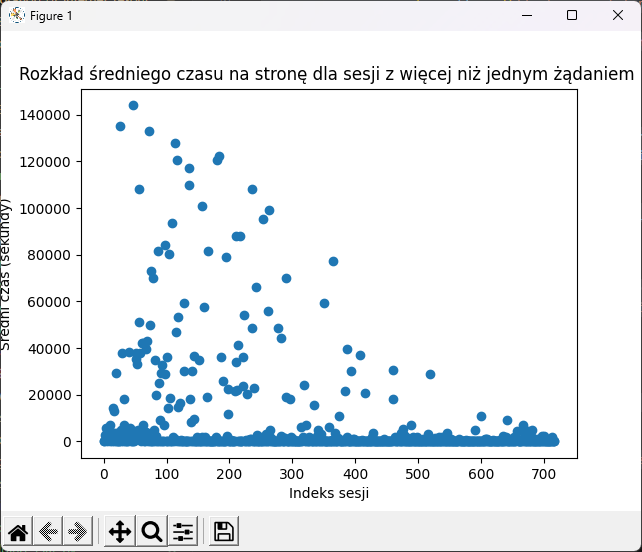
~ 0.0001637 sekund

**26. Średni czas przebywania na stronie:**

~ 73 minuty

**27. Jaki jest rozkład średniego czasu na stronę dla sesji składającej się z więcej niż jedno żądanie?**

Średni czas na stronę dla takich sesji: 7285s / 121min



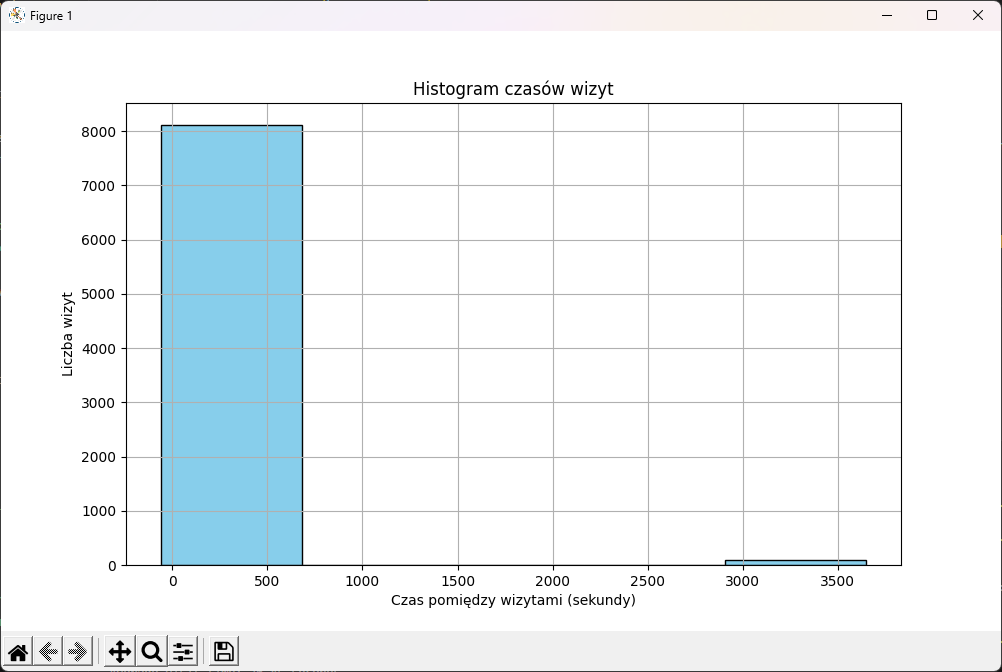
**28. Jaka jest całkowita liczba wizyt?**

8198

**29. Jaki jest średni czas wizyty?**

Średni czas wizyty: 36.45 sekund

**Zbuduj histogram czasów wizyt.**

****

**Jak się mają te czasy do czasów sesji?**

Sredni czas wizyt 36s – 500s sredni czas sesji

**30. Jaki jest minimalny czas wizyty?**

0 sekund ??

Np.

IP: 194.94.204.181

Session 1: 2015-05-17 18:05:26 - 2015-05-17 18:05:26

**31. Jak jest całkowita liczba różnych użytkowników?**

1666

**32. Jaka jest liczba użytkowników, jaka odwiedziła więcej niż jedną stronę, a ile takich którzy**

**mieli więcej niż jedną sesję danego dnia?**

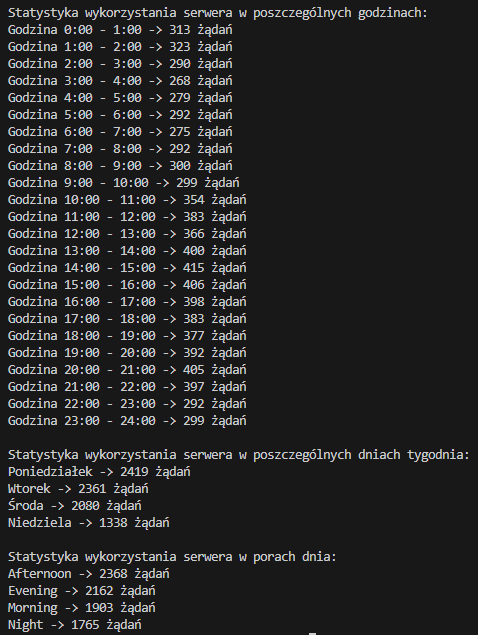
więcej niż jedną stronę: 2264

mieli więcej niż jedną sesję danego dnia: 946

**33. Jaka jest średnia liczba wizyt dla użytkownika?**

4.9208

**34. Sporządź statystykę wykorzystania serwera w poszczególnych dniach/tygodniach/etc.**

****

**35. Jaka jest liczba żądań/wywołać w sesji?**

Średnia liczba żądań w sesji: 0.2587