I have chosen 2 documents. They are

**1.) Desserts and Salads by Gesine Lemcke**

**2.) Woman's Institute Library of Cookery, Volume 3: Soup; Meat; Poultry and Game; Fish and Shell Fish**

These two books are about Cooking.

I am basically a foodie. I love to eat and cook different dishes as well. So I try out different recipes in my free time. So, this made me choose the cooking books.

The first book explains us the different methods of cooking different Deserts and Salads. The author uses different types of approach to make the sweets more tasty and delicious

The second book explain us about the cooking of Non-vegetarian dishes. Especially, meat and also includes poultry and sea food.

**“Deserts and salads” by Gesine Lemcke**

1. **List of top 50 words by frequency (normalized by the length of the document)**

sugar 0.0018326191202428159

cup 0.0014300929560421428

water 0.0013013345867481387

egg 0.0011875771925175136

cream 0.0011813267862411058

tablespoonful 0.0011438243485826578

cold 0.0010288168730967513

boil 0.0009425612664823213

stir 0.0008163030596988804

pudding 0.000790051353337967

minute 0.0007612994844664904

butter 0.0007600494032112087

sauce 0.0007412981843819849

serve 0.0007175466405316346

dish 0.0007175466405316346

apple 0.0007062959092341002

lemon 0.0006600429027886813

fire 0.0006550425777675549

white 0.0006487921714911469

form 0.0006325411151724862

pour 0.0006262907088960783

milk 0.0006200403026196703

pound 0.000602539165045728

teaspoonful 0.0005912884337481936

remove 0.0005912884337481936

ice 0.0005712871336636882

little 0.000556286158600309

jelly 0.0005412851835369299

flour 0.0005325346147499587

cake 0.000528784370984114

yolk 0.0005212838834524244

pint 0.0005200338021971428

saucepan 0.0004912819333256662

let 0.0004912819333256662

hour 0.0004862816083045398

boiling 0.0004612799831989079

cover 0.00045252941441193675

place 0.0004412786831144024

rice 0.0004362783580932761

done 0.00041877722051933376

bake 0.00041877722051933376

fruit 0.0004112767329876442

salt 0.00040002600169010985

vanilla 0.00039377559541370187

small 0.0003925255141584203

oven 0.00038627510788201235

hot 0.0003837749453714491

fine 0.00038002470160560436

lay 0.00038002470160560436

paste 0.0003762744578397596

1. **List of top 50 Bigrams by frequency**

(('cold', 'water'), 0.004788732394366197)

(('yolk', 'egg'), 0.004744255003706449)

(('tablespoonful', 'sugar'), 0.004462564862861379)

(('cup', 'sugar'), 0.0038398813936249073)

(('stiff', 'froth'), 0.0030244625648628612)

(('white', 'egg'), 0.0030244625648628612)

(('boil', 'minute'), 0.002787249814677539)

(('tablespoonful', 'butter'), 0.002742772424017791)

(('powdered', 'sugar'), 0.0024610822831727204)

(('boiling', 'water'), 0.002357301704966642)

(('remove', 'fire'), 0.002223869532987398)

(('let', 'stand'), 0.002090437361008154)

(('grated', 'rind'), 0.001927353595255745)

(('whipped', 'cream'), 0.0019125277983691623)

(('beaten', 'stiff'), 0.0018680504077094144)

(('egg', 'tablespoonful'), 0.0018383988139362491)

(('cup', 'milk'), 0.0017494440326167532)

(('cup', 'water'), 0.0017494440326167532)

(('teaspoonful', 'vanilla'), 0.0017049666419570052)

(('teaspoonful', 'salt'), 0.0016901408450704226)

(('juice', 'lemon'), 0.0016308376575240918)

(('sugar', 'cup'), 0.001586360266864344)

(('bread', 'crumb'), 0.0015567086730911787)

(('hot', 'oven'), 0.001497405485544848)

(('aside', 'cool'), 0.0014825796886582653)

(('pound', 'sugar'), 0.0014232765011119347)

(('serve', 'cold'), 0.0014084507042253522)

(('sugar', 'cream'), 0.0013787991104521869)

(('hot', 'water'), 0.0013343217197924388)

(('place', 'saucepan'), 0.0012750185322461082)

(('beat', 'white'), 0.0012601927353595255)

(('beaten', 'white'), 0.0012601927353595255)

(('rind', 'lemon'), 0.0012601927353595255)

