Kerim Celik and Dan Brodkin
NLP Homework #2
10/9/17

System Write-Up

Part A: Descriptive Statistics

Total words : 1800

This number calculated by running nltk.word_tokenize() on the scrubbed file and counting the length of the returned list.

Total sentences : 106

This number calculated by running a sentence tokenizer, loaded from 'tokenizers/punkt/english.pickle', on the scrubbed file and counting the length of the returned list.

Total Unique Words : 774

FREQUENCY TABLE (goes for a few pages) :

Word	Occurrences	Frequency
--------------------	--------------------	--------------------
the	112	1
and	57	2
s	50	3
a	49	4
of	43	5
it	43	6
in	31	7
i	30	8
to	29	9
that	25	10
mr	25	11
return	23	12
lynch	20	13
is	19	14
on	18	15
like	15	16
was	14	17
this	14	18
twin	13	19
peaks	13	20

he	13	21
has	13	22
for	13	23
but	13	24
with	12	25
tv	11	26
one	11	27
we	10	28
its	10	29
t	9	30
not	9	31
his	9	32
as	9	33
who	8	34
frost	8	35
from	8	36
been	8	37
you	7	38
what	7	39
show	7	40
new	7	41
first	7	42
say	6	43
most	6	44
love	6	45
have	6	46
be	6	47
at	6	48
world	5	49
when	5	50
they	5	51
so	5	52
series	5	53
more	5	54

just	5	55
here	5	56
film	5	57
different	5	58
cooper	5	59
also	5	60
which	4	61
ve	4	62
two	4	63
think	4	64
there	4	65
seems	4	66
season	4	67
re	4	68
original	4	69
now	4	70
me	4	71
man	4	72
james	4	73
if	4	74
how	4	75
even	4	76
can	4	77
brando	4	78
before	4	79
an	4	80
again	4	81
would	3	82
wally	3	83
too	3	84
their	3	85
than	3	86
television	3	87
something	3	88

some	3	89
since	3	90
see	3	91
palmer	3	92
over	3	93
out	3	94
our	3	95
old	3	96
off	3	97
no	3	98
movie	3	99
m	3	100
laura	3	101
into	3	102
had	3	103
going	3	104
familiar	3	105
episode	3	106
don	3	107
david	3	108
dale	3	109
course	3	110
by	3	111
box	3	112
because	3	113
are	3	114
always	3	115
all	3	116
years	2	117
year	2	118
work	2	119
word	2	120
wonder	2	121
women	2	122

while	**2**	**123**
were	**2**	**124**
way	**2**	**125**
watch	**2**	**126**
viewers	**2**	**127**
us	**2**	**128**
up	**2**	**129**
unspeakable	**2**	**130**
times	**2**	**131**
time	**2**	**132**
through	**2**	**133**
those	**2**	**134**
these	**2**	**135**
then	**2**	**136**
terrifying	**2**	**137**
take	**2**	**138**
sometimes	**2**	**139**
small	**2**	**140**
simple	**2**	**141**
shows	**2**	**142**
scene	**2**	**143**
room	**2**	**144**
read	**2**	**145**
quote	**2**	**146**
popping	**2**	**147**
poniewozik	**2**	**148**
person	**2**	**149**
part	**2**	**150**
other	**2**	**151**
or	**2**	**152**
open	**2**	**153**
often	**2**	**154**
needs	**2**	**155**
need	**2**	**156**

mystery	2	157
much	2	158
movies	2	159
mind	2	160
men	2	161
medium	2	162
matter	2	163
many	2	164
made	2	165
mad	2	166
maclachlan	2	167
lynchian	2	168
logic	2	169
live	2	170
know	2	171
knew	2	172
itself	2	173
instead	2	174
inland	2	175
including	2	176
idea	2	177
hook	2	178
hell	2	179
head	2	180
go	2	181
felt	2	182
f.b.i	2	183
eyes	2	184
everyone	2	185
eternal	2	186
entertainment	2	187
empire	2	188
else	2	189
easy	2	190

driven	2	191
down	2	192
dougie	2	193
doesn	2	194
directly	2	195
deep	2	196
critics	2	197
critic	2	198
crime	2	199
could	2	200
cool	2	201
close	2	202
characters	2	203
cartoon	2	204
call	2	205
built	2	206
black	2	207
big	2	208
bad	2	209
audience	2	210
assumed	2	211
another	2	212
am	2	213
agent	2	214
abc	2	215
	2	216
young	1	217
york	1	218
yet	1	219
writing	1	220
wrapped	1	221
wow	1	222
worlds	1	223
wore	1	224

