Language Modelling and Smoothing using Laplace Probability distribution (NLTK Library)

In [1]:
```
import nltk
```

Building model on the Mody Dick Corpus

Importing corpus via NLTK-Book

In [2]:
```
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

Frequency Distribution for text1 i.e Moby Dick

In [3]:
```
text1.vocab()
```

Out[3]:
```
FreqDist({'moon': 9,
          'horse': 26,
          'angels': 8,
          'IBID': 3,
          'truest': 3,
          'lantern': 12,
          'Gaining': 1,
          'THY': 1,
          'monotonously': 1,
          'Hill': 1,
          'understanding': 4,
          'thing': 188,
          'spinning': 4,
          'knotted': 5,
          'beginning': 19,
```

```
'success': 5,
'evil': 11,
'hollow': 20,
'abandoned': 7,
'snoring': 1,
'toed': 3,
'chase': 56,
'leisure': 3,
'sinew': 1,
'unscrew': 2,
'grindstone': 2,
'jingling': 2,
'compose': 1,
'garden': 6,
'hit': 14,
'SPLICE': 1,
'absurd': 3,
'mingling': 1,
'speechlessly': 1,
'isn': 1,
'certainty': 3,
'reprehensible': 1,
'occupied': 9,
'HISTORY': 4,
'nut': 3,
'joker': 2,
'sayest': 2,
'opposition': 1,
'scratch': 1,
'y': 2,
'exhaustion': 3,
'spear': 10,
'sweetly': 2,
'seen': 161,
'specific': 8,
'declared': 14,
'Noah': 8,
'appearance': 11,
'glistened': 3,
'mutations': 1,
'maned': 2,
'arboring': 1,
'hospitals': 2,
'gliding': 13,
'wit': 2,
'stamp': 3,
'nautical': 4,
'area': 4,
'Ganges': 1,
'Achilles': 1,
'noticed': 11,
'Io': 2,
'Stab': 1,
```

```
'afflictions': 1,
'ribbed': 7,
'unscathed': 1,
'mustering': 3,
'deserves': 3,
'delegated': 1,
'gases': 1,
'begets': 1,
'FILING': 1,
'dad': 1,
'BOUTON': 1,
'wedged': 3,
'sanity': 3,
'sorry': 6,
'Huron': 1,
'adult': 2,
'ado': 3,
'harm': 8,
'WAL': 1,
'substantiates': 1,
'discernible': 5,
'fairest': 1,
'eventually': 8,
'CAPTAIN': 2,
'anaconda': 2,
'header': 2,
'sneezing': 1,
'shod': 1,
'meet': 19,
'concernments': 1,
'grapple': 1,
'vengeance': 11,
'prize': 3,
'fortunes': 3,
'henceforth': 2,
'miserly': 1,
'specimen': 5,
'scramble': 1,
'moulder': 1,
'Crying': 1,
'Cancer': 1,
'distrusting': 2,
'overlooked': 2,
'awed': 2,
'whereto': 1,
'completely': 33,
'poorest': 1,
'halting': 3,
'misgrown': 1,
'respectful': 2,
'modern': 18,
'recur': 1,
'unicorns': 1,
```

```
'intercourse': 1,
'unretracing': 1,
'Hawaiian': 2,
'swallows': 4,
'excepting': 4,
'thirteenth': 1,
'striving': 10,
'diminish': 3,
'occasionally': 14,
'mayst': 1,
'graduates': 1,
'grievous': 1,
'whitenesses': 1,
'deliver': 3,
'microscopic': 2,
'standard': 3,
'quiver': 3,
'lowerings': 2,
'State': 4,
'Phidias': 1,
'preparing': 3,
'plainly': 39,
'sinful': 3,
'expired': 2,
'Michigan': 2,
'Cowper': 1,
'speech': 3,
'toiling': 4,
'corruption': 1,
'started': 26,
'HOLY': 1,
'buoyantly': 1,
'numbers': 6,
'hover': 2,
'vain': 30,
'king': 25,
'firmly': 11,
'Indiamen': 1,
'Enough': 2,
'tints': 2,
'ape': 2,
'monomaniac': 11,
'construction': 1,
'128': 1,
'searching': 2,
'invest': 6,
'Ellenborough': 2,
'jig': 3,
'distinguished': 7,
'studying': 4,
'or': 697,
'consent': 1,
'Dead': 1,
```

```
'ribbons': 1,
'Carefully': 1,
'barest': 1,
'Seychelle': 1,
'agent': 8,
'vivacious': 3,
'directed': 6,
'unfeatured': 1,
'bedfellow': 5,
'screwed': 7,
'131': 1,
'skittishly': 1,
'MOUTHED': 1,
'abbreviate': 1,
'engrossing': 1,
'Erroneous': 1,
'VESSEL': 1,
'oversight': 1,
'fought': 4,
'match': 9,
'seer': 2,
'winsome': 2,
'twos': 3,
'mended': 2,
'rag': 2,
'purposes': 3,
'Improving': 1,
'tying': 2,
'attentively': 5,
'maxim': 3,
'alleged': 4,
'background': 3,
'junks': 2,
'Dough': 16,
'Verdes': 1,
'slain': 12,
'palms': 15,
'Walking': 1,
'helped': 15,
'felt': 41,
'pace': 7,
'Eight': 3,
'surgeons': 2,
'relative': 2,
'garters': 1,
'goadings': 1,
'classification': 4,
'loiters': 1,
'wand': 2,
'rides': 2,
'vicinities': 1,
'vocation': 11,
'Smithfield': 1,
```

```
'reclines': 1,
'topic': 1,
'67': 1,
'hues': 3,
'attends': 1,
'allies': 4,
'exert': 2,
'tint': 1,
'decapitating': 1,
'Epitome': 1,
'averred': 2,
'unsettled': 1,
'Well': 56,
'happening': 4,
'STANDING': 3,
'D': 23,
'freely': 11,
'successive': 5,
'Three': 8,
'triumphs': 1,
'PURCHAS': 1,
'vernal': 2,
'send': 12,
'dams': 2,
'characteristically': 1,
'venture': 5,
'Julys': 1,
'serried': 1,
'overswarm': 1,
'sensibly': 1,
'revulsion': 1,
'labors': 2,
'oceans': 14,
'attend': 9,
'attack': 8,
'passenger': 6,
'stumbled': 1,
'gobbles': 1,
'indissoluble': 1,
'tainted': 1,
'glancing': 12,
'wood': 34,
'outstretched': 4,
'dart': 24,
'ought': 8,
'unexpected': 2,
'singing': 11,
'predicted': 1,
'densely': 1,
'embraced': 4,
'page': 8,
'MEANTIME': 1,
'lunacy': 3,
```

