

CS 851: Assignment #2

Due on Thursday, March 5, 2014

DR NELSON 4:20pm

VICTOR NWALA

Contents

Problem 1	3
Problem 2	6
CONCLUSION	9

Problem 1

Choose 100 URIs from A1

Generate WARC files of those URIs using: wget WARCreate Heritrix (stand-alone or via WAIL) we-brecreorder.io

Describe the resulting WARC files: quantitatively compare & contrast the results of the WARC files of the same URI as generated by different tools choose interesting examples.

Demonstrate playback of 2-3 WARCs in the (Wayback Machine (via WAIL or stand-alone) or pywb) and (webrecorder.io)

Listing 1: Script to download wget Warc files

```
import requests
import urllib2
import urllib
from urlparse import urlparse
5 import subprocess
import os, sys
import httplib
import re

10 fh = open("sample.txt", 'r')
count = 0
for line in fh:
    try:
        url=line

15         proc = subprocess.Popen(["wget --warc-file="+str(count)+" -p -l 1 "+url+
            " "], stdout=subprocess.PIPE, shell=True)
        (out, err) = proc.communicate()
        count = count + 1
    except BaseException, e:
20         print 'failed ',str(e)
```

I used the same 103 URIs for wget, warcreate, webrecorder. I could not use the same 100 URIs for WAIL because of the delimiter errors I experience while doing it, so I changed the URI mixes to minimise the errors I experienced. I combined the warc file using WARCmerge.

In order to quantitatively compare the different methods, I selected 4 URIs to generate warc files using the 4 methods, and I discovered this: Warcreate was over 33.5 MB, over because 1 URI failed to generate a warc file. WAIL was 3.71 MB, Web Recorder 4.6 MB, Wget 165.4 KB. Hence Warcreate warc files are the largest while wget warc files are the smallest.

Figure 1: Comparing file sizes for the different warc tools

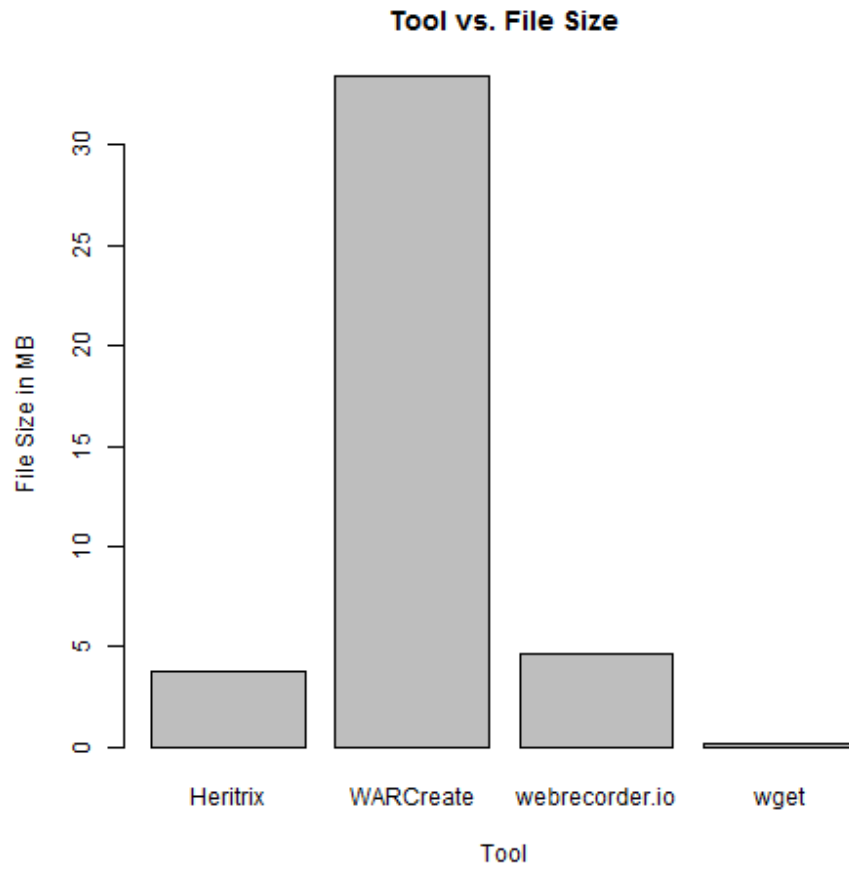


Figure 2: Replaying Warc file with webrecorder

WebRecorder.io REPLAY Expires In: 20:41

Note: Some or all of this archived data was not created in WebRecorder.io and its authenticity can not be verified by WebRecorder.io

Recorded Pages Url Search Total Archive Size: 24.74 MB

Below is WebRecorder.io best-guess on which urls are actual pages (up to 500 pages per archive). Some pages may not have been detected. Use the *Url Search* tab to lookup specific urls that may not be listed here.

