Natali Ayoub

Juan Sanchez

Task 1:

Generative process:

π ∼ Dirichlet(α)

for j = 1 to k do:

µj ∼ Dirichlet(β)

end for

for i = 1 to docCount do:

ci ∼ Categorical(π)

xi ∼ Multinomial(µci , li)

end for

From the generative process above, our PDF is:

Dirichlet(π | α) \* \*

Initialization:

* α = a vector/array of size k
* = a vector/array of size equal to dictionary size (20,000 in this assignment)
* π ∼ Dirichlet(α) = ; xj is a set of k-dim vector with probabilities of each categorical event where
* for j = 1 to k:

µj ∼ Dirichlet(β) = ; xw is a set of w-dim vector with probabilities of each word for each category where

* xi ∼ Multinomial(µci , li) = number of occurrences of each of the 20,000 words in each document
* ci ∼ Categorical(π) = randomly assign categories for each document by sampling j from the multinomial distribution using

Gibbs Sampling:

For count = 1 to 200:

Update π: Dirichlet(π | α) \* ∼ (using conjugate priors) Dirichlet(α + a) =

; xj is a set of k-dim vector with probabilities of each categorical event where

; a is a k-dim vector where each entry is the number of times each category was assigned

For j = 1 to k:

Update µj: \* ∼ (using conjugate priors) Dirichlet(β + b) =

; xw is a set of w-dim vector with probabilities of each word for each category where

; b is a 20,000-dim vector where each entry is the number of words in each document produced by category j

End for

For n = 1 to docCount:

Update ci:

= Pr [ci = j | π, µj, li and xi]

=

Then we update ci by sampling j from the above multinomial distribution using where and are the updated ones

End for

End for

Task 2 Output:

Number of times that each of the 100 most common dictionary words appear in document 20newsgroups/comp.graphics/37261:

[('subject', 1), ('lines', 1), ('date', 1), ('apr', 0), ('gmt', 1), ('would', 0), ('one', 2), ('writes', 0), ('article', 0), ('dont', 0), ('like', 0), ('people', 0), ('know', 0), ('get', 0), ('think', 0), ('also', 0), ('may', 1), ('use', 0), ('time', 0), ('good', 0), ('new', 0), ('could', 0), ('well', 0), ('even', 0), ('see', 0), ('way', 0), ('make', 0), ('say', 0), ('many', 0), ('much', 0), ('first', 0), ('two', 0), ('right', 0), ('distribution', 1), ('god', 0), ('system', 1), ('want', 0), ('world', 0), ('said', 0), ('anyone', 0), ('used', 0), ('need', 1), ('really', 0), ('work', 2), ('something', 0), ('please', 1), ('problem', 0), ('believe', 0), ('since', 0), ('still', 0), ('years', 0), ('back', 0), ('ive', 0), ('going', 0), ('maxaxaxaxaxaxaxaxaxaxaxaxaxaxax', 0), ('email', 0), ('find', 0), ('take', 0), ('might', 0), ('help', 0), ('point', 1), ('information', 2), ('using', 0), ('better', 0), ('question', 0), ('government', 0), ('file', 0), ('things', 0), ('last', 0), ('cant', 0), ('never', 0), ('thanks', 1), ('must', 1), ('without', 0), ('read', 0), ('windows', 0), ('sure', 0), ('another', 0), ('doesnt', 0), ('year', 0), ('number', 0), ('made', 0), ('etc', 0), ('bit', 0), ('someone', 0), ('got', 0), ('thing', 0), ('thats', 0), ('david', 2), ('university', 0), ('fact', 0), ('part', 0), ('look', 0), ('program', 0), ('available', 1), ('drive', 0), ('version', 0), ('come', 0), ('data', 0), ('anything', 0)]

Task 3 Output:

50 most important words in each of the 20 mixture components:

['stool', 'vaccine', 'armeniens', 'jeremy', 'uneasy', 'views', 'oneself', 'utter', 'disproved', 'wedding', 'murderers', 'asleep', 'alternative', 'demerscsucsdedu', 'irq', 'conclude', 'etxonssufsaericssonse', 'slice', 'pageview', 'gained', 'scenarios', 'kicker', 'aka', 'diegola', 'apologists', 'praying', 'bnr', 'fastball', 'soldered', 'rwrr', 'generalization', 'atmanager', 'typo', 'dobson', 'namely', 'cultural', 'dishes', 'antihistamines', 'cents', 'bow', 'concentrations', 'stanley', 'colons', 'unaware', 'gesture', 'presents', 'charlie', 'personalities', 'misrepresentation', 'substitution'],

['coma', 'heap', 'oliveira', 'keithleyapplecom', 'affiliate', 'suzuki', 'radically', 'fernandez', 'kay', 'honestly', 'innate', 'poverty', 'canvas', 'footnote', 'outright', 'saskatoon', 'educate', 'whats', 'hypocrisy', 'longitude', 'judges', 'phrasing', 'xlyxvaxcitcornelledu', 'tape', 'coercion', 'painting', 'computationally', 'fundamentalists', 'house', 'enter', 'regulars', 'blurb', 'plotting', 'het', 'nanosecond', 'keeper', 'gray', 'enclosure', 'behave', 'atf', 'workshop', 'reuter', 'falsehood', 'hurry', 'hispanics', 'achieved', 'jacket', 'loopback', 'nuclear', 'reducing'],

['biker', 'thick', 'weirdness', 'inquirer', 'unite', 'hoc', 'abusing', 'thy', 'deterministic', 'ics', 'feelings', 'cirrus', 'lam', 'fips', 'allowable', 'quattro', 'squeeze', 'fade', 'imagination', 'whaley', 'concatenated', 'nutrition', 'yesterday', 'santa', 'inserts', 'adda', 'optimized', 'rethought', 'uunetiscbrthorsteveh', 'shed', 'daigle', 'oconnor', 'brother', 'deshaies', 'learning', 'tvs', 'raise', 'plplot', 'linkage', 'symbol', 'sidewalk', 'assoc', 'reuter', 'indicating', 'awareness', 'cell', 'drastic', 'worthless', 'termination', 'particulars'],

['reusable', 'otoh', 'gca', 'cypriot', 'noems', 'outrage', 'geography', 'skeptics', 'obligations', 'amigaphysikunizhch', 'divider', 'pains', 'discharged', 'attested', 'ballots', 'nelsonpapollohpcom', 'pbmplus', 'broadspectrum', 'noyemidwayuchicagoedu', 'monogamous', 'maury', 'egyptians', 'marshalls', 'sang', 'costbenefit', 'weaken', 'resourceuc', 'maxaxaxaxaxaxaxaxaxaxaxaxaxaxaxq', 'altpsychoactives', 'yeltsin', 'deem', 'pgfsrlcacsusledulike', 'jarvis', 'armageddon', 'pwwspacsunriceedu', 'exports', 'pumps', 'jrutledgcsulowelledu', 'dietzcsrochesteredu', 'youll', 'smashing', 'parliamentary', 'curse', 'grieve', 'hardest', 'mechalas', 'acne', 'spafford', 'plaque', 'hilarious'],

['mayo', 'synthetic', 'archie', 'milton', 'constant', 'cooksonmbunixmitreorg', 'child', 'scouting', 'gives', 'bids', 'coli', 'egress', 'associates', 'helsingius', 'mexico', 'tshirts', 'invest', 'designers', 'neatly', 'maariv', 'lenses', 'illegally', 'biberdorf', 'coexist', 'undecidable', 'readingwriting', 'formulae', 'voices', 'rfc', 'ronning', 'fyi', 'vandenberg', 'oppression', 'snipers', 'ruthlessly', 'coward', 'roderick', 'vertex', 'suresh', 'pendleton', 'vandalize', 'receiver', 'brandon', 'bird', 'shells', 'laughed', 'pittsburgh', 'protestantism', 'chambers', 'ssto'],

