Data Intensive Lab1

Ali Shibli
Carina Wickström

September 2021

1 Results of Lab1

1.1 Question 1: Retrieve top 15 records

(aa,271_a.C,1,4675)
(aa,Category:User_th,1,4770)
(aa,Chiron_Elias_Krase,1,4694)
(aa,Dassault_rafaele,2,9372)
(aa,E.Desv,1,4662)
(aa,File:Wiktionary-logo-en.png,1,10752)
(aa,Indonesian_Wikipedia,1,4679)
(aa,Main_Page,5,266946)
(aa,Requests_for_new_languages/Wikipedia_Banyumasan,1,4733)
(aa,Special:Contributions/203.144.160.245,1,5812)
(aa,Special:Contributions/5.232.61.79,1,5805)
(aa,Special:Contributions/Ayarportugal,1,5808)
(aa,Special:Contributions/Born2bgratis,1,5812)
(aa,Special:ListFiles/Betacommand,1,5035)
(aa,Special:ListFiles/Bohdan_p,1,5036)

1.2 Question 2: get total number of records in the dataset

3324129

1.3 Question 3: Compute the min, max, and average page size.

(min,max,avg) = List(0, 141180155987, 132239)

1.4 Question 4: record(s) with the largest page size.

(en.mw,en,5466346,141180155987)

1.5 Question 5: record with the largest page size again. But now, pick the most popular.

(en.mw,en,5466346,141180155987)

1.6 Question 6: record(s) with the largest page title.

(zh,Special:e8b18ee6baafefbda5efbdbfe89cb7e6829fefbdbfe88b93e29980e89e9fefbda
9e89eb3efbda425636f256d6725736f257373256f38257373256f38257373256f38256b6d
73efbdaa256e6b256678256f6b2c687474703a2f2f7777772e65366231396665386135626
656f2d6f35393038636535626639376538383138616535361346139653561656134 2e636f2e
6d672e732e736f2e382e73736f386b2e6d2e372e73736f3873736f386b6d37332e752e622
e61616e6b66786f6b2e70772f2ce8b18ee6baafefbda5efbdbfe89cb7e6829fefbdbfe88b
93e29980e89e9fefbda9e89eb3efbda425636f256d6725736f257373256f38257373256f3
8257373256f38256b6d73efbdaa256e6b256678256f6b/,1,6043)

1.7 Question 7: create a new RDD with the records that have greater page size than the average.

Created it in code. Size of the RDD is 186819 records.

1.8 Question 8: total number of pageviews for each project

'sk − > 3887, mg.b − > 5, ht − > 528, ee.v − > 1, kr.mw − > 1, vec.d − >
26, pt.b − > 164, ace − > 143, kn.mw − > 1, gag − > 227, war − > 6165,
en.v − > 2243, is.q − > 8, eo.mw − > 1, usability.m − > 37, srn.mw − > 1,
vi.d − > 558, m.wd − > 4441, tr.d − > 2388, kn.d − > 827, da.v − > 1, pap
− > 86, ug.b − > 2, ru.s − > 7, qu.mw − > 1, mn.v − > 1, tyv.mw − > 1,
an.v − > 1, es.q − > 441, kl.mw − > 1, ie.d − > 37, dz.mw − > 1, als.b − >
3, ka.q − > 2, got.mw − > 1, mai.mw − > 1, kw.d − > 26, fa.d − > 112, ky.b
− > 1, sd.n − > 1, ro − > 7046, quality.m − > 3, test2 − > 82, cs.q − > 155,
dv.mw − > 1, zh.n − > 6, en.d − > 49466, mzn.mw − > 1, wuu.mw − > 1,
fj − > 37, pnb.d − > 1, ha.d − > 10, az.s − > 3, st − > 109, pag − > 139,
gn.v − > 1, da.d − > 263, ty − > 118, kv.mw − > 1, ee.mw − > 1, am − >
38, mn.d − > 20, av − > 33, pnt.v − > 1, an.d − > 44, cy.b − > 6, sn.d − >
2, tg − > 64, gn.d − > 7, ig − > 49, cs.b − > 20, am.q − > 1, ie.mw − > 1,
is − > 1018, ru.v − > 3, bjn − > 172, ch.d − > 4, bs.mw − > 1, simple − >
10113, ia.b − > 9, ro.s − > 74, sw − > 1141, az.d − > 144, ta − > 64, ia − >
623, or.mw − > 1, se − > 431, pt.n − > 58, zu.d − > 21, et.s − > 11, azb.v
− > 1, ia.mw − > 1, bat-smg.mw − > 1, zu − > 94, nostalgia − > 4, hak.mw
− > 1, bo.mw − > 1, oc − > 1316, commons.mw − > 1, vep.mw − > 1, mh.d
− > 2, dsb − > 106, zea − > 177, zh.q − > 18, ii.mw − > 1, id.q − > 61,
scn.mw − > 1, ny − > 44, ast.q − > 2, is.b − > 7, szl − > 143, nl.d − > 3351,
bm.mw − > 1, sa.b − > 1, ca.b − > 56, ga.b − > 2, tl.d − > 127, ru.d − >
2414, es.b − > 185, hu.d − > 762, war.v − > 2, ig.mw − > 1, he − > 761, sn
− > 178, bat-smg − > 203, new.mw − > 1, ka.b − > 1, ang.q − > 7, fiu-vro

