1. Choosing the data: either a) Choose existing large documents from NLTK or from the Gutenberg collection on the web, or b) [Counts as additional task] Collect your own data, by using your own documents or collecting data from other sources. Combine the text from these sources to make two documents for the corpora for the first task. Describe the method that you used to define and collect the data, including the difference between the documents. Note any limitations to the method or the text that you were able to find. Do preprocessing to get the text in a suitable format for processing and describe what you did.

=>

I am comparing two Sports Rule Book i.e. “Cricket Game Rule Book” and “Long tennis Rule Book”. My goal is to choose these sports rule as my corpora to study and compare two different type of rules which sports man have to follow while playing different sports. How sports rule deals with fitness of Players? And many more.

Data Collection from Sources:

- Downloaded latest rule book pdf files from internet and converted them to text files for using them in Python.
- Omitted front cover page, back cover page, all pictures.
- Omitted part or chapter numbers in between text and page numbers if any and any special design between or at the top or bottom of text and acknowledgement and preface pages.
- “CricketRule.txt” contains Cricket Game Rule
- “LongTennisRuleBook.txt” contains Long tennis Rule

Differences between Data:

Cricket Game Rule Book contains rule for playing Cricket game where as Long tennis Rule Book does the same for Long Tennis. Both of these rule book contains different key words related to specific sports, their emphasis on specific fields etc.

Processing of Text File:

- Used Regression Expression to remove numbers from rule books. Like re.sub("\d+", " " , filetext)
- Removing punctuations and tokenising the file by using Removed Cricket Country list from cricket rule book by writing a function to remove from start line to end line
  i.e. punctuation_tokenizer = RegexpTokenizer(r'\w+')

.

2. [Required task] Examine the text in the documents that you chose and decide how to process the words, i.e. decide on tokenization and whether to use all lower case, stopwords or lemmatization. Using the process developed in the lab, • list the top 50 words by frequency • list the top 50 bigrams by frequencies, and • list the top 50 bigrams by their Mutual Information scores. Note that you may wish to modify the stop word list, based on your question in Task 3. To complete this part: a) Briefly state why you chose the processing options that you did. b) Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams? How are the bigram frequency list and the bigram Mutual Information lists different?

=>

List the top 50 words by frequency [CricketRuleBook] :

shall	3986
law	1937
match	1870
icc	1396
umpire	1323
â	1124
time	1088
play	959
overs	933
player	933
apply	860
ball	859
playing	856
following	762
umpires	711
team	696
innings	657
code	655
may	625
b	563
field	554
cricket	541
conditions	537
one	530
referee	502
international	495
side	458
decision	450
bowler	433

third	432
pitch	420
conduct	414
number	398
subject	393
matches	368
day	362
anti	344
captain	344
bowling	343
article	329
support	328
minimum	327
offence	323
batting	321
personnel	321
players	316
ground	308
event	298
hearing	294
c	288

List the top 50 bigrams by frequencies without applying filter [CricketRuleBook]:

(('of',	'the'),	0.014787883957224952)
(('shall',	'be'),	0.008449457848666745)
(('to',	'the'),	0.007639162837311953)
(('in',	'the'),	0.004792468521104987)
(('the',	'icc'),	0.004360666574264604)
(('by',	'the'),	0.004307357691938631)
(('the',	'following'),	0.003582356892305396)
(('the',	'match'),	0.0033264742571407248)
(('shall',	'apply'),	0.003310481592442933)
(('law',	'shall'),	0.0031079078396042347)
(('on',	'the'),	0.003059929845510859)
(('at',	'the'),	0.002947981192626315)
(('for',	'the'),	0.002889341422067745)
(('with',	'the'),	0.002777392769183201)
(('the',	'umpires'),	0.0026174661222052816)

