Student certification:

Team member 1:
Print Name: Yousef Haggy
Date: 11/11/21
I have contributed by doing the following: 4.3
Signed:

Team member 2:
Print Name: Mahir Rahman
Date: 11/11/21
I have contributed by doing the following: 4.1 and 4.2
Signed:

Pr.4.1

a.) Provide the feature vector in your report

Feature Vector :

['on', 'review', 'mention', 'watch', 'oz', 'episod', 'youll', 'hook', 'right', 'exactli', 'happen', 'first', 'thing', 'struck', 'brutal', 'unflinch', 'scene', 'violenc', 'set', 'word', 'go', 'trust', 'show', 'faint', 'heart', 'timid', 'pull', 'punch', 'regard', 'drug', 'sex', 'hardcor', 'classic', 'us', 'call', 'nicknam', 'given', 'oswald', 'maximum', 'secur', 'state', 'penitentari', 'focus', 'mainli', 'emerald', 'citi', 'experiment', 'section', 'prison', 'cell', 'glass', 'front', 'face', 'inward', 'privaci', 'high', 'agenda', 'em', 'home', 'manyaryan', 'muslim', 'gangsta', 'latino', 'christian', 'italian', 'irish', 'moreso', 'scuffl', 'death', 'stare', 'dodgi', 'deal', 'shadi', 'agreement', 'never', 'far', 'awai', 'main', 'appeal', 'due', 'fact', 'goe', 'wouldnt', 'dare', 'forget', 'pretti', 'pictur', 'paint', 'mainstream', 'audienc', 'charm', 'romanceoz', 'doesnt', 'mess', 'around', 'saw', 'nasti', 'surreal', 'couldnt', 'readi', 'more', 'develop', 'tast', 'accustom', 'level', 'graphic', 'injustic', 'crook', 'guard', 'wholl', 'sold', 'out', 'nickel', 'inmat', 'kill', 'order', 'well', 'manner', 'middl', 'class', 'be', 'turn', 'bitch', 'lack', 'street', 'skill', 'experi', 'becom', 'comfort', 'uncomfort', 'viewingthat', 'touch', 'darker', 'side', 'wonder', 'littl', 'product', 'film', 'techniqu', 'veri', 'unassum', 'oldtimebbc', 'fashion', 'give', 'sometim', 'discomfort', 'sens', 'realism', 'entir', 'piec', 'actor', 'extrem', 'chosen', 'michael', 'sheen', 'polari', 'voic', 'down', 'pat', 'truli', 'see', 'seamless', 'edit', 'guid', 'refer', 'william', 'diari', 'entri', 'worth', 'terrificli', 'written', 'perform', 'master', 'great', 'comedi', 'life', 'realli', 'come', 'fantasi', 'tradit', 'dream', 'remain', 'solid', 'disappear', 'plai', 'knowledg', 'particularli', 'concern', 'orton', 'halliwel', 'flat', 'mural', 'decor', 'surfac', 'terribl', 'done', 'thought', 'wai', 'spend', 'time', 'hot', 'summer', 'weekend', 'sit', 'air', 'condit', 'theater', 'lightheart', 'plot', 'simplist', 'dialogu', 'witti', 'charact', 'likabl', 'even', 'bread', 'suspect', 'serial', 'killer', 'disappoint', 'realiz', 'match', 'point', 'risk', 'addict', 'proof', 'woodi', 'allen', 'still', 'fulli', 'control', 'style', 'mani', 'grown', 'love', 'id', 'laugh', 'year', 'decad', 'iv', 'impress', 'scarlet', 'johanson', 'manag', 'tone', 'sexi', 'imag', 'jump', 'averag', 'spirit', 'young', 'woman', 'crown', 'jewel', 'career', 'wittier', 'devil', 'wear', 'prada', 'interest', 'superman', 'friend', 'taut', 'organ', 'grip', 'edward', 'dmytryk', 'crossfir', 'distinct', 'suspens', 'thriller', 'unlik', 'messag', 'movi', 'look', 'devic', 'noir', 'cycl', 'bivouack', 'washington', 'dc', 'compani', 'soldier', 'cope', 'restless', 'hang', 'bar', 'three', 'end', 'up', 'stranger', 'apart', 'robert', 'ryan', 'drunk', 'belliger', 'beat', 'host', 'sam', 'leven', 'jewish', 'polic', 'detect', 'investig', 'help', 'mitchum', 'who', 'assign', 'outfit', 'suspicion', 'fall', 'second', 'georg', 'cooper', 'vanish', 'slai', 'third', 'buddi', 'steve', 'brodi', 'insur', 'silenc', 'befor', 'close', 'abet', 'superior', 'script', 'john', 'paxton', 'draw', 'precis', 'star', 'bob', 'natur', 'prototyp', 'angri', 'white', 'male', 'hilt', 'underplai', 'characterist', 'alert', 'nonchal', 'role', 'central', 'better', 'gloria', 'graham', 'fullyfledg', 'rendit', 'smartmouth', 'vulner', 'tramp', 'sad', 'sack', 'leech', 'paul', 'kelli', 'haunt', 'small', 'peripher', 'make', 'memor', 'polit', 'engag', 'perhap', 'inevit', 'succumb', 'sermon', 'much', 'confin', 'reminisc', 'grandfath', 'di', 'hand', 'bigot', 'centuri', 'earlier', 'thu', 'incident', 'stretch', 'chronolog', 'limit', 'there', 'attempt', 'render', 'explan', 'glib', 'hate', 'jew', 'hillbilli', 'curious', 'surviv', 'major', 'chang', 'wrought', 'upon', 'novel', 'base', 'richard', 'brook', 'brick', 'foxhol', 'dealt', 'gaybash', 'murder', 'homosexu', 'beyond', 'pale', 'new', 'holocaust', 'begun', 'emerg', 'ash', 'europ',