(('cool', 'place'), 0.0012305411415863602)

(('egg', 'cup'), 0.0012157153446997776)

(('hour', 'serve'), 0.0012157153446997776)

(('sugar', 'teaspoonful'), 0.0012008895478131949)

(('finely', 'chopped'), 0.0011860637509266123)

(('glass', 'dish'), 0.0011860637509266123)

(('cracked', 'ice'), 0.0011712379540400296)

(('sifted', 'flour'), 0.001156412157153447)

(('fire', 'boil'), 0.001111934766493699)

(('baking', 'powder'), 0.0010822831727205337)

(('pint', 'milk'), 0.0010674573758339512)

(('water', 'boil'), 0.0010674573758339512)

(('dish', 'pour'), 0.0010526315789473684)

(('ice', 'cream'), 0.0010526315789473684)

(('medium', 'hot'), 0.0010378057820607859)

(('small', 'piece'), 0.0010378057820607859)

(('stir', 'fire'), 0.0010378057820607859)

1. **List of Top 50 Mutual Information Scores (Using minimum frequency 5)**

(('gesine', 'lemcke'), 13.45656832200184)

(('complete', 'lining'), 13.234175900665392)

(('pate', 'chou'), 13.234175900665388)

(('kalte', 'schale'), 13.041530822722994)

(('public', 'domain'), 12.871605821280681)

(('rectified', 'spirit'), 12.871605821280681)

(('\_text', 'unchanged\_'), 12.778496416889197)

(('spirit', 'per'), 12.719602727835632)

(('muerber', 'teig'), 12.45656832200184)

(('knitting', 'needle'), 12.456568322001836)

(('tutti', 'frutti'), 12.412174202643385)

(('literary', 'archive'), 12.3410911045819)

(('slightly', 'salted'), 11.954067981472653)

(('per', 'cent'), 11.889527729277944)

(('united', 'state'), 11.649213399944234)

(('oyster', 'plant'), 11.582099204085695)

(('pierce', 'easily'), 11.54967772639332)

(('archive', 'foundation'), 11.397674632948268)

(('straw', 'pierce'), 11.3691054807515)

(('pastry', 'wheel'), 11.35970678274925)

(('gutenberg', 'literary'), 11.134640227114474)

(('bring', 'corner'), 11.079598863556514)

(('rolling', 'pin'), 10.904991989775228)

(('clarified', 'dripping'), 10.816564457722718)

(('french', 'mustard'), 10.772070147729764)

(('run', 'filtering'), 10.756128603860745)

(('text', 'index'), 10.756128603860745)

(('see', 'clarifying'), 10.719602727835632)

(('term', 'agreement'), 10.649213399944232)

(('+fromage', 'bavarois'), 10.615266068020896)

(('farcis', 'fromage'), 10.615266068020896)

(('celery', 'root'), 10.556103995552752)

(('greening', 'pippin'), 10.498147425753237)

(('day', 'old'), 10.48256353053478)

(('north', 'german'), 10.45656832200184)

(('string', 'bean'), 10.39767463294827)

(('+creme', 'francaise'), 10.389454126143299)

(('double', 'height'), 10.377616980607014)

(('keep', 'turning'), 10.348043865223667)

(('broad-bladed', 'knife'), 10.286643320559527)

(('wooden', 'spoon'), 10.23417590066539)

(('german', 'art'), 10.234175900665388)

(('golden', 'color'), 10.234175900665388)

(('german', 'style'), 10.134640227114476)

(('sharp', 'knife'), 10.134640227114474)

(('lady', 'finger'), 10.00863234220237)

(('draw', 'front'), 9.933006365944825)

(('creme', 'francaise'), 9.897140913387817)

(('au', 'chocolat.+'), 9.871605821280681)

(('au', 'chocolat'), 9.871605821280678)