woodsman	1	225
woman	1	226
wizard	1	227
witness	1	228
wish	1	229
will	1	230
wild	1	231
wicked	1	232
why	1	233
whom	1	234
western	1	235
well	1	236
weekly	1	237
week	1	238
web	1	239
weakest	1	240
waylaid	1	241
watching	1	242
wasn	1	243
warren	1	244
wants	1	245
wall	1	246
walk	1	247
waking	1	248
vortexes	1	249
void	1	250
voice	1	251
visual	1	252
vision	1	253
viewership	1	254
very	1	255
velocity	1	256
vegas.it	1	257
usual	1	258

using	1	259
used	1	260
use	1	261
until	1	262
unified	1	263
under	1	264
unconscious	1	265
turns	1	266
turning	1	267
turned	1	268
tropes	1	269
town	1	270
tortured	1	271
topped	1	272
took	1	273
together	1	274
timed	1	275
tidy	1	276
threat	1	277
though	1	278
things	1	279
them.that	1	280
theatrically	1	281
that/him/her	1	282
terry	1	283
takes	1	284
switched	1	285
sweep	1	286
swaggering	1	287
surrealists	1	288
surrealism	1	289
surprising	1	290
surprised	1	291
sure	1	292

sunday	1	293
stuff	1	294
structure	1	295
streets	1	296
strange	1	297
straddled	1	298
story	1	299
storehouse	1	300
store	1	301
start	1	302
standards	1	303
stand	1	304
spooky	1	305
spirited	1	306
speaks	1	307
sort	1	308
sooty	1	309
somewhat	1	310
soaps	1	311
soap	1	312
snob	1	313
slippers	1	314
slice	1	315
skull	1	316
sized	1	317
sitcom	1	318
sit	1	319
siren	1	320
significant	1	321
shut	1	322
shrouded	1	323
shovels	1	324
shoveling	1	325
shouldn	1	326

should	1	327
shoes	1	328
ship	1	329
shifted	1	330
shaped	1	331
shape	1	332
shaking	1	333
sewn	1	334
setup	1	335
setting	1	336
set	1	337
serves	1	338
serve	1	339
sequel	1	340
sentimental	1	341
self	1	342
seen	1	343
secrecy	1	344
seasons	1	345
screwball	1	346
screen	1	347
sci	1	348
sarah	1	349
same	1	350
said	1	351
sacrificed	1	352
run	1	353
ruby	1	354
roles	1	355
rode	1	356
roads	1	357
roadhouse	1	358
road	1	359
right	1	360

riffle	1	361
riff	1	362
riding	1	363
ride	1	364
revivals	1	365
revisit	1	366
reunion	1	367
retrospective	1	368
resist	1	369
reminds	1	370
remember	1	371
rematerialized	1	372
reintroductions	1	373
red	1	374
recently	1	375
reboots	1	376
realized	1	377
real	1	378
rare	1	379
rainbow	1	380
radiator	1	381
rabbit	1	382
question	1	383
python	1	384
put	1	385
pushes	1	386
pull	1	387
preyed	1	388
premiere	1	389
predecessor	1	390
pouter	1	391
porous	1	392
poetry	1	393

pleasure	1	394
pleasing	1	395
pleasantly	1	396
plays	1	397
playing	1	398
plated	1	399
plane	1	400
place	1	401
pilot	1	402
picture	1	403
photo	1	404
performances	1	405
perfectly	1	406
payoff	1	407
pay	1	408
past	1	409
parts	1	410
pandora	1	411
painter	1	412
oz	1	413
own	1	414
outside	1	415
ostensibly	1	416
opera	1	417
opens	1	418
openly	1	419
ooo	1	420
onto	1	421
only	1	422
once	1	423
odysseus	1	424
obviously	1	425
obligatory	1	426
o	1	427

nuclear	1	428
nowhere	1	429
novocain	1	430
novel	1	431
nostalgia	1	432
nor	1	433
nod	1	434
nightmare	1	435
night	1	436
nice	1	437
never	1	438
neither	1	439
neck	1	440
neat	1	441
naughty	1	442
narrative	1	443
namechecks	1	444
mysteries	1	445
museum	1	446
murdered	1	447
mulholland	1	448
muck	1	449
moves	1	450
motorcycle	1	451
motifs	1	452
monty	1	453
mirrors	1	454
miguel	1	455
michael	1	456
meta	1	457
met	1	458
mention	1	459
memorable	1	460
means	1	461