['dbnandrewcmuedu', 'ksec', 'travenpittedu', 'eugenebigelowebaysuncom', 'cssdaemaccsmcmasterca', 'bmclevelandfreenetedu', 'chongotoadcom', 'hyperhelp', 'obfuscateinfo', 'ntmi', 'mikeyccwfccutexasedu', 'vtwin', 'riggsdescartesetlarmymil', 'kingozcamelotbradleyedu', 'markagcxssdcsdharriscom', 'hightech', 'pautkvmutkedu', 'garroddynamoecnpurdueedu', 'dblvisualcom', 'windowmanager', 'jbhuxacsouiucedu', 'dosdosdosdosdosdosdosdosdosdosdosdosdos', 'warningplease', 'xpertexpolcsmitedu', 'pyeatttexacocom', 'chriscraytonozoneholecom', 'mpaulunledu', 'broadspectrum', 'benalialcorconcordiaca', 'wideband', 'goldbergoasysdtnavymil', 'indepth', 'wcc', 'prbaccessdigexnet', 'deanfringeraincom', 'netwisdom', 'uabdpodpouabedugila', 'russottoengumdedu', 'lutindex', 'fbibatf', 'ceejwellsfcaus', 'winfields', 'lowlevel', 'odwyersseie', 'gerhards', 'acoopermacalstredu', 'rra', 'seefeld', 'asalasdpaarf', 'israelpalestine'],

['armageddon', 'amehdisrchoneywellcom', 'appcontext', 'speedyengrlatechedu', 'everest', 'nitrogen', 'demanded', 'busting', 'als', 'ericsnetcomcom', 'pctcp', 'dictators', 'skipjacks', 'affiliate', 'caronninessiecsidethzch', 'eyewitnesses', 'flyback', 'subscriptions', 'ljonesutkvxutkedu', 'occupational', 'sqrt', 'persuasive', 'turgut', 'semiautomatic', 'theater', 'reconsidering', 'prbaccessdigexnet', 'previewer', 'perfection', 'utils', 'bowen', 'jesse', 'joslinpogoisppittedu', 'pmetzgersnarkshearsoncom', 'barring', 'complies', 'lafayette', 'spacerelated', 'winos', 'benny', 'figaro', 'maxx', 'settling', 'fission', 'ruff', 'sei', 'regulator', 'judea', 'semak', 'worshipped'],

['naslund', 'retirement', 'soap', 'clerk', 'conjugate', 'babes', 'dictatorship', 'wintach', 'coronary', 'diego', 'min', 'sad', 'pros', 'accidents', 'tripoli', 'mortals', 'awe', 'mfc', 'babble', 'relaxed', 'prob', 'gas', 'prompt', 'nationality', 'jonestown', 'markings', 'conception', 'succumb', 'defence', 'passenger', 'raz', 'interpretations', 'interpolation', 'philosophers', 'inability', 'lawyer', 'crux', 'purchase', 'mercy', 'castle', 'tap', 'rapists', 'barlas', 'bodys', 'torture', 'excessive', 'earnest', 'kills', 'time', 'calculators'],

['djwahemulnadakthse', 'appalling', 'xpertexpolcsmitedu', 'noemi', 'objectives', 'scalable', 'baptists', 'api', 'vogle', 'lavalliere', 'submission', 'singled', 'josephs', 'godhead', 'epistemology', 'peach', 'akyfnysuedu', 'fbis', 'kirschstafftcumnedu', 'orbits', 'coup', 'clarifications', 'richkgrebyncom', 'pinchhit', 'spline', 'unfounded', 'savage', 'quota', 'linden', 'cords', 'antifungal', 'thorny', 'hivinfected', 'specifying', 'infj', 'advise', 'micheal', 'incorrectly', 'unambiguous', 'pim', 'dialogue', 'anxiety', 'retrieval', 'collar', 'perth', 'sheffield', 'discern', 'binsh', 'similarities', 'sporanox'],