**“Woman's Institute Library of Cookery,** Volume 3: Soup; Meat; Poultry and Game; Fish and Shell Fish**”**

1. **List of top 50 words by frequency (normalized by the length of the document)**

meat 0.0013448175855260284

fish 0.0011124045020599349

soup 0.0009197201508102674

food 0.0007806695880527756

salt 0.0006694291378467822

fat 0.0006436054619061052

chicken 0.0006038767296896788

water 0.000566134434084074

piece 0.0005442836313650395

serve 0.0004906498428728641

dish 0.0004628397303213657

beef 0.0004409889276023313

pepper 0.0004409889276023313

sauce 0.00040920594182919026

small 0.00040523306860754766

cooking 0.00039927375877508374

hot 0.00039927375877508374

value 0.0003972873221642624

c. 0.00038338226588851323

stock 0.0003813958292776919

roast 0.0003774229560560493

make 0.00036947720961276405

place 0.0003615314631694788

made 0.0003615314631694788

well 0.0003515992801153722

preparation 0.00034961284350455094

steak 0.0003456399702829083

served 0.000343653533672087

flavor 0.000343653533672087

veal 0.0003416670970612657

butter 0.0003416670970612657

kind 0.00033968066045044436

cooked 0.0003317349140071591

oyster 0.0003277620407855165

illustration 0.0003257756041746952

poultry 0.00031981629434223125

method 0.0003098841112881247

remove 0.0003019383648448394

pork 0.00029995192823401815

bone 0.00028803330856909024

pan 0.00028803330856909024

tsp 0.000280087562125805

shell 0.00027015537907169846

tb 0.00027015537907169846

however 0.00026419606923923453

flour 0.0002602231960175919

cut 0.00025823675940677055

vegetable 0.0002502910129634853

cook 0.0002502910129634853

fried 0.0002463181397418427

1. **List of top 50 Bigrams by frequency**

(('salt', 'pepper'), 0.003122524118003449)

(('food', 'value'), 0.0025632660670177566)

(('tsp', 'salt'), 0.0015146572214195834)

(('project', 'gutenberg-tm'), 0.0013049354522999487)

(('sufficient', 'serve'), 0.0013049354522999487)

(('white', 'sauce'), 0.0012350281959267373)

(('tsp', 'pepper'), 0.0011884233583445963)

(('illustration', 'illustration'), 0.0011651209395535257)

(('salt', 'tsp'), 0.0011418185207624552)

(('shell', 'fish'), 0.001095213683180314)

(('tb', 'butter'), 0.0009787015892249616)

(('cold', 'water'), 0.0009320967516428206)

(('small', 'piece'), 0.0009320967516428206)

(('frying', 'pan'), 0.0009087943328517501)

(('bread', 'crumb'), 0.000862189495269609)

(('composition', 'food'), 0.000862189495269609)

(('deep', 'fat'), 0.0008388870764785384)

(('serve', 'hot'), 0.0008388870764785384)

(('cooking', 'meat'), 0.0008155846576874679)

(('hot', 'oven'), 0.0008155846576874679)

(('lamb', 'mutton'), 0.0008155846576874679)

(('boiling', 'water'), 0.0007689798201053269)

(('serve', 'six'), 0.0007689798201053269)

(('boiling', 'point'), 0.0007223749825231859)

(('hot', 'water'), 0.0006990725637321154)

(('project', 'gutenberg'), 0.0006990725637321154)

(('electronic', 'work'), 0.0006291653073589038)

(('method', 'cookery'), 0.0006291653073589038)

(('mutton', 'lamb'), 0.0006291653073589038)

(('season', 'salt'), 0.0006291653073589038)

(('tb', 'flour'), 0.0006291653073589038)

(('value', 'composition'), 0.0006291653073589038)

(('salt', 'pork'), 0.0005825604697767628)

(('strip', 'bacon'), 0.0005359556321946218)

(('chopped', 'parsley'), 0.0005126532134035513)

(('making', 'soup'), 0.0005126532134035513)

(('roasting', 'pan'), 0.0005126532134035513)

(('small', 'onion'), 0.0005126532134035513)

(('small', 'quantity'), 0.0005126532134035513)

(('soup', 'stock'), 0.0005126532134035513)

(('beaten', 'egg'), 0.0004893507946124808)

(('nutritive', 'value'), 0.0004893507946124808)

(('baking', 'dish'), 0.0004660483758214103)

(('c.', 'milk'), 0.0004660483758214103)

(('cracker', 'crumb'), 0.0004660483758214103)

(('piece', 'meat'), 0.0004660483758214103)

(('stewed', 'tomato'), 0.0004660483758214103)

(('timbale', 'case'), 0.0004660483758214103)

(('value', 'fish'), 0.0004660483758214103)

(('various', 'kind'), 0.0004660483758214103)

1. **List of Top 50 Mutual Information Scores (Using minimum frequency 5)**

(('ptomaine', 'poisoning'), 12.359413416645307)

(('human', 'being'), 12.219235758597048)

(('woman', 'institute'), 12.219235758597048)

