Data 620 - Week 10

April 12, 2024

1 Assignment 10

Shamecca Marshall

2 Project Overview

It can be useful to be able to classify new "test" documents using already classified "training" documents. A common example is using a corpus of labeled spam and ham (non-spam) e-mails to predict whether or not a new document is spam. Here is one example of such data: UCI Machine Learning Repository: Spambase Data Set

For this project, you can either use the above dataset to predict the class of new documents (either withheld from the training dataset or from another source such as your own spam folder).

For more adventurous students, you are welcome (encouraged!) to come up a different set of documents (including scraped web pages!?) that have already been classified (e.g. tagged), then analyze these documents to predict how new documents should be classified.

3 Choosing Documents for Classification

Let's look at available texts in the guttenberg corpus.

[1]:
```
import nltk
import random
random.seed(250)
import pandas as pd
pd.set_option('display.max_rows', 500)

nltk.corpus.gutenberg.fileids()
```

[1]:
```
['austen-emma.txt',
 'austen-persuasion.txt',
 'austen-sense.txt',
 'bible-kjv.txt',
 'blake-poems.txt',
 'bryant-stories.txt',
 'burgess-busterbrown.txt',
 'carroll-alice.txt',
```

```
 'chesterton-ball.txt',
 'chesterton-brown.txt',
 'chesterton-thursday.txt',
 'edgeworth-parents.txt',
 'melville-moby_dick.txt',
 'milton-paradise.txt',
 'shakespeare-caesar.txt',
 'shakespeare-hamlet.txt',
 'shakespeare-macbeth.txt',
 'whitman-leaves.txt']
```

We have 3 books by Jane Austen, Bible, 1 book by Blake, and so on. Each author writes using his/her own style. Can we use samples of their work to predict who wrote specific passage?

4 Austen vs Blake

5 Create texts

First we need to take all three of Austen's works and combine them to create one text. We will also remove punctuation and convert everything to lowercase to eliminate duplicate words. Then we can take that and create a list of text segments. Each segment will have a length of 1000 words.

[2]:
```
austen = nltk.corpus.gutenberg.words('austen-emma.txt')+nltk.corpus.gutenberg.
 ↪words('austen-persuasion.txt')+nltk.corpus.gutenberg.words('austen-sense.
 ↪txt')
austen = [word.lower() for word in austen if word.isalpha()]
austen1=[]
for i in range(366):
    austen1.append([austen[i*1000:(i+1)*1000],'au'])
len(austen)
```

[2]:
```
366454
```

[3]:
```
len(austen1)
```

[3]:
```
366
```

We now have a list of 432 1000-word segments of text written by Jane Austen.

We will skip the Bible since it was written by many different authors using many different styles, but let's take the next text in the guttenburg corpus, poems by Blake, and do the same thing we did with Austen.

[4]:
```
blake = nltk.corpus.gutenberg.words('blake-poems.txt')
blake = [word.lower() for word in blake if word.isalpha()]
blake1=[]
for i in range(7):
    blake1.append([blake[i*990:(i+1)*990],'bl'])
```

```
len(blake)
```

[4]: 6934

Since there are just shy of 7000 words total in the Blake text, we will make each segment 990 words in order to get 7 equal segments for Blake.

[5]:
```
len(blake1)
```

[5]: 7

We now have a list of seven 990-word segments of text written by William Blake.

6 Create Feature Extractor

Now let's take the two original lists of words and combine them to create one longer list and find the 2000 most frequent words, which we will later use to create a feature list for our classifier.

[6]:
```
ab=austen+blake
all_words = nltk.FreqDist(w.lower() for w in ab)
word_features = list(all_words)[:2000]

wlist = []
for i in range(0, 2000, 200):
    df = pd.DataFrame(word_features[i:(i+200)])
    df.columns=['200 words']
    wlist.append(df)

pd.concat(wlist, axis=1)
```

[6]:

	200 words	200 words	200 words	200 words	200 words	\
0	the	feelings	true	clay	purpose	
1	to	found	agreeable	benwick	assured	
2	and	few	taken	temper	extraordinary	
3	of	heart	state	isabella	write	
4	a	does	conversation	curiosity	ease	
5	i	going	dare	delighted	also	
6	her	perhaps	husband	sight	agitation	
7	in	believe	door	robert	distance	
8	was	fairfax	walked	admiral	welcome	
9	it	family	louisa	excellent	difficulties	
10	she	present	nobody	match	prevent	
11	not	myself	afterwards	bring	note	
12	be	colonel	things	resolution	lately	
13	that	half	received	sent	waiting	
14	he	yes	whose	happened	besides	
15	had	almost	neither	kept	thus	