(('match',	'referee'),	0.0025801499045771006)
(('and',	'the'),	0.0025161792457859327)
(('to',	'be'),	0.0023082746047146377)
(('playing',	'conditions'),	0.0021590097342019127)
(('shall',	'not'),	0.002116362628341134)
(('that',	'the'),	0.0020897081871781476)
(('third',	'umpire'),	0.0020683846342477584)
(('if',	'the'),	0.0020257375283869797)
(('of',	'a'),	0.0019990830872239932)
(('will',	'be'),	0.0019830904225262014)
(('the',	'umpire'),	0.001972428646061007)
(('subject',	'to'),	0.0019191197637350335)
(('the',	'third'),	0.001908457987269839)
(('icc',	'match'),	0.0018604799931764631)
(('of',	'conduct'),	0.0018018402226178925)
(('the',	'ball'),	0.0017645240049897113)
(('code',	'of'),	0.001759193116757114)
(('the',	'player'),	0.0017218768991289328)
(('umpire',	'shall'),	0.0016845606815007518)
(('number',	'of'),	0.001673898905035557)
(('of',	'overs'),	0.001657906240337765)
(('of',	'play'),	0.0016525753521051678)
(('player',	'support'),	0.0016259209109421812)
(('has',	'been'),	0.0015939355815465972)
(('support',	'personnel'),	0.001556619363918416)
(('the',	'pitch'),	0.0015459575874532214)
(('the',	'field'),	0.001487317816894651)
(('international',	'match'),	0.0014819869286620537)
(('is',	'not'),	0.0013913618287078992)
(('on',	'field'),	0.0013753691640101073)
(('in',	'addition'),	0.00137003827577751)
(('from',	'the'),	0.0013540456110797182)
(('be',	'replaced'),	0.0013273911699167315)
(('of',	'any'),	0.0013220602816841343)
(('under',	'the'),	0.0012687513993581612)

List the top 50 bigrams by frequencies by applying filter[CricketRuleBook]:

(('shall',	'apply'),	0.003310481592442933)
(('law',	'shall'),	0.0031079078396042347)

(('match',	'referee'),	0.0025801499045771006)
(('playing',	'conditions'),	0.0021590097342019127)
(('third',	'umpire'),	0.0020683846342477584)
(('icc',	'match'),	0.0018604799931764631)
(('umpire',	'shall'),	0.0016845606815007518)
(('player',	'support'),	0.0016259209109421812)
(('support',	'personnel'),	0.001556619363918416)
(('international',	'match'),	0.0014819869286620537)
(('playing',	'time'),	0.0011034938641476444)
(('apply',	'law'),	0.0010875011994498523)
(('anti',	'corruption'),	0.0010661776465194631)
(('home',	'board'),	0.0009542289936349195)
(('icc',	'code'),	0.0009329054407045302)
(('field',	'umpire'),	0.0009009201113089463)
(('match',	'playing'),	0.0009009201113089463)
(('judicial',	'commissioner'),	0.0008582730054481677)
(('umpires',	'shall'),	0.0008209567878199865)
(('apply',	'subject'),	0.000794302346657)
(('one',	'day'),	0.0007250007996332349)
(('corruption',	'code'),	0.0007143390231680402)
(('following',	'shall'),	0.0006716919173072617)
(('anti',	'racism'),	0.0006450374761442752)
(('powerplay',	'overs'),	0.0006397065879116779)
(('sq',	'inches'),	0.0006290448114464832)
(('day',	'international'),	0.0006237139232138859)
(('racism',	'code'),	0.0006130521467486912)
(('shall',	'call'),	0.0006130521467486912)
(('team',	'batting'),	0.0005970594820508994)
(('test',	'match'),	0.000591728593818302)
(('batting',	'side'),	0.0005810668173531074)
(('inches',	'cm\xe2\xb2'),	0.0005704050408879127)
(('extra',	'time'),	0.0005437505997249261)
(('short',	'pitched'),	0.0005277579350271342)
(('fielding',	'side'),	0.0005224270467945369)
(('national',	'cricket'),	0.0005224270467945369)
(('shall',	'take'),	0.0005224270467945369)
(('overs',	'per'),	0.0005170961585619396)
(('conditions',	'law'),	0.0005117652703293423)
(('fast',	'short'),	0.0004797799409337584)
(('icc',	'world'),	0.0004797799409337584)
(('time',	'lost'),	0.00046911816446856374)
(('batting',	'second'),	0.0004584563880033691)