(('distinguishing', 'feature'), 12.166768338702909)

(('public', 'domain'), 12.067232665151998)

(('examination', 'question'), 12.067232665151995)

(('finnan', 'haddie'), 12.026590680654653)

(('literary', 'archive'), 11.688721041898265)

(('mo', 'yr.'), 11.666694735568267)

(('curry', 'powder'), 11.652195165873152)

(('institute', 'library'), 11.54116385348441)

(('lake', 'trout'), 11.54116385348441)

(('baking-powder', 'biscuit'), 11.482270164430844)

(('along', 'spine'), 11.38916076003936)

(('trying', 'suet'), 11.318771432147962)

(('luncheon', 'supper'), 11.30169791878902)

(('october', 'to..'), 11.301697918789019)

(('red', 'snapper'), 11.067232665151996)

(('damp', 'cloth'), 11.067232665151995)

(('differ', 'materially'), 11.026590680654653)

(('tightly', 'closed'), 10.903733932869118)

(('strongly', 'flavored'), 10.816271091618777)

(('great', 'deal'), 10.745304570264636)

(('pin', 'feather'), 10.726195747316929)

(('clove', 'peppercorn'), 10.688721041898267)

(('bay', 'leaf'), 10.645768896713722)

(('per', 'cent'), 10.607801046514698)

(('suit', 'taste'), 10.581805837981754)

(('gutenberg', 'literary'), 10.48227016443084)

(('set', 'forth'), 10.48227016443084)

(('united', 'state'), 10.434964449652485)

(('ox', 'tail'), 10.434964449652483)

(('pierced', 'fork'), 10.34079773847796)

(('already', 'learned'), 10.273683542619422)

(('relative', 'nutritive'), 10.268145359077993)

(('principle', 'deep-fat'), 10.259877743094393)

(('archive', 'foundation'), 10.219235758597046)

(('fore', 'quarter'), 10.187526898869708)

(('internal', 'organ'), 10.158206325199487)

(('find', 'favor'), 10.08995274165208)

(('highly', 'seasoned'), 10.08995274165208)

(('forcemeat', 'ball'), 10.055260023485918)

(('simmer', 'gently'), 10.026590680654653)

(('change', 'usual'), 10.022838545793544)

(('hind', 'quarter'), 9.80146846656263)

(('freshly', 'killed'), 9.762721623342044)

(('turning', 'frequently'), 9.750121586562408)

(('biscuit', 'dough'), 9.745304570264633)

(('nicely', 'browned'), 9.572177136783978)

(('connective', 'tissue'), 9.564732324622813)

1. *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,*
2. *Briefly state why you chose the processing options that you did.*

**Answer**: The filters I used here are:

1. *converting all the text to Lower case*

Lower casing of all words in the texts will not create two different types of identity for same word (as different word). Hence lower casing will lead to more meaningful ratios

1. *Removal of the non-alphabetical characters from the text*

since there are more non-alpha characters present in the text which includes these non-alpha numeric will affect the top frequency lists

1. *Removal of the stopwords from the text*

what the stopword list should be will depend on the analysis that is using the word frequency list. Hence in this book we have more usage of stopwords like “‘n, ‘t, ‘s””

1. *Usage of WordNet Lemmatizer*

To reduce the derivationally related forms of a word to a common base form, I have used the wordnet lemmatizer.

1. *Applied filter to remove words that only occurred with a frequency of less than 5.*

when you apply the Mutual Information score to small documents or documents with rare words, the results don’t really make sense.

*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 top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?*

When I compare the top 50 bigrams with raw frequency with that of top 50 scored by Mutual information, I think that top 50 scored by mutual information have more meaningful words than that of the raw frequency one. Hence PMI Scores reveal more reliable probability of co-occurrence of the words than using raw freq. The more strongly connected 2 items are, the higher will

be their Mutual information value

*c) If you modify the stop word list, or expand the methods of filtering, describe that here*

For both the book in common, I have used stop words like ‘n, ‘t, ‘s. since we use them very much in common.