```
'seem': 85,
'milkiness': 1,
'remembrances': 1,
'Seas': 7,
'abstemious': 1,
'humored': 1,
'aghast': 4,
'indolent': 6,
'Blocksburg': 1,
'Thomas': 1,
'hero': 3,
'plain': 37,
'collapsed': 4,
'kind': 27,
'billows': 20,
'carcass': 1,
'bulky': 3,
'blindly': 7,
'Soothed': 1,
'curious': 53,
'tauntings': 1,
'shocks': 2,
'last': 277,
'oppose': 2,
'pointed': 29,
'didst': 8,
'Cherries': 1,
'troubled': 15,
'comber': 1,
'Tis': 7,
'raked': 1,
'home': 56,
'colder': 1,
'vases': 1,
'wick': 1,
'oft': 2,
'fictitiously': 1,
'museum': 1,
'morrow': 12,
'ignominious': 1,
'test': 4,
'Aside': 1,
'Usher': 2,
'faster': 4,
'prevailed': 3,
'vibrating': 6,
'amount': 6,
'weighing': 3,
'sufficit': 1,
'wordless': 1,
'blueness': 1,
'bamboo': 1,
'vision': 5,
```

```
'leaped': 13,
'23': 1,
'protection': 4,
'drags': 4,
'dam': 18,
'salutation': 3,
'mates': 39,
'Sunset': 1,
'Whalemen': 1,
'facilitate': 1,
'daughters': 4,
'"...': 1,
'codfish': 1,
'Delta': 1,
'wolves': 2,
'Capting': 4,
'reverberations': 1,
'scruples': 1,
'screamed': 1,
'omens': 2,
'ATTACK': 1,
'maintained': 5,
'casting': 6,
'dishes': 1,
'surely': 2,
'BLACK': 2,
'Tom': 3,
'peasants': 1,
'May': 10,
'flag': 18,
'poetic': 3,
'ay': 1,
'First': 24,
'tantalizing': 3,
'immaculate': 3,
'wee': 1,
'deliberated': 1,
'moose': 2,
'greenness': 3,
'unfailing': 1,
'cursing': 1,
'frothed': 1,
'Rhyme': 2,
'perfectly': 6,
'drinks': 2,
'thinkers': 1,
'confining': 1,
'purplish': 3,
'pile': 7,
'ropes': 13,
'invitingly': 1,
'refuses': 1,
'untouched': 3,
```

```
'Straits': 5,
'ribby': 1,
'symbolize': 2,
'ominous': 3,
'confinement': 2,
'immensity': 1,
'illustrate': 1,
'lulls': 1,
'stepped': 8,
'women': 11,
'obscurity': 1,
'invite': 1,
'guardian': 1,
'naked': 12,
'Outward': 1,
'branches': 5,
'meets': 3,
'father': 16,
'Coleman': 1,
'exhumed': 1,
'empties': 1,
'speedy': 5,
'shiningly': 1,
'Physiognomy': 1,
'hoofs': 3,
'rafted': 1,
'astronomical': 1,
'unstricken': 1,
'swimmer': 2,
'goes': 54,
'lasting': 1,
'warmer': 1,
'calculated': 7,
'Desolation': 1,
'privilege': 7,
'Step': 1,
'supposes': 1,
'mad': 36,
'snuffing': 3,
'meanwhile': 12,
'hoot': 1,
'Frederick': 4,
'FIFE': 1,
'unsurrendered': 1,
'pirouetting': 1,
'fragments': 3,
'unrighteous': 1,
'filliping': 1,
'streamed': 3,
'rehearsing': 2,
'TALBOT': 1,
'sanctuary': 1,
'asked': 14,
```

```
'promiscuously': 1,
'dictionary': 2,
'BEALE': 1,
'drops': 6,
'lactantem': 1,
'drug': 1,
'summit': 10,
'beached': 1,
'money': 12,
'discreditably': 1,
'afterwards': 24,
'country': 25,
'utter': 6,
'bruise': 1,
'unrifled': 1,
'blazing': 7,
'banqueter': 1,
'OPEN': 1,
'non': 2,
'humbly': 2,
'Societies': 1,
'gaunt': 5,
'capsizings': 1,
'summers': 2,
'minutest': 3,
'Persians': 1,
'author': 6,
'necessary': 7,
'crows': 2,
'Draws': 1,
'Napoleon': 3,
'Humpback': 1,
'woes': 1,
'ergo': 2,
'ripe': 6,
'route': 3,
'dilated': 2,
'coils': 9,
'EDMUND': 2,
'faculties': 1,
'compass': 17,
'unrolled': 1,
'suspects': 3,
'Thunder': 8,
'roared': 13,
'Bennett': 2,
'Tropics': 1,
'inquiring': 5,
'cried': 155,
'toilings': 2,
'spot': 23,
'undetected': 1,
'repentant': 1,
```

```
'yielded': 2,
'angle': 7,
'popular': 8,
'profession': 5,
'Scales': 1,
'portable': 3,
'SMITH': 1,
'incredulous': 5,
'elbow': 1,
'surging': 3,
'Blanco': 1,
'Savesoul': 2,
'antagonistic': 1,
'PRIMER': 1,
'wondrous': 41,
'controlling': 1,
'noted': 3,
'lamentable': 1,
'painful': 2,
'uncomfortableness': 1,
'reign': 3,
'straits': 11,
'veer': 1,
'get': 92,
'swearing': 4,
'shrines': 1,
'Tranque': 5,
'snap': 7,
'galliots': 1,
'appeals': 2,
'steadfastly': 6,
'rang': 1,
'come': 150,
'marble': 14,
'kings': 14,
'pagans': 5,
'secreted': 1,
'festival': 1,
'wolfish': 2,
'waiting': 3,
'felicities': 1,
'Stir': 1,
'invested': 21,
'robbed': 2,
'VISIT': 1,
'brothers': 2,
'slackened': 6,
'amen': 1,
'blindest': 1,
'pulled': 13,
'fore': 23,
'legal': 2,
'ears': 26,
```

```
'beset': 2,
'injunctions': 1,
'133': 1,
'maze': 2,
'crawled': 2,
'shore': 23,
'ASIDE': 4,
'comical': 5,
'asses': 1,
'globules': 2,
'submission': 1,
'persuading': 1,
',*': 3,
'capricious': 3,
'heels': 2,
'Western': 1,
'magnificence': 1,
'allaying': 1,
'Chilian': 2,
'arrangement': 3,
'disport': 1,
'seas': 80,
'nevertheless': 23,
'hazard': 3,
'Much': 5,
'allusions': 7,
'inducements': 2,
'elephants': 10,
'Thank': 3,
'mutineers': 1,
'propelled': 1,
'Yea': 5,
'floundered': 2,
'passports': 1,
'Ground': 4,
'horribles': 1,
'Actium': 1,
'latitudes': 15,
'faithfulness': 2,
'consuming': 3,
'risk': 5,
'Round': 3,
'Enderbys': 1,
'abided': 1,
'undress': 1,
'bluer': 1,
'tombstones': 1,
'crags': 1,
'Jenny': 3,
'Cachalot': 3,
'prophesies': 1,
'timorous': 1,
'heave': 16,
```