'hollywood', 'felt', 'embolden', 'regist', 'protest', 'against', 'antisemit', 'studio', 'alwai', 'quak', 'prospect', 'offend', 'potenti', 'ticket', 'buyer', 'homophobia', 'work', 'gener', 'specif', 'dont', 'fit', 'smoothli', 'victim', 'chat', 'lonesom', 'invit', 'back', 'odd', 'though', 'especi', 'girlfriend', 'tow', 'rais', 'question', 'whether', 'scenario', 'retain', 'inadvert', 'left', 'discreet', 'tipoff', 'origin', 'engin', 'rage', 'petter', 'mattei', 'monei', 'visual', 'stun', 'mr', 'offer', 'vivid', 'portrait', 'human', 'relat', 'seem', 'tell', 'power', 'success', 'peopl', 'differ', 'situat', 'encount', 'variat', 'arthur', 'schnitzler', 'same', 'theme', 'director', 'transfer', 'action', 'present', 'york', 'meet', 'connect', 'each', 'anoth', 'next', 'person', 'know', 'previou', 'contact', 'stylishli', 'sophist', 'luxuri', 'taken', 'live', 'world', 'habitat', 'get', 'soul', 'stage', 'loneli', 'inhabit', 'big', 'best', 'place', 'find', 'sincer', 'fulfil', 'discern', 'case', 'act', 'good', 'under', 'direct', 'buscemi', 'rosario', 'dawson', 'carol', 'kane', 'imperioli', 'adrian', 'grenier', 'rest', 'talent', 'cast', 'aliv', 'wish', 'luck', 'await', 'anxious', 'lost', 'didnt', 'begin', 'achingli', 'tediou', 'heroin', 'hous', 'actual', 'menac', 'forebod', 'creat', 'dure', 'appar', 'constant', 'thunderstorm', 'strang', 'heard', 'housegreat', 'doubl', 'glaze', 'few', 'mile', 'town', 'sever', 'hour', 'walk', 'girl', 'serv', 'purpos', 'except', 'provid', 'surprisingli', 'quick', 'gori', 'tedium', 'unbear', 'suggest', 'spate', 'throughout', 'area', 'ventur', 'bizarr', 'ritual', 'salt', 'pepper', 'sum', 'inher', 'add', 'lead', 'actress', 'cant', 'will', 'complet', 'irrelev', 'nude', 'shower', 'video', 'hope', 'follow', 'simpli', 'ban', 'uk', 's', 'mostli', 'final', 'over', 'extend', 'noth', 'curios', 'valu', 'daft', 'worryit', 'telegraph', 'ten', 'minut', 'wood', 'steep', 'upward', 'slope', 'obvious', 'struggl', 'halfwai', 'through', 'figur', 'top', 'dress', 'black', 'brandish', 'larg', 'scyth', 'slide', 'run', 'cours', 'stand', 'conveni', 'nice', 'upright', 'weapon', 'disclaim', 'seen', 'last', 'music', 'week', 'allow', 'judg', 'without', 'taint', 'wasnt', 'believ', 'dougla', 'quit', 'along', 'kasei', 'think', 'danc', 'part', 'worthwhil', 'addit', 'compar', 'dancer', 'sing', 'bigger', 'easier', 'light', 'expect', 'inde', 'deliv', 'song', 'common', 'whole', 'opinion', 'bad', 'obviou', 'cut', 'between', 'talk', 'dub', 'singer', 'portion', 'impecc', 'enjoy', 'amaz', 'fresh', 'innov', 'idea', 'brilliant', 'drop', 'funni', 'anymor', 'continu', 'declin', 'further', 'wast', 'todai', 'disgrac', 'fallen', 'write', 'painfulli', 'mildli', 'entertain', 'respit', 'guesthost', 'probabl', 'hard', 'creator', 'handselect', 'chose', 'band', 'hack', 'recogn', 'such', 'brillianc', 'replac', 'mediocr', 'respect', 'made', 'huge', 'now', 'aw', 'spoiler', 'real', 'familiar', 'stori', 'men', 'put', 'war', 'zone', 'gun', 'rifl', 'innoc', 'handl', 'fire', 'jimmi', 'davi', 'franchot', 'repeat', 'thousand', 'forc', 'take', 'arm', 'countri', 'want', 'kick', 'armi', 'encourag', 'stai', 'belt', 'mouth', 'fred', 'p', 'willi', 'spencer', 'traci', 'line', 'franc', 'unit', 'pin', 'german', 'machin', 'nest', 'singl', 'handedli', 'commiss', 'pick', 'half', 'dozen', 'safeti', 'nearbi', 'church', 'steepl', 'surrend', 'artilleri', 'shell', 'hit', 'serious', 'wound', 'recov', 'hospit', 'fell', 'volunt', 'nurs', 'rose', 'duffi', 'gladi', 'happi', 'lucki', 'despit', 'obnoxi', 'antic', 'toward', 'fight', 'western', 'later', 'marri', 'unexpectedli', 'french', 'station', 'sticki', 'both', 'alreadi', 'accept', 'propos', 'marriag', 'wwi', 'bitter', 'resent', 'man', 'accid', 'ran', 'discov', 'shock', 'surpris', 'meek', 'nonviol', 'knew', 'sent', 'european', 'smug', 'sure', 'himself', 'abil', 'shoot', 'mobster', 'underworld', 'found', 'wife', 'involv', 'law', 'abid', 'inoffens', 'adjust', 'crime', 'came', 'full', 'circl', 'secret', 'rat', 'prevent', 'execut', 'valentin', 'dai', 'massacr', 'gang', 'member', 'cop', 'rival', 'trial', 'admit', 'guilt', 'sentenc', 'togeth', 'hear', 'rumor', 'fellow', 'convict', 'have', 'affair', 'behind', 'broke', 'fugit', 'circu', 'manger', 'barker', 'true', 'sudden', 'ad', 'decid', 'let', 'track', 'job', 'dresser', 'onc', 'ago', 'try', 'splash', 'god', 'two', 'leav', 'speechless', 'albert', 'finnei', 'tom', 'courtenai', 'less', 'sir', 'ag', 'shakespearean', 'norman', 'sort', 'valet', 'king', 'lear', 'blitz', 'london', 'ii',