$- > 124$, pl.d $- > 3741$, no.s $- > 436$, meta.mw $- > 1$, ik.mw $- > 1$, szl.mw $- > 1$, vi.s $- > 35$, roa-rup $- > 93$, fo.mw $- > 1$, arc $- > 13$, sh.mw $- > 1$, sv.s $- > 248$, br.s $- > 109$, mg.mw $- > 1$, sn.mw $- > 1$, lv.mw $- > 1$, sr.mw $- > 1$, xh.b $- > 1$, mi.mw $- > 1$, tk.q $- > 1$, gu $- > 18$, hif $- > 455$, bo.v $- > 1$, en.voy $- > 5477$, www $- > 2$, udm $- > 13$, lt.q $- > 63$, ang.b $- > 44$, cs $- > 11710$, yo.mw $- > 1$, no.d $- > 1350$, vo.d $- > 50$, map-bms.mw $- > 1$, bo.d $- > 2$, th.b $- > 3$, ast.b $- > 5$, sv.mw $- > 1$, fy $- > 744$, wo.q $- > 3$, pt.q $- > 136$, lo.v $- > 1$, strategy.mw $- > 1$, ca.n $- > 42$, mk.mw $- > 1$, mn $- > 36$, bn $- > 105$, pt $- > 59803$, et $- > 6287$, fy.mw $- > 1$, ro.d $- > 366$, ak.d $- > 3$, fo.d $- > 20$, ilo.mw $- > 1$, tr.s $- > 55$, lo.d $- > 117$, lbe.mw $- > 1$, hr.s $- > 112$, st.mw $- > 1$, sl.mw $- > 1$, frp $- > 52$, ss.v $- > 1$, cho $- > 20$, li.s $- > 19$, gl $- > 2381$, fi.s $- > 122$, www.wd $- > 140853$, to.d $- > 1$, als.q $- > 2$, lb.q $- > 3$, na $- > 187$, id.b $- > 125$, ca $- > 9851$, zh.b $- > 21$, kab.mw $- > 1$, simple.q $- > 49$, sd.mw $- > 1$, mus.mw $- > 1$, fr.s $- > 2415$, ak.v $- > 1$, su.b $- > 7$, ky $- > 30$, nl.s $- > 177$, fi.d $- > 1518$, kg $- > 95$, ca.q $- > 53$, te.q $- > 1$, he.q $- > 1$, nv $- > 25$, cv $- > 120$, be.q $- > 1$, uk $- > 2148$, fr.v $- > 230$, hu.s $- > 24$, eml.mw $- > 1$, kr.q $- > 1$, bi.d $- > 1$, mh $- > 13$, xh $- > 1206$, war.mw $- > 1$, bh $- > 14$, csb.d $- > 6$, simple.b $- > 74$, gv.v $- > 1$, br.d $- > 171$, jbo.d $- > 24$, fr.d $- > 14413$, crh.mw $- > 1$, nn.q $- > 18$, krc.mw $- > 1$, lt.b $- > 43$, sr.q $- > 9$, zh-min-nan.mw $- > 1$, mdf $- > 8$, lmo $- > 755$, uk.voy $- > 1$, lez.mw $- > 1$, chr $- > 64$, tk.b $- > 1$, fiu-vro.mw $- > 1$, ast $- > 711$, pih.mw $- > 1$, ti.d $- > 7$, nah.d $- > 43$, pl.s $- > 871$, hi.d $- > 44$, br.v $- > 1$, fa $- > 230$, sm.d $- > 54$, hr.d $- > 223$, pt.mw $- > 1$, pdc $- > 154$, de.q $- > 278$, fi.v $- > 31$, nl.voy $- > 86$, sv.d $- > 2298$, ee $- > 51$, ur.q $- > 1$, li.d $- > 1206$, en $- > 1689466$, or $- > 24$, sv.v $- > 8$, ug.d $- > 34$, my.v $- > 1$, xmf $- > 7$, tr $- > 24188$, fo.s $- > 1$, eo.n $- > 31$, ay.v $- > 1$, uk.mw $- > 1$, hsb.mw $- > 1$, mi.b $- > 6$, ay.d $- > 44$, ny.mw $- > 1$, wo $- > 62$, mk $- > 298$, lo $- > 10$, species.m $- > 8672$, mt $- > 410$, ht.mw $- > 1$, chr.d $- > 411$, tpi.mw $- > 1$, ti $- > 13$, ii $- > 8$, mg.d $- > 4758$, mdf.mw $- > 1$, ug.mw $- > 1$, ka $- > 268$, de.b $- > 1111$, af $- > 2040$, ky.d $- > 20$, sl.q $- > 22$, gl.q $- > 27$, cy $- > 1544$, m.w $- > 538$, sd $- > 16$, hr.mw $- > 1$, az.mw $- > 1$, kj $- > 16$, sa.s $- > 1$, ks $- > 11$, ba.mw $- > 1$, he.b $- > 2$, ng $- > 25$, lad $- > 194$, es.s $- > 507$, ne.b $- > 1$, be.b $- > 1$, mwl.mw $- > 1$, zh-yue $- > 256$, no.mw $- > 1$, my.d $- > 382$, su.q $- > 17$, tum $- > 54$, mwl $- > 172$, bn.b $- > 1$, myv $- > 7$, koi.mw $- > 1$, be.mw $- > 1$, bug.mw $- > 1$, pfl.mw $- > 1$, sg.d $- > 10$, bi.mw $- > 1$, cy.v $- > 1$, sm $- > 46$, ar.mw $- > 1$, sv $- > 28487$, cy.d $- > 255$, tt.mw $- > 1$, ang $- > 213$, cs.s $- > 74$, udm.mw $- > 1$, ml.q $- > 1$, is.s $- > 3$, ku.q $- > 13$, rue.mw $- > 1$, eu.q $- > 17$, zh-min-nan $- > 297$, bg.mw $- > 1$, bpy $- > 14$, ang.mw $- > 1$, av.mw $- > 1$, ks.d $- > 2$, ro.mw $- > 1$, iu $- > 31$, pcd $- > 104$, scn.v $- > 1$, ru.mw $- > 1$, es.voy $- > 78$, li $- > 515$, nov.mw $- > 1$, eu.b $- > 14$, ms.v $- > 1$, tw.v $- > 1$, scn.d $- > 179$, ss.d $- > 16$, oc.b $- > 5$, ga.v $- > 1$, cs.d $- > 850$, bug $- > 132$, ml.b $- > 1$, is.d $- > 65$, rm.mw $- > 1$, lt.mw $- > 1$, lad.mw $- > 1$, ko.q $- > 1$, ln.mw $- > 1$, eo.q $- > 22$, ka.v $- > 1$, zh.s $- > 37$, ar.n $- > 1$, kbd $- > 16$, sr.n $- > 27$, kv $- > 36$, ar $- > 880$, be $- > 240$, co.q $- > 4$, wa.d $- > 94$, ka.d $- > 31$, hy $- > 201$,