```
'manage': 2,
'narrating': 2,
'blade': 8,
'procure': 2,
'political': 2,
'hoop': 1,
'wish': 13,
'travels': 3,
'wandereth': 1,
'groove': 1,
'imperative': 1,
'elegant': 2,
'forgotten': 9,
'excellently': 1,
'grinding': 1,
'disrated': 1,
'mortal': 39,
'poncho': 1,
'maintaining': 3,
'alarms': 2,
'fourths': 2,
'ducat': 1,
'gambol': 1,
'tailed': 3,
'canted': 3,
'twain': 6,
'nailest': 1,
'Perhaps': 7,
'belie': 2,
'salamander': 1,
'enemies': 2,
'dolphin': 4,
'device': 2,
'anointed': 2,
'summoned': 4,
'energies': 1,
'felled': 1,
'CHRONICLER': 1,
'altitude': 3,
'THOMAS': 3,
'Pacifics': 2,
'homeward': 6,
'contracted': 4,
'unbecomingness': 1,
'packed': 5,
'passionateness': 1,
'chaps': 5,
'walruses': 1,
'BROWN': 1,
'hammock': 33,
'nimble': 2,
'infantileness': 1,
'Crusaders': 1,
```

```
'suggestions': 2,
'marriage': 2,
'genial': 4,
'Mapple': 9,
'retreated': 4,
'bars': 2,
'tawny': 4,
'watched': 5,
'unbelief': 1,
'tormenting': 2,
'awry': 1,
'eve': 4,
'pocketing': 1,
'squash': 1,
'inequality': 1,
'1807': 2,
'bashful': 1,
'shared': 2,
'Timor': 2,
'skelter': 2,
'canal': 12,
'COMES': 1,
'facilitating': 1,
'tellin': 1,
'conception': 2,
'Rope': 2,
'ballroom': 1,
'plugged': 1,
'departments': 2,
'strenuous': 1,
'jammed': 3,
'Sermon': 1,
'.--"': 1,
'taking': 50,
'wandered': 1,
'battalions': 2,
'latest': 3,
'Mills': 1,
'plumes': 1,
'block': 11,
'goats': 1,
'Sheet': 1,
'luck': 13,
'defyingly': 2,
'harems': 1,
'TRUE': 1,
'swaller': 1,
'rumours': 1,
'drawbacks': 1,
'Forecastle': 2,
'WHALEMAN': 1,
'crane': 3,
'PREFACE': 1,
```

```
'sullenly': 4,
'almighty': 2,
'dawning': 1,
'conceives': 1,
'tip': 3,
'everythings': 1,
'tellest': 2,
'timber': 6,
'show': 32,
'extras': 1,
'graven': 1,
'--"': 95,
'tokens': 5,
'rostrated': 1,
'narrative': 8,
'--\'"': 1,
'Angelo': 1,
'section': 1,
'capture': 14,
'beards': 3,
'rigid': 2,
'gestures': 4,
'insensible': 1,
'wooden': 26,
'easy': 38,
'Depend': 1,
'rescue': 5,
'eatable': 1,
'fiend': 8,
'cones': 1,
'abbreviation': 1,
'scornful': 3,
'multiplied': 2,
'consistency': 1,
'afoam': 1,
'stake': 2,
'Venetian': 4,
'assailing': 1,
'SUMMER': 1,
'Titans': 1,
'ulceration': 1,
'spoiled': 4,
'83': 1,
'beaks': 2,
'farces': 1,
'Krusenstern': 2,
'artful': 1,
'Nathan': 1,
'freebooters': 1,
'propensity': 2,
'dentistical': 1,
'126': 1,
'cooling': 1,
```

```
'bleed': 1,
'Crack': 2,
'overspread': 1,
'flies': 7,
'array': 3,
'thrown': 25,
'madden': 1,
'intently': 9,
'allude': 2,
'Sebastian': 9,
'thundered': 1,
'Holland': 5,
'brand': 2,
'dust': 10,
'one': 889,
'ached': 1,
'concussions': 1,
'peppered': 1,
'homes': 1,
'Saturday': 5,
'unharmed': 5,
'dumpling': 1,
'commodores': 1,
'haze': 2,
'Quohog': 9,
'dreams': 8,
'willed': 2,
'tiled': 1,
'Flukes': 1,
'ladders': 1,
'200th': 1,
'Son': 3,
'musically': 1,
'amounted': 1,
'blows': 23,
'noise': 12,
'Back': 8,
'foreshortened': 1,
'cracks': 5,
'swamping': 1,
'Bays': 1,
'probed': 1,
'drilled': 1,
'streams': 4,
'animate': 1,
'legislative': 1,
'syllable': 2,
'spins': 1,
'East': 10,
'staid': 2,
'subservient': 1,
'traits': 2,
'like': 624,
```

```
'traps': 2,
'Australian': 1,
'debel': 1,
'incidents': 1,
'OF': 49,
'peculiarity': 7,
'Receiving': 2,
'infinite': 8,
'Coast': 2,
'wanted': 9,
'quarters': 5,
'candle': 7,
'Winds': 1,
'saddle': 1,
'bowed': 10,
'metropolis': 1,
'savageness': 1,
'jingle': 1,
'Saul': 1,
'twists': 1,
'case': 69,
'SIZED': 1,
'may': 230,
'warmth': 3,
'brook': 2,
'Boy': 15,
'Leaning': 1,
'pine': 9,
'ravens': 1,
'panelled': 2,
'supplants': 1,
'multitudinous': 2,
'confoundedly': 1,
'sung': 2,
'seethe': 3,
'shinbones': 2,
'COILS': 1,
'snake': 2,
'mint': 1,
'enemy': 5,
'Quito': 2,
'dies': 6,
'Dauphine': 1,
'vanish': 1,
'absorbing': 2,
'endless': 15,
'riding': 3,
'halt': 1,
'elasticity': 3,
'notched': 1,
'Hampshire': 3,
'perpendicularly': 8,
'interregnum': 1,
```

```
'plates': 7,
'charge': 10,
'disastrous': 3,
'Presbyterians': 1,
'vero': 1,
'nasty': 1,
'proclaimed': 1,
'presuming': 1,
'Kentucky': 1,
'admirably': 2,
'Dinting': 1,
'extorting': 1,
'drawings': 4,
'Caesarian': 1,
'crusts': 1,
'Maker': 1,
'toothless': 1,
'begun': 4,
'greybeards': 1,
'attainable': 1,
'courageous': 2,
'circulate': 1,
'Wild': 2,
'lanes': 1,
'an': 582,
'accountable': 1,
'door': 45,
'regardless': 1,
'sanguinary': 1,
'clothes': 11,
'approaches': 1,
'stolen': 3,
'unconquering': 1,
'inactive': 1,
'vagueness': 1,
'hanging': 19,
'perfidious': 2,
'gabled': 1,
'hatches': 10,
'Better': 3,
'concentric': 3,
'Careful': 2,
'triune': 1,
'counterpart': 3,
'tub': 14,
'motives': 4,
'research': 4,
'conjure': 1,
'sceptical': 2,
'coming': 50,
'outfits': 2,
'weakling': 1,
'experimental': 1,
```