16	you	jennings	farther	ten	attentions
17	as	down	continued	lost	join
18	for	off	yourself	spent	calling
19	but	hear	object	pity	trying
20	with	both	spoke	spoken	daughters
21	his	churchill	marriage	late	breakfast
22	have	pleasure	means	whatever	tears
23	is	take	full	occasion	mention
24	at	once	attachment	equally	girls
25	s	letter	fortune	lyme	conviction
26	very	wentworth	thinking	twenty	actually
27	all	those	directly	danger	immediate
28	so	cried	making	understanding	don
29	him	willoughby	sense	view	agreed
30	could	left	wonder	second	forced
31	on	knew	appeared	delightful	disappointment
32	by	frank	giving	aunt	play
33	been	replied	ladies	invitation	joined
34	would	world	talking	meaning	dancing
35	no	certainly	usual	morrow	required
36	my	together	natural	lived	become
37	mr	feel	afraid	times	pride
38	which	evening	doing	sensible	boy
39	they	nor	real	greatest	pleasing
40	mrs	tell	call	beauty	elegant
41	were	party	company	delight	summer
42	from	kind	particular	surprise	information
43	any	least	added	easy	leaving
44	me	between	middleton	knowing	trouble
45	this	life	particularly	difference	expectation
46	what	less	bear	creature	lord
47	do	looked	satisfied	style	avoid
48	them	seen	strong	silence	dixon
49	more	subject	equal	ago	tea
50	or	speak	exactly	followed	observe
51	their	these	matter	answered	rain
52	will	people	son	hands	command
53	there	acquaintance	fine	thank	unhappy
54	an	visit	bad	turn	moments
55	must	lucy	words	distress	elegance
56	such	therefore	power	warm	send
57	said	whole	entirely	wishes	occurred
58	one	ought	engaged	regret	d
59	than	old	nature	evil	dreadful
60	if	back	stay	merely	disappointed
61	much	possible	different	tried	indifference
62	miss	told	pleased	high	wishing

63	every	whom	settled	persuaded	cousin
64	when	hardly	used	cole	rich
65	your	three	cold	ma	sigh
66	emma	others	open	friendship	pay
67	are	word	likely	loved	beautiful
68	am	smith	new	disposed	spend
69	only	short	kindness	looks	service
70	well	happiness	girl	influence	light
71	should	spirits	believed	quiet	indifferent
72	did	manner	expected	opportunity	quick
73	who	friends	elizabeth	praise	obliging
74	think	side	london	plan	value
75	how	often	gentleman	low	attempt
76	being	gone	extremely	countenance	consideration
77	little	doubt	degree	future	claims
78	own	immediately	men	everything	thee
79	we	brother	use	judgment	intimacy
80	good	since	behaviour	amiable	hers
81	never	deal	sat	goddard	odd
82	now	far	dinner	interesting	uncle
83	might	known	impossible	seems	youth
84	know	person	four	yesterday	compliment
85	herself	right	hours	confidence	formed
86	elinor	given	case	honour	picture
87	time	want	face	instantly	hundred
88	can	charles	ask	law	judge
89	thing	years	country	news	listened
90	some	idea	randalls	perfect	returning
91	nothing	leave	met	mentioned	writing
92	before	towards	round	except	promised
93	too	part	question	spirit	alarm
94	say	wife	week	scarcely	remain
95	has	sort	read	remained	musgroves
96	marianne	under	air	park	period
97	other	hartfield	cottage	croft	attached
98	great	suppose	barton	event	servant
99	see	coming	imagine	gratitude	reflection
100	soon	ill	speaking	aware	amusement
101	most	next	smile	understood	guess
102	though	opinion	early	expect	shew
103	harriet	find	five	fixed	strange
104	anne	gave	sometimes	arrival	gentlemen
105	again	best	common	resolved	safe
106	without	else	street	especially	delay
107	man	hour	taste	favourite	offer
108	about	let	asked	mine	journey
109	may	general	acquainted	decided	fellow

110	quite	obliged	remember	difficulty	eldest
111	first	perfectly	society	admiration	death
112	always	each	meeting	smiling	warmth
113	out	because	got	instead	declared
114	after	comfort	small	companion	hill
115	two	eyes	circumstances	pray	clever
116	ever	passed	martin	became	miles
117	weston	able	t	enjoyment	papa
118	made	bates	handsome	led	placed
119	day	russell	months	wait	comes
120	thought	brought	proper	smallest	suspect
121	lady	situation	conduct	reached	important
122	then	name	walking	highly	considerable
123	shall	attention	allow	age	six
124	dear	brandon	respect	observed	respectable
125	sure	began	appearance	please	terms
126	like	walter	meet	pass	standing
127	up	mary	keep	weeks	stood
128	sister	children	alone	vain	couple
129	mother	put	forward	women	hair
130	knightley	whether	greater	window	during
131	make	account	silent	money	quarter
132	elton	pretty	large	tone	sudden
133	into	through	hoped	otherwise	stand
134	young	why	meant	turning	tender
135	father	against	need	promise	invited
136	however	affection	perry	cheerful	hint
137	give	wanted	superior	excuse	around
138	indeed	longer	consequence	charming	neighbourhood
139	come	interest	necessary	bed	suspicion
140	himself	end	help	forget	observation
141	away	carriage	turned	norland	god
142	house	get	pain	disposition	declare
143	long	looking	truth	loss	particulars
144	upon	either	eye	proof	painful
145	its	voice	anxious	generally	liked
146	better	course	comfortable	says	considering
147	sir	daughter	sisters	parties	compassion
148	over	chapter	sit	fair	fire
149	many	wished	supposed	chance	hurry
150	oh	talked	hearing	repeated	proved
151	even	return	beginning	justice	laughing
152	seemed	set	convinced	mistaken	importance
153	go	answer	probably	behind	income
154	while	musgrove	worth	effect	vanity
155	room	ferrars	uppercross	misery	above
156	friend	went	palmer	fanny	fortnight