['escrowed', 'illnesses', 'harper', 'theism', 'lox', 'selfdefence', 'kbps', 'safeguarding', 'candidates', 'dumb', 'documentary', 'collective', 'liblibxso', 'forum', 'cruise', 'stripped', 'humbly', 'kah', 'karabakh', 'scouting', 'medication', 'aligned', 'calmed', 'romdasuclinkberkeleyedu', 'earthquake', 'ocean', 'lilleyvcgumccacuk', 'oliveira', 'diverted', 'saskatoon', 'ball', 'henderson', 'cowboy', 'wintach', 'accomplishments', 'adams', 'biases', 'rvmcmscsuohioedu', 'pkk', 'spectrometer', 'suffice', 'developments', 'alarms', 'punisher', 'curriculum', 'lookalike', 'plead', 'viruses', 'chord', 'written'],

['scrollbar', 'remotes', 'selfhating', 'occurrence', 'pagans', 'retreating', 'coincide', 'snes', 'dialins', 'mov', 'cstkocrsvdelcoelectcom', 'supper', 'corpses', 'luna', 'plaza', 'rubbing', 'xray', 'provocation', 'youre', 'martial', 'haaretz', 'fiberglass', 'herzog', 'fistiscpbellcorecom', 'amesarcnasagov', 'lossfsececmuedu', 'consultations', 'chant', 'categorized', 'maximus', 'stevebaccessdigexcom', 'vermont', 'coastal', 'echl', 'casually', 'congressional', 'ibmpc', 'kolodny', 'leon', 'fergusoncsrochesteredu', 'prompted', 'silliness', 'questioned', 'answerfax', 'aforementioned', 'welty', 'diaspora', 'compromised', 'contraceptive', 'nazareth'],

['personally', 'pixie', 'episodes', 'girls', 'mod', 'israeli', 'decs', 'erosion', 'imam', 'mens', 'cause', 'trim', 'vince', 'gideon', 'sub', 'drums', 'yassin', 'challenged', 'llicscmuedu', 'regression', 'preferable', 'deed', 'inappropriate', 'armies', 'defect', 'optimum', 'behalf', 'slightly', 'settings', 'analyses', 'controlling', 'highlight', 'luriemallegedumichael', 'cdn', 'booklet', 'disclaimer', 'inflamed', 'wedlock', 'richman', 'woe', 'kai', 'gary', 'netters', 'folks', 'hence', 'tracer', 'spanhepnet', 'ala', 'recurrence', 'testimonial'],

['categorized', 'kmh', 'tao', 'propose', 'stampfli', 'shalt', 'also', 'lausanne', 'bourque', 'supplement', 'boudrie', 'nords', 'aunt', 'garaged', 'hypocrites', 'swear', 'ongoing', 'balloon', 'highlevel', 'bumpy', 'dimensions', 'tagged', 'arthurpnetctscom', 'eloquent', 'math', 'onsite', 'faxes', 'lucas', 'drifted', 'scholarship', 'rumors', 'cruncher', 'abpsoft', 'exterminated', 'ega', 'locale', 'coordinated', 'accomplishing', 'ide', 'alfa', 'genuine', 'specify', 'coil', 'zijdenbos', 'radically', 'inquiring', 'newbie', 'smells', 'flight', 'finite'],

['ibmpc', 'amd', 'twin', 'broken', 'newest', 'believable', 'dialogue', 'oxford', 'issue', 'breathe', 'piston', 'homo', 'sprites', 'destroying', 'implement', 'upwards', 'allied', 'sauyet', 'southern', 'combo', 'schaefer', 'became', 'pythagorean', 'routine', 'governmental', 'marxhausen', 'applying', 'dam', 'arisen', 'prophetic', 'forgiveness', 'drops', 'interception', 'dsc', 'putting', 'kolstadcaewiscedu', 'guess', 'usable', 'annoyed', 'says', 'recovering', 'conflict', 'injure', 'subsistence', 'eden', 'macweek', 'forced', 'typo', 'jackson', 'sack'],