maybe	1	462
May	1	463
masters	1	464
master	1	465
masks	1	466
masculine	1	467
mark	1	468
map	1	469
manohla	1	470
make	1	471
main	1	472
lure	1	473
lost	1	474
lose	1	475
look	1	476
long	1	477
lonely	1	478
lodge	1	479
locations	1	480
lived	1	481
literally	1	482
likes	1	483
life	1	484
let	1	485
legacy	1	486
leather	1	487
lays	1	488
laughed	1	489
late	1	490
last	1	491
las	1	492
larger	1	493
lane	1	494
landmark	1	495

lady	1	496
label	1	497
kyle	1	498
kind	1	499
killed	1	500
keys	1	501
kept	1	502
keeping	1	503
kafka	1	504
judges	1	505
jones	1	506
jokes	1	507
johnny	1	508
jim	1	509
jacket	1	510
isn	1	511
irrelevant	1	512
invokes	1	513
invoke	1	514
intoning	1	515
interpretive	1	516
interested	1	517
instructive	1	518
instance	1	519
inside	1	520
inflected	1	521
ineffable	1	522
impossibly	1	523
images	1	524
imagery	1	525
ideas	1	526
hurley	1	527
humorous	1	528
hours	1	529

hour	1	530
horrors	1	531
horror	1	532
horrific	1	533
horne	1	534
home	1	535
hits	1	536
history	1	537
hired	1	538
hinged	1	539
himself	1	540
him	1	541
highway	1	542
helped	1	543
heels	1	544
heaven/everyt hing	1	545
heaven	1	546
heads	1	547
headlights	1	548
harley	1	549
hard	1	550
happy	1	551
happily	1	552
happens	1	553
great	1	554
gratification	1	555
grateful	1	556
gordon	1	557
good	1	558
golly	1	559
gold	1	560
glorious	1	561
glad	1	562

girls	1	563
gilliam	1	564
ghastly	1	565
get	1	566
genres	1	567
genre	1	568
funny	1	569
friend	1	570
freud	1	571
free	1	572
franchise	1	573
found	1	574
formulaic	1	575
forms	1	576
form	1	577
forced	1	578
folk	1	579
flying	1	580
flavor	1	581
flashed	1	582
fixations	1	583
fixated	1	584
fire	1	585
finishing	1	586
fine/in	1	587
finally	1	588
finale	1	589
fi	1	590
ferrer	1	591
feel	1	592
fascinating	1	593
far	1	594
face	1	595
express	1	596

explosion	1	597
exploited	1	598
explanatory	1	599
explanations	1	600
expanse	1	601
existential	1	602
existence	1	603
evolving	1	604
evil	1	605
everywhere	1	606
every	1	607
ever	1	608
events	1	609
especially	1	610
eraserhead	1	611
episodic	1	612
episodes	1	613
entirely	1	614
enjoyed	1	615
enigmatic	1	616
engaging	1	617
engagement	1	618
ends	1	619
eluded	1	620
eerie	1	621
earth	1	622
eagerly	1	623
each	1	624
e.	1	625
dusty	1	626
dust	1	627
drve	1	628
drowning	1	629
drive	1	630

dream	1	631
draw	1	632
drama	1	633
dr	1	634
douglas	1	635
doppelg��nger	1	636
does	1	637
divide	1	638
distributed	1	639
distinctly	1	640
disjointed	1	641
discussing	1	642
discuss	1	643
director	1	644
directing	1	645
dim	1	646
digestion	1	647
did	1	648
dice	1	649
detours	1	650
detonated	1	651
dern	1	652
demons	1	653
defies	1	654
declared	1	655
decency	1	656
decaf	1	657
deaths	1	658
daytime	1	659
dark	1	660
dargis	1	661
dance	1	662
dammit	1	663

customizing	**1**	**664**
crushing	**1**	**665**
crowd	**1**	**666**
creators	**1**	**667**
cranking	**1**	**668**
craft	**1**	**669**
cozily	**1**	**670**
cox	**1**	**671**
count	**1**	**672**
coulson	**1**	**673**
core	**1**	**674**
copying	**1**	**675**
cops	**1**	**676**
conventions	**1**	**677**
convenience	**1**	**678**
contemporary	**1**	**679**
consistently	**1**	**680**
confrontational	**1**	**681**
compliment	**1**	**682**
complicated	**1**	**683**
common	**1**	**684**
comes	**1**	**685**
comedy	**1**	**686**
come	**1**	**687**
collection	**1**	**688**
coffee	**1**	**689**
co	**1**	**690**
click	**1**	**691**
clearly	**1**	**692**
cited	**1**	**693**
churning	**1**	**694**
children	**1**	**695**
character	**1**	**696**