157	woodhouse	talk	living	fact	warmly
158	last	seeing	satisfaction	knows	getting
159	way	near	drawing	concern	angry
160	happy	night	fancy	altogether	spring
161	captain	glad	knowledge	share	tired
162	jane	rest	taking	completely	rooms
163	enough	hand	sitting	music	constant
164	just	character	appear	eager	dance
165	felt	business	anything	finding	cousins
166	having	head	harville	ashamed	surprised
167	mind	minutes	pleasant	secret	hayter
168	done	highbury	advantage	fear	steele
169	still	town	former	campbell	nearly
170	moment	mean	consider	taylor	dine
171	same	ready	henrietta	horses	scheme
172	elliot	change	self	whenever	melancholy
173	us	married	past	written	ideas
174	home	beyond	health	form	thinks
175	came	days	live	itself	complete
176	hope	understand	table	produced	surprize
177	really	bath	ah	duty	reasonable
178	rather	child	notice	pounds	telling
179	here	feeling	seem	work	cause
180	saw	manners	entered	suffering	imagined
181	cannot	took	thousand	easily	humour
182	heard	among	kellynch	due	try
183	place	called	allowed	inclination	earnest
184	wish	returned	joy	donwell	severe
185	edward	regard	considered	wanting	simple
186	woman	marry	serious	exceedingly	reply
187	morning	saying	fond	caught	latter
188	another	care	thoughts	o	recollect
189	poor	point	circumstance	marrying	pause
190	love	engagement	none	begin	listen
191	our	reason	sake	struck	necessity
192	john	determined	wrong	anxiety	stopped
193	where	assure	worse	hopes	paid
194	body	sorry	certain	absolutely	opened
195	till	themselves	letters	staying	suffer
196	something	within	spite	arrived	human
197	dashwood	walk	sweet	suffered	hall
198	yet	already	favour	consciousness	wallis
199	look	year	weather	depend	gentle

	200 words	200 words	200 words	200 words	200 words
0	smiled	estate	companions	sensibility	morton
1	thoroughly	run	suit	heartily	chiefly

2	enscombe	totally	fast	cruel	selfishness
3	desirable	shewed	pressed	relation	turns
4	seat	line	dark	buildings	animated
5	thrown	settle	highest	existence	faults
6	concerned	venture	altered	third	shook
7	close	advice	convince	black	wherever
8	encouragement	possibility	observing	christmas	triumph
9	opinions	piece	niece	belong	comfortably
10	credit	confusion	seated	paying	improve
11	approbation	habit	excepting	visitor	missed
12	speech	profession	lives	loves	comforts
13	confess	prevented	succeeded	divided	residence
14	drew	cut	subjects	instance	establishment
15	minute	white	miserable	goodness	owed
16	wrote	servants	sunk	seek	request
17	intended	astonished	england	chosen	continue
18	laughed	post	interrupted	apparent	engaging
19	deep	camden	suddenly	horror	shewing
20	sea	sorrow	crofts	suited	farm
21	receive	ways	charlotte	condition	notion
22	laugh	happily	weep	visitors	careful
23	intelligence	claim	rank	thorough	ignorant
24	fully	matters	powers	hurt	plainly
25	thou	circle	recommended	introduction	submit
26	desire	worst	reach	fifty	places
27	margaret	illness	happier	foot	cared
28	dashwoods	probability	endeavour	nearer	hesitation
29	seven	truly	born	forty	striking
30	recollection	compliments	produce	wants	grace
31	increased	carried	secured	twelve	stopt
32	absence	cheerfulness	eat	failed	entreat
33	smiles	arrangement	frequently	memory	unlike
34	winter	undoubtedly	moved	ignorance	corner
35	plain	contrary	confined	frightened	addressing
36	strength	inconvenience	wholly	excited	assurances
37	maid	henry	steady	conceal	safety
38	shepherd	consolation	charm	decision	fearful
39	mere	nice	gives	exquisite	intelligible
40	peace	sick	supposing	honourable	cares
41	fresh	imagination	points	desired	previously
42	express	heaven	likeness	takes	experience
43	presently	parting	deceived	certainty	partiality
44	clear	dress	offence	forgiven	weak
45	exertion	blessing	occur	declined	involved
46	solicitude	shut	visited	alteration	sincerity
47	ball	twice	induced	autumn	assisted
48	tenderness	pianoforte	attentive	noise	lips