['voltaire', 'overlook', 'stamps', 'arrow', 'performs', 'respectability', 'wall', 'region', 'personaly', 'mpu', 'outlawing', 'scroll', 'geo', 'subsequently', 'dig', 'exempt', 'atkins', 'apacgw', 'perlmanqsocoloradoedu', 'seannabnrca', 'hitchings', 'faqs', 'organizational', 'owns', 'authority', 'afterwards', 'knee', 'sciviguy', 'hicolor', 'cullen', 'dlbtamutstamuedu', 'valiant', 'welcome', 'posing', 'turk', 'michael', 'babble', 'scripter', 'elabdellaoui', 'briandanasazicom', 'microcontroller', 'asserted', 'radios', 'ballgame', 'economically', 'suggesting', 'tract', 'affiliate', 'encompasses', 'relevance'],

['monty', 'premenstrual', 'defrag', 'jaynemmaltguildorg', 'elliott', 'seeker', 'schooling', 'playback', 'patricia', 'obesity', 'commandcom', 'astronomers', 'clarinet', 'disappointing', 'bart', 'donot', 'grandparents', 'judgestoadcom', 'jgecsuncedu', 'questionnaire', 'peaceloving', 'nyr', 'projectile', 'auction', 'safetybelt', 'gummint', 'mbeavingbnrca', 'krs', 'arent', 'calmed', 'antimatter', 'vinge', 'reinstalling', 'homeopathic', 'metroareas', 'gtcprismgatechedu', 'ultralong', 'hes', 'icol', 'numb', 'rcstageurctuenl', 'nntppostinghost', 'tyrant', 'partisan', 'johnhmacadammpcemqeduau', 'elegant', 'curtains', 'deviation', 'piggyback', 'discourse'],

['chorion', 'oxalic', 'ssdnetworking', 'shellgatellouupsicom', 'shoots', 'crosses', 'mcgrawhill', 'wholly', 'spout', 'holmes', 'maxentry', 'xmgr', 'goering', 'contemplate', 'jimzisfeinfactorycom', 'launchpaduncedu', 'bumpy', 'dewey', 'arrange', 'pressured', 'davison', 'shuts', 'diamonds', 'lube', 'guberniia', 'tvs', 'arthurnoguerolavosstratuscom', 'safeguarding', 'deluded', 'posit', 'scoreless', 'jrogoffscottskidmoreedu', 'thfkimbarkuchicagoedu', 'dcooksonmitreorg', 'jargon', 'xon', 'fuess', 'videotaped', 'outreach', 'crepslateranucsindianaedu', 'specialty', 'caere', 'economically', 'sirtf', 'acidophilus', 'prk', 'rashid', 'amandainterconcom', 'floppyburn', 'thedoge'],

['subsequent', 'porn', 'heath', 'camino', 'gestapo', 'canonical', 'observe', 'vechorik', 'nutters', 'disposed', 'oname', 'exyugoslavia', 'trodwell', 'ratings', 'pirates', 'svoboda', 'increased', 'mcdonnell', 'friendship', 'gord', 'pale', 'callisonuokmaxecnuoknoredu', 'herring', 'hormone', 'usenetpadeccom', 'newsnimaster', 'extends', 'fiasco', 'supra', 'keyseach', 'misclegal', 'ping', 'countless', 'goexe', 'steves', 'outlets', 'cyprus', 'arising', 'headache', 'retract', 'drdos', 'acceptance', 'archimedes', 'overwhelm', 'assaulted', 'lump', 'sick', 'tenant', 'woodbury', 'suspend'],

['hunters', 'conducive', 'laser', 'roby', 'suspiciously', 'stall', 'smt', 'connor', 'stock', 'macinnis', 'symmetric', 'anybodys', 'conference', 'madison', 'rope', 'contingent', 'brock', 'pvasilion', 'oppressive', 'integral', 'strategy', 'kent', 'busses', 'prefix', 'respective', 'keysym', 'nagarnokarabagh', 'oppinions', 'urbana', 'sinning', 'wednesday', 'rogoff', 'nicfunetfi', 'lionessmaplecircaufledu', 'handle', 'las', 'screamed', 'evolved', 'dumping', 'darken', 'interact', 'ejected', 'imaging', 'ownership', 'square', 'polygon', 'omnipotence', 'nasas', 'hernlem', 'hwang'],