Search:

Showing 1 to 274 of 274 entries

Page	Recorded At
http://fashionablycrohns.blogspot.co.uk/p/about-me.html	3/4/2015, 6:14:49 PM
http://pinkmario.tumblr.com/analytics.html	3/4/2015, 6:14:49 PM
http://sassitude.ca/products/bustier-top	3/4/2015, 6:14:49 PM
https://farm6.staticflickr.com/	3/4/2015, 6:14:49 PM
http://tooxclusive.com/deferredfunctions.js	3/4/2015, 6:14:49 PM
https://farm1.staticflickr.com/	3/4/2015, 6:14:49 PM
http://glamdoodle.tumblr.com/analytics.html	3/4/2015, 6:14:49 PM

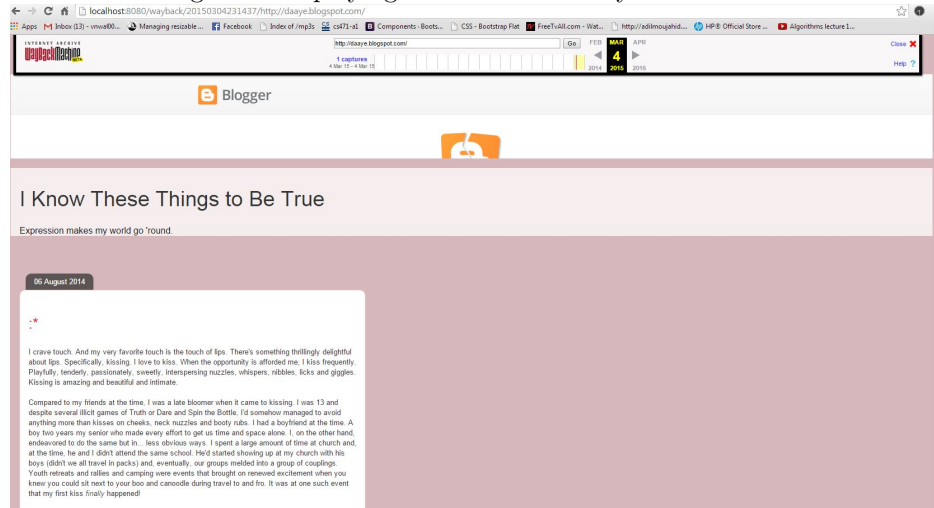
For any questions, comments, inquiries or feature requests,
contact: info@webrecorder.io

Donations graciously accepted: [Gratipay](#)