(('law',	'b'),	0.00043713283507297986)
(('player',	'review'),	0.00043713283507297986)
(('alleged',	'offence'),	0.00042647105860778523)
(('cricket',	'league'),	0.00042647105860778523)
(('bowling',	'action'),	0.0004158092821425906)
(('take',	'place'),	0.00041047839390999326)

List the top 50 bigrams by their Mutual Information scores without applying filter [CricketRuleBook]

(('adduction\xe2',	'curves'),	17.517192634463576)
(('adrian',	'griffith'),	17.517192634463576)
(('afghanistan',	'austria'),	17.517192634463576)
(('ahm',	'mustafa'),	17.517192634463576)
(('aids',	'abets'),	17.517192634463576)
(('al',	'journal'),	17.517192634463576)
(('ambush',	'marketing'),	17.517192634463576)
(('animation',	'data\xe2'),	17.517192634463576)
(('appealpursuanttoarticle',	'shallnot'),	17.517192634463576)
(('arabia',	'seychelles'),	17.517192634463576)
(('argentina',	'belgium'),	17.517192634463576)
(('assists',	'encourages'),	17.517192634463576)
(('austria',	'bahamas'),	17.517192634463576)
(('avenue',	'benoni'),	17.517192634463576)
(('awritten',	'reasoneddecisioninaccordance'),	17.517192634463576)
(('bahamas',	'bahrain'),	17.517192634463576)
(('bahrain',	'belize'),	17.517192634463576)
(('bandaraya',	'kuala'),	17.517192634463576)
(('bandula',	'warnapura'),	17.517192634463576)
(('belize',	'bhutan'),	17.517192634463576)
(('bhutan',	'brazil'),	17.517192634463576)
(('brazil',	'brunei'),	17.517192634463576)
(('brdb',	'jalan'),	17.517192634463576)
(('brunei',	'bulgaria'),	17.517192634463576)
(('bukit',	'bandaraya'),	17.517192634463576)
(('bulgaria',	'cameroon'),	17.517192634463576)
(('cameroon',	'chile'),	17.517192634463576)
(('campbell',	'jamieson'),	17.517192634463576)
(('cassim',	'suliman'),	17.517192634463576)
(('chile',	'china'),	17.517192634463576)

(('china',	'cook'),	17.517192634463576)
(('chris',	'moller'),	17.517192634463576)
(('clive',	'hitchcock'),	17.517192634463576)
(('clock',	'tower'),	17.517192634463576)
(('closely',	'reflects'),	17.517192634463576)
(('companieswhich',	'shallremain'),	17.517192634463576)
(('compensatory',	'awards'),	17.517192634463576)
(('concord',	'gate'),	17.517192634463576)
(('costa',	'rica'),	17.517192634463576)
(('crest',	'mascot'),	17.517192634463576)
(('croatia',	'cuba'),	17.517192634463576)
(('cross',	'validated'),	17.517192634463576)
(('cuba',	'cyprus'),	17.517192634463576)
(('custom',	'built'),	17.517192634463576)
(('cyprus',	'czech'),	17.517192634463576)
(('czech',	'republic'),	17.517192634463576)
(('denmark',	'fiji'),	17.517192634463576)
(('der',	'bijl'),	17.517192634463576)
(('devmanager',	'asiancricket'),	17.517192634463576)
(('doors',	'mirrors'),	17.517192634463576)

List the top 50 bigrams by their Mutual Information scores with applying filter [CricketRuleBook]