cs.v − > 41, fa.mw − > 1, et.mw − > 1, km − > 35, xmf.v − > 1, pdc.mw − > 1, lij − > 148, ms.d − > 66, yi.mw − > 1, rw.d − > 32, sl.b − > 19, gl.b − > 22, fur.mw − > 1, rw.mw − > 1, tw.d − > 2, ckb − > 24, map-bms − > 249, lb.mw − > 1, outreach.m − > 169, sa.d − > 14, uz − > 8367, gu.b − > 2, fi.mw − > 1, ku.b − > 2, roa-rup.d − > 16, hak − > 50, de.n − > 546, ia.d − > 26, id.s − > 127, es.v − > 98, arc.mw − > 1, ia.v − > 1, es.d − > 2981, mhr.mw − > 1, sg − > 21, tl − > 2389, pnt.mw − > 1, dz.d − > 1, gd.d − > 16, ja.mw − > 1, kab − > 124, ang.d − > 44, gn − > 70, pl.mw − > 1, it.mw − > 1, co.mw − > 1, pfl − > 59, id.d − > 1515, xal.mw − > 1, xh.d − > 8, fr − > 128730, af.b − > 34, am.d − > 4, gv.d − > 19, cy.s − > 18, to − > 87, pt.s − > 129, io − > 1598, srn − > 54, fi − > 21978, cy.mw − > 1, gan.mw − > 1, it.voy − > 264, rw − > 97, vls.mw − > 1, sah.v − > 1, nah − > 196, zh-yue.mw − > 1, haw − > 64, incubator.m − > 1119, frr − > 194, id.v − > 3, lt.s − > 8, ace.mw − > 1, rn − > 46, sd.d − > 10, cu.mw − > 1, hsb.d − > 14, ce.mw − > 1, th.d − > 75, m − > 28, zh.d − > 1721, ht.s − > 4, fr.voy − > 161, zh.voy − > 17, lrc.mw − > 1, eo.b − > 21, ar.q − > 8, ss − > 1173, za.b − > 3, ru.voy − > 1, om.d − > 42, cs.mw − > 1, vi.mw − > 1, nso − > 94, zh − > 12952, km.d − > 14, ckb.mw − > 1, vo.mw − > 1, bcl − > 318, co.b − > 2, ha − > 122, sa − > 1, ar.b − > 9, als.d − > 1, te.s − > 1, chr.mw − > 1, diq.mw − > 1, arz − > 62, af.mw − > 1, pt.v − > 45, uz.b − > 7, min.mw − > 1, tn.mw − > 1, test2.mw − > 1, csb − > 58, fur − > 114, ltg − > 23, nds-nl.mw − > 1, an.mw − > 1, yo − > 371, gu.mw − > 1, ab.mw − > 1, tk.d − > 9, zh-min-nan.d − > 32, tr.mw − > 1, diq.v − > 1, krc.v − > 1, qu − > 572, na.mw − > 1, th.mw − > 1, tl.mw − > 1, mt.mw − > 1, tt.d − > 30, ng.mw − > 1, ca.s − > 68, el.q − > 1, lv.b − > 2, co − > 342, no − > 19368, ms − > 4802, hy.q − > 2, ta.n − > 1, roa-rup.mw − > 1, beta.mw − > 1, zu.mw − > 1, pt.d − > 1180, bs.n − > 36, nrm − > 209, kbd.mw − > 1, mn.mw − > 1, mr.mw − > 1, ne.mw − > 1, yo.v − > 1, lij.mw − > 1, incubator.mw − > 1, tum.mw − > 1, cbk-zam − > 120, lt.d − > 625, rm.b − > 2, bar.mw − > 1, en.mw − > 1, kg.mw − > 1, lez − > 3, or.v − > 1, de.d − > 4682, nds.b − > 1, bs.q − > 74, nn.d − > 51, beta.v − > 280, vep − > 129, fa.q − > 1, bn.d − > 1, sl.s − > 102, roa-tara.mw − > 1, ur.d − > 6, de − > 166511, tpi − > 191, aa.d − > 4, fo − > 319, fy.b − > 14, te.d − > 103, ga.d − > 32, he.d − > 18, yi.d − > 2, za.q − > 1, sq.n − > 22, ln.d − > 8, da.q − > 16, gom.mw − > 1, pa − > 53, wuu − > 27, kw.mw − > 1, kk.mw − > 1, nn.v − > 2, ang.s − > 3, zea.mw − > 1, ky.mw − > 1, kw.q − > 1, pcd.mw − > 1, hy.b − > 1, tn.d − > 5, de.voy − > 6670, la.q − > 28, af.q − > 29, nds − > 509, sr.d − > 418, ko.mw − > 1, kl − > 216, ku − > 661, ca.d − > 572, csb.mw − > 1, ku.mw − > 1, ki.mw − > 1, yi − > 38, ps − > 15, cr − > 48, ne.d − > 2, xal − > 19, xh.mw − > 1, ff − > 29, rn.d − > 2, nap.mw − > 1, bm − > 39, ca.v − > 1, uk.b − > 1, de.v − > 263, ja.q − > 2, az − > 1311, bpy.v − > 1, km.mw − > 1, el.mw − > 1, bxr.mw − > 1, ks.mw − > 1, es − > 122503, la.b − > 24, nds-nl − > 404, tw − > 254, ml.d − > 192, uz.mw − > 1, ku.d − > 657, su.d − > 17, mg − > 874, pms.mw − > 1, roa-tara − > 141, mzn − > 4, lt − > 4367, eml − > 152, sq.q − > 8, my − > 182, cu − > 11, jbo.mw − > 1, kl.v − > 2, te − > 48, az.q − > 9, tn − > 144,