```
'thro': 1,
'pony': 1,
'cables': 4,
'others': 37,
'2ND': 1,
'visitors': 3,
'believers': 1,
'Cretan': 2,
'savor': 4,
'casement': 1,
'incommunicable': 3,
'duskier': 1,
'EZEKIEL': 1,
'Herod': 1,
'Run': 3,
'posture': 5,
'cause': 22,
'appellative': 2,
'Paean': 1,
'obvious': 11,
'maketh': 5,
'methodization': 1,
'muttering': 4,
'Further': 2,
'edifices': 1,
'Santa': 2,
'Lamatins': 1,
'booming': 2,
'commands': 3,
'illumination': 1,
'indebted': 1,
'viciously': 2,
'prow': 21,
'sped': 3,
'raises': 6,
'excitedly': 1,
'1652': 1,
'103': 1,
'utterly': 13,
'FIRMLY': 1,
'phenomenon': 5,
'Mufti': 1,
'SKY': 1,
'flowered': 1,
'tens': 3,
'writ': 2,
'ANOTHER': 1,
'Senators': 1,
'Arch': 2,
'Tash': 5,
'presto': 2,
'unanswerable': 1,
'design': 4,
```

```
'decks': 21,
'live': 60,
'Tormentoto': 1,
'reveal': 4,
'hen': 3,
'influenced': 1,
'unbidden': 2,
'snatch': 3,
'sullen': 6,
'Weep': 1,
'superlative': 2,
'primeval': 1,
'judging': 2,
'shriek': 2,
'oneself': 1,
'protect': 2,
'suggest': 1,
'mutters': 2,
'command': 28,
'horizontal': 15,
'intertangled': 1,
'respects': 8,
'forewarnings': 1,
'grandmother': 1,
'THEY': 3,
'fortune': 2,
'ourselves': 15,
'frost': 8,
'Iroquois': 2,
'induced': 4,
'examining': 1,
...})
```

Verifying validity of the importes corpus through concordance

In [4]:
```
text1.concordance("monstrous")
```

```
Displaying 11 of 11 matches:
ong the former , one was of a most monstrous size . ... This came towa
rds us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork
we have r
ll over with a heathenish array of monstrous clubs and spears . Some w
ere thick
d as you gazed , and wondered what monstrous cannibal and savage could
ever hav
that has survived the flood ; most monstrous and most mountainous ! Th
at Himmal
they might scout at Moby Dick as a monstrous fable , or still worse an
d more de
th of Radney .'" CHAPTER 55 Of the Monstrous Pictures of Whales . I sh
all ere l
ing Scenes . In connexion with the monstrous pictures of whales , I am
strongly
ere to enter upon those still more monstrous stories of them which are
to be fo
ght have been rummaged out of this monstrous cabinet there is no telli
ng . But
of Whale - Bones ; for Whales of a monstrous size are oftentimes cast
up dead u
```

In [5]:
```
len(text1)
```

Out[5]:
```
260819
```

In [6]:
```
fdist1 = FreqDist(text1)
```

In [7]:
```
print(fdist1)
```

```
<FreqDist with 19317 samples and 260819 outcomes>
```

Finding the most common 50 most used words in the corpus

In [8]:
```
fdist1.most_common(50)
```

Out[8]:
```
[(',', 18713),
 ('the', 13721),
 ('.', 6862),
 ('of', 6536),
 ('and', 6024),
 ('a', 4569),
 ('to', 4542),
 (';', 4072),
 ('in', 3916),
 ('that', 2982),
 ("'", 2684),
 ('-', 2552),
 ('his', 2459),
 ('it', 2209),
 ('I', 2124),
 ('s', 1739),
 ('is', 1695),
 ('he', 1661),
 ('with', 1659),
 ('was', 1632),
 ('as', 1620),
 ('"', 1478),
 ('all', 1462),
 ('for', 1414),
 ('this', 1280),
 ('!', 1269),
 ('at', 1231),
 ('by', 1137),
 ('but', 1113),
 ('not', 1103),
 ('--', 1070),
 ('him', 1058),
 ('from', 1052),
 ('be', 1030),
 ('on', 1005),
 ('so', 918),
 ('whale', 906),
 ('one', 889),
 ('you', 841),
 ('had', 767),
 ('have', 760),
 ('there', 715),
 ('But', 705),
 ('or', 697),
 ('were', 680),
 ('now', 646),
 ('which', 640),
 ('?', 637),
 ('me', 627),
 ('like', 624)]
```

Finding most common bigrams

In [9]:
```
text1.collocations(100)
```

```
Sperm Whale; Moby Dick; White Whale; old man; Captain Ahab; sperm
whale; Right Whale; Captain Peleg; New Bedford; Cape Horn; cried Ahab;
years ago; lower jaw; never mind; Father Mapple; cried Stubb; chief
mate; white whale; ivory leg; one hand; thou art; would fain; well
known; cried Starbuck; forty years; 000 lbs; Good Hope; Captain
Bildad; thus far; must needs; Samuel Enderby; New Zealand; seven
hundred; Heidelburgh Tun; whaling voyage; said Stubb; would seem; one
side; dost thou; three years; South Sea; every one; three days; good
deal; something like; ever since; upper part; present day; steering
oar; Deacon Deuteronomy; New England; young man; Greenland whale;
centuries ago; SPERM WHALE; ere long; Thou art; one knows; thousand
miles; Sperm Whales; Aunt Charity; thou hast; New York; three boats;
art thou; good luck; Frederick Cuvier; poor Queequeg; five feet; four
boats; poor fellow; Captain Sleet; drew nigh; may possibly; Huzza
Porpoise; slouched hat; four years; closed eyes; Give way; NANTUCKET
SAILOR; Low Dutch; Saturday night; sperm whales; Sag Harbor; Eight
bells; OLD MANX; drawing nigh; feet long; Sir Clifford; twenty feet;
inclined plane; one hundred; Ahab stood; mortal man; little negro;
open air; Lord Warden; one single; full grown; Pequod Meets
```

In [10]:
```
fdist1 = FreqDist(text1);
top25 = fdist1.most_common(25)
```

In [11]:
```
bigrams = (tuple(nltk.bigrams(text1,pad_left=True, pad_right=True)))
```

In [12]:
```
bigr = nltk.bigrams(text1,pad_left=True, pad_right=True)
```

Finding frequency distribution for all bigrams in the language

In [13]:
```
freq_dist = nltk.FreqDist(bigr)
freq_dist
```

Out[13]:
```
FreqDist({('is', 'true'): 5,
          ('comrade', 'on'): 1,
          ('singularly', 'common'): 1,
          ('mediocrity', 'in'): 1,
          ('do', 'entreat'): 1,
          ('noticed', 'the'): 1,
          ('present', 'hour'): 1,
          ('it', 'they'): 1,
          ('--', 'head'): 1,
          ('INTO', 'ASIA'): 1,
          ('we', 'set'): 1,
          ('-', 'kettles'): 1,
          ('father', "'"): 6,
          ('mate', 'there'): 1,
          ('thumb', ';'): 1,
          ('jiffy', ';"'): 1,
          ("'", 'so'): 1,
          ('himself', 'was'): 4,
          ('ding', ','): 1,
```

In [14]:
```
frequencyDist = nltk.ConditionalFreqDist(bigrams)
```

In [15]:
```
frequencyDist
```

Out[15]:
```
ConditionalFreqDist(nltk.probability.FreqDist,
                    {'moon': FreqDist({',': 4,
                              '-': 1,
                              '.': 1,
                              'shows': 1,
                              'why': 1,
                              'would': 1}),
                     'horse': FreqDist({"'": 1,
                              ',': 4,
                              '-': 9,
                              '.': 4,
                              ';': 2,
                              'but': 1,
                              'he': 1,
                              'instead': 1,
                              'is': 2,
                              'walks': 1}),
                     'angels': FreqDist({',': 3,
                              '.': 1,
```

In [16]:
```
frequencyDist["Whale"].keys()
```

Out[16]:
```
dict_keys(['should', 'were', 'Watch', 'which', 'as', 'rolls', 'so', 'h
ave', 'be', 'Fisheries', 'into', '!"', 'lay', 'did', '--', ',"', 'at',
'grounds', 'to', 'presents', 'fishery', ')', ':', 'Fleet', 'will', 'fl
oats', 'than', 'strains', 'shall', 'HAS', 'was', 'gives', ',', 'only',
'had', '"', 'now', 'really', "'", '?"', 'swam', 'made', 'dashed', 'fis
hermen', 'might', ';', 'alongside', '-', 'the', 'Fishery', 'close', 'i
n', 'been', 'are', 'when', 'I', 'of', 'cruising', 'not', 'Cruising', '
spouts', 'ever', 'its', 'churning', 'drawings', '!', 'is', 'has', '.',
'Porpoise', 'embraces', 'designated', 'must', 'thus', '?--', 'first',
'stove', 'sometimes', '?', 'darted', 'tossed', 'anywhere', 'largely',
'and', 'can', 'most', 'blows', 'on', "!'", 'fully', 'found'])
```

In [17]:
```
frequencyDist["cried"].most_common(20)
```

Out[17]:
```
[('Ahab', 32),
 ('Stubb', 23),
 ('the', 17),
 ('Starbuck', 17),
 ('out', 8),
 (',', 7),
 ('a', 7),
 ('Queequeg', 4),
 ('Peleg', 3),
 ('Daggoo', 3),
 ('Don', 3),
 ('Flask', 3),
 ('.', 2),
 ('to', 2),
 ('Bildad', 2),
 ('Steelkilt', 2),
 ('Captain', 2),
 ('--"', 1),
 ('all', 1),
 ('stationary', 1)]
```

Generating probabilistic distribution of the bigrams after smoothing

Smoothing by Conditional probability distribution using Laplace

In [20]:
```
probabilityDist = nltk.ConditionalProbDist(frequencyDist, nltk.LaplacePro
```

In [21]:
```
for i in probabilityDist.keys():
    for j in probabilityDist[i].samples():
        print ('{0}'.format(i) + " " + '{0}'.format(j) + " " + '{0}'.form
```