Figure 3: Replaying Warc file with webrecorder

https://...replay/ x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19 x20 x21 x22 x23 x24 x25 x26 x27 x28 x29 x30 x31 x32 x33 x34 x35 x36 x37 x38 x39 x40 x41 x42 x43 x44 x45 x46 x47 x48 x49 x50 x51 x52 x53 x54 x55 x56 x57 x58 x59 x60 x61 x62 x63 x64 x65 x66 x67 x68 x69 x70 x71 x72 x73 x74 x75 x76 x77 x78 x79 x80 x81 x82 x83 x84 x85 x86 x87 x88 x89 x90 x91 x92 x93 x94 x95 x96 x97 x98 x99 x100 x101 x102 x103 x104 x105 x106 x107 x108 x109 x110 x111 x112 x113 x114 x115 x116 x117 x118 x119 x120 x121 x122 x123 x124 x125 x126 x127 x128 x129 x130 x131 x132 x133 x134 x135 x136 x137 x138 x139 x140 x141 x142 x143 x144 x145 x146 x147 x148 x149 x150 x151 x152 x153 x154 x155 x156 x157 x158 x159 x160 x161 x162 x163 x164 x165 x166 x167 x168 x169 x170 x171 x172 x173 x174 x175 x176 x177 x178 x179 x180 x181 x182 x183 x184 x185 x186 x187 x188 x189 x190 x191 x192 x193 x194 x195 x196 x197 x198 x199 x200 x201 x202 x203 x204 x205 x206 x207 x208 x209 x210 x211 x212 x213 x214 x215 x216 x217 x218 x219 x220 x221 x222 x223 x224 x225 x226 x227 x228 x229 x230 x231 x232 x233 x234 x235 x236 x237 x238 x239 x240 x241 x242 x243 x244 x245 x246 x247 x248 x249 x250 x251 x252 x253 x254 x255 x256 x257 x258 x259 x260 x261 x262 x263 x264 x265 x266 x267 x268 x269 x270 x271 x272 x273 x274 x275 x276 x277 x278 x279 x280 x281 x282 x283 x284 x285 x286 x287 x288 x289 x290 x291 x292 x293 x294 x295 x296 x297 x298 x299 x300 x301 x302 x303 x304 x305 x306 x307 x308 x309 x310 x311 x312 x313 x314 x315 x316 x317 x318 x319 x320 x321 x322 x323 x324 x325 x326 x327 x328 x329 x330 x331 x332 x333 x334 x335 x336 x337 x338 x339 x340 x341 x342 x343 x344 x345 x346 x347 x348 x349 x350 x351 x352 x353 x354 x355 x356 x357 x358 x359 x360 x361 x362 x363 x364 x365 x366 x367 x368 x369 x370 x371 x372 x373 x374 x375 x376 x377 x378 x379 x380 x381 x382 x383 x384 x385 x386 x387 x388 x389 x390 x391 x392 x393 x394 x395 x396 x397 x398 x399 x400 x401 x402 x403 x404 x405 x406 x407 x408 x409 x410 x411 x412 x413 x414 x415 x416 x417 x418 x419 x420 x421 x422 x423 x424 x425 x426 x427 x428 x429 x430 x431 x432 x433 x434 x435 x436 x437 x438 x439 x440 x441 x442 x443 x444 x445 x446 x447 x448 x449 x450 x451 x452 x453 x454 x455 x456 x457 x458 x459 x460 x461 x462 x463 x464 x465 x466 x467 x468 x469 x470 x471 x472 x473 x474 x475 x476 x477 x478 x479 x480 x481 x482 x483 x484 x485 x486 x487 x488 x489 x490 x491 x492 x493 x494 x495 x496 x497 x498 x499 x500 x501 x502 x503 x504 x505 x506 x507 x508 x509 x510 x511 x512 x513 x514 x515 x516 x517 x518 x519 x520 x521 x522 x523 x524 x525 x526 x527 x528 x529 x530 x531 x532 x533 x534 x535 x536 x537 x538 x539 x540 x541 x542 x543 x544 x545 x546 x547 x548 x549 x550 x551 x552 x553 x554 x555 x556 x557 x558 x559 x560 x561 x562 x563 x564 x565 x566 x567 x568 x569 x570 x571 x572 x573 x574 x575 x576 x577 x578 x579 x580 x581 x582 x583 x584 x585 x586 x587 x588 x589 x590 x591 x592 x593 x594 x595 x596 x597 x598 x599 x600 x601 x602 x603 x604 x605 x606 x607 x608 x609 x610 x611 x612 x613 x614 x615 x616 x617 x618 x619 x620 x621 x622 x623 x624 x625 x626 x627 x628 x629 x630 x631 x632 x633 x634 x635 x636 x637 x638 x639 x640 x641 x642 x643 x644 x645 x646 x647 x648 x649 x650 x651 x652 x653 x654 x655 x656 x657 x658 x659 x660 x661 x662 x663 x664 x665 x666 x667 x668 x669 x670 x671 x672 x673 x674 x675 x676 x677 x678 x679 x680 x681 x682 x683 x684 x685 x686 x687 x688 x689 x690 x691 x692 x693 x694 x695 x696 x697 x698 x699 x700 x701 x702 x703 x704 