49	grove	entrance	supply	alike	approve
50	affectionate	calm	passage	doubted	approach
51	hold	box	madam	houses	preparation
52	sleep	dalrymple	discourse	preference	re
53	useful	office	abroad	happiest	bow
54	civil	mutual	cure	nephew	leading
55	sad	employment	presence	suspicions	remark
56	forgotten	choice	indulgence	violent	housekeeper
57	drawn	admired	gay	forgive	scene
58	safely	regular	readily	hawkins	orders
59	assistance	free	song	shop	playing
60	acknowledge	amused	changed	kindly	convenience
61	card	order	bright	described	step
62	garden	blush	frederick	peculiar	jealous
63	history	personal	concerns	shocked	accommodation
64	lay	sufficient	habits	wonderful	attraction
65	happen	borne	village	infant	embarrassment
66	growing	direction	spared	ay	ireland
67	bringing	lines	connected	mamma	blame
68	civility	sentiments	inquiries	middletons	injured
69	affair	propriety	dependence	beloved	grey
70	relief	accept	behaved	arose	ford
71	astonishment	conscience	shame	bye	rejoice
72	folly	trust	entering	lane	tongue
73	burst	proceeded	expense	gallantry	chaise
74	instrument	selfish	established	silly	valued
75	soul	fall	addition	break	apples
76	single	conscious	families	invite	concealment
77	fortunate	ended	superiority	mortification	dream
78	attended	month	felicity	pains	lamb
79	several	chair	treated	unless	agony
80	ground	dining	encouraging	liberal	believing
81	naturally	ourselves	watched	seeming	exeter
82	book	stop	figure	soft	board
83	receiving	reproach	expecting	forming	hills
84	liberty	calmness	patience	intentions	smart
85	following	heavy	pretend	ventured	combe
86	sooner	communication	learn	independent	harley
87	attending	spare	size	encouraged	thel
88	unfortunate	caution	lead	merry	er
89	act	along	assurance	force	youngest
90	attend	politeness	tolerable	awkward	composed
91	thy	instant	delicate	carrying	played
92	passing	motive	direct	design	peculiarly
93	inquiry	removal	reserve	fate	solitude
94	fears	distant	suspense	proposal	november
95	capable	impatient	natured	higher	regrets

96	consent	stairs	grew	variety	troublesome
97	mistress	unable	fix	resemblance	quarrel
98	prospect	connection	sought	tall	interference
99	rate	cleveland	hastily	precious	succeeding
100	comprehend	disagreeable	eltons	censure	cease
101	education	domestic	satisfy	hitherto	churchills
102	surprized	james	message	explained	wealth
103	friendly	lucky	remaining	ceremony	accomplished
104	public	draw	impression	inclined	prospects
105	grateful	proud	quietly	extreme	source
106	deserve	neighbours	prove	wondered	partial
107	dislike	parlour	restored	effects	season
108	follow	evidently	opposition	syllable	recovering
109	fit	books	wild	rise	parish
110	exclaimed	arranged	quit	conclusion	simplicity
111	expressed	suspected	broken	communicated	principles
112	escape	finished	died	reserved	formerly
113	likewise	pardon	humoured	escaped	sweetness
114	address	perceive	persons	weymouth	pleasantly
115	move	unpleasant	luck	interval	introduce
116	views	putting	active	carry	supper
117	utmost	eagerness	bore	wind	describe
118	arm	boys	objects	wise	sink
119	rose	clock	merits	thanks	precisely
120	acknowledged	affections	features	colour	bred
121	explanation	excessively	watch	sincerely	wondering
122	recovered	engagements	chose	steps	blushed
123	eagerly	campbells	seldom	uncertain	sex
124	story	admire	sincere	imprudence	preparing
125	yours	description	scruple	whisper	grow
126	nay	dared	length	concluded	cross
127	emotion	everybody	watching	supported	gain
128	anybody	devonshire	touch	harvilles	falling
129	thirty	remembrance	depended	allenham	animation
130	intimate	informed	waited	authority	progress
131	occupied	anywhere	stronger	evils	keeping
132	merit	somebody	sun	afforded	employ
133	chuse	tolerably	indignation	connexions	throw
134	introduced	willing	wit	square	entreaties
135	fell	nonsense	master	shocking	agitated
136	charge	properly	uneasiness	judging	paused
137	overcome	visits	charade	heir	expressions
138	quitted	opening	affliction	belonging	grieved
139	begun	proposed	earth	separation	sentence
140	lively	united	ran	nerves	bestow
141	abbey	church	begged	measure	plans
142	persuasion	unwilling	rendered	preferred	remembered