```
Joe , 7.668094210205467e-06
horse he 7.667359284788726e-06
horse is 1.1501038927183089e-05
horse - 3.8336796423943627e-05
horse . 1.9168398211971813e-05
horse ' 7.667359284788726e-06
horse instead 7.667359284788726e-06
horse , 1.9168398211971813e-05
horse ; 1.1501038927183089e-05
horse walks 7.667359284788726e-06
horse but 7.667359284788726e-06
angels that 7.667888416887757e-06
angels mobbing 7.667888416887757e-06
angels in 7.667888416887757e-06
angels . 7.667888416887757e-06
angels indeed 7.667888416887757e-06
angels , 1.5335776833775514e-05
SNEEZES )-- 1.5335776833775514e-05
SNEEZES ) 1.5335776833775514e-05
```

In [22]:
```
probabilityDist.conditions()
```

Out[22]:
```
['Joe',
 'horse',
 'angels',
 'SNEEZES',
 'blind',
 'Touching',
 'which',
 'IBID',
 'traitors',
 'truest',
 'pilau',
 'Friesland',
 'IV',
 'lantern',
 'Gaining',
 'Crappoes',
 'interfusing',
 'envelope',
 'monotonously',
```

In [23]:
```
probabilityDist["whale"].samples()
```

Out[23]:
```
dict_keys(['should', 'they', 'which', 'all', 'as', 'author', 'also', '
so', 'bone', 'have', 'hunters', 'be', 'must', 'flashes', 'tore', 'more
', 'draughtsmen', 'then', 'carries', '!"', 'round', 'or', 'by', 'windi
ng', 'differ', 'principal', ',"', 'statements', 'possibly', 'agent', '
now', 'fishery', 'cannot', 'yet', 'obliquely', 'struck', 'bears', 'alo
ne', 'rushed', 'than', 'contains', 'shall', 'yaw', 'book', 'dead', 'co
mmanders', 'came', 'he', 'little', 'tribe', 'relaxed', '"', 'wheeled',
'really', 'immortal', '?"', 'does', 'dallied', 'supplies', '),', 'migh
t', 'alongside', 'upon', 'the', 'was', 'that', 'belong', 'in', 'wounde
d', 'I', 'mentioned', 'oil', 'thus', 'being', 'obtains', 'boat', 'ceme
teries', 'vertically', 'till', 'it', 'stays', 'broke', 'drew', '!', 'a
ttacked', 'is', 'has', '.', 'shoots', 'surgeons', 'eat', 'line', 'hunt
er', 'his', 'myself', 'there', 'whose', 'sunwards', 'shakes', 'when',
"'", 'like', 'referred', 'for', 'plunged', 'looks', 'sideways', '.*',
'beyond', 'captains', 'authors', 'did', '(', '!--', 'may', 'would', '-
-', 'could', 'face', 'swimming', 'furnishes', 'host', 'will', 'fleet',
'always', 'at', 'rose', 'head', 'precisely', 'with', 'starts', 'previo
usly', 'a', ',', 'only', 'had', 'eluded', 'soon', 'taken', 'goes', 'fl
ew', 'grounded', 'something', 'affords', 'hunt', 'to', ';', '-', 'seem
s', 'towing', 'spout', 'and', 'are', 'of', 'betakes', 'can', 'seemed',
'before', 'remains', 'ship', 'somewhat', 'almost', 'looked', 'within',
'belonged', 'became', 'ships', 'just', 'once', 'averages', 'rolling',
'eye', 'abated', 'lies', 'keeps', 'both', 'escaping', 'fell', 'sometim
es', 'shed', '?', 'ran', 'started', 'from', 'research', 'stranded', 'n
ot', 'himself', 'heave', 'ivory', '."', 'on', 'outlast', 'again', 'cau
ght'])
```

Generating Laplace smoothed corpus through NLTK.LaplaceProbDist

In [24]:
```
laplaceProbabilityDist = nltk.LaplaceProbDist(freq_dist, bins=freq_dist.N
```

In [25]:
```
bins=freq_dist.B()
bins
```

Out[25]:
```
118750
```

In [26]:
```
samples = laplaceProbabilityDist.samples()
```

In [27]:
```
listOfSamples = list(laplaceProbabilityDist.samples())
```

In [28]:
```
laplaceProbabilityDist.max()
```

Out[28]:
```
(',', 'and')
```

In [29]:
```
(',', 'and') in laplaceProbabilityDist.samples()
```

Out[29]:
```
True
```

In [30]:
```
d = dict(laplaceProbabilityDist.samples())
```

In [31]:
```
d['s']
```

Out[31]:
```
'sail'
```

In [32]:
```
laplaceProbabilityDist
```

Out[32]:
```
<LaplaceProbDist based on 260820 samples>
```

In [33]:
```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(ngram_range=(2,2))
analyzer = vectorizer.build_analyzer()
analyzer(str(list(text1)))
```

Out[33]:
```
['moby dick',
 'dick by',
 'by herman',
 'herman melville',
 'melville 1851',
 '1851 etymology',
 'etymology supplied',
 'supplied by',
 'by late',
 'late consumptive',
 'consumptive usher',
 'usher to',
 'to grammar',
 'grammar school',
 'school the',
 'the pale',
 'pale usher',
 'usher threadbare',
 'threadbare in',
 'in coat',
 'coat heart',
 'heart body',
 'body and',
 'and brain',
 'brain see',
 'see him',
 'him now',
 'now he',
 'he was',
 'was ever',
 'ever dusting',
 'dusting his',
 'his old',
```