['unnoticed', 'receiving', 'dear', 'fashion', 'orientations', 'slant', 'romantic', 'joan', 'shattered', 'listing', 'kamloops', 'confusing', 'suited', 'octane', 'theism', 'admitted', 'exactly', 'extremes', 'shadows', 'onethird', 'sft', 'canisters', 'cultures', 'albany', 'deficits', 'shutouts', 'modmask', 'itar', 'gassing', 'andrewidacomhpcom', 'monophysites', 'audi', 'myron', 'dawns', 'supplied', 'gerard', 'priori', 'sales', 'foot', 'industrial', 'encapsulated', 'swap', 'felony', 'flu', 'frank', 'educators', 'oversampling', 'building', 'observational', 'enormous']

Task 4 Output:

The number of documents assigned to each cluster:

[(2, 3), (3, 3184), (5, 7716), (6, 1595), (8, 1697), (9, 23), (10, 1700), (13, 1), (15, 2100), (16, 1648), (17, 330)]

Top 3 categories assigned to each cluster and percentage of documents that are assigned to the cluster that fall in that category:

[],

[],

[(2, 66.66666666666666, 'misc.forsale'), (2, 33.33333333333333, 'rec.sport.baseball')],

[(3, 15.10678391959799, 'misc.forsale'), (3, 7.192211055276382, 'comp.graphics'), (3, 7.035175879396985, 'comp.sys.mac.hardware')],

[],

[(5, 7.179885951270087, 'soc.religion.christian'), (5, 6.972524624157595, 'talk.politics.guns'), (5, 6.544841886988078, 'talk.politics.misc')],

[(6, 7.3354231974921635, 'talk.religion.misc'), (6, 7.2727272727272725, 'alt.atheism'), (6, 6.583072100313479, 'misc.forsale')],

[],

[(8, 8.36770771950501, 'alt.atheism'), (8, 8.014142604596348, 'rec.motorcycles'), (8, 7.483794932233353, 'talk.religion.misc')],

[(9, 30.434782608695656, 'rec.sport.hockey'), (9, 13.043478260869565, 'rec.autos'), (9, 8.695652173913043, 'sci.electronics')],

[(10, 10.941176470588236, 'rec.motorcycles'), (10, 6.0588235294117645, 'sci.crypt'), (10, 5.764705882352941, 'comp.sys.mac.hardware')],

[],

[],

[(13, 100.0, 'sci.med')],

[],

[(15, 7.0476190476190474, 'rec.autos'), (15, 7.000000000000001, 'comp.sys.ibm.pc.hardware'), (15, 7.000000000000001, 'sci.med')],

[(16, 12.439320388349515, 'soc.religion.christian'), (16, 11.650485436893204, 'comp.os.ms-windows.misc'), (16, 11.04368932038835, 'comp.windows.x')],

[(17, 13.636363636363635, 'comp.sys.ibm.pc.hardware'), (17, 12.727272727272727, 'comp.windows.x'), (17, 11.818181818181818, 'comp.sys.mac.hardware')],

[],

[]]

Although it is not 100% correct, we can see our clustering algorithm works by looking at the top 3 categories for each cluster. For example, if we look at cluster 5 the top 3 categories are 'soc.religion.christian', 'talk.politics.guns' and 'talk.politics.misc' and it makes sense to put these 3 categories into one cluster. However, some clusters have some inconsistency in the categories. For example, cluster 16 the top 3 categories are 'soc.religion.christian', 'comp.os.ms-windows.misc' and 'comp.windows.x' and we can note that religion is not related to windows.

Perhaps we could get better presentation if we kept the more generalized form of each real category. For example, if we only look at ‘religion’ from 'soc.religion.christian' and 'talk.religion.misc' we might see that each cluster will have a higher prevalence of several related categories.