143	school	sending	dangerous	widow	provided
144	admitted	number	dull	distinction	cheeks
145	recommend	raise	related	middle	glow
146	makes	employed	bless	recommendation	tempted
147	feared	complexion	deeply	affairs	trifling
148	reading	dread	collected	sing	sufferings
149	glance	improvement	concert	cards	rights
150	slight	infinitely	fallen	trusted	windows
151	mistake	deserved	wedding	compared	perfection
152	judged	continual	mile	later	extent
153	private	hurried	advantages	doubtful	c
154	aye	language	daily	vicarage	arms
155	increase	improved	separate	uncomfortable	ordered
156	assist	admit	accepted	deny	ear
157	william	relations	nervous	appears	lodgings
158	composure	greatly	frequent	angel	reflections
159	evident	paper	rising	unjust	heat
160	maple	proceed	remarkably	lover	previous
161	interested	affected	becoming	difficult	accepting
162	rational	setting	advise	mama	sacrifice
163	removed	drive	water	civilities	offended
164	hard	enjoy	humble	shade	resolve
165	dearest	sensations	gratified	strongly	respects
166	lose	nurse	valuable	absent	younger
167	connexion	shewn	sentiment	offering	roused
168	comparison	perceived	questions	foolish	tuesday
169	leisure	green	fancying	uneasy	unwell
170	expression	engage	birth	resist	abilities
171	earnestly	grounds	forgot	submitted	recovery
172	fancied	ceased	grown	fail	started
173	inferior	grief	blind	throat	shock
174	parted	intercourse	decidedly	tells	spread
175	busy	differently	hints	unfortunately	neglect
176	afford	unnecessary	attempted	moreover	positive
177	road	catch	refuse	moving	february
178	grave	independence	doors	carriages	exert
179	objection	success	tree	increasing	court
180	persuade	worthy	endure	continually	quitting
181	material	harm	principal	learnt	won
182	delicacy	offered	asking	catching	valley
183	beg	favourable	conveyed	discovery	tomorrow
184	wretched	report	result	amazement	bride
185	prepared	lovely	horse	continuing	taught
186	enter	spot	gradually	softened	attach
187	seriously	raised	refused	principle	amuse
188	sound	listening	intention	notions	various
189	crown	closed	acting	coles	talents

190	delaford	laid	snow	distinguished	goes
191	steeles	exercise	apology	openly	flow
192	fault	esteem	resentment	courage	dreadfully
193	support	forth	dignity	risk	constantly
194	agree	gained	fortitude	across	blessed
195	eight	held	performance	distressed	endeavouring
196	possibly	probable	furniture	explain	idle
197	secure	add	earlier	filled	pursuits
198	hearted	arise	musical	cloud	anger
199	property	possession	inn	berkeley	scruples

We will use the function in the Natural Language Processing with Python textbook on page 228 to create a feature generator that uses the 2000 most frequent words list and indicates whether or not each word is present in the text as a feature.

[7]:
```
def document_features(document):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = (word in document_words)
    return features
```

Let's test it on the full Blake text

[8]:
```
features = document_features(blake)
list(features.items())[:20]
```

[8]:
```
[('contains(the)', True),
 ('contains(to)', True),
 ('contains(and)', True),
 ('contains(of)', True),
 ('contains(a)', True),
 ('contains(i)', True),
 ('contains(her)', True),
 ('contains(in)', True),
 ('contains(was)', True),
 ('contains(it)', True),
 ('contains(she)', True),
 ('contains(not)', True),
 ('contains(be)', True),
 ('contains(that)', True),
 ('contains(he)', True),
 ('contains(had)', True),
 ('contains(you)', True),
 ('contains(as)', True),
 ('contains(for)', True),
 ('contains(but)', True)]
```

7 Create Test Train Dataset

Now we need to create a list of all text segments from both Austen and Blake and shuffle them to create the text corpus that we will use to train and test our classifier model.

[9]:
```
documents=austen1+blake1
```

[10]:
```
import random
random.shuffle(documents)
featuresets = [(document_features(d), c) for (d,c) in documents]
len(featuresets)
```

[10]: 373

Next we split our dataset into test and train sections, train our classifier on the training set, and check the accuracy of our model on the test set.

[11]:
```
train_set, test_set = featuresets[:100], featuresets[100:]
classifier = nltk.NaiveBayesClassifier.train(train_set)
```

[12]:
```
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.9926739926739927
```

It is very easy to for NLKT to distinguish between Austen and Blake. Let's try more authors.

8 Adding Bryant

[14]:
```
bryant = nltk.corpus.gutenberg.words('bryant-stories.txt')
bryant = [word.lower() for word in bryant if word.isalpha()]
bryant1=[]
for i in range(46):
    bryant1.append([bryant[i*1000:(i+1)*1000],'br'])
len(bryant)
```

[14]: 46611

[15]:
```
abb=austen+blake+bryant
all_words = nltk.FreqDist(w.lower() for w in abb)
word_features = list(all_words)[:2000]

documents=austen1+blake1+bryant1
```

[16]:
```
random.shuffle(documents)
featuresets = [(document_features(d), c) for (d,c) in documents]
len(featuresets)
```

[16]: 419

[18]:
```
train_set, test_set = featuresets[:100], featuresets[100:]
classifier = nltk.NaiveBayesClassifier.train(train_set)
```

[19]:
```
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.9937304075235109
```

We still get even better results in distinguishing between Austen, Blake, and Braynt.