'depend', 'helpless', 'aid', 'cajol', 'wheedl', 'bulli', 'onstag', 'th', 'vicari', 'need', 'anywai', 'characterdriven', 'secondari', 'interact', 'requir', 'highest', 'calib', 'bring', 'old', 'sick', 'petul', 'hiss', 'fume', 'theyr', 'bow', 'convinc', 'minc', 'mother', 'elderli', 'employ', 'wrong', 'term', 'although', 'technic', 'relationship', 'employe', 'coupl', 'ye', 'other', 'marvel', 'ronald', 'harwood', 'arent', 'fine', 'notabl', 'eileen', 'atkin', 'longsuff', 'madg', 'desir', 'regret', 'rememb', 'nomin', 'five', 'academi', 'award', 'peter', 'yate', 'adapt', 'screenplai', 'prepar', 'mesmer', 'exampl', 'view', 'tobe', 'hooper', 'gem', 'crocodil', 'collegecrocodil', 'nich', 'exploitationmonst', 'genr', 'forward', 'wayward', 'produc', 'sequel', 'delight', 'bonbon', 'camp', 'ed', 'subtl', 'flair', 'digniti', 'remark', 'room', 'monke', 'special', 'effect', 'comput', 'wed', 'crocki', 'fodder', 'russ', 'meyer', 'breast', 'ren', 'hoek', 'pector', 'implant', 'opu', 'referenc', 'blood', 'surf', 'dish', 'bunch', 'chum', 'bucket', 'past', 'reveng', 'nerd', 'allusionshomagesripoff', 'jaw', 'templ', 'doom', 'indiana', 'jone', 'crusad', 'convent', 'godzilla', 'jame', 'bond', 'readyfortv', 'fade', 'editor', 'gave', 'stock', 'crock', 'sotto', 'voce', 'tenor', 'soliloqui', 'environmentalismor', 'appreci', 'quasicaptain', 'ahab', 'tour', 'de', 'speach', 'gallop', 'shootout', 'golden', 'sunset', 'hopefulli', 'monkei', 'flush', 'toilet', 'intern', 'space', 'midget', 'enjoi', 'exploit', 'waltz', 'zerog', 'monkeymidgetcrocodil', 'bloodsh', 'allinal', 'whammi', 'irk', 'im', 'fan', 'boll', 'again', 'postal', 'mayb', 'bought', 'cry', 'long', 'game', 'itself', 'finsish', 'merc', 'infiltr', 'research', 'lab', 'locat', 'tropic', 'island', 'warn', 'someth', 'scheme', 'legion', 'schmuck', 'feel', 'lonelei', 'countrymen', 'player', 'name', 'til', 'schweiger', 'udo', 'kier', 'ralf', 'moeller', 'self', 'biz', 'tale', 'jack', 'carver', 'hail', 'bratwurst', 'eat', 'dude', 'badass', 'complain', 'he', 'perspect', 'storylin', 'dement', 'evil', 'mad', 'scientist', 'dr', 'krieger', 'geneticallymutatedsoldi', 'gm', 'topsecret', 'remind', 'vancouv', 'reason', 'that', 'palm', 'tree', 'here', 'instead', 'rich', 'lumberjackwood', 'havent', 'gone', 'start', 'meheh', 'wanna', 'shenanigan', 'mean', 'suck', 'impli', 'boat', 'until', 'cromedalbino', 'squad', 'enter', 'everyth', 'reek', 'scheiss', 'poop', 'simpleton', 'wiff', 'ahead', 'btw', 'annoi', 'sidekick', 'screen', 'horribl', 'monster', 'chanc', 'busi', 'sword', 'emot', 'attach', 'destroi', 'blatantli', 'stolen', 'lotr', 'matrix', 'ghost', 'yoda', 'obe', 'vader', 'spider', 'frodo', 'attack', 'return', 'elijah', 'waitit', 'hypnot', 'sting', 'wrap', 'upuh', 'hello', 'vs', 'matrixor', 'termin', 'someon', 'nazi', 'juvenil', 'rush', 'conclus', 'children', 'adult', 'save', 'die', 'glut', 'cash', 'gui', 'concept', 'cliffhang', 'mountain', 'rescu', 'sly', 'stop', 'mom', 'stallon', 'nitpick', 'those', 'expert', 'climb', 'basejump', 'aviat', 'facial', 'express', 'excus', 'dismiss', 'overblown', 'pile', 'junk', 'outact', 'hors', 'nonsens', 'lovabl', 'undeni', 'romp', 'plenti', 'thrill', 'unintention', 'youv', 'lithgow', 'sneeri', 'tick', 'box', 'baddi', 'perman', 'harass', 'hapless', 'turncoat', 'agent', 'rex', 'linn', 'traver', 'henri', 'rooker', 'noteworthi', 'cringeworthi', 'hal', 'insist', 'constantli', 'shriek', 'pain', 'disbelief', 'captor', 'hurt', 'anybodi', 'whilst', 'ralph', 'wait', 'frank', 'grin', 'plummet', 'former', 'burn', 'craig', 'fairbrass', 'brit', 'cropper', 'footbal', 'bit', 'judgement', 'care', 'lower', 'volum', 'your', 'qaulen', 'helicopt', 'fate', 'walter', 'sparrow', 'possess', 'mysteri', 'eeri', 'similar', 'number', 'unfold', 'fiction', 'book', 'eventu', 'intrigu', 'premis', 'undon', 'weak', 'fail', 'held', 'worst', 'fill', 'silli', 'sequenc', 'laughabl', 'dialog', 'mood', 'screenwrit', 'nineti', 'low', 'twist', 'joel', 'schumach', 'respons', 'redeem', 'phone', 'booth', 'capabl', 'stinker', 'drench', 'focu', 'move', 'clunki', 'slow', 'pace', 'switch', 'realiti', 'what', 'quickli', 'titl', 'stuck', 'listen', 'carrei', 'narrat', 'bore', 'tear', 'solv', 'tension', 'impati', 'reach', 'unconvinc', 'celebr', 'finish', 'forgett', 'jim', 'clearli', 'sleepwalk', 'wooden', 'insid', 'virginia', 'madsen', 'logan', 'lerman', 'bland', 'danni',