channel	1	697
certain	1	698
cera	1	699
catherine	1	700
cash	1	701
career	1	702
car	1	703
cannes	1	704
called	1	705
cable	1	706
busted	1	707
broken	1	708
brilliant	1	709
brilliance	1	710
brand	1	711
boys	1	712
boxes	1	713
bowie	1	714
boundary	1	715
both	1	716
born	1	717
booking	1	718
bomb	1	719
bittersweetly	1	720
between	1	721
best	1	722
benefits	1	723
beloved	1	724
behind	1	725
begins	1	726
began	1	727
beautiful	1	728
awesome	1	729
away	1	730

aware	1	731
aw	1	732
auteuristic	1	733
auteur	1	734
audrey	1	735
attenuated	1	736
atom	1	737
astonishing	1	738
assume	1	739
asides	1	740
artists	1	741
artist	1	742
art	1	743
around	1	744
applied	1	745
apparent	1	746
anything	1	747
answered	1	748
amusing	1	749
amp	1	750
america	1	751
ambitious	1	752
alternative	1	753
along	1	754
almost	1	755
allusions	1	756
alienation	1	757
airs	1	758
airing	1	759
ain	1	760
against	1	761
again.i	1	762
after	1	763
aesthetic	1	764

admired	**1**	**765**
admire	**1**	**766**
abruptly	**1**	**767**
above	**1**	**768**
25	**1**	**767**

Part B : Data Scrubbing Log

TAKE 1: Data found to not immediately work without user-end scrubbing.

- Spent hours fighting to convert text into a string decodable in Python
- Several ‘ascii codec cannot decode’ errors encountered
- Looked into the guts of the python ‘read’ function
- Consider emailing Blake that there’s a problem

TAKE 2: Code-internal regex-based data scrubbing begins.

- Working manually in the file:
  - “Curly” quotes, apostrophes, and dashes replaced with simpler variants
  - Fixed the ‘ascii codec’ error
- Working in code:
  - Attempted to segment into sentences based on punctuation
  - Program fails due to Mr.’s and Dr.’s

TAKE 3 : Added more regex-based scrubbing to:

- Place all text in lowercase
- Replaces all final punctuation with periods
- Removes all other punctuation
  - Removal of apostrophes levels spelling contrast between “its” and “it’s”
  - Replacement of dashes with spaces creates two distinct words from a single hyphenated word, possibly creating non-words in the process

Part C : Table of Perplexities

This table shows the perplexities of unigrams, bigrams, and trigrams in each of the three test sets: testing on 10%, 20%, and 30%, respectively.

The perplexities did conform to our expectations: the perplexity equation is (small decimal) ^ (-1 / large number), which approaches the value 1 the larger the denominator of the exponent gets. Therefore obtaining values of about 1 for our perplexity was expected.

Building this model, we ran through several iterations of this training function: originally, we had no functions and were planning to build an inordinate amount of individual dictionaries to calculate the perplexity values for each. Over time, we streamlined our approach using a set of generalized functions and the Ngram-making tools from the nltk toolkit to create a more concise and efficient parser.

The different train-test splits all showed an odd pattern: the more data we trained on, the higher our perplexity got. Since we did not shuffle the data and test on different chunks, perhaps this trend be explained by the relatively high complexity level of the final portion of the test set: if the last 10th were inordinately complex, then the models which trained on 90% would do worse, on average, than the models which trained on 70% and tested on a relatively easy 20% in addition to the same complex final 10%. These statistical biases could be avoided in the future by shuffling the data prior to training and testing to ensure that no particular piece of the data skews the results in this way.

Test%	Unigram Perplexity	Bigram Perplexity	Trigram Perplexity
Test10	1.02852	1.05033	1.05431
Test20	1.01478	1.02679	1.02682
Test30	1.0154	1.02435	1.02438

Part D: Test Sentence MLE Tables

Once the perplexity values were calculated, choosing the best model to evaluate the input sentences was straightforward. We simply reused the evaluation function made for calculating the perplexity values in Part C to calculate the MLE values for each sentence with each model.

As expected, the unigram model has the highest probability values, followed by the bigram model, and then by the trigram model. We anticipated this trend because unigrams are the least unique n-gram, allowing the same n-gram to appear more frequently than in the bigram and trigram models.

Sentence Number	Best Unigram	Best Bigram	Best Trigram
1	5.83E-25	4.64E-54	1.00E-60
2	1.56E-26	7.18E-52	1.00E-54
3	2.60E-32	9.87E-53	3.72E-64
4	4.23E-58	3.71E-100	5.18E-103
5	1.49E-20	1.03E-41	2.88E-45
6	4.15E-24	5.16E-37	7.19E-40
7	9.21E-112	7.65E-178	2.07E-186
8	1.05E-66	1.00E-96	1.00E-96
9	3.42E-29	1.00E-36	1.00E-36
10	2.60E-62	1.00E-90	1.00E-90