Installation

Programming language: Python

Source code location: /home/sbillah/nlp2/

Programming language: Statistical methods implemented:

1. **chi-square test**
2. **likelihood ratio**

Documentation:

```
sbillah@turing:~/nlp2$ ./NLPEngine.py -h

usage: NLPEngine.py [-h] [-o DST_DIR] [-s SRC_DIR] [-nlp COMMANDS]


This is Syed's NLP program for HW02

optional arguments:

  -h, --help            show this help message and exit

  -o DST_DIR, --output DST_DIR

                        The directory where output goes

  -s SRC_DIR, --source SRC_DIR

                        The directory where raw files reside

  -nlp COMMANDS, --nlp COMMANDS

                     index of all bigrams in src directory, or compute

                        colocaiton. e.g, -nlp bigrams, -nlp colocation
```

Here is a complete command line example:

```
sbillah@turing:~/nlp2$ ./NLPEngine.py -s
/home/sgauch/public_html/5013IR/files/ -o parsed/ -nlp colocations

Done!
```

Algorithm

I use an in-memory Hash-table to count all bigrams. The pseudo code is given bellow:

```
// I assume all <bigram, freq> pairs are in a file named 'bigrams.txt'
// And all <token, freq> pairs are in a file named 'all_tokens.txt'

load tokens in a hash-table tokens_ht<string, int>

# chi-square test
load bigrams in a hash-table bigragm_ht<(w0, w1), int>
foreach key <w0,w1> in bigragm_ht:
      chi = chi_squere_test(bigragm_ht[<w0,w1>], tokens_ht[w0],
                                        tokens_ht[w1], num_bigrams);
      bigragm_ht[<w0,w1>] = chi;

sort bigragm_ht;
write bigragm_ht to file;

# likelihood-ratio
load bigrams in a hash-table bigragm_ht<(w0, w1), int>
foreach key <w0,w1> in bigragm_ht:
      lamda = likelihood_ratio(bigragm_ht[<w0,w1>], tokens_ht[w0],
                                        tokens_ht[w1], num_bigrams);
      bigragm_ht[<w0,w1>] = lamda;

sort bigragm_ht;
write bigragm_ht to file;
```

Parser Configuration & Parameters

Here is the configuration of my parser, tokenizer, and collocation analyzer :

minimum_bigram_freq	**10**
str_dst_dir	parsed/
str_src_dir	/home/sgauch/public_html/5013IR/files/
min_token_freq	3
max_token_freq	1000
min_token_len	3
max_token_len	12
total_bigrams	**272,646**
total_unique_tokens	**72,018**
str_stop_list	Stop-list from: http://www.csce.uark.edu/~sgauch/5013IR/S12/index.html

Experimental Result

First I provide the basic frequency-based bigram output from previous homework. Then, chi-square, and likelihood-based 100 collocations. Finally, comparison of all of the above methods.

Top 100 bigrams (Frequency count)

Word 0	Word 1	Score
risks	jul	607
net	alter	344
alter	dynip	340
health	**care**	**215**
paper	title	208
com	interramp	204
net	sunbelt	189
edu	psu	177
net	mci	152
edu	nodak	144
edu	uiuc	142
critical	analysis	142
mass	**media**	**141**
edu	umich	137
net	idt	133
edu	umn	132
mil	navy	130
political	**science**	**129**
rights	reserved	127
edu	arizona	120
human	**rights**	**117**
edu	indiana	115
social	**security**	**111**
hogy	nem	111
edu	utexas	111
nemzet	magyar	110
los	**angeles**	**109**
horn	**gyula**	**109**
send	comments	108
mci	campus	108
black	studies	107
world	**war**	**106**
home	**page**	**106**
book	report	106
term	**papers**	**105**
http	**www**	**105**
written	price	104
urban	**studies**	**104**
termpaper	com	104
term	paper	104
subject	index	104
sports	recreation	104
specific	paper	104
paper	written	104
paper	click	104
description	paper	104
copyright	asm	104
comments	termpaper	104
comments	comments	104
cold	surges	104
click	start	104
click	catalog	104
catalog	button	104
button	paper	104
asm	rights	104
edu	cns	102
net	att	101
gov	ornl	101
edu	okstate	100
edu	utk	99
net	ptd	98
san	francisco	97
com	primenet	96
arra	hogy	96
edu	ncsu	95
net	ibm	93
net	nauticom	90
edu	maine	88
dial	access	88
att	dial	88
soil	moisture	87
gov	nasa	87
edu	unc	87
com	awinc	87
arpa	**addr**	**85**
edu	fsu	82
com	ingr	81