x705 x706 x707 x708 x709 x710 x711 x712 x713 x714 x715 x716 x717 x718 x719 x720 x721 x722 x723 x724 x725 x726 x727 x728 x729 x730 x731 x732 x733 x734 x735 x736 x737 x738 x739 x740 x741 x742 x743 x744 x745 x746 x747 x748 x749 x750 x751 x752 x753 x754 x755 x756 x757 x758 x759 x760 x761 x762 x763 x764 x765 x766 x767 x768 x769 x770 x771 x772 x773 x774 x775 x776 x777 x778 x779 x780 x781 x782 x783 x784 x785 x786 x787 x788 x789 x790 x791 x792 x793 x794 x795 x796 x797 x798 x799 x800 x801 x802 x803 x804 x805 x806 x807 x808 x809 x810 x811 x812 x813 x814 x815 x816 x817 x818 x819 x820 x821 x822 x823 x824 x825 x826 x827 x828 x829 x830 x831 x832 x833 x834 x835 x836 x837 x838 x839 x840 x841 x842 x843 x844 x845 x846 x847 x848 x849 x850 x851 x852 x853 x854 x855 x856 x857 x858 x859 x860 x861 x862 x863 x864 x865 x866 x867 x868 x869 x870 x871 x872 x873 x874 x875 x876 x877 x878 x879 x880 x881 x882 x883 x884 x885 x886 x887 x888 x889 x890 x891 x892 x893 x894 x895 x896 x897 x898 x899 x900 x901 x902 x903 x904 x905 x906 x907 x908 x909 x910 x911 x912 x913 x914 x915 x916 x917 x918 x919 x920 x921 x922 x923 x924 x925 x926 x927 x928 x929 x930 x931 x932 x933 x934 x935 x936 x937 x938 x939 x940 x941 x942 x943 x944 x945 x946 x947 x948 x949 x950 x951 x952 x953 x954 x955 x956 x957 x958 x959 x960 x961 x962 x963 x964 x965 x966 x967 x968 x969 x970 x971 x972 x973 x974 x975 x976 x977 x978 x979 x980 x981 x982 x983 x984 x985 x986 x987 x988 x989 x990 x991 x992 x993 x994 x995 x996 x997 x998 x999 x1000 x1001 x1002 x1003 x1004 x1005 x1006 x1007 x1008 x1009 x1010 x1011 x1012 x1013 x1014 x1015 x1016 x1017 x1018 x1019 x1020 x1021 x1022 x1023 x1024 x1025 x1026 x1027 x1028 x1029 x1030 x1031 x1032 x1033 x1034 x1035 x1036 x1037 x1038 x1039 x1040 x1041 x1042 x1043 x1044 x1045 x1046 x1047 x1048 x1049 x1050 x1051 x1052 x1053 x1054 x1055 x1056 x1057 x1058 x1059 x1060 x1061 x1062 x1063 x1064 x1065 x1066 x1067 x1068 x1069 x1070 x1071 x1072 x1073 x1074 x1075 x1076 x1077 x1078 x1079 x1080 x1081 x1082 x1083 x1084 x1085 x1086 x1087 x1088 x1089 x1090 x1091 x1092 x1093 x1094 x1095 x1096 x1097 x1098 x1099 x1100 x1101 x1102 x1103 x1104 x1105 x1106 x1107 x1108 x1109 x1110 x1111 x1112 x1113 x1114 x1115 x1116 x1117 x1118 x1119 x1120 x1121 x1122 x1123 x1124 x1125 x1126 x1127 x1128 x1129 x1130 x1131 x1132 x1133 x1134 x1135 x1136 x1137 x1138 x1139 x1140 x1141 x1142 x1143 x1144 x1145 x1146 x1147 x1148 x1149 x1150 x1151 x1152 x1153 x1154 x1155 x1156 x1157 x1158 x1159 x1160 x1161 x1162 x1163 x1164 x1165 x1166 x1167 x1168 x1169 x1170 x1171 x1172 x1173 x1174 x1175 x1176 x1177 x1178 x1179 x1180 x1181 x1182 x1183 x1184 x1185 x1186 x1187 x1188 x1189 x1190 x1191 x1192 x1193 x1194 x1195 x1196 x1197 x1198 x1199 x1200 x1201 x1202 x1203 x1204 x1205 x1206 x1207 x1208 x1209 x1210 x1211 x1212 x1213 x1214 x1215 x1216 x1217 x1218 x1219 x1220 x1221 x1222 x1223 x1224 x1225 x1226 x1227 x1228 x1229 x1230 x1231 x1232 x1233 x1234 x1235 x1236 x1237 x1238 x1239 x1240 x1241 x1242 x1243 x1244 x1245 x1246 x1247 x1248 x1249 x1250 x1251 x1252 x1253 x1254 x1255 x1256 x1257 x1258 x1259 x1260 x1261 x1262 x1263 x1264 x1265 x1266 x1267 x1268 x1269 x1270 x1271 x1272 x1273 x1274 x1275 x1276 x1277 x1278 x1279 x1280 x1281 x1282 x1283 x1284 x1285 x1286 x1287 x1288 x1289 x1290 x1291 x1292 x1293 x1294 x1295 x1296 x1297 x1298 x1299 x1300 x1301 x1302 x1303 x1304 x1305 x1306 x1307 x1308 x1309 x1310 x1311 x1312 x1313 x1314 x1315 x1316 x1317 x1318 x1319 x1320 x1321 x1322 x1323 x1324 x1325 x1326 x1327 x1328 x1329 x1330 x1331 