(('champions',	'trophy'),	14.932230133742419)
(('disc',	'measuring'),	14.932230133742419)
(('metal',	'disc'),	14.932230133742419)
(('sight',	'screens'),	14.932230133742419)
(('de',	'novo'),	14.932230133742415)
(('invalid',	'unenforceable'),	14.932230133742415)
(('clean',	'catches'),	14.517192634463576)
(('uniformly',	'calibrated'),	14.517192634463576)
(('continuous',	'painted'),	14.517192634463573)
(('common',	'sense'),	14.347267633021264)
(('dimension',	'smaller'),	14.347267633021264)
(('news',	'crews'),	14.347267633021264)
(('comfort',	'break\xe2'),	14.34726763302126)
(('external',	'blow'),	14.34726763302126)
(('irregularly',	'inspect'),	14.34726763302126)
(('paramount',	'importance'),	14.34726763302126)
(('personally',	'informs'),	14.34726763302126)
(('pulled',	'muscle'),	14.34726763302126)

(('spiked',	'footwear'),	14.34726763302126)
(('decimal',	'places'),	14.34726763302126)
(('initially',	'consulted'),	14.294800213127125)
(('south',	'africa'),	14.294800213127125)
(('travel',	'accommodation'),	14.224410885235727)
(('anticipated',	'testimony'),	14.195264539576211)
(('absolute',	'impartiality'),	14.195264539576211)
(('functions',	'assigned'),	14.195264539576211)
(('ignore',	'fractions'),	14.195264539576211)
(('originally',	'allotted'),	14.10215513518473)
(('landmark',	'achievement'),	13.932230133742419)
(('long',	'sleeved'),	13.932230133742419)
(('negative',	'tactic'),	13.932230133742419)
(('sound',	'link'),	13.932230133742419)
(('varied',	'hereunder'),	13.932230133742415)
(('remove',	'water'),	13.905757922381225)
(('satisfactory',	'passing'),	13.905757922381225)
(('thereby',	'obstructed'),	13.81675291632248)
(('perceived',	'comes'),	13.780227040297365)
(('heat',	'transfer'),	13.709837712405973)
(('correctly',	'reflect'),	13.709837712405971)
(('layout',	'basic'),	13.709837712405971)
(('communication',	'devices'),	13.709837712405967)
(('sense',	'approach'),	13.610302038855057)
(('concedes',	'defeat'),	13.610302038855053)
(('single',	'dimension'),	13.610302038855053)
(('compelling',	'justification'),	13.517192634463576)
(('meter',	'readings'),	13.517192634463576)
(('larger',	'cylindrical'),	13.517192634463576)
(('slightly',	'larger'),	13.517192634463576)
(('marks',	'insignia'),	13.517192634463573)
(('sole',	'judge'),	13.517192634463573)

List the top 50 words by frequency [LongTennisRuleBook] :

shall	463
player	243
grand	182
tournament	182

slam	173
may	138
person	112
draw	111
match	108
gsc	104
players	98
umpire	85
violation	85
covered	79
play	79
qualifying	79
referee	77
corruption	74
chair	72
event	72
medical	72
tennis	70
section	67
set	67
time	66
offense	62
court	61
main	61
â	60
must	55
hearing	54
aho	53
director	52
one	51
within	51
entry	48
unless	47
code	44
competition	44
doubles	44
appeal	43
b	43
conduct	43
c	42
first	42
line	42

singles	42
subject	42
prior	41
tournaments	41

List the top 50 bigrams by frequencies without applying filter [LongTennisRuleBook]:

(('of',	'the'),	0.01272666777574447)
(('shall',	'be'),	0.009524205623024456)
(('to',	'the'),	0.008941939777075362)
(('grand',	'slam'),	0.007195142239228082)
(('by',	'the'),	0.005573115954084179)
(('in',	'the'),	0.004533355514889369)
(('with',	'the'),	0.004034270504075861)
(('a',	'player'),	0.003410414240558975)
(('for',	'the'),	0.0032024621527200134)
(('the',	'player'),	0.0031192813175844286)
(('covered',	'person'),	0.002952919647313259)
(('of',	'a'),	0.002952919647313259)
(('the',	'referee'),	0.002828148394609882)
(('of',	'this'),	0.0027449675594742973)
(('the',	'tournament'),	0.0027033771419065046)
(('at',	'the'),	0.00262019630677092)
(('and',	'the'),	0.002495425054067543)
(('corruption',	'offense'),	0.002495425054067543)
(('on',	'the'),	0.002495425054067543)
(('chair',	'umpire'),	0.002412244218931958)
(('the',	'gsc'),	0.002412244218931958)
(('main',	'draw'),	0.0023290633837963733)
(('slam',	'tournament'),	0.0023290633837963733)
(('the',	'chair'),	0.0021627017135252037)
(('violation',	'of'),	0.0020379304608218267)
(('after',	'the'),	0.001996340043254034)
(('the',	'director'),	0.001996340043254034)
(('the',	'grand'),	0.001996340043254034)
(('to',	'a'),	0.001996340043254034)
(('may',	'be'),	0.001954749625686242)
(('the',	'aho'),	0.0019131592081184496)
(('a',	'match'),	0.0018299783729828648)
(('if',	'the'),	0.0018299783729828648)
(('to',	'be'),	0.0017883879554150724)

(('director',	'gsc'),	0.00174679753784728)
(('shall',	'not'),	0.0015804358675761104)
(('accordance',	'with'),	0.001538845450008318)
(('any',	'other'),	0.001538845450008318)
(('in',	'accordance'),	0.001538845450008318)
(('qualifying',	'competition'),	0.001538845450008318)
(('and',	'or'),	0.0014972550324405256)
(('not',	'be'),	0.0014972550324405256)
(('prior',	'to'),	0.0014972550324405256)
(('slam',	'tournaments'),	0.0014972550324405256)
(('that',	'the'),	0.0014556646148727334)
(('the',	'qualifying'),	0.0014556646148727334)
(('this',	'section'),	0.001414074197304941)
(('will',	'be'),	0.001414074197304941)
(('the',	'first'),	0.0013724837797371486)
(('the',	'same'),	0.0013724837797371486)

List the top 50 bigrams by frequencies by applying filter [LongTennisRuleBook]:

(('grand',	'slam'),	0.007195142239228082)
(('covered',	'person'),	0.002952919647313259)
(('corruption',	'offense'),	0.002495425054067543)
(('chair',	'umpire'),	0.002412244218931958)
(('main',	'draw'),	0.0023290633837963733)
(('slam',	'tournament'),	0.0023290633837963733)
(('director',	'gsc'),	0.00174679753784728)
(('qualifying',	'competition'),	0.001538845450008318)
(('slam',	'tournaments'),	0.0014972550324405256)
(('set',	'forth'),	0.0013308933621693562)
(('section',	'shall'),	0.001206122109465979)
(('player',	'shall'),	0.0011229412743303943)
(('athletic',	'trainer'),	0.000998170021627017)
(('physiotherapist',	'athletic'),	0.000998170021627017)
(('medical',	'condition'),	0.0009565796040592248)
(('lucky',	'loser'),	0.0009149891864914324)
(('hereinafter',	'set'),	0.0007902179337880552)
(('person',	'shall'),	0.0007902179337880552)
(('prize',	'money'),	0.0007902179337880552)
(('gsc',	'shall'),	0.0007486275162202628)
(('line',	'umpire'),	0.0007486275162202628)
(('shall',	'subject'),	0.0007486275162202628)

(('players',	'shall'),	0.0007070370986524705)
(('point',	'penalty'),	0.0007070370986524705)
(('slam',	'chief'),	0.0007070370986524705)
(('slam',	'committee'),	0.0007070370986524705)
(('shall',	'also'),	0.0006654466810846781)
(('tournament',	'doctor'),	0.0006654466810846781)
(('unless',	'otherwise'),	0.0006654466810846781)
(('lucky',	'losers'),	0.0006238562635168857)
(('match',	'including'),	0.0006238562635168857)
(('penalty',	'schedule'),	0.0006238562635168857)
(('related',	'person'),	0.0006238562635168857)
(('tournament',	'site'),	0.0006238562635168857)
(('aho',	'shall'),	0.0005822658459490933)
(('slam',	'code'),	0.0005822658459490933)
(('aggravated',	'behaviour\xe2'),	0.0005406754283813009)
(('draw',	'shall'),	0.0005406754283813009)
(('inches',	'sq'),	0.0005406754283813009)
(('major',	'offence'),	0.0005406754283813009)
(('medical',	'time'),	0.0005406754283813009)
(('sq',	'cm'),	0.0005406754283813009)
(('square',	'inches'),	0.0005406754283813009)
(('tournament',	'shall'),	0.0005406754283813009)
(('tournament',	'support'),	0.0005406754283813009)
(('umpire',	'shall'),	0.0005406754283813009)
(('player',	'may'),	0.0004990850108135085)
(('treatable',	'medical'),	0.0004990850108135085)
(('wild',	'cards'),	0.0004990850108135085)
(('medical',	'treatment'),	0.0004574945932457162)