'huston', 'overal', 'rate', 'mario', 'fond', 'memori', 'super', 'kid', 'brought', 'galaxi', 'latest', 'instal', 'franchis', 'keep', 'intact', 'greatest', 'element', 'notic', 'receiv', 'letter', 'princess', 'peach', 'castl', 'mushroom', 'kingdom', 'arriv', 'bowser', 'son', 'jr', 'airship', 'kidnap', 'lift', 'midst', 'land', 'unknown', 'planet', 'luma', 'float', 'leader', 'rosalina', 'scatter', 'univers', 'adventur', 'fly', 'consist', 'multipl', 'travel', 'amongst', 'via', 'retriev', 'lose', 'graviti', 'environ', 'stuff', 'wiimot', 'uniqu', 'shake', 'remot', 'spin', 'primari', 'activ', 'pointer', 'enemi', 'object', 'wii', 'describ', 'minor', 'gripe', 'upsid', 'matter', 'restart', 'problem', 'asid', 'superb', 'highli', 'challeng', 'type', 'weve', 'perfect', 'outstand', 'driven', 'obligatori', 'romanc', 'endless', 'car', 'chase', 'dandi', 'aidan', 'quinn', 'terrorist', 'naval', 'offic', 'recruit', 'elimin', 'rare', 'tier', 'import', 'carri', 'usual', 'alist', 'donald', 'sutherland', 'moral', 'ambigu', 'somewhat', 'creepi', 'ben', 'kingslei', 'isra', 'complex', 'converg', 'gradual', 'paranoia', 'claustrophobia', 'captur', 'portrai', 'frighten', 'intens', 'soundtrack', 'abov', 'qualiti', 'dumb', 'andor', 'meaningless', 'loud', 'villain', 'ham', 'repeatedli', 'costum', 'cater', 'assuredli', 'releas', 'revolt', 'zombi', 'prove', 'revamp', 'recycl', 'necessarili', 'lightn', 'strike', 'twice', 'halperin', 'brother', 'horror', 'trite', 'garbag', 'mere', 'popular', 'closeup', 'lugosi', 'ey', 'court', 'battl', 'owner', 'everyon', 'appear', 'victor', 'uninterest', 'insult', 'debacl', 'dead', 'mindcontrol', 'crack', 'egg', 'includ', 'lame', 'jealou', 'send', 'spiral', 'tri', 'advantag', 'win', 'isnt', 'spit', 'youd', 'blind', 'stupid', 'racial', 'insensit', 'bill', 'dvd', 'atmospher', 'masterpiec', 'comparison', 'atom', 'scifi', 'alien', 'chees', 'fest', 'invis', 'invad', 'seriou', 'drama', 'ball', 'ridicul', 'melodramat', 'korean', 'twitch', 'south', 'known', 'melodrama', 'credit', 'list', 'iron', 'favorit', 'whose', 'idiot', 'recommend', 'decent', 'remak', 'gosh', 'ms', 'english', 'assembl', 'experienc', 'writerdirector', 'formula', 'accord', 'imdb', 'bio', 'exclus', 'televis', 'glaringli', 'none', 'read', 'kept', 'joke', 'terrif', 'tv', 'camera', 'usag', 'awkward', 'chop', 'miniseri', 'sitcom', 'cinema', 'sadli', 'translat', 'cring', 'embarrass', 'writer', 'shrivel', 'meg', 'perki', 'cute', 'plastic', 'surgeri', 'recreat', 'stereotyp', 'annett', 'bene', 'motion', 'poor', 'women', 'caricatur', 'substanc', 'updat', 'subtleti', 'assist', 'frustrat', 'yorker', 'cartoon', 'yesterdai', 'appropri', 'exec', 'caption', 'wors', 'orwel', 'prophet', 'total', 'confus', 'cover', 'hearsai', 'quot', 'educ', 'plu', 'cinematographi', 'burton', 'wonderfulli', 'grim', 'desol', 'prostitut', 'fantast', 'dark', 'propaganda', 'explain', 'bandi', 'eurasia', 'etc', 'winston', 'report', 'food', 'canteen', 'drink', 'ill', 'brainwash', 'sister', 'father', 'scrub', 'essenti', 'myself', 'arthous', 'guess', 'disjoint', 'badli', 'chatter', 'swamp', 'harder', 'artist', 'choic', 'nuditi', 'gratuit', 'thrown', 'coverag', 'step', 'imagin', 'reli', 'liter', 'understand', 'deni', 'predict', 'societi', 'date', 'bibl', 'novelist', 'gape', 'hole', 'bui', 'copi', 'shot', 'keira', 'knightlei', 'elizabeth', 'bennet', 'wander', 'field', 'dawn', 'invok', 'clich', 'address', 'phenomenon', 'strongmind', 'rebelli', 'femal', 'period', 'joe', 'wright', 'regrett', 'misapprehens', 'jane', 'austen', 'nuanc', 'conduct', 'sparkl', 'delic', 'social', 'eighteenth', 'drawingroom', 'ucertif', 'wuther', 'height', 'treat', 'darci', 'outsid', 'inappropri', 'rug', 'sceneri', 'pour', 'rain', 'particular', 'passion', 'sexual', 'strategi', 'negoti', 'stultif', 'ignor', 'rambuncti', 'chaotic', 'everybodi', 'shout', 'underwear', 'chair', 'pig', 'happili', 'rowdi', 'reload', 'danceorgi', 'slightest', 'proprieti', 'geniu', 'li', 'explor', 'void', 'nobodi', 'overwhelm', 'tragic', 'predica', 'aris', 'misunderstand', 'miscommun', 'enabl', 'gap', 'factor', 'function', 'eras', 'nowher', 'int', 'sacrif', 'favour', 'overwrought', 'jar', 'materi', 'humour', 'pride', 'prejudic', 'methodolog', 'suppress', 'pofac', 'clumsili', 'narr', 'weightier', 'intertwin', 'embed', 'bare', 'bone', 'heavyhand', 'mysticalnumin', 'fauxbrow', 'bennett', 'suppos', 'matur', 'sensibl', 'clearsight',

