|  |  |
| --- | --- |
| Lang-Wise-Eval | Details of Fetched Web Pages & Accuracy of Detector Algorithms |
| Language(s) | # URLs | # have valid charset in HTTP header | # fetched successfully | icu4jStripped | jUniCharDet | tika | iustWithJUniChar | iustWithoutMarkupElim | iust |
| Arabic (.iq, ...) | 1904 | 1175 | 1168 | 603 = %52 | 806 = %69 | 743 = %64 | 931 = %80 | 1107 = %95 | 1136 = %97 |
| korean (.kr) | 1604 | 736 | 735 | 622 = %85 | 616 = %84 | 632 = %86 | 655 = %89 | 704 = %96 | 700 = %95 |
| Turkish (.tr) | 2545 | 1569 | 1561 | 467 = %30 | 1209 = %78 | 1282 = %82 | 1350 = %87 | 1466 = %94 | 1498 = %96 |
| Vietnamese (.vn) | 2721 | 1467 | 1463 | 968 = %66 | 1233 = %84 | 1265 = %87 | 1305 = %89 | 1446 = %99 | 1448 = %99 |
| Chinese (.cn) | 10086 | 3870 | 3855 | 1612 = %42 | 3411 = %89 | 3452 = %90 | 3555 = %92 | 3748 = %97 | 3725= %97 |
| French (.fr) | 13156 | 8730 | 8713 | 1889 = %22 | 6555 = %75 | 7039 = %81 | 7197 = %83 | 8266 = %95 | 8076 = %93 |
| Japanese (.jp) | 20339 | 7747 | 7737 | 3483 = %45 | 6455 = %83 | 6753 = %87 | 6771 = %88 | 7623 = %99 | 7619 = %99 |
| Spanish (.es) | 10208 | 6304 | 6287 | 1571 = %25 | 4745 = %76 | 4914 = %78 | 5308 = %84 | 6033 = %96 | 5956 = %95 |
| Italian (.it) | 11512 | 7398 | 7372 | 1540 = %21 | 5646 = %77 | 4690 = %64 | 6201 = %84 | 7074 = %96 | 6970 = %95 |
| Indian, English (.in) | 12236 | 6174 | 6157 | 995 = %16 | 3190 = %52 | 1607 = %26 | 3559 = %58 | 5987 = %97 | 5973 = %97 |
| Persian (.ir) | 7396 | 4014 | 4000 | 3752 = %94 | 3738 = %94 | 3878 = %97 | 3899 = %98 | 3975 = %99 | 3982 = %100 |
| English (.uk, .us) | 21565 | 13610 | 13583 | 2653 = %20 | 7551 = %56 | 4870 = %36 | 8886 = %65 | 13167 = %97 | 13025 = %96 |
| Russian (.ru) | 39453 | 28289 | 28229 | 22406 = %79 | 24796 = %88 | 25449 = %90 | 25638 = %91 | 22813 = %81 | 27563 = %98 |
| German (.de) | 35318 | 24914 | 24891 | 5618 = %23 | 15655 = %63 | 17261 = %69 | 18321 = %74 | 23552 = %95 | 23095 = %93 |
| All Above (.all!) | 190043 | 115997 | 115751 | 48179 = %42 | 85606 = %74 | 83835 = %72 | 93576 = %81 | 106961 = %92 | 110493 = %96 |