List the top 50 bigrams by their Mutual Information scores without applying filter [LongTennisRuleBook]

(('accepts',	'wagers'),	14.55338930472162)
(('aise',	'de'),	14.55338930472162)
(('angle',	'slope'),	14.55338930472162)
(('avenue',	'gordon'),	14.55338930472162)
(('a\xe2',	'scale'),	14.55338930472162)
(('baseline',	'sideline'),	14.55338930472162)
(('bennett',	'victoria'),	14.55338930472162)
(('canvas',	'plastic'),	14.55338930472162)

(('care',	'provider'),	14.55338930472162)
(('careful',	'deliberation'),	14.55338930472162)
(('chain',	'continuum'),	14.55338930472162)
(('church',	'road'),	14.55338930472162)
(('cloth',	'canvas'),	14.55338930472162)
(('commercials',	'encouraging'),	14.55338930472162)
(('completely',	'familiar'),	14.55338930472162)
(('computers',	'hard'),	14.55338930472162)
(('croquet',	'club'),	14.55338930472162)
(('current',	'postal'),	14.55338930472162)
(('degree',	'angle'),	14.55338930472162)
(('derogatory',	'insulting'),	14.55338930472162)
(('disclosed',	'according'),	14.55338930472162)
(('drinks',	'cups'),	14.55338930472162)
(('d\xe3',	'ration'),	14.55338930472162)
(('easily',	'adjustable'),	14.55338930472162)
(('emblem',	'logo'),	14.55338930472162)
(('familiarise',	'thoroughly'),	14.55338930472162)
(('fran\xe3',	'aise'),	14.55338930472162)
(('french',	'german'),	14.55338930472162)
(('f\xe3',	'd\xe3'),	14.55338930472162)
(('gordon',	'bennett'),	14.55338930472162)
(('hard',	'drives'),	14.55338930472162)
(('hat',	'headband'),	14.55338930472162)
(('health',	'care'),	14.55338930472162)
(('heels',	'ribs'),	14.55338930472162)
(('housing',	'meals'),	14.55338930472162)
(('illegal',	'drugs'),	14.55338930472162)
(('impact',	'improperly'),	14.55338930472162)
(('implies',	'dishonesty'),	14.55338930472162)
(('inspected',	'sufficiently'),	14.55338930472162)
(('intravenous',	'infusions'),	14.55338930472162)
(('kinetic',	'chain'),	14.55338930472162)
(('linesmen',	'boxes'),	14.55338930472162)
(('logo',	'trademark'),	14.55338930472162)
(('masculine',	'gender'),	14.55338930472162)
(('mr',	'william'),	14.55338930472162)
(('negligent',	'disregard'),	14.55338930472162)
(('normal',	'course'),	14.55338930472162)
(('off\xe2',	'switch'),	14.55338930472162)
(('operator',	'employee'),	14.55338930472162)
(('owner',	'operator'),	14.55338930472162)

List the top 50 bigrams by their Mutual Information scores with applying filter [LongTennisRuleBook]