'emptyhead', 'giggl', 'schoolgirl', 'wit', 'comb', 'verbal', 'exchang', 'quintessenti', 'folli', 'strength', 'composur', 'clearsighted', 'lot', 'distanc', 'head', 'their', 'scream', 'provoc', 'genuin', 'hissi', 'integr', 'observ', 'within', 'boundari', 'sustain', 'impeach', 'whatsoev', 'furthermor', 'barefoot', 'mud', 'version', 'establish', 'doubt', 'therefor', 'astonishingli', 'unsubtl', 'quest', 'abli', 'matthew', 'macfayden', 'ineffectu', 'befuddl', 'wick', 'detach', 'expens', 'fascin', 'lizzi', 'fool', 'depriv', 'impact', 'binglei', 'longer', 'amiabl', 'wellmean', 'retard', 'cheap', 'wildli', 'inconsist', 'anyth', 'veer', 'verbatim', 'chunk', 'clumsi', 'contemporan', 'languag', 'modern', 'romant', 'bbc', 'yourself', 'heartach', 'oneandahalf', 'torment', 'viewer', 'bloodi', 'basic', 'horrormovi', 'per', 'se', 'spectacl', 'grotesqu', 'enough', 'ya', 'theyv', 'blair', 'whitch', 'project', 'handheldcamera', 'witch', 'hei', 'yall', 'sound', 'depress', 'hyster', 'sai', 'effort', 'soundwis', 'wise', 'vari', 'eurohous', 'grungi', 'hardrock', 'advis', 'circumst', 'failur', 'convei', 'hopeless', 'shameless', 'environment', 'selfright', 'suv', 'promot', 'anim', 'carel', 'hypothet', 'mark', 'piti', 'comic', 'relief', 'wanda', 'syke', 'frequent', 'absolut', 'bruce', 'almighti', 'evan', 'blow', 'recordbreak', 'budget', 'advanc', 'build', 'construct', 'ark', 'learn', 'vessel', 'gag', 'conclud', 'flood', 'gather', 'statement', 'influenc', 'heath', 'ledger', 'heartthrob', 'deform', 'naomi', 'watt', 'item', 'spent', 'longest', 'orlando', 'bloom', 'scraggli', 'beard', 'deerintheheadlight', 'agre', 'rachel', 'griffith', 'fabul', 'geoffrei', 'sorri', 'bankrobb', 'butch', 'cassidi', 'clicheridden', 'hilari', 'frontier', 'hotel', 'dy', 'presenc', 'armor', 'knight', 'monti', 'python', 'holi', 'grail', 'bite', 'yer', 'leg', 'howl', 'laughter', 'warp', 'paid', 'disast', 'sneak', 'preview', 'certainli', 'free', 'cost', 'leonard', 'maltin', 'dread', 'bomb', 'vengeanc', 'mount', 'beauti', 'photograph', 'color', 'open', 'grab', 'attent', 'tip', 'welldon', 'satir', 'spaghetti', 'homag', 'fairbank', 'strip', 'famou', 'showdown', 'ugli', 'edd', 'byrn', 'hilton', 'gilbert', 'roland', 'brilliantli', 'poorli', 'undeserv', 'miss', 'meant', 'kudo', 'belong', 'ludicr', 'angel', 'yearold', 'annakin', 'whini', 'brat', 'somehow', 'amidala', 'senior', 'jedi', 'warrior', 'hero', 'slaughter', 'framework', 'exist', 'crazi', 'preciou', 'b', 'ludicros', 'squar', 'pai', 'unbeliev', 'pervert', 'obiwan', 'kenobi', 'kind', 'anchor', 'seri', 'climax', 'lava', 'suffer', 'anyon', 'plausibl', 'motiv', 'oh', 'yeah', 'cgi', 'cool', 'wwii', 'british', 'latterdai', 'peer', 'respectfulli', 'confluenc', 'near', 'dive', 'descend', 'admir', 'horatio', 'nelson', 'student', 'aspect', 'warfar', 'favor', 'depict', 'sub', 'north', 'atlant', 'unacquaint', 'target', 'prioriti', 'warship', 'event', 'submarin', 'opportun', 'similarli', 'deliber', 'typic', 'eal', 'rank', 'britishgaumont', 'frankli', 'prefer', 'quieter', 'cerebr', 'approach', 'overproduc', 'powel', 'pressburg', 'parallel', 'thank', 'powerfulli', 'persuas', 'eric', 'portman', 'mine', 'mill', 'smaller', 'font', 'shine', 'utterli', 'gainsborough', 'precgi', 'festiv', 'night', 'shorter', 'mcnalli', 'summari', 'content', 'nossit', 'deftli', 'blend', 'wine', 'wider', 'globalis', 'homogenis', 'mass', 'media', 'capit', 'divers', 'handheld', 'dv', 'offput', 'catch', 'occas', 'possibl', 'equip', 'sprawl', 'sharp', 'parad', 'dog', 'undercut', 'interviewe', 'comment', 'contradictori', 'suffici', 'rope', 'themselv', 'degre', 'evok', 'moor', 'recent', 'oper', 'root', 'marcel', 'ophul', 'sorrow', 'earth', 'peasant', 'montil', 'pere', 'et', 'fil', 'lff', 'answer', 'afterward', 'disord', 'bravo', 'hubert', 'excel', 'implic', 'ourselv', 'organis', 'funnyy', 'ground', 'weeksdespit', 'storyabout', 'robot', 'famili', 'pizza', 'kurt', 'edison', 'finger', 'threw', 'rememberto', 'cousin', 'edis', 'therejust', 'fought', 'worship', 'starewicz', 'strangest', 'short', 'toi', 'search', 'orang', 'protect', 'devilish', 'nightclub', 'featur', 'scari', 'stuf', 'cat', 'scurvi', 'mascot', 'sync', 'mixtur', 'stopmot', 'puppetri', 'feet', 'puppet', 'concret', 'sidewalk', 'honk', 'cri', 'vendor', 'shift', 'costli', 'util', 'technolog', 'hat', 'contribut', 'club', 'twigss', 'newspap', 'shred', 'skeleton', 'bird', 'lai',