Ex: wouldn’t, shouldn’t

*d) tri-gram lists*

trigrams for deserts

(('beaten', 'stiff', 'froth'), 0.0018235730170496664)

(('egg', 'tablespoonful', 'sugar'), 0.0012305411415863602)

(('yolk', 'egg', 'tablespoonful'), 0.0012305411415863602)

(('grated', 'rind', 'lemon'), 0.0010674573758339512)

(('cup', 'sugar', 'cup'), 0.0010378057820607859)

(('medium', 'hot', 'oven'), 0.001022979985174203)

(('white', 'egg', 'beaten'), 0.0009191994069681246)

(('beaten', 'white', 'egg'), 0.0008895478131949593)

(('egg', 'beaten', 'stiff'), 0.0008895478131949593)

(('sugar', 'cup', 'water'), 0.0008747220163083766)

(('white', 'beaten', 'stiff'), 0.000859896219421794)

(('stir', 'tablespoonful', 'butter'), 0.0008005930318754633)

(('teaspoonful', 'vanilla', 'extract'), 0.0008005930318754633)

(('pint', 'whipped', 'cream'), 0.0007857672349888806)

(('let', 'stand', 'hour'), 0.000770941438102298)

(('white', 'stiff', 'froth'), 0.0007412898443291327)

(('beat', 'white', 'stiff'), 0.00072646404744255)

(('rinse', 'cold', 'water'), 0.0007116382505559674)

(('sugar', 'yolk', 'egg'), 0.0007116382505559674)

(('teaspoonful', 'baking', 'powder'), 0.0007116382505559674)

(('bake', 'medium', 'hot'), 0.0006671608598962194)

(('sugar', 'teaspoonful', 'vanilla'), 0.0006375092661230541)

(('tablespoonful', 'butter', 'tablespoonful'), 0.0006375092661230541)

(('tablespoonful', 'powdered', 'sugar'), 0.0006375092661230541)

(('boil', 'hour', 'serve'), 0.0006226834692364715)

(('nearly', 'boiling', 'remove'), 0.0006078576723498888)

(('yolk', 'egg', 'cup'), 0.0005930318754633062)

(('cream', 'yolk', 'egg'), 0.0005782060785767235)

(('let', 'stand', 'minute'), 0.0005782060785767235)

(('boil', 'cup', 'sugar'), 0.0005633802816901409)

(('butter', 'tablespoonful', 'sugar'), 0.0005633802816901409)

(('cup', 'boiling', 'water'), 0.0005633802816901409)

(('cup', 'powdered', 'sugar'), 0.0005337286879169756)

(('tablespoonful', 'butter', 'cream'), 0.0005337286879169756)

(('tablespoonful', 'sugar', 'teaspoonful'), 0.0005337286879169756)

(('fire', 'boil', 'minute'), 0.0005189028910303929)

(('cold', 'water', 'return'), 0.0005040770941438103)

(('boil', 'minute', 'drain'), 0.00047442550037064493)

(('boiling', 'water', 'stir'), 0.00047442550037064493)

(('colander', 'rinse', 'cold'), 0.00047442550037064493)

(('finely', 'chopped', 'almond'), 0.00047442550037064493)

(('return', 'rice', 'saucepan'), 0.00047442550037064493)

(('soak', 'ounce', 'gelatine'), 0.00047442550037064493)

(('sprinkle', 'bread', 'crumb'), 0.00047442550037064493)

(('turn', 'pudding', 'onto'), 0.00047442550037064493)

(('beat', 'white', 'egg'), 0.0004595997034840623)

(('cup', 'cold', 'water'), 0.0004595997034840623)

(('egg', 'cup', 'sugar'), 0.0004595997034840623)

(('little', 'cold', 'water'), 0.0004595997034840623)

(('cup', 'sifted', 'flour'), 0.00044477390659747963)

Trigram for womens volume 3

(('tsp', 'salt', 'tsp'), 0.001095213683180314)

(('salt', 'tsp', 'pepper'), 0.0009787015892249616)

(('composition', 'food', 'value'), 0.0008155846576874679)

(('sufficient', 'serve', 'six'), 0.0007689798201053269)

(('illustration', 'illustration', 'illustration'), 0.0006524677261499744)

(('season', 'salt', 'pepper'), 0.0006058628885678333)

(('sprinkle', 'salt', 'pepper'), 0.0004427459570303398)

(('fry', 'deep', 'fat'), 0.0004194435382392692)

(('gutenberg-tm', 'electronic', 'work'), 0.0004194435382392692)

(('project', 'gutenberg-tm', 'electronic'), 0.0004194435382392692)

(('sufficient', 'serve', 'four'), 0.0003262338630749872)

(('tb', 'butter', 'tb'), 0.0003262338630749872)

(('butter', 'tb', 'flour'), 0.00030293144428391667)

(('fish', 'table', 'showing'), 0.00030293144428391667)

