Unsupervised POS-Tagging Employing Efficient Graph Clustering

Chris Biemann, University of Leipzig, Coling/ACL SRW 2006
biem@informatik.uni-leipzig.de

Tagset for ENGLISH

This tagset was automatically induced by merely presenting large amounts of raw text to the algorithm. Numbers following hashmarks # correspond to tag labels, numbers in brackets give the number of words per tag in the lexicon. Maximally, four randomly chosen entries per tag are given. Special characters are transformed for technical reasons, e.g. _ESENTS_ is the full stop, _CLOSEBRAC_ is a closing bracket etc.

#1(19144):levity, kingship, generalisation, measles #2(4203):proven, scrapped, counted, invalided #3(4990):inept, unsure, inconsequential, perverted #4(1038):misrepresenting, sacrificing, substituting, differentiating #5(1330):maximise, refill, recover, employ #6(683):Holidays, Kingdom, Bowl, Components #7(1039):Realism, globalization, Parachute, Monuments #8(3522):Dell, Butts, Beaumont, Papworth #9(398):recover, employ, inorder, references #10(2239):Miranda, Diana, Jehan, Benedict #11(859):Yugoslavian, Bohemian, Premier, Sikh #12(1862):Bremen, Southwold, Otago, Maidenhead #13(912):defiantly, rarely, satisfactorily, slavishly #14(293):absorbs, ensures, eliminates, reminds #15(1208):postsynaptic, shorthand, artistic, algebraic #16(115):farted, quipped, cooed, mumbled #17(255):26_000, 11_500, 20_000_000, thirty-four #18(77):crikey, oops, aye, ho, ah #19(35):northeastern, southern, northwest, north #20(355):Vanuatu, Castile, CNN, Algeria #21(308):Obviously, Window, Animals, Tip #22(99):Microsoft, Eurotunnel, AST, Farr-Jones #23(80):retching, whistling, snoring, growling #24(4):thousands, hundreds, millions, billions #25(5):loads, dozens, hundreds, plenty #26(31):Thousand, Six, Twenty-eight, Zero #27(1410):AT, AMERICA, LORD, BOTTOM #28(71):Rule, Precedent, Sections, Article #29(94):perinatal, ventricular, anaerobic, microbial #30(159):33rd, 40th, _APOS_91, 27th #31(6):ii, viii, iii, vii #32(14):mainly, purely, solely, partly #33(71):earliest, 13-year-old, elder, eleven-year-old #34(14):somehow, nevertheless, fortunately, unknowingly #35(21):ye, thy, anybody, ya #36(11):insteadof, albeit, including, while #37(279):potentially, eminently, manifestly, understandably #38(1):_APOS_d #39(2):friend, colleague #40(4):Mr, Dr, Miss, Ms #41(2):p, r #42(1):who #43(1):now #44(9):Council, Minister, Committee, Society #45(1):found #46(2):safety, rehabilitation #47(1):%N% #48(5):man, person, woman, lady, gentleman #49(1):work #50(7):Oct., Jan., Sept., Nov. #51(5):Yeah, Mm, Mhm, Aha #52(1):to #53(1):only #54(6):=, No., Page, Morethan, CHAPTER, Example #55(2):proposals, guidelines #56(3):Ireland, Africa, Asia #57(2):my, your #58(1):can #59(12):reducing, improving, identifying, promoting #60(5):May, April, March, June #61(27):sense, range, list, choice #62(3):perhaps, maybe, presumably #63(1):we #64(2):everyone, everybody #65(1):go #66(1):come #67(2):animal, insect #68(2):A, An #69(5):head, eyes, arms, mouth #70(2):C, D #71(1):so #72(1):and #73(1):do #74(4):Right, Okay, Aye, Alright #75(2):When, Assoonas #76(2):er, erm #77(1):going #78(103):Smith, Jones, Brown, Wilson #79(1):need #80(2):tomorrow, tonight #81(1):how #82(1):no #83(2):Owen, Jay #84(2):morethan, lessthan #85(4):York, Kong, Angeles, Francisco #86(2):larger, smaller #87(2):Fig, c. #88(2):wore, wears #89(1):little #90(2):processor, recorder #91(2):results, findings #92(1):off #93(2):My, Your #94(3):top, bottom, rear #95(1):_quot_ #96(1):In #97(2):detailed, systematic #98(4):kind, sort, type, lack #99(1):This #100(2):He, She #101(2):often, usually #102(1):too #103(1):But #104(2):games, matches #105(1):So #106(2):_KOM_, _mdash_ #107(1):them #108(2):demand, search #109(3):write, learn, submit #110(3):total, maximum, minimum #111(3):son, daughter, mistress #112(1):_APOS_ll #113(1):put #114(1):never #115(1):been #116(2):smile, grin #117(2):If, Unless #118(3):Hence, Inside, Outside #119(2):Eh, Pardon #120(2):appointed, elected #121(2):minutes, seconds #122(3):One, Much, ONE #123(1):did #124(1):both #125(2):regarded, classified #126(1):it #127(2):No, Yes #128(2):extra, additional #129(1):then #130(1):There #131(2):think, reckon #132(11):schools, hospitals, hotels, libraries #133(2):miles, kilometres #134(1):not #135(1):is #136(7):October, September, January, December #137(1):use #138(2):His, Her #139(1):have #140(1):about #141(2):using, incorporating #142(1):That #143(2):School, College #144(1):could #145(1):_APOS_ve #146(2):yeah, yes #147(2):customers, clients #148(2):Zealand, Lanka #149(1):me #150(2):Dear, Yours #151(16):J., C., M., R. #152(1):two #153(2):_pound_20, _pound_25 #154(1):_APOS_s #155(2):Their, Its #156(4):his, their, its, our #157(2):removed, withdrawn #158(2):wife, girlfriend #159(1):here #160(1):_SEMICOL_ #161(1):_OPENBRAC_ #162(1):way #163(3):San, Santa, El #164(3):families, bodies, voices #165(1):What #166(1):that #167(2):_pound_10, _pound_5 #168(2):output, revenue #169(1):_APOS_re #170(1):all #171(1):still #172(1):And #173(1):time #174(1):well #175(1):something #176(2):wish, intend #177(2):_APOS_m, am #178(2):mean, suppose #179(1):does #180(2):Though, Eventhough #181(1):between #182(1):him #183(1):these #184(2):application, object #185(2):hospital, prison