'hatch', 'chick', 'pan', 'rock', 'zoom', 'accomplish', 'speed', 'filmmak', 'demand', 'mutil', 'selfmutil', 'abus', 'childhood', 'andi', 'copp', 'commentari', 'avail', 'prior', 'wont', 'clue', 'gorehound', 'lure', 'promis', 'harsh', 'splatter', 'unsettl', 'reallif', 'footag', 'unless', 'pretenti', 'headacheinduc', 'theyll', 'chore', 'imageri', 'accompani', 'dischord', 'incomprehens', 'mindnumbingli', 'drivel', 'test', 'saniti', 'marbl', 'rubbish', 'awar', 'bark', 'ladi', 'estat', 'garden', 'stumbl', 'own', 'smart', 'pollut', 'evid', 'commit', 'caus', 'miracul', 'sabotag', 'hire', 'yai', 'confront', 'backup', 'danger', 'rosemari', 'thyme', 'insignific', 'lowbrain', 'haha', 'thirteen', 'consid', 'amus', 'mind', 'tacki', 'atroci', 'paragraph', 'brain', 'quietamerican', 'adolesc', 'diner', 'coolkid', 'school', 'cherri', 'tvguru', 'blake', 'ador', 'petbrain', 'nationwid', 'label', 'independ', 'thinker', 'giant', 'cheesi', 'wave', 'meadowval', 'teenrebel', 'funniest', 'monstrou', 'extraterrestri', 'meh', 'background', 'unfortun', 'overs', 'sockpuppet', 'buff', 'crew', 'hunt', 'barri', 'pearson', 'birthdai', 'guilti', 'pleasur', 'plagu', 'resign', 'industri', 'unquestion', 'david', 'gale', 'forev', 'reanim', 'christin', 'kossak', 'nudityfactor', 'she', 'repertoir', 'debut', 'runawai', 'model', 'babi', 'flick', 'doppelgang', 'astronaut', 'crash', 'counterearth', 'opposit', 'sun', 'cold', 'totalitarian', 'vibe', 'pilot', 'sank', 'trace', 'perfectli', 'adequ', 'cameron', 'mitchel', 'glenn', 'corbett', 'saxon', 'individualist', 'pose', 'threat', 'foundat', 'counter', 'energi', 'design', 'firmli', 'tether', 'launch', 'pad', 'templat', 'dopplegang', 'vehicl', 'ship', 'appal', 'stagger', 'helplessli', 'smoke', 'wind', 'smack', 'overcom', 'faceless', 'yell', 'loudspeak', 'strand', 'seat', 'buzzer', 'blur', 'blackout', 'twilight', 'outer', 'took', 'madefortv', 'essenc', 'imaginationjust', 'token', 'gestur', 'boi', 'enthusiasm', 'cv', 'press', 'recal', 'regardless', 'public', 'consumpt', 'group', 'camcord', 'septemb', 'attend', 'jon', 'satejowski', 'donnybrook', 'rough', 'safe', 'compet', 'push', 'record', 'static', 'unimpress', 'reduc', 'modestli', 'grant', 'shown', 'engross', 'steven', 'soderbergh', 'videotap', 'immedi', 'unfocus', 'mend', 'aspir', 'n', 'roll', 'impromptu', 'casual', 'strum', 'guitar', 'told', 'gig', 'mode', 'creativ', 'glam', 'era', 'hed', 'earli', 'dean', 'meantim', 'random', 'thread', 'slim', 'abandon', 'alarm', 'frequenc', 'ie', 'subplot', 'terri', 'haphazard', 'napolean', 'dynamit', 'ask', 'speak', 'choppi', 'collect', 'insight', 'anger', 'grunt', 'overthetop', 'outburst', 'behav', 'normal', 'ration', 'easi', 'al', 'hudson', 'do', 'imit', 'elliott', 'spot', 'correct', 'unwil', 'necessari', 'fix', 'itll', 'vision', 'unabl', 'critic', 'selfindulg', 'misguid', 'rob', 'schneider', 'highschoolset', 'knockoff', 'zach', 'braff', 'combin', 'equal', 'velvet', 'goldmin', 'worthi', 'task', 'compliment', 'ambit', 'distribut', 'delud', 'grandeur', 'hold', 'associ', 'pleas', 'file', 'lesson', 'liken', 'iii', 'undoubtedli', 'violent', 'downbeat', 'pousoi', 'cheang', 'sleazi', 'lurid', 'sensationalist', 'earn', 'pack', 'gritti', 'hardedg', 'chen', 'pang', 'cambodian', 'hitman', 'hong', 'kong', 'assassin', 'lee', 'ruthless', 'determin', 'whatev', 'ensur', 'escapeuntil', 'yue', 'illeg', 'immigr', 'escap', 'relentlessli', 'score', 'newcom', 'pei', 'hardhit', 'asian', 'hyperviol', 'chanwook', 'park', 'trilogi', 'stab', 'merciless', 'regularli', 'caught', 'unflinchingli', 'destin', 'unhappi', 'tragedi', 'unintent', 'moment', 'cross', 'overdramat', 'lock', 'pregnant', 'fraca', 'dii', 'ceasarean', 'slice', 'excess', 'deliri', 'ott', 'stylish', 'seek', 'h', 'maci', 'fargo', 'hasnt', 'deceiv', 'archetyp', 'hide', 'lawrenc', 'newman', 'loyal', 'hardwork', 'stiff', 'harbour', 'handicap', 'mccarthyism', 'resourc', 'accident', 'descent', 'pair', 'improv', 'eyesight', 'simpl', 'repercuss', 'gertrud', 'hart', 'laura', 'dern', 'unravel', 'akin', 'disturb', 'racist', 'fairli', 'chill', 'guardian', 'oliv', 'mindset', 'thoroughli', 'dampen', 'trailer', 'water', 'flashback', 'kevin', 'costner', 'subject', 'waterworld', 'ashton', 'kutcher', 'butterfli', 'simian', 'approxim', 'fear', 'subsid', 'hesit', 'slip', 'throw', 'tens', 'mission', 'tighter', 'kenni', 'roger', 'briefli', 'bristl', 'initi', 'sunglass', 'tough',

'toothpick', 'sportin', 'smirk', 'thatd', 'cloonei', 'proud', 'strong', 'soften', 'jab', 'darn', 'ap', 'effici', 'mix', 'excit', 'humor', 'easili', 'quibbl', 'overcook', 'sappi', 'train', 'contrari', 'session', 'coast', 'swimmer', 'inform', 'introduc', 'underappreci', 'disorient', 'exhaust', 'hypothermia', 'oxygen', 'panick', 'decis', 'kutcherkevin', 'tribut', 'breed', 'custom', 'gist', 'moviego', 'embark', 'ocean', 'wouldv', 'gotten', 'biggest', 'principl', 'sara', 'selfsacrif', 'minchin', 'punish', 'vulgar', 'disneyfi', 'trot', 'indian', 'improb', 'puhleez', 'heartno', 'friendship', 'becki', 'adopt', 'lotti', 'variou', 'core', 'flung', 'poverti', 'suddenli', 'younger', 'feelgood', 'shame', 'theyd', 'harp', 'alex', 'sander', 'sic', 'select', 'desecr', 'dramat', 'horrorscifi', 'predat', 'scifiact', 'messabout', 'blame', 'versu', 'grade', 'connoisseur', 'site', 'ridden', 'ruin', 'overrunnot', 'overrun', 'ok', 'went', 'result', 'alienpred', 'hybrid', 'realis', 'sooner', 'coher', 'strictli', 'suspend', 'door', 'damag', 'calcul', 'thick', 'woodland', 'alon', 'hug', 'hugger', 'movement', 'irrespons', 'redneck', 'muppet', 'edgi', 'thrillertyp', 'excon', 'met', 'emotionless', 'dull', 'bu', 'feebl', 'crap', 'slasherhorror', 'supposedli', 'nerdi', 'undesir', 'beaten', 'jock', 'american', 'sportsman', 'scottish', 'cutenot', 'clever', 'avoid', 'shite', 'revers', 'pc', 'riplei', 'credenti', 'introduct', 'husband', 'child', 'smile', 'demis', 'storytel', 'brightest', 'alienridden', 'poster', 'post', 'late', 'predatoralien', 'disgust', 'alright', 'scare', 'titil', 'pervers', 'shag', 'sauci', 'teenag', 'depth', 'disinterest', 'tarnish', 'gain', 'worri', 'ps', 'amateur', 'civilian', 'overlong', 'swiss', 'cowrit', 'vinegar', 'oil', 'boot', 'addition', 'realist', 'liberti', 'legal', 'system', 'citizen', 'eas', 'abscond', 'fast', 'speech', 'exlov', 'worn', 'preposter', 'lover', 'nutti', 'divert', 'revert', 'ismael', 'primarili', 'nora', 'detriment', 'viewpoint', 'psychiatr', 'epilogu', 'equival', 'snow', 'hell', 'glare', 'fault', 'pepe', 'le', 'moko', 'charl', 'boyer', 'crimin', 'mastermind', 'casbah', 'algier', 'reluct', 'pari', 'custodi', 'arrest', 'oscar', 'classi', 'letdown', 'local', 'command', 'commission', 'janvier', 'inspector', 'sliman', 'joseph', 'calleia', 'difficult', 'harden', 'swept', 'gabi', 'hedi', 'lamarr', 'populac', 'ironi', 'freedom', 'imposs', 'trap', 'lesser', 'extent', 'bowl', 'heartfelt', 'karmic', 'eman', 'ordinari', 'commut', 'inspir', 'protagonist', 'assort', 'smooth', 'floor', 'prologu', 'japan', 'symbol', 'sociolog', 'sight', 'dignifi', 'diminish', 'imperfect', 'lent', 'synchron', 'geek', 'teen', 'shall', 'asi', 'linnett', 'connor', 'st', 'molli', 'sign', 'contract', 'stipul', 'silent', 'emphasi', 'don', 'amech', 'alic', 'fai', 'notch', 'tune', 'extra', 'restor', 'emphas', 'singin', 'dick', 'van', 'dykecarl', 'reiner', 'bygon', 'eraalthough', 'includedin', 'gloriou', 'bw', 'pristin', 'buster', 'keaton', 'uphil', 'superbl', 'input', 'histori', 'cavalcad', 'mack', 'sennett', 'legaci', 'posit', 'rework', 'mabel', 'normand', 'sinnott', 'pie', 'bath', 'keyston', 'turpincameo', 'rosco', 'fatti', 'arbucklebodi', 'pastur', 'categori', 'fortun', 'instrument', 'evolut', 'semi', 'retir', 'charli', 'chaplin', 'swanson', 'bing', 'crosbi', 'wc', 'harri', 'langdon', 'arbuckl', 'roi', 'rogersin', 'thara', 'path', 'lifetim', 'eg', 'coedswher', 'oyster', 'soup', 'williesi', 'road', 'nostalgia', 'lane', 'particip', 'knack', 'guest', 'etern', 'welk', 'radio', 'rereleas', 'triumph', 'tilli', 'punctur', 'honor', 'autobiographi', 'companion', 'publish', 'abbott', 'costello', 'cameo', 'ac', 'kop', 'compil', 'pioneer', 'banquet', 'tabl', 'silver', 'hair', 'rise', 'subdu', 'mental', 'individu', 'owen', 'mammaonli', 'six', 'laughoutloud', 'devito', 'billi', 'crystal', 'ann', 'ramsei', 'momma', 'shallow', 'tack', 'gutsplit', 'ride', 'emploi', 'chemistri', 'separ', 'avid', 'untal', 'block', 'professor', 'donner', 'exwif', 'stole', 'teach', 'bud', 'middleag', 'clancytyp', 'captain', 'upholsteri', 'salesman', 'pinski', 'laughsperminutesonscreen', 'ascotwear', 'weirdo', 'literatur', 'pork', 'pupil', 'teacher', 'unusu', 'manchild', 'overbear', 'lardass', 'endear', 'sentiment', 'cruelti', 'onto', 'identifi', 'lark', 'gambit', 'enlist', 'plan', 'hawaii', 'ex', 'swap', 'hitchcock',