x1332 x1333 x1334 x1335 x1336 x1337 x1338 x1339 x1340 x1341 x1342 x1343 x1344 x1345 x1346 x1347 x1348 x1349 x1350 x1351 x1352 x1353 x1354 x1355 x1356 x1357 x1358 x1359 x1360 x1361 x1362 x1363 x1364 x1365 x1366 x1367 x1368 x1369 x1370 x1371 x1372 x1373 x1374 x1375 x1376 x1377 x1378 x1379 x1380 x1381 x1382 x1383 x1384 x1385 x1386 x1387 x1388 x1389 x1390 x1391 x1392 x1393 x1394 x1395 x1396 x1397 x1398 x1399 x1400 x1401 x1402 x1403 x1404 x1405 x1406 x1407 x1408 x1409 x1410 x1411 x1412 x1413 x1414 x1415 x1416 x1417 x1418 x1419 x1420 x1421 x1422 x1423 x1424 x1425 x1426 x1427 x1428 x1429 x1430 x1431 x1432 x1433 x1434 x1435 x1436 x1437 x1438 x1439 x1440 x1441 x1442 x1443 x1444 x1445 x1446 x1447 x1448 x1449 x1450 x1451 x1452 x1453 x1454 x1455 x1456 x1457 x1458 x1459 x1460 x1461 x1462 x1463 x1464 x1465 x1466 x1467 x1468 x1469 x1470 x1471 x1472 x1473 x1474 x1475 x1476 x1477 x1478 x1479 x1480 x1481 x1482 x1483 x1484 x1485 x1486 x1487 x1488 x1489 x1490 x1491 x1492 x1493 x1494 x1495 x1496 x1497 x1498 x1499 x1500 x1501 x1502 x1503 x1504 x1505 x1506 x1507 x1508 x1509 x1510 x1511 x1512 x1513 x1514 x1515 x1516 x1517 x1518 x1519 x1520 x1521 x1522 x1523 x1524 x1525 x1526 x1527 x1528 x1529 x1530 x1531 x1532 x1533 x1534 x1535 x1536 x1537 x1538 x1539 x1540 x1541 x1542 x1543 x1544 x1545 x1546 x1547 x1548 x1549 x1550 x1551 x1552 x1553 x1554 x1555 x1556 x1557 x1558 x1559 x1560 x1561 x1562 x1563 x1564 x1565 x1566 x1567 x1568 x1569 x1570 x1571 x1572 x1573 x1574 x1575 x1576 x1577 x1578 x1579 x1580 x1581 x1582 x1583 x1584 x1585 x1586 x1587 x1588 x1589 x1590 x1591 x1592 x1593 x1594 x1595 x1596 x1597 x1598 x1599 x1600 x1601 x1602 x1603 x1604 x1605 x1606 x1607 x1608 x1609 x1610 x1611 x1612 x1613 x1614 x1615 x1616 x1617 x1618 x1619 x1620 x1621 x1622 x1623 x1624 x1625 x1626 x1627 x1628 x1629 x1630 x1631 x1632 x1633 x1634 x1635 x1636 x1637 x1638 x1639 x1640 x1641 x1642 x1643 x1644 x1645 x1646 x1647 x1648 x1649 x1650 x1651 x1652 x1653 x1654 x1655 x1656 x1657 x1658 x1659 x1660 x1661 x1662 x1663 x1664 x1665 x1666 x1667 x1668 x1669 x1670 x1671 x1672 x1673 x1674 x1675 x1676 x1677 x1678 x1679 x1680 x1681 x1682 x1683 x1684 x1685 x1686 x1687 x1688 x1689 x1690 x1691 x1692 x1693 x1694 x1695 x1696 x1697 x1698 x1699 x1700 x1701 x1702 x1703 x1704 x1705 x1706 x1707 x1708 x1709 x1710 x1711 x1712 x1713 x1714 x1715 x1716 x1717 x1718 x1719 x1720 x1721 x1722 x1723 x1724 x1725 x1726 x1727 x1728 x1729 x1730 x1731 x1732 x1733 x1734 x1735 x1736 x1737 x1738 x1739 x1740 x1741 x1742 x1743 x1744 x1745 x1746 x1747 x1748 x1749 x1750 x1751 x1752 x1753 x1754 x1755 x1756 x1757 x1758 x1759 x1760 x1761 x1762 x1763 x1764 x1765 x1766 x1767 x1768 x1769 x1770 x1771 x1772 x1773 x1774 x1775 x1776 x1777 x1778 x1779 x1780 x1781 x1782 x1783 x1784 x1785 x1786 x1787 x1788 x1789 x1790 x1791 x1792 x1793 x1794 x1795 x1796 x1797 x1798 x1799 x1800 x1801 x1802 x1803 x1804 x1805 x1806 x1807 x1808 x1809 x1810 x1811 x1812 x1813 x1814 x1815 x1816 x1817 x1818 x1819 x1820 x1821 x1822 x1823 x1824 x1825 x1826 x1827 x1828 x1829 x1830 x1831 x1832 x1833 x1834 x1835 x1836 x1837 x1838 x1839 x1840 x1841 x1842 x1843 x1844 x1845 x1846 x1847 x1848 x1849 x1850 x1851 x1852 x1853 x1854 x1855 x1856 x1857 x1858 x1859 x1860 x1861 x1862 x1863 x1864 x1865 x1866 x1867 x1868 x1869 x1870 x1871 x1872 x1873 x1874 x1875 x1876 x1877 x1878 x1879 x1880 x1881 x1882 x1883 x1884 x1885 x1886 x1887 x1888 x1889 x1890 x1891 x1892 x1893 x1894 x1895 x1896 x1897 x1898 x1899 x1900 x1901 x1902 x1903 x1904 x1905 x1906 x1907 x1908 x1909 x1910 