9 Adding Burgess

[20]:
```
burgess = nltk.corpus.gutenberg.words('burgess-busterbrown.txt')
burgess = [word.lower() for word in burgess if word.isalpha()]
burgess1=[]
for i in range(16):
    burgess1.append([burgess[i*1000:(i+1)*1000],'bu'])
len(burgess)
```

[20]:
```
16327
```

[21]:
```
abbb=austen+blake+bryant+burgess
all_words = nltk.FreqDist(w.lower() for w in abbb)
word_features = list(all_words)[:2000]

documents=austen1+blake1+bryant1+burgess1
```

[22]:
```
random.shuffle(documents)
featuresets = [(document_features(d), c) for (d,c) in documents]
len(featuresets)
```

[22]:
```
435
```

[23]:
```
train_set, test_set = featuresets[:100], featuresets[100:]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.9522388059701492
```

Accuracy declined a bit, but we are still in the mid 90's.

Let's see what features are most important in training our model...

[24]:
```
classifier.show_most_informative_features(25)
```

```
Most Informative Features
          contains(much) = False              bl : au     =     52.8 : 1.0
        contains(breath) = True               bu : au     =     44.0 : 1.0
        contains(chosen) = True               bu : au     =     44.0 : 1.0
           contains(eat) = True               bu : au     =     44.0 : 1.0
```

```
   contains(feet) = True              bu : au     =     44.0 : 1.0
 contains(helped) = True              bu : au     =     44.0 : 1.0
 contains(jumped) = True              bu : au     =     44.0 : 1.0
     contains(ll) = True              bu : au     =     44.0 : 1.0
contains(terrible) = True             bu : au     =     44.0 : 1.0
   contains(tree) = True              bu : au     =     44.0 : 1.0
    contains(sun) = True              br : au     =     42.4 : 1.0
 contains(bright) = True              bl : au     =     41.1 : 1.0
    contains(cry) = True              bl : au     =     41.1 : 1.0
    contains(fly) = True              bl : au     =     41.1 : 1.0
   contains(gold) = True              bl : au     =     41.1 : 1.0
  contains(grass) = True              bl : au     =     41.1 : 1.0
  contains(shade) = True              bl : au     =     41.1 : 1.0
   contains(wild) = True              bl : au     =     41.1 : 1.0
  contains(earth) = True              bl : au     =     31.7 : 1.0
   contains(been) = False             bl : au     =     29.3 : 1.0
 contains(fields) = True              bl : au     =     29.3 : 1.0
   contains(free) = True              bl : au     =     29.3 : 1.0
   contains(grey) = True              bl : au     =     29.3 : 1.0
   contains(king) = True              br : au     =     29.3 : 1.0
   contains(soft) = True              bl : au     =     29.3 : 1.0
```

It appears that a text that does not contain the word 'much' is 52 times more likely to be by Blake than by Austen, while a text that contains the word "eat", "below", "chosen", "stout" or "becomes" are each 44 times more likely to be by Burgess than by Austen. Texts that contain the word 'free', 'youthful', 'soft' or "tear" are each 29 times more likely to be by Blake than by Austen.

10 Adding Carroll

[25]:
```
carroll = nltk.corpus.gutenberg.words('carroll-alice.txt')
carroll = [word.lower() for word in carroll if word.isalpha()]
carroll1=[]
for i in range(27):
    carroll1.append([carroll[i*1000:(i+1)*1000],'ca'])
len(carroll)
```

[25]:
```
27333
```

[26]:
```
abbbc=austen+blake+bryant+burgess+carroll
all_words = nltk.FreqDist(w.lower() for w in abbbc)
word_features = list(all_words)[:2000]

documents=austen1+blake1+bryant1+burgess1+carroll1
```

[27]:
```
random.shuffle(documents)
featuresets = [(document_features(d), c) for (d,c) in documents]
len(featuresets)
```

[27]: 462

[28]:
```
train_set, test_set = featuresets[:100], featuresets[100:]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.9475138121546961
```

By adding Carroll accuracy declined significantly. This is probably because we now are using a smaller percent (100/462) of our corpus for training and we are adding more complexity by adding more categories to classify into. Let's keep adding more authors and see what happens.