```
'old lexicons',
'lexicons and',
'and grammars',
'grammars with',
'with queer',
'queer handkerchief',
'handkerchief mockingly',
'mockingly embellished',
'embellished with',
'with all',
'all the',
'the gay',
'gay flags',
'flags of',
'of all',
'all the',
'the known',
'known nations',
'nations of',
'of the',
'the world',
'world he',
'he loved',
'loved to',
'to dust',
'dust his',
'his old',
'old grammars',
'grammars it',
'it somehow',
'somehow mildly',
'mildly reminded',
'reminded him',
'him of',
'of his',
'his mortality',
'mortality while',
'while you',
'you take',
'take in',
'in hand',
'hand to',
'to school',
'school others',
'others and',
'and to',
'to teach',
'teach them',
'them by',
'by what',
'what name',
'name whale',
'whale fish',
```

```
'fish is',
'is to',
'to be',
'be called',
'called in',
'in our',
'our tongue',
'tongue leaving',
'leaving out',
'out through',
'through ignorance',
'ignorance the',
'the letter',
'letter which',
'which almost',
'almost alone',
'alone maketh',
'maketh the',
'the signification',
'signification of',
'of the',
'the word',
'word you',
'you deliver',
'deliver that',
'that which',
'which is',
'is not',
'not true',
'true hackluyt',
'hackluyt whale',
'whale sw',
'sw and',
'and dan',
'dan hval',
'hval this',
'this animal',
'animal is',
'is named',
'named from',
'from roundness',
'roundness or',
'or rolling',
'rolling for',
'for in',
'in dan',
'dan hvalt',
'hvalt is',
'is arched',
'arched or',
'or vaulted',
'vaulted webster',
'webster dictionary',
```

```
'dictionary whale',
'whale it',
'it is',
'is more',
'more immediately',
'immediately from',
'from the',
'the dut',
'dut and',
'and ger',
'ger wallen',
'wallen walw',
'walw ian',
'ian to',
'to roll',
'roll to',
'to wallow',
'wallow richardson',
'richardson dictionary',
'dictionary ketos',
'ketos greek',
'greek cetus',
'cetus latin',
'latin whoel',
'whoel anglo',
'anglo saxon',
'saxon hvalt',
'hvalt danish',
'danish wal',
'wal dutch',
'dutch hwal',
'hwal swedish',
'swedish whale',
'whale icelandic',
'icelandic whale',
'whale english',
'english baleine',
'baleine french',
'french ballena',
'ballena spanish',
'spanish pekee',
'pekee nuee',
'nuee nuee',
'nuee fegee',
'fegee pekee',
'pekee nuee',
'nuee nuee',
'nuee erromangoan',
'erromangoan extracts',
'extracts supplied',
'supplied by',
'by sub',
'sub sub',
```

```
'sub librarian',
'librarian it',
'it will',
'will be',
'be seen',
'seen that',
'that this',
'this mere',
'mere painstaking',
'painstaking burrower',
'burrower and',
'and grub',
'grub worm',
'worm of',
'of poor',
'poor devil',
'devil of',
'of sub',
'sub sub',
'sub appears',
'appears to',
'to have',
'have gone',
'gone through',
'through the',
'the long',
'long vaticans',
'vaticans and',
'and street',
'street stalls',
'stalls of',
'of the',
'the earth',
'earth picking',
'picking up',
'up whatever',
'whatever random',
'random allusions',
'allusions to',
'to whales',
'whales he',
'he could',
'could anyways',
'anyways find',
'find in',
'in any',
'any book',
'book whatsoever',
'whatsoever sacred',
'sacred or',
'or profane',
'profane therefore',
'therefore you',
```

```
'you must',
'must not',
'not in',
'in every',
'every case',
'case at',
'at least',
'least take',
'take the',
'the higgledy',
'higgledy piggledy',
'piggledy whale',
'whale statements',
'statements however',
'however authentic',
'authentic in',
'in these',
'these extracts',
'extracts for',
'for veritable',
'veritable gospel',
'gospel cetology',
'cetology far',
'far from',
'from it',
'it as',
'as touching',
'touching the',
'the ancient',
'ancient authors',
'authors generally',
'generally as',
'as well',
'well as',
'as the',
'the poets',
'poets here',
'here appearing',
'appearing these',
'these extracts',
'extracts are',
'are solely',
'solely valuable',
'valuable or',
'or entertaining',
'entertaining as',
'as affording',
'affording glancing',
'glancing bird',
'bird eye',
'eye view',
'view of',
'of what',
```

```
'what has',
'has been',
'been promiscuously',
'promiscuously said',
'said thought',
'thought fancied',
'fancied and',
'and sung',
'sung of',
'of leviathan',
'leviathan by',
'by many',
'many nations',
'nations and',
'and generations',
'generations including',
'including our',
'our own',
'own so',
'so fare',
'fare thee',
'thee well',
'well poor',
'poor devil',
'devil of',
'of sub',
'sub sub',
'sub whose',
'whose commentator',
'commentator am',
'am thou',
'thou belongest',
'belongest to',
'to that',
'that hopeless',
'hopeless sallow',
'sallow tribe',
'tribe which',
'which no',
'no wine',
'wine of',
'of this',
'this world',
'world will',
'will ever',
'ever warm',
'warm and',
'and for',
'for whom',
'whom even',
'even pale',
'pale sherry',
'sherry would',
```

```
'would be',
'be too',
'too rosy',
'rosy strong',
'strong but',
'but with',
'with whom',
'whom one',
'one sometimes',
'sometimes loves',
'loves to',
'to sit',
'sit and',
'and feel',
'feel poor',
'poor devilish',
'devilish too',
'too and',
'and grow',
'grow convivial',
'convivial upon',
'upon tears',
'tears and',
'and say',
'say to',
'to them',
'them bluntly',
'bluntly with',
'with full',
'full eyes',
'eyes and',
'and empty',
'empty glasses',
'glasses and',
'and in',
'in not',
'not altogether',
'altogether unpleasant',
'unpleasant sadness',
'sadness give',
'give it',
'it up',
'up sub',
'sub subs',
'subs for',
'for by',
'by how',
'how much',
'much the',
'the more',
'more pains',
'pains ye',
'ye take',
```

'take to',
'to please',
'please the',
'the world',
'world by',
'by so',
'so much',
'much the',
'the more',
'more shall',
'shall ye',
'ye for',
'for ever',
'ever go',
'go thankless',
'thankless would',
'would that',
'that could',
'could clear',
'clear out',
'out hampton',
'hampton court',
'court and',
'and the',
'the tuileries',
'tuileries for',
'for ye',
'ye but',
'but gulp',
'gulp down',
'down your',
'your tears',
'tears and',
'and hie',
'hie aloft',
'aloft to',
'to the',
'the royal',
'royal mast',
'mast with',
'with your',
'your hearts',
'hearts for',
'for your',
'your friends',
'friends who',
'who have',
'have gone',
'gone before',
'before are',
'are clearing',
'clearing out',
'out the',

