NLP Assignment :02

Anupam Pandey
20162118

Baum Welch Algorithm

The data set used was the brown data set from nltk.

Implementation Issues:
• For longer observations, we need to scale forward probabilities and Backward probabilities otherwise they will become 0.
• alphas sum can be zero ad hence divide by zero error can come.
• Initial values for transition matrix should sum to 1.
• Initial values for emission matrix should sum to 1.
Observations:
• We found that some of the words comes in different tags multiple no. of times.
Output:
Tag: tag1
['exaggeration', 'tales', 'boasts', 'enforced', 'abetted', 'gallery', 'stirring', 'garrisoned', '6-1', 'rogues','uncommitted','quickie', 'acted', 'assailed', 'ballards', 'denial', "builder's", 'congregation', 'utilize', 'cloudless', 'surrealist', 'represent', 'catchy', 'woodsmoke', 'peterson', 'extended', 'barn', 'wells', 'trophy', 'technician', '26-year-old', 'handle', 'infected', 'materialize', 'masterpiece', 'lundeen', "brooklyn's", 'arrington', 'dictators', "taxpayer's", 'wangemans', 'flames', 'amy', 'intelligible', 'u.n.', 'unsigned', 'cramp', 'boonton', 'short-skirted', 'consequences', 'forces', 'country', 'pleasure-boat', "tamiris'", 'luther', "segal's", "monday's", 'swelling', 'fledermaus', 'closely', 'twenty-first', 'fables', 'earlier', 'clip', 'hearty', 'contemporary', 'household-type', 'firefighting', '27-year-old', 'listings', 'taruffi', "season's", 'credit', "blanchard's", '45-degree', 'pastime', 'tommie', 'ambition', 'minor', 'chinese', 'responsibilities', 'bat', 'stanton', 'affirmative', 'life', 'expenditure', 'few', 'porpoises', 'self-crimination', 'vegetable', 'rabbit', 'fold', 'vices', 'exception', 'insofar', 'ambassador', 'precede', 'castanets', 'adversary', 'baron']
Tag: tag2
['festive', 'taxi','unenviable', 'tuesday', 'prior', 'devils', 'knock-down', 'scene', 'colonialist', 'cherry-flavored',"bultmann's", 'rapidly', 'full-bodied', 'rink', 'eliminate', 'cosmology', 'unsuspecting', 'attracted','lid', 'actual', 'sympathetically', 'columnist', "brassbound's", 'bond', '2', 'palm', 'newcomer', 'simultaneously', 'incentive', 'handy', 'chest', 'abstract', 'traditions', 'unimpaired', 'rigged', 'useless', 'house', 'pro', 'mind', 'misguided', 'intolerant', 'incinerator', 'gins', 'yearbook', "woman's", 'dinh', 'governed', 'strafaci', 'shattered', 'recede', 'bianco', 'motors', 'debut', 'churning', 'circulation', 'offensive', 'logical', 'syndicate', 'longing', 'prevent', 'gifted', 'civic', 'impossibly', 'hostile', 'faster', 'weather', 'welton', 'drizzly', 'loaded', '$451,500', 'whitfield', 'conceded', 'seems', 'divertimento', 'emotional', 'ike', '1962', 'welled', 'sholom', 'bultmann', 'ceramic', 'mouth', 'chow', 'con', 'portraits', '50%', 'chansons', "blanchard's", 'praises', 'outspoken', 'irreversible', 'awoke', 'assisting', 'morning', 'conceptualization', 'football', 'fit', 'disturbing', 'leland', 'tift']
Tag: tag3
["government's",'snacks', 'sherman', 'losses', 'tosses', 'wartime', 'relentless', 'twin', 'bills', 'lawns', "o'clock",'manifest', '$3,000', 'infantile', 'formidably', 'navy', 'voice', 'weaken', 'sanantonio', 'extension','inaccuracies', 'counties', 'terror', 'plainest', 'stretches', 'reformers', 'beans', 'indoors', 'frequent','neutralists', 'musts', 'rev.', 'minors', 'assuming', 'understood', 'responded', 'geographic', 'glories', 'resourcefully', 'thunder', 'mennen', 'cosmic', "farmers'", 'builder', "latter's", 'giacometti', 'army', 'noting', 'brainwashing', "yanks'", 'melvin', 'overt', 'charts', 'budgeting', 'rosella', 'mirror', 'long-sought', 'idols', '102', 'hurried', 'pastime', 'nikolais', 'partisan', 'speech', 'axes', 'columns', 'numerous', 'bolsheviks', 'home-run', 'tatian', "tourists'", 'transporting', 'muscular', 'sibly', 'apotheosis', 'detract', 'bouffe', 'no', 'unpopular', 'neutralism', 'scampini', 'females', 'make', "ginner's", 'tile', 'calypso', '$11.50', 'doltish', 'unconditional', 'whiplash', 'nori', 'deliverance', 'attributed', 'begins', 'farrell', 'greer', 'leni', 'uptown',