(('muscle',	'cramping'),	11.383464303279307)
(('audible',	'obscenity'),	11.290354898887824)
(('particularly',	'injurious'),	11.093957686084321)
(('singularly',	'egregious'),	10.956454162334385)
(('clear',	'mistake'),	10.898037476109064)
(('inches',	'sq'),	10.852949586580525)
(('sq',	'cm'),	10.852949586580525)
(('highest',	'ranked'),	10.700946493135477)
(('square',	'inches'),	10.6464987091131)
(('support',	'personnel'),	10.6464987091131)
(('vacancy',	'created'),	10.60853085891408)
(('twenty',	'seconds'),	10.553389304721618)
(('governing',	'body'),	10.465926463471279)
(('indirectly',	'solicit'),	10.383464303279307)
(('aggravated',	'behaviour\xe2'),	10.383464303279306)
(('best',	'efforts'),	10.103356384086574)
(('wild',	'card'),	10.029827348664607)
(('physiotherapist',	'athletic'),	9.96842680400046)
(('non',	'appearance'),	9.909533114946896)
(('additional',	'penalties'),	9.880963962750124)
(('re',	'warm'),	9.852949586580527)
(('athletic',	'trainer'),	9.852949586580525)
(('unavoidable',	'circumstances'),	9.830923280250527)
(('wild',	'cards'),	9.807434927328156)
(('major',	'offence'),	9.794397404225414)
(('written',	'submission'),	9.630557165244078)
(('official',	'opponent'),	9.614789849385762)
(('also',	'constitute'),	9.576109381221702)
(('prize',	'money'),	9.553389304721618)
(('late',	'withdrawal'),	9.535467396724357)
(('penalties',	'hereinafter'),	9.453853631170706)
(('money',	'benefit'),	9.415885780971685)
(('next',	'highest'),	9.39351796794323)
(('ten',	'days'),	9.383464303279307)
(('business',	'days'),	9.360744226779225)

(('lucky',	'losers'),	9.34393593909267)
(('thirty',	'days'),	9.33099688338517)
(('computer',	'ranking'),	9.324570614225738)
(('lucky',	'loser'),	9.279805601672953)
(('unsportsmanlike',	'conduct'),	9.12712455001952)
(('forth',	'therefor'),	9.09395768608432)
(('even',	'number'),	9.017336404481409)
(('among',	'seeds'),	8.82003496410779)
(('penalty',	'schedule'),	8.727264578644176)
(('schedule',	'hereinafter'),	8.707899253777244)
(('point',	'penalty'),	8.52450818473449)
(('original',	'entry'),	8.482999976830222)
(('line',	'umpires'),	8.460632163801767)
(('conduct',	'contrary'),	8.449052644906883)
(('set',	'forth'),	8.442905994905393)

Q. Briefly state why you chose the processing options that you did

The results are pretty intriguing. From a frequency standpoint, the most common bigrams are either related to health rules and prizes or are common phrases. In turn, PMI probability shows that the most probable pairs are more complex and specific. This makes sense, because the higher the PMI probability, the more likely both words in a bigram will only occur with one another. The more specific the phrase, the more likely it is to occur only with its companion. In short, PMI and frequency contrast each other nicely, one showing commonality and the other uniqueness.

For word frequencies:

1. Calculated frequencies of words on original corpus.
2. Secondly, removed stopwords using lambda filter and again calculated frequencies of words.

I used this filter to make sure that I did not get most common regularly used words in frequency list of rule books which I want to compare. Thus, giving us more information about words used in rule book that really explains their contents.

For bigrams list using frequencies scores:

- Applied filter using lambda function

  I used this filter to remove ‘shall’, ’a’, ‘b’ and ‘c’ from corpus, as we can see from frequency list of original frequency they are very high in frequency.

- Applied filter to remove non-alphabetical tokens

I used this filter so that I can only focus on text words used in rule and not the numbers or special characters used in rule.

- Applied stopwords filter using lambda function as well as from Google Stop Project filter.

  I used this filter to remove common stopwords from rule book, as they can have high frequency in original books.

For bigrams list using Mutual Information Scores:
We used this so that we can effectively count bigram frequencies using mutual information scores.