x1911 x1912 x1913 x1914 x1915 x1916 x1917 x1918 x1919 x1920 x1921 x1922 x1923 x1924 x1925 x1926 x1927 x1928 x1929 x1930 x1931 x1932 x1933 x1934 x1935 x1936 x1937 x1938 x1939 x1940 x1941 x1942 x1943 x1944 x1945 x1946 x1947 x1948 x1949 x1950 x1951 x1952 x1953 x1954 x1955 x1956 x1957 x1958 x1959 x1960 x1961 x1962 x1963 x1964 x1965 x1966 x1967 x1968 x1969 x1970 x1971 x1972 x1973 x1974 x1975 x1976 x1977 x1978 x1979 x1980 x1981 x1982 x1983 x1984 x1985 x1986 x1987 x1988 x1989 x1990 x1991 x1992 x1993 x1994 x1995 x1996 x1997 x1998 x1999 x2000 x2001 x2002 x2003 x2004 x2005 x2006 x2007 x2008 x2009 x2010 x2011 x2012 x2013 x2014 x2015 x2016 x2017 x2018 x2019 x2020 x2021 x2022 x2023 x2024 x2025 x2026 x2027 x2028 x2029 x2030 x2031 x2032 x2033 x2034 x2035 x2036 x2037 x2038 x2039 x2040 x2041 x2042 x2043 x2044 x2045 x2046 x2047 x2048 x2049 x2050 x2051 x2052 x2053 x2054 x2055 x2056 x2057 x2058 x2059 x2060 x2061 x2062 x2063 x2064 x2065 x2066 x2067 x2068 x2069 x2070 x2071 x2072 x2073 x2074 x2075 x2076 x2077 x2078 x2079 x2080 x2081 x2082 x2083 x2084 x2085 x2086 x2087 x2088 x2089 x2090 x2091 x2092 x2093 x2094 x2095 x2096 x2097 x2098 x2099 x2100 x2101 x2102 x2103 x2104 x2105 x2106 x2107 x2108 x2109 x2110 x2111 x2112 x2113 x2114 x2115 x2116 x2117 x2118 x2119 x2120 x2121 x2122 x2123 x2124 x2125 x2126 x2127 x2128 x2129 x2130 x2131 x2132 x2133 x2134 x2135 x2136 x2137 x2138 x2139 x2140 x2141 x2142 x2143 x2144 x2145 x2146 x2147 x2148 x2149 x2150 x2151 x2152 x2153 x2154 x2155 x2156 x2157 x2158 x2159 x2160 x2161 x2162 x2163 x2164 x2165 x2166 x2167 x2168 x2169 x2170 x2171 x2172 x2173 x2174 x2175 x2176 x2177 x2178 x2179 x2180 x2181 x2182 x2183 x2184 x2185 x2186 x2187 x2188 x2189 x2190 x2191 x2192 x2193 x2194 x2195 x2196 x2197 x2198 x2199 x2200 x2201 x2202 x2203 x2204 x2205 x2206 x2207 x2208 x2209 x2210 x2211 x2212 x2213 x2214 x2215 x2216 x2217 x2218 x2219 x2220 x2221 x2222 x2223 x2224 x2225 x2226 x2227 x2228 x2229 x2230 x2231 x2232 x2233 x2234 x2235 x2236 x2237 x2238 x2239 x2240 x2241 x2242 x2243 x2244 x2245 x2246 x2247 x2248 x2249 x2250 x2251 x2252 x2253 x2254 x2255 x2256 x2257 x2258 x2259 x2260 x2261 x2262 x2263 x2264 x2265 x2266 x2267 x2268 x2269 x2270 x2271 x2272 x2273 x2274 x2275 x2276 x2277 x2278 x2279 x2280 x2281 x2282 x2283 x2284 x2285 x2286 x2287 x2288 x2289 x2290 x2291 x2292 x2293 x2294 x2295 x2296 x2297 x2298 x2299 x2300 x2301 x2302 x2303 x2304 x2305 x2306 x2307 x2308 x2309 x2310 x2311 x2312 x2313 x2314 x2315 x2316 x2317 x2318 x2319 x2320 x2321 x2322 x2323 x2324 x2325 x2326 x2327 x2328 x2329 x2330 x2331 x2332 x2333 x2334 x2335 x2336 x2337 x2338 x2339 x2340 x2341 x2342 x2343 x2344 x2345 x2346 x2347 x2348 x2349 x2350 x2351 x2352 x2353 x2354 x2355 x2356 x2357 x2358 x2359 x2360 x2361 x2362 x2363 x2364 x2365 x2366 x2367 x2368 x2369 x2370 x2371 x2372 x2373 x2374 x2375 x2376 x2377 x2378 x2379 x2380 x2381 x2382 x2383 x2384 x2385 x2386 x2387 x2388 x2389 x2390 x2391 x2392 x2393 x2394 x2395 x2396 x2397 x2398 x2399 x2400 x2401 x2402 x2403 x2404 x2405 x2406 x2407 x2408 x2409 x2410 x2411 x2412 x2413 x2414 x2415 x2416 x2417 x2418 x2419 x2420 x2421 x2422 x2423 x2424 x2425 x2426 x2427 x2428 x2429 x2430 x2431 x2432 x2433 x2434 x2435 x2436 x2437 x2438 x2439 x2440 x2441 x2442 x2443 x2444 x2445 x2446 x2447 x2448 x2449 x2450 x2451 x2452 x2453 x2454 x2455 x2456 x2457 x2458 x2459 x2460 x2461 x2462 x2463 x2464 x2465 x2466 x2467 x2468 x2469 x2470 x2471 x2472 x2473 x2474 x2475 x2476 x2477 x2478 x2479 x2480 x2481 x2482 x2483 x2484 x2485 x2486 x2487 x248