```
'the seven',
'seven storied',
'storied heavens',
'heavens and',
'and making',
'making refugees',
'refugees of',
'of long',
'long pampered',
'pampered gabriel',
'gabriel michael',
'michael and',
'and raphael',
'raphael against',
'against your',
'your coming',
'coming here',
'here ye',
'ye strike',
'strike but',
'but splintered',
'splintered hearts',
'hearts together',
'together there',
'there ye',
'ye shall',
'shall strike',
'strike unsplinterable',
'unsplinterable glasses',
'glasses extracts',
'extracts and',
'and god',
'god created',
'created great',
'great whales',
'whales genesis',
'genesis leviathan',
'leviathan maketh',
'maketh path',
'path to',
'to shine',
'shine after',
'after him',
'him one',
'one would',
'would think',
'think the',
'the deep',
'deep to',
'to be',
'be hoary',
'hoary job',
'job now',
```

'now the',
'the lord',
'lord had',
'had prepared',
'prepared great',
'great fish',
'fish to',
'to swallow',
'swallow up',
'up jonah',
'jonah jonah',
'jonah there',
'there go',
'go the',
'the ships',
'ships there',
'there is',
'is that',
'that leviathan',
'leviathan whom',
'whom thou',
'thou hast',
'hast made',
'made to',
'to play',
'play therein',
'therein psalms',
'psalms in',
'in that',
'that day',
'day the',
'the lord',
'lord with',
'with his',
'his sore',
'sore and',
'and great',
'great and',
'and strong',
'strong sword',
'sword shall',
'shall punish',
'punish leviathan',
'leviathan the',
'the piercing',
'piercing serpent',
'serpent even',
'even leviathan',
'leviathan that',
'that crooked',
'crooked serpent',
'serpent and',
'and he',

'he shall',
'shall slay',
'slay the',
'the dragon',
'dragon that',
'that is',
'is in',
'in the',
'the sea',
'sea isaiah',
'isaiah and',
'and what',
'what thing',
'thing soever',
'soever besides',
'besides cometh',
'cometh within',
'within the',
'the chaos',
'chaos of',
'of this',
'this monster',
'monster mouth',
'mouth be',
'be it',
'it beast',
'beast boat',
'boat or',
'or stone',
'stone down',
'down it',
'it goes',
'goes all',
'all incontinently',
'incontinently that',
'that foul',
'foul great',
'great swallow',
'swallow of',
'of his',
'his and',
'and perisheth',
'perisheth in',
'in the',
'the bottomless',
'bottomless gulf',
'gulf of',
'of his',
'his paunch',
'paunch holland',
'holland plutarch',
'plutarch morals',
'morals the',

```
'the indian',
'indian sea',
'sea breedeth',
'breedeth the',
'the most',
'most and',
'and the',
'the biggest',
'biggest fishes',
'fishes that',
'that are',
'are among',
'among which',
'which the',
'the whales',
'whales and',
'and whirlpooles',
'whirlpooles called',
'called balaene',
'balaene take',
'take up',
'up as',
'as much',
'much in',
'in length',
'length as',
'as four',
'four acres',
'acres or',
'or arpens',
'arpens of',
'of land',
'land holland',
'holland pliny',
'pliny scarcely',
'scarcely had',
'had we',
'we proceeded',
'proceeded two',
'two days',
'days on',
'on the',
'the sea',
'sea when',
'when about',
'about sunrise',
'sunrise great',
'great many',
'many whales',
'whales and',
'and other',
'other monsters',
'monsters of',
```

```
'of the',
'the sea',
'sea appeared',
'appeared among',
'among the',
'the former',
'former one',
'one was',
'was of',
'of most',
'most monstrous',
'monstrous size',
'size this',
'this came',
'came towards',
'towards us',
'us open',
'open mouthed',
'mouthed raising',
'raising the',
'the waves',
'waves on',
'on all',
'all sides',
'sides and',
'and beating',
'beating the',
'the sea',
'sea before',
'before him',
'him into',
'into foam',
'foam tooke',
'tooke lucian',
'lucian the',
'the true',
'true history',
'history he',
'he visited',
'visited this',
'this country',
'country also',
'also with',
'with view',
'view of',
'of catching',
'catching horse',
'horse whales',
'whales which',
'which had',
'had bones',
'bones of',
'of very',
```

```
'very great',
'great value',
'value for',
'for their',
'their teeth',
'teeth of',
'of which',
'which he',
'he brought',
'brought some',
'some to',
'to the',
'the king',
'king the',
'the best',
'best whales',
'whales were',
'were catched',
'catched in',
'in his',
'his own',
'own country',
'country of',
'of which',
'which some',
'some were',
'were forty',
'forty eight',
'eight some',
'some fifty',
'fifty yards',
'yards long',
'long he',
'he said',
'said that',
'that he',
'he was',
'was one',
'one of',
'of six',
'six who',
'who had',
'had killed',
'killed sixty',
'sixty in',
'in two',
'two days',
'days other',
'other or',
'or octher',
'octher verbal',
'verbal narrative',
'narrative taken',
```

'taken down',
'down from',
'from his',
'his mouth',
'mouth by',
'by king',
'king alfred',
'alfred 890',
'890 and',
'and whereas',
'whereas all',
'all the',
'the other',
'other things',
'things whether',
'whether beast',
'beast or',
'or vessel',
'vessel that',
'that enter',
'enter into',
'into the',
'the dreadful',
'dreadful gulf',
'gulf of',
'of this',
'this monster',
'monster whale',
'whale mouth',
'mouth are',
'are immediately',
'immediately lost',
'lost and',
'and swallowed',
'swallowed up',
'up the',
'the sea',
'sea gudgeon',
'gudgeon retires',
'retires into',
'into it',
'it in',
'in great',
'great security',
'security and',
'and there',
'there sleeps',
'sleeps montaigne',
'montaigne apology',
'apology for',
'for raimond',
'raimond sebond',
'sebond let',

```
'let us',
'us fly',
'fly let',
'let us',
'us fly',
'fly old',
'old nick',
'nick take',
'take me',
'me if',
'if is',
'is not',
'not leviathan',
'leviathan described',
'described by',
'by the',
'the noble',
'noble prophet',
'prophet moses',
'moses in',
'in the',
'the life',
'life of',
'of patient',
'patient job',
'job rabelais',
'rabelais this',
'this whale',
'whale liver',
'liver was',
'was two',
'two cartloads',
'cartloads stowe',
'stowe annals',
'annals the',
'the great',
'great leviathan',
'leviathan that',
'that maketh',
'maketh the',
'the seas',
'seas to',
'to seethe',
'seethe like',
'like boiling',
'boiling pan',
'pan lord',
'lord bacon',
'bacon version',
'version of',
'of the',
'the psalms',
'psalms touching',
```