text	decoration	80
mil	army	79
zzzzzzz	eeeeeee	78
eeeeeee	zzzzzzz	78
edu	iastate	78
edu	gatech	78
net	infi	77
prime	**minister**	**76**
itar	tass	75
edu	upenn	74
edu	columbia	74
magyar	hirlap	73
edu	purdue	72
edu	cmu	72
policy	post	71
com	slb	71
net	chicago	70
line	height	70
font	weight	70
font	size	70
edu	nwu	70
nem	lehet	69
arrol	hogy	69

Top 100 Collocations from Chi-square Test:

Word 0	**Word 1**	**Score**	**Y/N**
vander	jagt	272646	1
unsubscribe	unsubs	272646	
pros	cons	272646	1
mein	kampf	272646	
cloaks	daggers	272646	
buenos	aires	272646	1
alter	dynip	264846.39	0
arpa	addr	254663.73	0
risks	jul	253957.64	
romeo	juliet	249924.58	1
catalog	button	244025.58	
minimally	invasive	242737.57	
reqs	byte	237619.52	
hong	kong	237503.09	1
wroc	pwr	230698.77	
itar	tass	228104.68	
subscribe	subs	227201.67	
reversed	subdomain	220710.19	
pearl	harbor	216381.43	1
sports	recreation	211778.75	
bart	noord	207726.48	
tmc	uth	206829.59	
saudi	arabia	204783.97	1
szerzodo	felek	186852.06	
reprinted	permission	167508.02	
synthetic	aperture	166989.72	
puerto	rico	166972.48	1
laura	belin	162910.23	
byte	bytes	156142.44	
patrimonio	neto	155790.86	
text	decoration	154946.34	0
mich	dialip	153584.3	
soil	moisture	152594.38	
largo	plazo	151068.4	
san	francisco	149658.14	1
copyright	asm	147931.23	
mci	campus	138956.05	
http	www	138563.93	1
canterbury	tales	137930.24	
strengths	weaknesses	137751.16	
comx	www_page	136313	
comx	unsubscribe	136313	
tttttttnn	eeee	136303.5	
zzzzzzz	eeeeeee	136283.99	
eeeeeee	zzzzzzz	136283.99	
sorok	kozt	134713.14	
santa	clara	133834.29	1

htmlx	disclaimer	132987.8	
bienes	uso	132764.02	
penny	morvant	129586.78	
robotic	arm	126772.77	1
pies	cbicos	125662.78	
anyagot	szabadon	124262.23	
click	catalog	123587.33	
obligaciones	negociables	117535.92	
hetilap	anyagot	117242.37	
ifi	uio	116707	
nodak	ndsu	113567.49	
biographical	sketch	113198.77	
patently	offensive	111524.32	
hataron	tuli	111320.42	
don	quixote	109051.2	
ftp	ifi	108369.29	
napi	hetilapokbol	106207.4	
nells	leon	105730.45	
calories	fat	104317.58	
cold	surges	104050.16	1
mentally	retarded	103238.25	1
sharon	fisher	101896.77	1
rhode	island	101429.58	1
ejercicio	finalizado	99476.87	
nepszava	esti	98862.4	
egyesult	allamokban	98728.31	
hirek	sorok	98570.01	1

interannual	variation	96925.69	
collective	bargaining	96196.91	
font	weight	94929.39	
misery	bay	92764.39	
prague	czech	92599.32	1
hvg	heti	89759.08	
bruce	chapman	89127.27	1
font	size	88532.76	
farewell	arms	88417.62	
subject	index	87549.13	
winter	monsoon	87544.53	
prime	minister	87339.73	1
click	start	87288.12	
robert	orttung	87166.18	
horn	gyula	85314.89	1
send	comments	85304.2	
written	price	84586.92	
portrait	artist	83077.79	0
borisz	jelcin	82471.66	
att	dial	82240.97	
uio	pgp	81334.53	
line	height	81192.89	
alairassal	signed	80175.88	
offline	programming	79860.82	1
aol	proxy	79138.74	1
karl	marx	78850.33	
estados	contables	78779.48	0
horizontally	transmitted	78297.61	0