Figure 5: Replaying Warc file with Wayback Machine



I noticed for wayback machine I could only search for 1 URL at a time even if I have a warc file with several URIs, while for webrecorder I could see all URIs in a warc file, then you can click on the link you want.

Problem 2

Ingest the 100 URIs from their resulting WARC files into a SOLR instance see the code + tutorial at: <https://github.com/ukwa/webarchive-discovery> Demonstrate several functioning queries on the files (a full front-end is not required) describe the configuration choices you made in setting up SOLR and processing the documents.

The configuration choices I made are still the same as the default settings. Some of which are; Maximum payload size allowed to be kept wholly in RAM: 10M. Maximum payload size that will be serialised out to disk instead of held in RAM: 100M. URLs to skip: NONE. Use the hash+url as the ID for the documents. Do not check SOLR for duplicates during indexing. Solr document batch size for submissions: 500.

Figure 6: Embedding URIs in SOLR

```

Terminal - vnwala@template: ~/webarchive-discovery/warc-indexer
File Edit View Terminal Tabs Help

vnwala@template:~/webarchive-discovery/warc-indexer$ java -jar warc-indexer-2.0.
1-20150116.110435-2-jar-with-dependencies.jar -s http://localhost:8080/discover
y -t /home/vnwala/merge/*.warc
2015-03-05 21:27:38 INFO WARCIndexer:153 - Extract text = true
2015-03-05 21:27:38 INFO WARCIndexer:156 - Store text = true
2015-03-05 21:27:38 INFO WARCIndexer:158 - hashUrlId = true
2015-03-05 21:27:38 INFO WARCIndexer:198 - Hashing & Caching thresholds are: <
10485760 in memory, < 104857600 on disk.
2015-03-05 21:27:38 INFO WARCIndexer:201 - Setting up analysers...
2015-03-05 21:27:38 INFO WARCPayloadAnalysers:85 - Image feature extraction = t
rue
2015-03-05 21:27:42 INFO TikaExtractor:116 - Config: MIME exclude list: [x-tar,
x-gzip, bz, lz, compress, zip, javascript, css, octet-stream]
2015-03-05 21:27:42 INFO TikaExtractor:119 - Config: Parser timeout (ms) 300000
2015-03-05 21:27:42 INFO TikaExtractor:122 - Config: Maximum length of text to
extract (characters) 524288
2015-03-05 21:27:42 INFO TikaExtractor:126 - Config: extractAllMetadata false
2015-03-05 21:27:42 INFO TikaExtractor:129 - Config: useBoilerpipe false
2015-03-05 21:27:42 INFO HTMLAnalyser:59 - HTML - Extract resource links false
2015-03-05 21:27:42 INFO HTMLAnalyser:61 - HTML - Extract host links true
2015-03-05 21:27:42 INFO HTMLAnalyser:63 - HTML - Extract domain links true
2015-03-05 21:27:42 INFO HTMLAnalyser:65 - HTML - Extract elements used true
2015-03-05 21:27:42 INFO ImageAnalyser:69 - Image - detect faces = true
2015-03-05 21:27:42 INFO ImageAnalyser:71 - Image - max size in bytes 1048576
2015-03-05 21:27:42 INFO ImageAnalyser:74 - Image sample rate 0.1
2015-03-05 21:27:42 INFO FaceDetectionParser:90 - Face detection enabled.
2015-03-05 21:27:42 INFO FaceDetectionParser:92 - Dominant colour extraction en
abled.
2015-03-05 21:27:43 INFO LanguageAnalyser:51 - Constructing...
2015-03-05 21:27:45 INFO WARCIndexer:206 - Initialisation of WARCIndexer comple
te.
Parsing Archive File [1/69]:/home/vnwala/merge/100.warc
Parsing Archive File [2/69]:/home/vnwala/merge/101.warc
Parsing Archive File [3/69]:/home/vnwala/merge/102.warc
Parsing Archive File [4/69]:/home/vnwala/merge/103.warc
Parsing Archive File [5/69]:/home/vnwala/merge/34.warc
Parsing Archive File [6/69]:/home/vnwala/merge/35.warc
2015-03-05 21:32:00 ERROR TikaExtractor:405 - ParseRunner.run() Exception: Write
OutContentHandler.WriteLimitReachedException: Your document contained more than
524288 characters, and so your requested limit has been reached. To receive the
full text of the document, increase your limit. (Text up to the limit is however
available).
Parsing Archive File [7/69]:/home/vnwala/merge/36.warc
Parsing Archive File [8/69]:/home/vnwala/merge/37.warc
2015-03-05 21:32:27 ERROR WARCPayloadAnalysers:162 - java.lang.IllegalArgumentEx
ception: Illegal character in path at index 10: file:///.[KAI RISM]+140917,+1409
20-140921,+140928+KAI_32P.rar: Illegal character in path at index 10: file:///.[
KAI RISM]+140917,+140920-140921,+140928+KAI_32P.rar;dd; http://www.mediafire.com
/download/5nm8xqxql48qdin/%5BKAI RISM%5D+140917%2C+140920-140921%2C+140928+KAI_32
P.rar @3121
2015-03-05 21:32:27 ERROR AbstractPayloadAnalyser:66 - uk.bl.wa.tika.parser.imag
efeatures.FaceDetectionParser.parse(): org.apache.sanselan.ImageReadException: C
an't parse this format.
2015-03-05 21:32:27 ERROR WARCPayloadAnalysers:162 - java.lang.IllegalArgumentEx
ception: Illegal character in path at index 10: file:///.[KAI RISM]+140917,+1409
20-140921,+140928+KAI_32P.rar: Illegal character in path at index 10: file:///.[