see next page

#186(2):responsible, reserved #187(1):The #188(1):more #189(4):e.g., i.e., eg, ie #190(3):morning, evening, afternoon #191(3):_ESENTS_, _ESENTQ_, _ESENTE_ #192(1):down #193(1):very #194(1):most #195(15):changes, developments, involvement, cuts #196(1):_APOS_ #197(2):data, material #198(1):They #199(1):thought #200(32):business, service, light, language, trade, capital, space, property, design, colour, security, peace, stock, traffic, debt, welfare, painting, youth, household, rent, drama, routine, currency, discipline, bone, craft, grain, narrative, seed, ritual, comedy, romance #1(2):Big, Golden #202(2):Mrs, Aunt #203(1):_hellip_ #204(1):the #205(2):up, out #206(1):another #207(3):plus, comprising, featuring #208(1):of #209(2):senior, junior #210(2):English, Ulster #211(2):chief, Chief #212(1):want #213(1):made #214(3):year, week, month #215(2):aware, unaware #216(2):higher, greater #217(2):difference, gap #218(2):achieved, exercised #219(1):own #220(1):It #221(2):ability, inability #222(2):ASIA, EAST #223(1):has #224(1):just #225(2):Sun, Pope #226(2):he, she #227(2):add, attach #228(1):got #229(1):said #230(1):know #231(1):Oh #232(1):much #233(1):be #234(2):responsibility, justification #235(12):potatoes, beans, cakes, nuts #236(4):behind, infrontof, beside, nextto #237(2):player, weapon #238(2):concentrate, rely #239(2):ICI, BT #240(3):S, R, L #241(2):pulling, pushing #242(3):Although, While, Whilst #243(1):even #244(1):had #245(2):Parliament, parliament #246(2):interim, ongoing #247(1):some #248(1):one #249(2):case, event #250(1):life #251(1):this #252(1):her #253(1):than #254(2):bloody, fucking #255(2):last, next #256(4):Some, Many, Most, None #257(1):they #258(2):Come, Go #259(2):Award, Primary #260(2):King, Queen #261(2):match, contest #262(224):men, others, services, staff #263(2):are, were #264(1):people #265(1):n_APOS_t #266(2):review, strip #267(1):many #268(2):motivated, constrained #269(1):long #270(2):interested, engaged #271(2):Suddenly, Atlast #272(1):us #273(1):like #274(1):when #275(1):other #276(4):For, With, Within, Using #277(1):say #278(1):or #279(2):Is, Was #280(1):I #281(2):dangerous, frightening #282(1):as #283(1):right #284(2):ready, destined #285(1):also #286(17):followed, caused, affected, supported #287(2):effect, impact #288(27):experience, success, theory, performance #289(4):Accordingto, Unlike, Asfor, Intermsof #290(2):Health, Youth #291(3):highly, relatively, comparatively #292(2):a, an #293(2):strategy, facility #294(1):You #295(1):those #296(2):campaign, struggle #297(4):Chapter, Figure, Table, Fig. #298(11):Eliot, Shakespeare, Hitler, Freud #299(1):_COL_ #300(1):being #301(1):which #302(1):any #303(1):_CLOSEBRAC_ #304(5):_pound_100, _pound_1,000, _pound_500, _pound_50 #305(2):famous, notorious #306(3):produced, created, adopted #307(1):such #308(2):God, Christ #309(2):Saudi, Sri #310(2):hardly, scarcely #311(2):Today, Yesterday #312(6):will, would, may, should, must, might #313(1):there #314(1):what #315(1):before #316(2):advice, guidance #317(2):St, St. #318(12):religion, poetry, medicine, fiction #319(1):back #320(2):solution, barrier #321(2):television, TV #322(1):_bquo_ #323(1):see #324(3):years, months, hours #325(2):side, slopes #326(3):seemed, seems, appears #327(11):glass, gift, photograph, bowl #328(2):eachother, oneanother #329(1):_equo_ #330(2):complex, subtle #331(2):particularly, especially #332(1):was #333(3):ca, wo, ai #334(2):De, Van #335(1):you #336(2):if, aslongas #337(2):Erm, Er #338(6):b, i, c, d #339(1):We #340(2):asif, asthough #341(4):with, after, via, inresponseto #342(1):where #343(3):standing, sitting, lying