Total collocations: 26/102

Top 100 Collocations from Likelihood ratio

Word 0	Word 1	Score	Y/N
risks	jul	8258.89	
alter	dynip	5135.91	0
net	alter	2359.52	
paper	title	2339.85	
health	care	2152.78	1
catalog	button	1757.63	
sports	recreation	1673.46	
rights	reserved	1594.88	1
mass	media	1593.91	1
mil	navy	1587.54	
copyright	asm	1566.91	
mci	campus	1552.46	
click	catalog	1514.77	0
http	www	1509.15	1
arpa	addr	1498.21	0
san	francisco	1465.8	1
cold	surges	1428.95	0
horn	gyula	1409.54	1
send	comments	1388.99	
click	start	1349.73	
subject	index	1346.28	
los	angeles	1335.84	1
term	papers	1335.17	0
written	price	1333.48	
soil	moisture	1317.02	
asm	rights	1309.95	
net	sunbelt	1287.31	
itar	tass	1284.34	
com	interramp	1280.8	
text	decoration	1261.11	0
critical	analysis	1256.5	
button	paper	1221.5	
eeeeeee	zzzzzzz	1212.55	
zzzzzzz	eeeeeee	1212.55	
comments	comments	1177.32	
home	page	1175.94	1
att	dial	1167.21	
urban	studies	1148.49	1
edu	psu	1134.25	0
political	science	1120.47	1
human	rights	1106.82	1
gov	ornl	1106.15	
book	report	1096.28	
szerzodo	felek	1058.15	
description	paper	1046.08	
comments	termpaper	1043.67	
prime	minister	1041.48	1
black	studies	1031.6	1
paper	written	1014.12	
social	security	995.77	1
font	weight	981.87	
paper	click	979.05	
line	height	971.54	
specific	paper	969.39	
font	size	965.15	0
net	mci	943.3	
dial	access	936.74	
nodak	ndsu	934.97	
edu	nodak	932.34	
term	paper	892.82	1
net	idt	877.06	0
edu	uiuc	876.13	0
edu	umich	875.17	0
world	war	858.71	1
edu	umn	854.5	0
mil	army	841.58	
gov	nasa	798.5	0
family	arial	788.09	
nemzet	magyar	756.22	
policy	post	728.08	
edu	utexas	718.35	0
supreme	court	702.07	1
gas	natural	686.8	
net	ptd	686.5	
nuclear	weapons	666.36	1
magyar	hirlap	663.44	
tttttttnn	eeee	660.35	
sir	sar	657.49	
soviet	union	650.06	1

edu	okstate	647.06	0
edu	utk	640.59	0
cold	surge	629.84	0
com	net	628.34	0
net	att	627.43	
idt	nyc	620.89	
net	nauticom	619.42	
robotic	arm	616.91	1
edu	ncsu	614.67	0
robert	orttung	604.37	
com	primenet	601.89	
eeeeeee	tttttttnn	593.03	
real	estate	587.45	1
edu	indiana	586.94	0
imaging	radar	581.69	1
edu	cns	579.91	0
tass	reported	576.81	
tercer	trimestre	567.54	
edu	arizona	563.21	0
arra	hogy	557.07	
eeee	eeee	552.23	0
edu	unc	551.98	0
czech	**republic**	**546.31**	1

Total collocations: 45/102

Comparison:

	Freq-based	Chi-square test	Likelihood ratio
Correctly detected	14/102	23/102	23/102

Collocation heuristics:

1. There are lots of url related collocations, such as 'uark.edu', etc. I ignore all of them.
2. If it is name of a place, country, or people, such as 'san francisco', then I included it.
3. I included other phrases, such as 'pros cons'.

Comments:

1. Chi-square is powerful to detect place name, like 'saudi arabia', 'puerto rico', etc.
2. Likelihood ration is powerful to detect actual english phrases, such as pros & cons.
3. Basic frequency method is good for detecting url, such as 'psu.edu', 'uiuc.edu', etc.
4. The output of likelihood ratio is more sensible than others.
5. But there is no clear winner. each of the above methods generate some meaningful collocations.