```


Figure 7: Query for movies

Request-Handler (qt) /select

q *MOVIES*

fq

sort

start.rows 0 10

fl

df

Raw Query Parameters key1=val1&key2=val2

wt json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☐ facet

☐ spatial

☐ spellcheck

http://localhost:8080/discovery/select?q=*MOVIES*&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1459,
    "params": {
      "indent": "true",
      "q": "*MOVIES*",
      "_": "1425614947599",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 13,
    "start": 0,
    "docs": [
      {
        "source_file_s": "20150228053738771.warc@45750",
        "url": "https://www.familyvideo.com/catalog/product_used.php?products_id=442805&source=webgains&siteid=80806",
        "content_type_ext": "php",
        "host": "www.familyvideo.com",
        "domain": "familyvideo.com",
        "public_suffix": "com",
        "server": {
          "apache/2.2.22 (EL)"
        },
        "content_type_served": "text/html; charset=ISO-8859-1",
        "content_length": 94081,
        "id": "sha1:0K4B71A7UAXJYVJPUJTHP2UMBNPF/0uR0Jgrd50Iky104E2Pew==",
        "hash": [
          "sha1:0K4B71A7UAXJYVJPUJTHP2UMBNPF"
        ],
        "crawl_date": "2015-02-28T05:37:38Z",
        "crawl_year": "2015",
        "wayback_date": "20150228053738",
        "content": [
          "Buy Guardians of the Galaxy Blu-ray Disc (USED)All Movie Categories\\Cuddle-Close\\ Horror Movies 3D Movies Act"
        ],
        "content_text_length": 10224,
        "content_type": {

```

Figure 8: Query for domain

Request-Handler (qt) /select

q *domain*

fq

sort

start.rows 0 10

fl

df

Raw Query Parameters key1=val1&key2=val2

wt json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☐ facet

☐ spatial

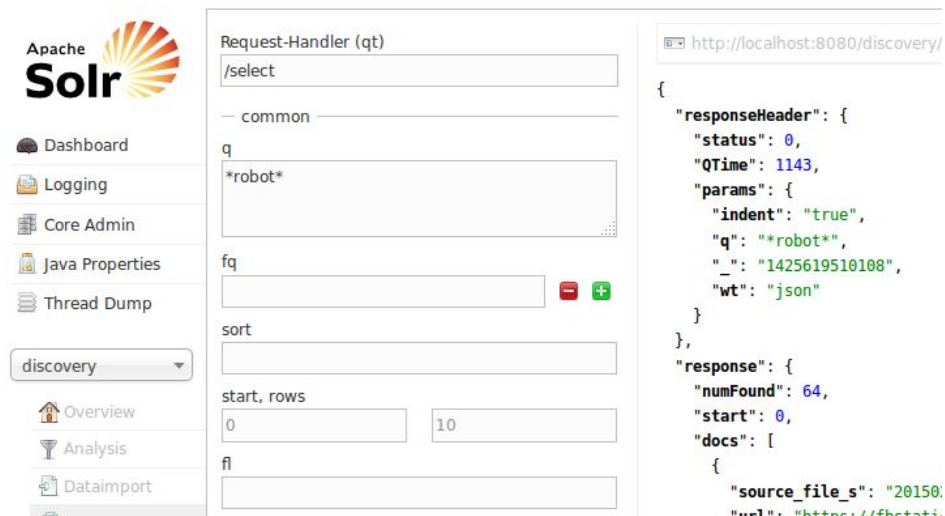
☐ spellcheck

http://localhost:8080/discovery/select?q=*domain*&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 410,
    "params": {
      "indent": "true",
      "q": "*domain*",
      "_": "1425615896402",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 71,
    "start": 0,
    "docs": [
      {
        "source_file_s": "20150228055711445.warc@5088753",
        "url": "https://abs.twimg.com/a/1424828248/css/tl/twitter_more_1.bundle.css",
        "content_type_ext": "css",
        "host": "abs.twimg.com",
        "domain": "twimg.com",
        "public_suffix": "com",
        "server": {
          "tsa_b"
        },
        "content_length": 266091,
        "id": "sha1:7T7ZE0QSKU6PSDFBG6DTQD6NQPMNV7FB/0u3QVnA2dohfSUz4e701YQ==",
        "hash": [
          "sha1:7T7ZE0QSKU6PSDFBG6DTQD6NQPMNV7FB"
        ],
        "crawl_date": "2015-02-28T05:57:11Z",
        "crawl_year": "2015",
        "wayback_date": "20150228055711",
        "content": [
          ".tooltip{position:absolute;z-index:5001;display:block;visibility:visible;pad"
        ],

```


Figure 9: Query for files with robot.txt or the word robot, 64 documents with robot.txt files



The screenshot displays the Apache Solr Admin UI. On the left is a sidebar with navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, and a dropdown menu currently showing 'discovery'. Below these are links for Overview, Analysis, and Dataimport. The main panel is titled 'Request-Handler (qt)' and contains several input fields: a path field with '/select', a 'common' section, a query field 'q' containing '*robot*', a field 'fq' with a red minus and green plus icon, a 'sort' field, a 'start, rows' section with '0' and '10' in input boxes, and a 'fl' field. On the right, a JSON response is shown for the URL 'http://localhost:8080/discovery/'. The JSON includes a 'responseHeader' with status 0, QTime 1143, and parameters for indent, q, _ (id), and wt. The main 'response' object shows 'numFound': 64, 'start': 0, and 'docs': an array of documents, with the first document having 'source_file_s': '20150'.

CONCLUSION

I could not upload my warc files on github because of their size, I also noticed some URIs did not produce warc files in either Warcreate or Web Recorder. I produced 1 warc file for Heritrix with 100 URIs, while for the rest I did it one at a time and combined them. I am still learning how to query SOLR, hence my querying methods were not really complex. I did not change the configuration of SOLR because I had already uploaded my warc files.