[29]:
```
classifier.show_most_informative_features(25)
```

```
Most Informative Features
         contains(brook) = True              bu : au     =     47.8 : 1.0
          contains(crow) = True              bu : au     =     47.8 : 1.0
        contains(beside) = True              bl : au     =     41.0 : 1.0
        contains(breath) = True              bl : au     =     41.0 : 1.0
          contains(dead) = True              bl : au     =     41.0 : 1.0
        contains(forgot) = True              bl : au     =     41.0 : 1.0
           contains(had) = False             bl : au     =     41.0 : 1.0
            contains(ll) = True              ca : au     =     41.0 : 1.0
         contains(mouth) = True              bl : au     =     41.0 : 1.0
         contains(queen) = True              bl : au     =     41.0 : 1.0
         contains(river) = True              bl : au     =     41.0 : 1.0
          contains(sing) = True              bl : au     =     41.0 : 1.0
        contains(spring) = True              bl : au     =     41.0 : 1.0
        contains(tongue) = True              bl : au     =     41.0 : 1.0
     contains(trembling) = True              bl : au     =     41.0 : 1.0
           contains(by) = False             bu : au     =     34.2 : 1.0
       contains(fishing) = True              bu : au     =     34.2 : 1.0
           contains(re) = True              ca : au     =     31.9 : 1.0
         contains(brown) = True              bu : au     =     28.7 : 1.0
         contains(green) = True              bu : au     =     28.7 : 1.0
         contains(birth) = True              bl : au     =     24.6 : 1.0
        contains(bright) = True              bl : au     =     24.6 : 1.0
         contains(could) = False             bl : au     =     24.6 : 1.0
          contains(deep) = True              bl : au     =     24.6 : 1.0
         contains(earth) = True              bl : au     =     24.6 : 1.0
```

Common words that indicate that a text is more likely to have been written by Blake are "youthful", "forgot", "mild", "sing", "glass", "walks", and "gently" which each indicate a text is 41 times more likely to have been written by Blake than Austen. For Burgess, indicator words are "eaten" and "black", and for Austen, "had", "by", "could" and "would".

11 Adding Chesterson

[30]:
```
chesterson = nltk.corpus.gutenberg.words('chesterton-ball.txt')+nltk.corpus.
 ↪gutenberg.words('chesterton-brown.txt')+nltk.corpus.gutenberg.
 ↪words('chesterton-thursday.txt')
chesterson = [word.lower() for word in chesterson if word.isalpha()]
chesterson1=[]
for i in range(214):
    chesterson1.append([chesterson[i*1000:(i+1)*1000],'ch'])
len(chesterson)
```

[30]:
```
214692
```

[31]:
```
abbbcc=austen+blake+bryant+burgess+carroll+chesterson
all_words = nltk.FreqDist(w.lower() for w in abbbcc)
word_features = list(all_words)[:2000]

documents=austen1+blake1+bryant1+burgess1+carroll1+chesterson1
```

[32]:
```
random.shuffle(documents)
featuresets = [(document_features(d), c) for (d,c) in documents]
len(featuresets)
```

[32]:
```
676
```

Since we now have 676 texts, let's increase our training set to keep it at about 25% of the corpus.

[33]:
```
train_set, test_set = featuresets[:170], featuresets[170:]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.958498023715415
```

Still at about 95%.

[34]:
```
classifier.show_most_informative_features(25)
```

```
Most Informative Features
         contains(brown) = True              bu : au     =     53.1 : 1.0
            contains(ll) = True              ca : au     =     49.3 : 1.0
        contains(bright) = True              bl : au     =     45.5 : 1.0
         contains(cloud) = True              bl : au     =     45.5 : 1.0
          contains(coat) = True              bl : au     =     45.5 : 1.0
         contains(crime) = True              bl : au     =     45.5 : 1.0
           contains(cry) = True              bl : au     =     45.5 : 1.0
     contains(dangerous) = True              bl : au     =     45.5 : 1.0
         contains(devil) = True              bl : au     =     45.5 : 1.0
         contains(drink) = True              bl : au     =     45.5 : 1.0
       contains(flowers) = True              bl : au     =     45.5 : 1.0
```

```
    contains(fly) = True                bl : au     =     45.5 : 1.0
   contains(gold) = True                bl : au     =     45.5 : 1.0
   contains(iron) = True                bl : au     =     45.5 : 1.0
 contains(rising) = True                bl : au     =     45.5 : 1.0
   contains(sing) = True                bl : au     =     45.5 : 1.0
contains(streets) = True                bl : au     =     45.5 : 1.0
  contains(takes) = True                bl : au     =     45.5 : 1.0
   contains(thou) = True                bl : au     =     45.5 : 1.0
   contains(wind) = True                bl : au     =     45.5 : 1.0
 contains(fields) = True                br : au     =     33.4 : 1.0
    contains(eat) = True                bu : au     =     31.9 : 1.0
  contains(green) = True                bu : au     =     31.9 : 1.0
contains(admired) = True                bl : ch     =     30.5 : 1.0
   contains(bore) = True                bl : ch     =     30.5 : 1.0
```

12 Adding the rest of the authors

[35]:
```
edgeworth = nltk.corpus.gutenberg.words('edgeworth-parents.txt')
edgeworth = [word.lower() for word in edgeworth if word.isalpha()]
edgeworth1=[]
for i in range(170):
    edgeworth1.append([edgeworth[i*1000:(i+1)*1000],'ed'])
len(edgeworth)
```