su.v − > 1, nov − > 138, ca.mw − > 1, lb − > 1197, ab − > 17, ja.b − > 6, bg − > 766, ceb − > 6910, ve.mw − > 1, om.mw − > 1, ak − > 62, iu.d − > 25, usability.mw − > 1, dz − > 40, meta.m − > 5454, os.mw − > 1, pnb.mw − > 1, oc.mw − > 1, m.s − > 43, pl.voy − > 41, sc.d − > 1, vo − > 1337, ko − > 1011, bs.b − > 28, sw.b − > 7, gu.v − > 1, pag.mw − > 1, fa.b − > 1, na.b − > 10, sl.d − > 119, uk.n − > 1, nl − > 39549, lbe − > 1, pam.mw − > 1, uz.q − > 10, qu.d − > 12, frr.mw − > 1, eu.d − > 351, nrm.mw − > 1, ja − > 7073, bm.q − > 1, gu.d − > 2, it.n − > 218, zh-classical.mw − > 1, sco.mw − > 1, hy.mw − > 1, nap − > 254, bn.mw − > 1, pa.b − > 1, da.b − > 50, ur.mw − > 1, pms − > 655, frp.mw − > 1, mn.b − > 1, id.mw − > 1, sl.v − > 12, cdo − > 43, qu.v − > 1, bh.d − > 1, br.mw − > 1, got − > 12, ast.d − > 160, min − > 403, eo.d − > 309, sq.mw − > 1, sa.mw − > 1, ar.s − > 8, sk.q − > 54, bg.q − > 3, se.mw − > 1, ml.mw − > 1, sr.s − > 9, mh.mw − > 1, nds.q − > 1, sq.b − > 26, commons.m − > 232824, az.b − > 7, mo.d − > 2, sm.mw − > 1, io.d − > 1540, su.mw − > 1, za.mw − > 1, jv − > 1137, si.mw − > 1, en.n − > 2587, bxr − > 12, so.mw − > 1, kaa.mw − > 1, lb.d − > 61, sah.mw − > 1, fr.mw − > 1, bs − > 2965, species.mw − > 1, it.q − > 1004, th − > 375, als.mw − > 1, de.s − > 6857, sco − > 1537, arz.mw − > 1, ko.d − > 1219, ba − > 19, sk.mw − > 1, sw.mw − > 1, sg.mw − > 1, su − > 687, hu − > 8623, ga.mw − > 1, vi − > 9574, zu.b − > 3, be-tarask − > 160, ki − > 46, et.q − > 23, an − > 725, ln − > 59, kr − > 10, stq − > 208, fj.mw − > 1, so.d − > 33, co.d − > 31, ts.d − > 9, sc − > 476, ss.mw − > 1, sc.mw − > 1, sl − > 4440, bm.b − > 3, tr.n − > 15, mus − > 22, io.v − > 1, be.s − > 1, ja.n − > 1, wo.v − > 1, vo.q − > 1, he.s − > 7, simple.d − > 461, donate − > 1, wo.d − > 32, bn.s − > 2, ilo − > 599, rue − > 82, si.v − > 1, ho − > 10, vi.q − > 3, tr.q − > 586, kn.q − > 1, test.wd − > 1, et.b − > 21, diq − > 234, jv.mw − > 1, sv.voy − > 35, mi.d − > 16, de.mw − > 1, da.mw − > 1, bjn.mw − > 1, si.d − > 2, pt.voy − > 30, simple.mw − > 1, ga − > 1425, tet.mw − > 1, tk − > 162, kk.b − > 1, jv.d − > 79, jp − > 1, dsb.mw − > 1, vo.b − > 1, pi − > 3, outreach.mw − > 1, mr.d − > 7, wa.mw − > 1, ar.d − > 214, sk.b − > 7, or.d − > 1, lv.d − > 83, cr.d − > 3, hif.mw − > 1, ten.mw − > 1, en.q − > 3886, mk.b − > 2, glk − > 1, be-tarask.mw − > 1, sah − > 65, jv.v − > 1, be.d − > 15, bpy.mw − > 1, nah.mw − > 1, ceb.mw − > 1, ug − > 23, rm.d − > 3, ru.n − > 5, wo.mw − > 1, ka.mw − > 1, so − > 388, ik − > 54, it.b − > 1185, tt − > 64, ar.v − > 2, it − > 91324, pl.n − > 90, lv.v − > 1, ro.voy − > 20, gl.s − > 24, myv.mw − > 1, ro.b − > 7, ak.b − > 2, xmf.mw − > 1, gd − > 602, strategy.m − > 103, chy.mw − > 1, ru.q − > 28, hi.mw − > 1, vec.mw − > 1, el.d − > 646, nah.b − > 1, el − > 762, kl.d − > 58, pl − > 63685, ro.n − > 8, hr − > 6430, eu − > 4619, ho.mw − > 1, tr.b − > 90, gn.b − > 1, nv.mw − > 1, ie − > 171, si − > 48, hi − > 42, hr.b − > 31, glk.mw − > 1, pl.q − > 1525, hu.mw − > 1, oc.d − > 162, he.mw − > 1, as.mw − > 1, chy − > 61, gl.d − > 300, rm − > 173, gan − > 16, mrj − > 9, cbk-zam.mw − > 1, gv − > 329, tw.mw − > 1, krc − > 16, el.v − > 2, li.b − > 1, ksh − > 687, ha.mw − > 1, fr.b − > 290, bh.mw − > 1, sr − > 1295, hu.q − > 34, ie.b − > 7, ts.mw − > 1, hy.d − > 295, nn.mw − > 1, ty.mw − > 1, kaa − > 98, tg.d − > 36, ay.mw − > 1, eo.s − > 30, bs.s