```
'touching that',
'that monstrous',
'monstrous bulk',
'bulk of',
'of the',
'the whale',
'whale or',
'or ork',
'ork we',
'we have',
'have received',
'received nothing',
'nothing certain',
'certain they',
'they grow',
'grow exceeding',
'exceeding fat',
'fat insomuch',
'insomuch that',
'that an',
'an incredible',
'incredible quantity',
'quantity of',
'of oil',
'oil will',
'will be',
'be extracted',
'extracted out',
'out of',
'of one',
'one whale',
'whale ibid',
'ibid history',
'history of',
'of life',
'life and',
'and death',
'death the',
'the sovereignest',
'sovereignest thing',
'thing on',
'on earth',
'earth is',
'is parmacetti',
'parmacetti for',
'for an',
'an inward',
'inward bruise',
'bruise king',
'king henry',
'henry very',
'very like',
'like whale',
```

```
'whale hamlet',
'hamlet which',
'which to',
'to secure',
'secure no',
'no skill',
'skill of',
'of leach',
'leach art',
'art mote',
'mote him',
'him availle',
'availle but',
'but to',
'to returne',
'returne againe',
'againe to',
'to his',
'his wound',
'wound worker',
'worker that',
'that with',
'with lowly',
'lowly dart',
'dart dinting',
'dinting his',
'his breast',
'breast had',
'had bred',
'bred his',
'his restless',
'restless paine',
'paine like',
'like as',
'as the',
'the wounded',
'wounded whale',
'whale to',
'to shore',
'shore flies',
'flies thro',
'thro the',
'the maine',
'maine the',
'the faerie',
'faerie queen',
'queen immense',
'immense as',
'as whales',
'whales the',
'the motion',
'motion of',
'of whose',
```

```
'whose vast',
'vast bodies',
'bodies can',
'can in',
'in peaceful',
'peaceful calm',
'calm trouble',
'trouble the',
'the ocean',
'ocean til',
'til it',
'it boil',
'boil sir',
...]
```

Using laplaceProbabilityDist to generate smotthing for (a,a) type bigrams which don't exist in corpus

In [ ]:
```
d = {}
values=(set(text1))
for i in values:
    for j in values:
        temp = ()
        if(temp in freq_dist):
            d[temp] = freq_dist.get(temp) + 1;
        else:
            print(temp)
            d[temp] =1;
```

In [35]:
```
d = {}
sz=len(set(text1))
for i in top25:
    for j in top25:
        temp = (i[0],j[0])
        if(temp in freq_dist):
            d[temp] = laplaceProbabilityDist.prob(temp);
        else:
            d[temp] =1/(i[1] + sz)
```

In [36]:
```
d
```

Out[36]:
```
{('"', '"'): 4.8088482808367394e-05,
 ('"', "'"): 4.8088482808367394e-05,
 ('"', ','): 4.8088482808367394e-05,
 ('"', '-'): 4.8088482808367394e-05,
 ('"', '.'): 4.8088482808367394e-05,
 ('"', ';'): 4.8088482808367394e-05,
 ('"', 'I'): 0.0001763668430335097,
 ('"', 'a'): 7.668123610152596e-06,
 ('"', 'all'): 7.668123610152596e-06,
 ('"', 'and'): 1.533624722030519e-05,
 ('"', 'as'): 3.834061805076298e-06,
 ('"', 'for'): 4.8088482808367394e-05,
 ('"', 'he'): 2.4921401732995937e-05,
 ('"', 'his'): 4.8088482808367394e-05,
 ('"', 'in'): 7.668123610152596e-06,
 ('"', 'is'): 7.668123610152596e-06,
 ('"', 'it'): 9.585154512690744e-06,
 ('"', 'of'): 7.668123610152596e-06,
 ('"', 's'): 4.8088482808367394e-05,
 ('"', 'that'): 2.875546253807333e-05
```

In [37]:
```
d = {}
sz=len(set(text1))
for i in top25:
    for j in top25:
        temp = (i[0],j[0])
        if(temp in freq_dist):
            d[temp] = (freq_dist.get(temp) + 1)/(i[1] + sz);
        else:
            d[temp] =1/(i[1] + sz)
```

In [38]:
```
d
```

Out[38]:
```
{('"', '"'): 4.8088482808367394e-05,
 ('"', "'"): 4.8088482808367394e-05,
 ('"', ','): 4.8088482808367394e-05,
 ('"', '-'): 4.8088482808367394e-05,
 ('"', '.'): 4.8088482808367394e-05,
 ('"', ';'): 4.8088482808367394e-05,
 ('"', 'I'): 0.004424140418369801,
 ('"', 'a'): 0.00019235393123346957,
 ('"', 'all'): 0.00019235393123346957,
 ('"', 'and'): 0.00038470786246693915,
 ('"', 'as'): 9.617696561673479e-05,
 ('"', 'for'): 4.8088482808367394e-05,
 ('"', 'he'): 0.0006251502765087762,
 ('"', 'his'): 4.8088482808367394e-05,
 ('"', 'in'): 0.00019235393123346957,
 ('"', 'is'): 0.00019235393123346957,
 ('"', 'it'): 0.00024044241404183698,
 ('"', 'of'): 0.00019235393123346957,
 ('"', 's'): 4.8088482808367394e-05,
 ('"', 'that'): 0.00721327242125511
```

In [39]:
```
temp = ('all','of')
freq_dist.get(temp)
```

Out[39]:
```
26
```

In [40]:
```
(freq_dist.get(temp)+1)/(fdist1.get('all')+len(set(text1)))
```

Out[40]:
```
0.0012993888060060638
```

In [41]:
```
from collections import defaultdict

nesteddict = defaultdict(dict)
for key,value in d.items():
     names = key
     name1 = names[0]
     name2 = names[1]
     nesteddict[name1][name2] = value
```

Applying laplace smoothing of text using imperical formulas

In [42]:
```
import numpy
mat = numpy.zeros((25, 25))
row=0
col=0
matrix = []
for i in top25:
    matrow = []
    for j in top25:
        matrow.append(d.get((i[0],j[0])))
        col = col + 1
    matrix.append(matrow)
    row =  row + 1
```

In [43]:
```
len(top25)
```

Out[43]:
```
25
```

In [44]:
```
matrix
```

Out[44]:
```
[[2.6295030239284775e-05,
  0.02390218248750986,
  2.6295030239284775e-05,
  0.0016565869050749408,
  0.0685774388640547,
  0.005101235866421246,
  0.005916381803839074,
  2.6295030239284775e-05,
  0.010596897186431765,
  0.015382592689981593,
  0.000815145937417828,
  2.6295030239284775e-05,
  0.0020247173284249275,
  0.006284512227189061,
  0.012148303970549567,
  2.6295030239284775e-05,
  0.0037864843544570077,
  0.011753878516960295,
  0.0068630028924533265,
```

Outputting the 25 most commonly used bigrams to csv file

In [45]:
```
import csv
with open('dict2.csv', 'w') as csv_file:
    writer = csv.writer(csv_file)
    for i in range(0,25,1):
        writer.writerow(matrix[i])
```