[35]:
```
170737
```

[36]:
```
melville = nltk.corpus.gutenberg.words('melville-moby_dick.txt')
melville = [word.lower() for word in melville if word.isalpha()]
melville1=[]
for i in range(218):
    melville1.append([melville[i*1000:(i+1)*1000],'me'])
len(melville)
```

[36]:
```
218361
```

[37]:
```
shakespeare = nltk.corpus.gutenberg.words('shakespeare-caesar.txt')+nltk.corpus.
 ↪gutenberg.words('shakespeare-hamlet.txt')+nltk.corpus.gutenberg.
 ↪words('shakespeare-macbeth.txt')
shakespeare = [word.lower() for word in shakespeare if word.isalpha()]
shakespeare1=[]
for i in range(69):
    shakespeare1.append([shakespeare[i*1000:(i+1)*1000],'sh'])
len(shakespeare)
```

[37]:
```
69340
```

[38]:
```
whitman = nltk.corpus.gutenberg.words('whitman-leaves.txt')
whitman = [word.lower() for word in whitman if word.isalpha()]
whitman1=[]
for i in range(126):
    whitman1.append([whitman[i*1000:(i+1)*1000],'wh'])
len(whitman)
```

[38]: 126276

[39]:
```
abbbccemsw=austen+blake+bryant+burgess+carroll+chesterson+edgeworth+melville+shakespeare+whitm
all_words = nltk.FreqDist(w.lower() for w in abbbccemsw)
word_features = list(all_words)[:2000]

documents=austen1+blake1+bryant1+burgess1+carroll1+chesterson1+edgeworth1+melville1+shakespear
```

[40]:
```
random.shuffle(documents)
featuresets = [(document_features(d), c) for (d,c) in documents]
len(featuresets)
```

[40]: 1259

[41]:
```
train_set, test_set = featuresets[:320], featuresets[320:]
classifier = nltk.NaiveBayesClassifier.train(train_set)
print(nltk.classify.accuracy(classifier, test_set))
```

```
0.9478168264110756
```

[42]:
```
classifier.show_most_informative_features(40)
```

```
Most Informative Features
          contains(thou) = True              sh : au     =     61.7 : 1.0
        contains(farmer) = True              bu : au     =     58.9 : 1.0
           contains(her) = False             bu : au     =     58.9 : 1.0
        contains(beside) = True              bl : au     =     56.1 : 1.0
         contains(earth) = True              bl : au     =     56.1 : 1.0
        contains(fields) = True              bl : au     =     56.1 : 1.0
          contains(very) = False             bl : au     =     56.1 : 1.0
         contains(mouth) = True              bu : au     =     42.1 : 1.0
          contains(wide) = True              bu : au     =     42.1 : 1.0
          contains(have) = False             sh : au     =     39.3 : 1.0
         contains(river) = True              br : au     =     37.0 : 1.0
       contains(herself) = True              ca : ch     =     36.7 : 1.0
           contains(mrs) = True              au : ch     =     36.2 : 1.0
          contains(boat) = True              me : au     =     35.7 : 1.0
          contains(exit) = True              sh : au     =     35.5 : 1.0
            contains(st) = True              sh : au     =     35.5 : 1.0
         contains(whale) = True              me : ch     =     34.0 : 1.0
          contains(bare) = True              bl : au     =     33.7 : 1.0
```

contains(broke) = True	bl : au	=	33.7 : 1.0
contains(divine) = True	bl : au	=	33.7 : 1.0
contains(feet) = True	bl : au	=	33.7 : 1.0
contains(filled) = True	bl : au	=	33.7 : 1.0
contains(floor) = True	bl : au	=	33.7 : 1.0
contains(fly) = True	bl : au	=	33.7 : 1.0
contains(gold) = True	bl : au	=	33.7 : 1.0
contains(ha) = True	bl : au	=	33.7 : 1.0
contains(human) = True	bl : au	=	33.7 : 1.0
contains(mighty) = True	bl : au	=	33.7 : 1.0
contains(neck) = True	bl : au	=	33.7 : 1.0
contains(seek) = True	bl : au	=	33.7 : 1.0
contains(sing) = True	bl : au	=	33.7 : 1.0
contains(sky) = True	bl : au	=	33.7 : 1.0
contains(joy) = True	bl : ch	=	33.3 : 1.0
contains(lovely) = True	bl : ch	=	33.3 : 1.0
contains(shade) = True	bl : ch	=	33.3 : 1.0
contains(sleep) = True	bl : ch	=	33.3 : 1.0
contains(sorrow) = True	bl : ch	=	33.3 : 1.0
contains(sweet) = True	bl : ch	=	33.3 : 1.0
contains(tears) = True	bl : ch	=	33.3 : 1.0
contains(had) = False	wh : au	=	31.5 : 1.0