$- > 12$, en.b $- > 5457$, tet $- > 63$, no.n $- > 23$, hsb $- > 247$, fy.d $- > 94$, pih $- > 84$, nl.q $- > 51$, nl.mw $- > 1$, hu.b $- > 15$, ja.d $- > 892$, got.b $- > 2$, bs.v $- > 2$, br $- > 1455$, lv $- > 1799$, tyv $- > 4$, za $- > 50$, sv.n $- > 96$, gom $- > 5$, lg.mw $- > 1$, no.q $- > 18$, sw.d $- > 149$, ff.mw $- > 1$, vec $- > 434$, ltg.mw $- > 1$, eu.mw $- > 1$, la.mw $- > 1$, tpi.d $- > 8$, ru.b $- > 17$, fi.n $- > 38$, ksh.mw $- > 1$, fa.voy $- > 1$, fr.n $- > 86$, ps.d $- > 62$, ta.v $- > 2$, nl.b $- > 232$, za.d $- > 1$, pl.b $- > 146$, na.d $- > 16$, li.mw $- > 1$, bi $- > 87$, mi $- > 269$, ta.d $- > 737$, es.mw $- > 1$, os $- > 17$, my.b $- > 7$, la.v $- > 1$, af.v $- > 1$, bs.d $- > 52$, da $- > 9946$, ne $- > 5$, vls $- > 362$, nn $- > 2577$, zh-classical $- > 21$, la.d $- > 112$, ten $- > 107$, af.d $- > 122$, fj.d $- > 112$, lmo.mw $- > 1$, koi $- > 15$, mr $- > 17$, rn.mw $- > 1$, ja.v $- > 4$, tl.b $- > 48$, ch.b $- > 1$, ce $- > 24$, pa.v $- > 1$, lo.mw $- > 1$, ro.q $- > 232$, eo $- > 5897$, pa.d $- > 4$, mhr $- > 38$, scn $- > 1404$, www.w $- > 3813$, sh.d $- > 254$, nl.n $- > 46$, kk $- > 671$, nso.mw $- > 1$, ps.mw $- > 1$, lad.v $- > 2$, uz.d $- > 139$, br.q $- > 11$, sv.q $- > 18$, fr.q $- > 202$, mrj.mw $- > 1$, iu.mw $- > 1$, ts $- > 50$, cdo.mw $- > 1$, ml $- > 254$, li.q $- > 5$, no.b $- > 13$, bo.b $- > 1$, haw.mw $- > 1$, pnb $- > 36$, ch.mw $- > 1$, uk.s $- > 2$, uz.v $- > 1$, ext.mw $- > 1$, dv.d $- > 6$, crh $- > 156$, hu.n $- > 3$, it.s $- > 1383$, pa.mw $- > 1$, new $- > 11$, bcl.mw $- > 1$, hr.q $- > 17$, cv.mw $- > 1$, io.mw $- > 1$, cr.mw $- > 1$, gag.mw $- > 1$, ch $- > 56$, sq.d $- > 169$, bm.d $- > 2$, dv $- > 26$, sk.s $- > 1$, pnt $- > 1$, pi.mw $- > 1$, lg $- > 51$, fi.q $- > 47$, ms.b $- > 21$, is.mw $- > 1$, ay $- > 151$, pap.mw $- > 1$, om $- > 92$, m.f $- > 251$, dv.v $- > 1$, hi.b $- > 3$, kn.s $- > 1$, ta.mw $- > 1$, it.d $- > 5695$, pam $- > 296$, tg.mw $- > 1$, wa $- > 331$, bi.b $- > 20$, ja.s $- > 6$, bo $- > 96$, am.mw $- > 1$, sv.b $- > 