'complement', 'sceneset', 'accent', 'bold', 'unreal', 'toe', 'stu', 'quotabl', 'patter', 'size', 'oldsmobil', 'womansh', 'ripper', 'context', 'wrote', 'entireti', 'theatric', 'attract', 'patti', 'panicki', 'slam', 'forehead', 'groan', 'loonei', 'directori', 'pathet', 'scienc', 'justifi', 'religionphilosophi', 'quantum', 'physic', 'plain', 'theori', 'support', 'eastern', 'religion', 'illus', 'subatom', 'definit', 'particl', 'affect', 'pass', 'barrier', 'accur', 'assert', 'wall', 'poppycock', 'belief', 'remotest', 'astronom', 'misrepres', 'marle', 'maitlan', 'neg', 'selfhat', 'truth', 'string', 'nowadai', 'dreck', 'rent', 'eleg', 'nova', 'pb', 'newton', 'witten', 'm', 'hourlong', 'mechan', 'fog', 'metaphys', 'rariti', 'popcorn', 'bottl', 'vodka', 'honest', 'mo', 'ogrodnik', 'cinemat', 'wolfgang', 'photographi', 'collabor', 'reveal', 'peen', 'beetl', 'bang', 'violet', 'butt', 'porn', 'subsequ', 'smash', 'awesom', 'nyu', 'uptown', 'hammer', 'skull', 'bottom', 'straight', 'sheer', 'amount', 'dido', 'thirti', 'pacifi', 'elephantther', 'secretli', 'underag', 'cobain', 'lookalik', 'store', 'ohsonatur', 'heartshap', 'delightfulli', 'deriv', 'lolita', 'chinup', 'pornograph', 'intellectu', 'contempl', 'parti', 'ogrodnikrip', 'pornographi', 'auteur', 'somebodi', 'refillpe', 'pee', 'sobrieti', 'bodi', 'sip', 'underus', 'hulahoop', 'magic', 'throughlin', 'beyotch', 'whenev', 'internet', 'ripe', 'hawaiian', 'smirnoff', 'perpetu', 'azumi', 'fruition', 'ultim', 'deeper', 'rampag', 'mercilessli', 'saga', 'poetic', 'heavi', 'dagger', 'credibl', 'aya', 'ueto', 'blooddriven', 'previous', 'shortcom', 'sluggish', 'uninspir', 'omit', 'stealth', 'tactic', 'logic', 'swift', 'quicker', 'slightli', 'tweak', 'obstacl', 'plant', 'warlord', 'displai', 'charisma', 'chiaki', 'foolishli', 'shelv', 'minu', 'disguis', 'epic', 'adversari', 'flashi', 'flashier', 'overpow', 'satisfi', 'bijomaru', 'achiev', 'knock', 'samurai', 'rule', 'broken', 'enrich', 'resolut', 'wide', 'varieti', 'lie', 'neutral', 'flawless', 'richer', 'expand', 'bumpi', 'journei', 'desper', 'peac', 'unwav', 'courag', 'shakespear', 'four', 'jacobi', 'hamlet', 'patrick', 'stewart', 'claudiu', 'ceas', 'clear', 'length', 'bard', 'tinker', 'structur', 'olivi', 'concentr', 'indecis', 'gibson', 'ambiti', 'upset', 'uncl', 'usurp', 'throne', 'furiou', 'royal', 'kin', 'solidifi', 'claim', 'wittenburg', 'impot', 'damnat', 'overreach', 'form', 'hubri', 'undo', 'poloniu', 'rosencrantz', 'guildenstern', 'ophelia', 'laert', 'thiev', 'highwai', 'git', 'ghastli', 'firstli', 'warmer', 'borrow', 'walker', 'healthi', 'occasion', 'whimper', 'hooker', 'palat', 'nearli', 'beg', 'chicken', 'adorn', 'chef', 'cook', 'bail', 'consequ', 'indulg', 'secondli', 'lewi', 'gere', 'seduc', 'yep', 'moron', 'suit', 'lotu', 'esprit', 'turbo', 'necklac', 'diamond', 'limo', 'where', 'piano', 'shop', 'eddi', 'importantli', 'like', 'hangin', 'round', 'merci', 'pardon', 'sell', 'meand', 'youill', 'woke', 'wake', 'brave', 'slumber', 'bank', 'teller', 'provinci', 'stunningli', 'gorgeou', 'wild', 'girltodiefor', 'nadia', 'nicol', 'kidman', 'email', 'russia', 'belov', 'glitch', 'calm', 'refund', 'polici', 'servic', 'dictionari', 'commun', 'process', 'henceforth', 'decentlypaid', 'safefeel', 'clerk', 'decisionmak', 'sublimin', 'likewis', 'cassel', 'kassovitz', 'team', 'russian', 'indistinguish', 'slight', 'nativ', 'cultur', 'captiv', 'soon', 'advic', 'below', 'rude', 'awaken', 'artifici', 'routin', 'wheel', 'keen', 'pink', 'floyd', 'welcom', 'bet', 'sophia', 'whereev', 'authent', 'percent', 'rudimentari', 'helen', 'curti', 'harrington', 'whoever', 'slew', 'aunti', 'roo', 'scriptwrit', 'farrel', 'hushhushsweet', 'charlott', 'charg', 'becam', 'enchant', 'horrif', 'swing', 'tapdanc', 'catapult', 'sublim', 'wardrob', 'lifeimprison', 'adel', 'debbi', 'reynold', 'shellei', 'winter', 'flee', 'california', 'herself', 'millionair', 'sink', 'downward', 'insan', 'rambl', 'radioevangelist', 'slowli', 'tendenc', 'flaw', 'tango', 'unrel', 'omin', 'benefic', 'downright', 'unpredict', 'introvert', 'snap', 'petrifi', 'freddi', 'krueger', 'jason', 'voorhe', 'myer', 'underdevelop', 'michal', 'macliammir', 'cocki', 'elocut', 'agn',