- Applied filter using lambda function

  I used this filter to remove 'shall', 'a', 'b' and 'c' from corpus, as we can see from frequency list of original frequency they are very high in frequency.

- Applied filter to remove non-alphabetical tokens

  I used this filter so that I can only focus on text words used in rule and not the numbers or special characters used in rule.
- Applied frequency filter for words with frequency higher than 5

I used this so that we can effectively count bigram frequencies using mutual information scores.
- Applied stopwords filter using lambda function as well as from Google Stop Project filter

Q. Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams?
As we can see in the above bigram lists nltk tokenization and bigram generator functions does not handle " *\xe2" . We can use some other custom filters using lambda functions to handle these cases or we can also use regular expressions for this purpose.

Q. How are the bigram frequency list and the bigram Mutual Information lists different?
We can see from above two bigrams lists using frequency scores and mutual information scores that those two are different, as bigrams with frequency depends on the probability of second word following first word in whole corpus and bigrams with mutual information depends on probability of two words occurring in sequence.

When you apply the Mutual Information score to small documents, the results don't really make sense. In particular, here all our scores are the same. So, by running the PMI scorer with a minimum frequency of 5, which will make more sense on very large documents.

3. [Required task] How to characterize the style between two authors or two works of different classes. Another example would be to compare the informal text in blogs with more formal text. Or you can do a topic related comparison that selects words (as in the SOTU speeches example). You could also make a comparison of similar text but at two different times
=>
How both rules are written?
Is top PMI related to type of specific rule book?
How does both talk about rules related to specific field?
How both books talks about prize?

Both these rules book uses “Shall” mostly to describe their content which is described by its highest number of frequency count in both of the rule books.

Both of these book talk about fitness issues that can be concluded by
PMI Score from Cricket rule books (('pulled', 'muscle'), 14.34726763302126) as well as from
PMI Score from Long Tennis books (('muscle', 'cramping'), 11.383464303279307).

Along with this both books focus on their respective world series. Like, in Long Tennis we can find Top PMI scores (('grand', 'slam'), 8. 794397404225414) for whereas in Cricket Top PMI scores are (('champions', 'trophy'), 14.932230133742419).

Both of these books talk about their specific rules. Like, in Long Tennis we can find Top PMI scores for (('wild', 'card'), 10.029827348664607); (('major', 'offence'), 9.794397404225414) ; (('point', 'penalty'), 8.52450818473449) whereas in Cricket Top PMI scores are (('clean', 'catches'), 14.517192634463576) ; (('metal', 'disc'), 14.932230133742419) ; (('sole', 'judge'), 13.517192634463573).

4. [Additional task] Now answer the question you have chosen by giving a discussion of the comparison of the texts. Using one or more of the types of measures that you ran in the first task, i.e. word frequencies, bigram frequencies, or bigram mutual information, make a comparison of the two documents to answer the problem or question. For this analysis, you will want to choose or to revise data that will be applicable for your question. You may wish to hand pick out particular examples of word frequencies, bigram frequencies or mutual information scores that contribute evidence for your comparison, or combine examples into categories. If your documents in task 1 required a lot of preprocessing steps, you may give a short discussion here.

=>

As both rule books, Cricket rule books and Long Tennis rule book, talk about rules so they use “shall” to describe their content which can be cleared by highest frequency that is “shall 463” in case of Long tennis and “shall 3986” in case of Cricket. Along with this we observe that Long Tennis talk about grand slam (('grand', 'slam'), 0.007195142239228082) whereas Cricket rule book talks about power play (('powerplay', 'overs'), 0.0006397065879116779). As these are sports writing, both books has high PMI score with fitness Cricket rule books (('pulled', 'muscle'), 14.34726763302126) and Long Tennis books (('muscle', 'cramping'), 11.383464303279307). These things make sense considering different sports rule books within sport categories.

In short, the higher the PMI probability, the more likely both words in a bigram will only occur with one another. The more specific the phrase, the more likely it is to occur only with its companion. In short, PMI and frequency contrast each other nicely, one showing commonality and the other uniqueness.