23$, cy.q $- > 110$, gl.mw $- > 1$, la.s $- > 211$, sk.d $- > 47$, te.mw $- > 1$, kw $- > 267$, mk.d $- > 18$, tk.mw $- > 1$, en.s $- > 14087$, fi.b $- > 76$, la $- > 3410$, ms.mw $- > 1$, aa $- > 36$, bar $- > 728$, mo $- > 11$, azb.mw $- > 1$, my.mw $- > 1$, ru $- > 19618$, ky.q $- > 1$, it.v $- > 1847$, gd.mw $- > 1$, vi.voy $- > 17$, ak.mw $- > 1$, gv.mw $- > 1$, kk.d $- > 42$, ur $- > 40$, mt.d $- > 14$, jbo $- > 174$, to.mw $- > 1$, zh.mw $- > 1$, st.d $- > 24$, rmy $- > 71$, es.n $- > 102$, ti.mw $- > 1$, hz $- > 7$, pam.v $- > 1$, id $- > 16200$, kk.v $- > 1$, test $- > 25$, vec.s $- > 80$, da.s $- > 128$, arbcom-en $- > 16$, et.v $- > 1$, bg.d $- > 13$, fa.s $- > 1$, as $- > 38$, ext $- > 164$, stq.mw $- > 1$, new.v $- > 1$, nds.d $- > 101$, rmy.mw $- > 1$, sh $- > 7966$, sq $- > 1862$, et.d $- > 772$, cs.n $- > 33$, ve $- > 32$, ug.q $- > 1$, mo.mw $- > 1$, kn $- > 9$, uk.v $- > 1$, gn.mw $- > 1$, uk.d $- > 63$, nds.mw $- > 1$, ik.d $- > 1$, ast.mw $- > 1$, als $- > 939$'

sum of values = 3324129

1.9 Question 9: 10 most popular pageviews of all projects, sorted by the total number of hits.

(en.mw,en,5466346,141180155987)

(es.mw,es,695531,12261337515)

(ja.mw,ja,611443,15021588551)

(de.mw,de,572119,9523069696)

(fr.mw,fr,536978,11752030020)

(ru.mw,ru,466742,11847816616)
(it.mw,it,400297,8176042087)
(en,Main_Page,257915,4289970372)
(pt.mw,pt,196160,4029404403)
(pl.mw,pl,176059,2782453516)

1.10 Question 10: Determine the number of page titles that start with the article "The". How many of those page titles are not part of the English project (Pages that are part of the English project have "en" as the first field)?

The number of page titles that start with the article "The": 45020 Not part of the English project: 10292

1.11 Question 11: percentage of pages that have only received a single page view

76.96247648632169 %

1.12 Question 12: number of unique terms appearing in the page titles.

1910691

1.13 Question 13: the most frequently occurring page title term in this dataset.

(of,192573)