'heathenish', 'sponsorship']

Tag: tag4
['deficiency','aged-care', 'harlingen', 'self-confident', 'delicacies', 'dreamed', 'cohesion', 'gangs',
'boeing','grappelly', 'swept', 'streamer', 'original', 'raoul', 'ultra-violet', 'arnold', 'opened', 'streetcar',
'zeising', 'metropolitan', '$278,877,000', 'mixed', 'reds', 'useful', 'positively', 'blockade',
'biggest', 'slope', 'mcelvaney', 'legitimate', 'authentically', 'tunnels', 'overpowered', 'sprinted',
'secondary', 'aircraft', 'theaters', 'seals', 'reinterpret', 'hopes', '$10', 'brandeis',
'topics', '1960', 'bertrand', 'bellows', 'dragonetti', 'orcutt', 'bolted', 'curtains', 'nocturne', '1:48',
'twisty', 'intensive', 'self', 'underdeveloped', 'greeted', 'parson', 'disappointments', 'neusteters',
'modernists', 'f.b.i.', 'pick', 'varies', 'amicable', 'angel', 'violin', 'heretofore', 'breath', 'horace',
'attorney', 'inflow', 'jussel', 'compassionate', 'june', 'naturalistic', 'insoluble', 'wing', 'intolerant',
'nominally', 'arguing', 'lemonade', 'mccarthy', 'hyperbolically', 'lullaby', 'continuum',
'abstractions', 'carolyn', 'realty', 'choruses', 'lows', 'separable', 'dequindre', 'clarification',
'minority', 'improvisations', 'goat', 'aloud', 'hard-liquor', '75th']
Tag: tag5
['tradition', 'l', 'fewer','diplomat', 'lawn', 'beat', 'mood', 'centers', 'come', 'corrosion', 'publishes',
'shakespeare',
'mysterious', 'tenth', 'lucille', 'disaster', 'pageant', 'discrimination', 'base-runner',
'circumstances', 'groups', 'embodiment', 'billions', 'disciples', 'r-warren', 'subbing', 'engines',
'self-rule', 'dignity', 'links', 'repealed', 'sidechairs', 'rescue', 'confines',
'earl', 'phenomena', 'remanding', 'speculated', 'prohibiting', 'confirming', 'forgetting', 'process',
'reluctantly', 'overactive', 'officers', 'fausto', 'utilizing', 'radio-tv', 'festival', 'drowning',
'ailerons', 'triple', 'hultberg', 'commotion', 'soberly', 'kings', 'speculative', 'choreographic',
'shortage', 'newsmen', 'breathing', 'wondrously', 'ltd.', 'francis', 'shy', 'courses', '24th',
'arrangement', '$3', 'delegation', 'measure', 'ind', 'nationwide', 'hardwick', 'extern', 'typhoon',