'moorehead', 'priestess', 'timothi', 'carei', 'obtrus', 'visitor', 'gruesom', 'grand', 'guignol', 'fanat', 'breathtak']

Pr.4.2

Using the feature vector generated in the first task, write a program that generates the Term Document Matrix (TDM) for ALL of the paragraphs in "Project4_paragraphs.txt", similar to TDM below. (5 pts)

a) **Provide the TDM in your report. (3 pts)**

-Preview

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Keyword Set	on	review	mention	watch	oz	episod	youll	hook	right	exactli	happen	first	thing	struck	brutal	unflinch	scene	violenc	set	word	go
2	Paragraph 1	1	1	1	3	5	2	1	1	2	1	1	2	1	2	1	1	1	4	1	2	1
3	Paragraph 2	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
4	Paragraph 3	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
5	Paragraph 4	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
6	Paragraph 5	6	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
7	Paragraph 6	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
8	Paragraph 7	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
9	Paragraph 8	1	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
10	Paragraph 9	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	2
11	Paragraph 10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	Paragraph 11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
13	Paragraph 12	1	0	1	0	0	0	0	0	2	0	0	2	1	0	0	0	2	0	1	0	4
14	Paragraph 13	4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3	0	0	0	0
15	Paragraph 14	3	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
16	Paragraph 15	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
17	Paragraph 16	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1
18	Paragraph 17	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Paragraph 18	2	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	3	0
20	Paragraph 19	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
21	Paragraph 20	0	0	1	2	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	2	3

-Excel File Attached

b) **For each of the text mining steps (A to H), explain the purpose of each step and what sort of information is lost while applying each of those text-mining steps. (2 pts)**

Tokenize paragraphs: Addresses each sentence in a paragraph separately and obtain per paragraphs metrics. Contextual information that spans multiple sentences or even paragraphs is generally lost.

Remove Punctuation: Helps focus more strictly on words. We lose some grammatical information of the sentences, but note that it turns into noise if taken out of context.

Remove Numbers: Just like punctuations, this helps the program focus on words. Numbers have no meaning when taken out of context though we could possibly lose some precision that would be useful in information retrieval.

Convert upper to lowercase: Removes extraneous detail that has no useful information out of context. This information can be useful for noun detection.

Remove stop words: Removes words that are extremely common in the English language and therefore have no meaning when taken out of context.

Perform Stemming: Reduces words down to their base form which carry the most useful information. This allows us to avoid alternate versions of the same words skewing any learning. The information lost here would be useful for things such as tense recognition.

Combining Stemmed Words: We take the stemmed words and combine them back into a pseudo-sentence for further processing and analytics among the sentences, to build the term document matrix. No information loss.

Extract the most frequent words: Readers benefit from keywords because they can judge more quickly whether the text is worth reading. Or grouping similar content is possible too by topics. Reduces the dimensionality of text to the most important features.

Pr.4.3

Write a program implementing the clustering algorithm of your choice (Kohonen WTA or FCAN). Apply that algorithm to TDM to group similar paragraphs together. (6 pts)

FCAN Chosen

a) **How many clusters/topics have you identified? (2 pts)**

8

b) **What drives the dimensionality of TDM? What can you do to reduce that dimensionality? Does the order of data being fed to the algorithm matter? (2 pts)**

The dimensionality of the TDM is determined by the number of words included in the feature vector. We can reduce the dimensionality by reducing the number of words considered in TDM. For example, the removal of stopwords reduces the dimensionality of the TDM. We could potentially also remove words that are found in all documents an equal amount of times or other tweaks like that to remove dimensionality.

For the FCAN algorithm, the order of the patterns fed does matter. As the order determines the initial cluster locations.

c) **Show and comment on the results. (2 pts)**

With Max Radius 40 and Alpha of 6:

Grouped Paragraphs (index off by one, EX: 0 indicates Paragraph 1 in TDM)

[0, 1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 19, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 48, 51]

[8, 10]

[15, 16]
[17, 18]
[20, 21, 47]
[45, 46]
[49, 50]
[52]

It seems that with the above parameters, most of the paragraphs fell under a single cluster. However, the other 14 paragraphs could be divided into 7 other clusters. This is expected as while looking at the paragraphs as a human, you can tell that they are all in the same domain and have similar styles and vocabularies since they are all movie reviews. It makes sense that most fall under a single bucket