(('gutenberg', 'literary', 'archive'), 0.00030293144428391667)

(('literary', 'archive', 'foundation'), 0.00030293144428391667)

(('project', 'gutenberg', 'literary'), 0.00030293144428391667)

(('serve', 'six', 'c.'), 0.00030293144428391667)

(('tb', 'chopped', 'parsley'), 0.00030293144428391667)

(('food', 'value', 'fish'), 0.00027962902549284616)

(('flour', 'salt', 'pepper'), 0.00025632660670177565)

(('salt', 'pepper', 'serve'), 0.00025632660670177565)

(('salt', 'smoked', 'fish'), 0.00025632660670177565)

(('soup', 'sufficient', 'serve'), 0.00025632660670177565)

(('butter', 'tsp', 'salt'), 0.00023302418791070514)

(('c.', 'stewed', 'tomato'), 0.00023302418791070514)

(('crumb', 'fry', 'deep'), 0.00023302418791070514)

(('high', 'food', 'value'), 0.00023302418791070514)

(('name', 'season', 'us'), 0.00023302418791070514)

(('project', 'gutenberg-tm', 'work'), 0.00023302418791070514)

(('tb', 'flour', 'c.'), 0.00023302418791070514)

(('tsp', 'pepper', 'tb'), 0.00023302418791070514)

(('well', 'salt', 'pepper'), 0.00023302418791070514)

(('bacon', 'salt', 'pork'), 0.0002097217691196346)

(('bake', 'hot', 'oven'), 0.0002097217691196346)

(('fish', 'shell', 'fish'), 0.0002097217691196346)

(('fried', 'deep', 'fat'), 0.0002097217691196346)

(('heat', 'thoroughly', 'serve'), 0.0002097217691196346)

(('hot', 'oven', 'bake'), 0.0002097217691196346)

(('medium', 'white', 'sauce'), 0.0002097217691196346)

(('place', 'hot', 'oven'), 0.0002097217691196346)

(('showing', 'name', 'season'), 0.0002097217691196346)

(('table', 'showing', 'name'), 0.0002097217691196346)

(('baked', 'broiled', 'fried'), 0.0001864193503285641)

(('butter', 'frying', 'pan'), 0.0001864193503285641)

(('easily', 'pierced', 'fork'), 0.0001864193503285641)

(('food', 'value', 'shell'), 0.0001864193503285641)

(('project', 'gutenberg-tm', 'license'), 0.0001864193503285641)

(('salt', 'pepper', 'place'), 0.0001864193503285641)

(('season', 'well', 'salt'), 0.0001864193503285641)

***3) Describe a problem or question that is based on the difference between the two documents.***

1. How is “Deserts and Salads” book similar or different to the book “Woman's Institute Library of Cookery” in the use of the language, based on your results?

As I mentioned, the Deserts and Salads mainly concentrate on preparing different types of sweets and deserts that can be made. It provides different varieties of sweets and directions to cook them deliciously

Whereas on the other hand, in Woman's Institute Library of Cookery, we can see different types of recipe prepared by using different types of meats, poultry, and sea foods. This book deals with complete non-vegetarian dishes, which makes it easier for us to differentiate between these two books.

Secondly in the book Deserts and Salads, we most frequently use words like sweet, sugar, milk, bake, water etc

In Woman's Institute Library of Cookery, we use words like beef, chicken, fish, salt etc

In first , the author is giving the brief explanations about the directions to prepare deserts.

Where as, in the second, the author is giving a detailed explanation. The author is creating separate categories in the book for the preparation of soups, meat etc. the author is creating separate columns for ingrediets and another column for directions.

1. How are the options similar or different for the two analysis tasks?

First, as mentioned above language usage and the sytle of presentation is different. Apart from standard stopwords of NLTK there was need to add few extra stopwords for both.

For book, Desserts and Salads by Gesine Lemcke , I have used stop words like “add, put, till, set, cut, mix”

For book, Woman's Institute Library of Cookery, Volume 3: Soup; Meat; Poultry and Game; Fish and Shell Fish, I have used stop words like “may, cut, fig, add, way, one, used”.

1. What is the average number of characters per word, average number of words per sentence and the average occurrence of each vocabulary in both the books.

This can be calculated by using

Deserts and Salads

1 1 22

Woman's Institute Library of Cookery, Volume 3: Soup; Meat; Poultry and Game; Fish and Shell Fish

1 1 10