'expressing', 'ruth', 'boland', 'gold', 'opposition', 'above-water', 'statistical', 'tuck', 'principle',
'broke', 'purposely', 'trailers', 'objects', 'masses', 'correction', 'payoff', 'geology', 'outlawry',
'ruler', 'spaces', 'sulcer', 'ahrens', 'undermine', 'rostagnos']
Tag: tag6
['polities','outfought', 'han', 'encores', 'categorize', "system's", 'unhappiest', 'provincial', '50th',
'corps',
'leg', 'chief', 'indicate', 'learn', 'spite', 'pointing', 'cloudless', 'explain', 'tearle', 'exempt',
'midnight', 'infield', 'religions', 'dismaying', 'socks', 'trawler', 'logically', 'flourish',
'robbins', 'rehabilitation', 'correctness', 'alluring', 'question', 'farce', 'urn', "network's", 'warn',
'fathers', 'russians', 'responds', 'ind', 'tense', 'quincy', 'glamour', 'cat', 'firm', 'husbands', "hours'",
'stetson', 'revisited', 'regulated', 'remarque', 'bouvier', 'major-market', 'fluently', 'clay-mining',
'helpful', 'pastors', '3-hitter', 'toll-road', 'relative', 'endowed', 'ncta', 'sledding', 'juice', 'suezhungary',
'stumbling-block', 'laureate', 'aeronautics', 'higher', 'harrity', 'occupies', 'academy',
'bookshelf', 'visits', 'marshall', 'winds', 'inches', 'give-and-take', 'trades', 'perfectly', 'linking',
'turner', 'overshot', "thurber's", 'devised', 'threads', 'complained', 'suspensions', 'achieving',
"pal's", 'enchantingly', 'work-weary', 'globetrotter', 'years', 'begging', 'beauty', 'seventy',
'gaudy', 'criminals']
Tag: tag7
['curtain','arrogance', 'non-public', 'zones', 'functional', 'relevant', 'coordinating', 'fla.', 'home-
owners',
'tomb', "leonato's", 'businesses', 'crispin', 'integration', 'federation', 'thomas', 'enchant',
'compensation', 'aggregations', 'suspected', 'stengel', 'musing', 'perturbed', 'hash', 'baby',
'knobby-knuckled', 'itinerary', 'scraps', 'fats', 'extraordinarily', '10-year-old',
'explanations', 'aiken', 'mcn.', 'amounts', 'tax-exempt', '$590,000', 'diversions', "leningrad's",
'200-megaton', 'slaves', 'handling', 'equally', 'unification', 'robbery', 'tabit', 'possessing',
'commies', 'salsich', 'mourn', 'sells', 'cigarette', 'biological', 'itself', 'destroyers', 'physicians',

'gift', 'loud', 'ciudad', 'chromatic', 'allegory', 'unfortunately', ''landis'', 'flat', 'slob', 'kershbaum',
'5000', 'beronio', 'thrusting', 'waved', 'caressing', 'build', 'three-fifths', 'unforgivable', 'tatras',
'phoenix', 'debuting', 'uncle', 'mass.', 'hard-liquor', 'satire', 'correctly', 'liberating', 'erroneously',
'usually', 'oafs', 'throw-rug', 'vacant', 'wayne', 'emperor', 'werner', 'olney', 'estrangement',
'mined', 'politics', '12-month', ''hendricks'', 'mcalester', 'pegged', 'attorneys']

Tag: tag8

['cynics', 'spain', 'blast', 'weak', 'harry', 'nevah', 'billie', 'drift', 'expansive', 'mortal',
'lilac', 'knock','"cipriani's'', 'dystrophy', 'proselytizing', 'monotony', 'freeman', 'stately', 'dali', 'lacking',
'candid', 'would-be', ''play's'', 'wanders', 'raft', 'britten', 'tv', 'grist', 'herrmann', 'rivalry', 'marty',
'glides', ''stone's'', 'communicating', 'bal', 'kaplan', 'bodies', 'affectingly', 'trick', 'banquet',
'misleading', 'randall', 'allege', 'market', 'banker', 'double-talk', 'poverty', '9-1/2', 'persons',
'specific', 'yet', 'suites', 'short-skirted', 'pine', 'sievers', 'strauss', 'electrons', 'norborne', 'lazy',
'wingback', 'interrogation', 'tortuous', 'habib', 'phrases', 'sciences', 'sandman', 'temple',
'characters', 'raw', 'destiny', 'bass', 'shock', 'handles', 'y.', 'usn.', 'subordinated', 'open-meeting',
'functioning', 'stumbled', 'proportionate', 'root', 'liberation', 'backward', 'mulligatawny',
'selectively', 'fellow', 'exported', 'educate', 'topical', ''gladden's'', 'russell', 'shareholders',
'triandos', 'dixieland', 'wears', 'grief', ''allison's'', 'lehman', 'nations', 'west']

Tag: tag9

['accidents', 'sp-44002', 'wine', 'highways', 'lena', 'neckline', 'tired',
'seventy-fourth', 'stepping', 'dynamically', 'batters', 'shoulders', 'malapropism', 'unmindful',
'infectious', 'wilted', 'tasteful','swc', 'duels', 'tsvetkov', 'ambiguous', 'incapacity', 'rapprochement',
'preserved', 'jnr',
'extractors', 'safe', 'northern', 'sleep', 'key', ''students'', 'bankruptcy', 'huddling', 'hard-hearted',
'manipulate', 'contract', 'linking', 'diehards', 'clay', 'kel', 'distressed', 'ucla', 'fireworks', 'intake',
'chapels', 'quartet', 'layers', ''you've'', 'wichita', 'copland', 'speedily', 'albacore', 'modified', 'cbs',
'ambassador-designate', '6:30', 'politically', 'tass', 'fulfilled', 'induce', 'much', 'appeared',
'reunion', 'publicists', 'fritz', 'breathing', 'greyhound', 'minutes', 'revived', 'businesses',
'glamorous', '1845', 'father', ''jackson's'', 'tries', 'catastrophe', 'purchase', 'caltech', 'formby',
'authorizing', 'o.', 'kicked', 'nice', 'run', 'menuhin', 'refusal', 'charlayne', 'ignite', 'unshakable',
'polling', 'menilmontant', 'apparent', 'obviously', 'behan', 'slave', 'hysterical', 'hardship', '$60',
'belvidere', 'beer']

Tag: tag10

['28th', 'disrepute', 'expenses', 'formidably', 'pixies', 'sixty', 'businessman', 'rain',
'delay', 'tucking', 'compensatory', 'glimpsed', 'sturdy', 'skimmed', 'imposing', 'equitably',
'trimmings', 'enthralling', 'revival', 'argued', 'callas', 'shiny', 'orange', 'billie', 'ceremonies',
'tales','ld060', 'pitchers', 'ciphers', 'devastated', 'hiroshima', '3.28', 'signaling', 'pooling', 'primarily',
''karns'', 'animals', 'encampment', 'setup', 'lawful', 'blackout', 'custom', 'gates', 'addresses',
'culmone', 'huddling', 'fills', '12-month', 'buffets', 'wrigley', 'projection', 'fill', 'bo', 'inquiring',
'melcher', 'large-scale', 'reform', 'evensong', 'ten', 'detailed', 'musial', 'tells', 'plus-one', '$120',
'desiring', 'paganini', 'performances', 'fatima', 'counterbalanced', 'uhles', 'saw', 'rushing',
'requests', 'petipa', 'directions', 'countries', 'knife', 'beer', 'nine', 'soviets', 'darling', 'yorkers',
'toronto', 'lumumba', '16-year-old', 'dewitt', 'alleghenies', 'kill', 'heather', 'moderately',
'irreparable', 'supervisor', 'ankles', 'insurgents', 'conceded', 'developers', 'industrialized',
'weeping', 'polaris', 'impressed']