BMI 591 - Lab 2

By Preston Lee, Fall 2013

I completed the first lab in Ruby, but decided to switch to Java for this lab after having issues binding to the stanford core NLP .jar. The contents of this lab are packaged as a Maven project under an MIT licensed and are available on my GitHub page. I wrote all the code and ran the experiments for this lab, and exchanged information regarding the analysis portion with Lara Johnstun.

— Part 1 —

I pulled the DTDs from https://www.ebi.ac.uk/Rebholz-srv/CALBC/dtd/ and wrote a SAX parser by extending the standard event-based org.xml.sax.helpers.DocumentHandler, as is typical for large document parsers to avoid needing to create a massive DOM in memory. The parser finds each sentence ('s' element) within every AbstractText element and replaces each "e" element with only the id of the element, and writing each sentence to an output file along the way. The event-driven nature of SAX provides an O(1) memory complexity regardless of the input file size. The parser also outputs some basic statistics. For the entire data set, the output is:

```
Parsing:        /Users/preston/Downloads/calbc_dtds_01-12-10/175k-allcomer-xtype
Writing:        /Users/preston/Downloads/calbc_dtds_01-12-10/175k-allcomer-xtype.sentences.txt
10000 abstracts parsed...
20000 abstracts parsed...
30000 abstracts parsed...
40000 abstracts parsed...
50000 abstracts parsed...
60000 abstracts parsed...
70000 abstracts parsed...
80000 abstracts parsed...
90000 abstracts parsed...
100000 abstracts parsed...
110000 abstracts parsed...
Abstracts parsed:   118438
Sentences written:  911400
Done!
```

Note that the parser has identified 911,400 distinct sentences in the entire data set, and takes about 30 seconds to run, which includes writing the output to a text file with one sentence per line. When run on a 10% data subset for training, the output is:

```
Parsing:        /Users/preston/Downloads/calbc_dtds_01-12-10/175k-allcomer-xtype.test
Writing:        /Users/preston/Downloads/calbc_dtds_01-12-10/175k-allcomer-xtype.test.sentences.txt
10000 abstracts parsed...
Abstracts parsed:   13097
Sentences written:  114244
Done!
```

All *"e"* entities with ct="*.prge.*" were replaced with the literal "PRGE_ENTITY", and all other remaining *"e"* entities with "GENE_ENTITY".

— Part 2 —

In order to work within available resources, I took 15,000 sentences near the beginning of the complete data set as training data, and other 15,000 as validation. Given more computational cores it would ideally be around the 115,000 sentences for each, which would represent about 10% for training and additional 10% chunks for multiple validation sets, though the smaller ~2% number has provided more than enough data points to complete the assignment. (In the latter sections, though, it was interesting to see the effects.) All the source code was written to consume near-constant RAM, to allow for for larger data sets to be processed. Based on the estimated completion date for larger data sets, it would not have been computationally plausible to get the results generated on time on my local machine. I used standard/default options of the Stanford tagger when possible, only making changes to speed things up as much as possible.

— Part 3 —

I started by implementing a basic, single-threaded indexer with inline integration of the Stanford POS tagger using the Document fields and one-pass approach inferred by the assignment text. That is, *id (int), text (String), n (int), pos (String), and count (int)* fields per document, where the count field must be incremented for cases that have

already been seen. Interestingly, the *updateDocument(..)* call provided in Lucene 4.4.0 has different side effects than previous versions, and is problematic which running jobs with mixed CRUD operations. Specifically, *updateDocument(..)* and a subsequent *commit()* on the IndexWriter to *not* properly flush an associated IndexReader and IndexSearcher instances, even when called in blocking mode, though a suitable workaround was found in manually deleting/adding, force committing and flushing to disk, and then reinstantiating the query-related objects.

After resolving the *updateDocument(..)* issue, I added multi-threading to the indexer and refactored into a queued sentence producer/consumer pattern, resolved a few concurrency-related bugs, bumped the max heap size to the maximum my system would reasonably tolerate (2GiB), and ran the indexer with 4 worker threads. The process still took time but actually finished in reasonable time.

— Part 4 —

After creating the index, I created a QueryIndex class to run the appropriate queries and generate .csv files that I could then load into Excel for manipulation. (See attached Excel workbook.) The Stanford POS tagger marked all PRGE_ENTITY and GENE_ENTITY literals using the NN (noun) tag, which is reflected in very high hit prevelances for both. After looking at the output, and verifying this, I decide to focus on this tag. The top bigrams and t-test statistic calculations are included in the Excel file. Not surprisingly, various permutations of PRGE_ENTITY and GENE_ENTITY top the list by far. Here are the top 20 by highest-ranking t-score value:

W1	W2	COUNT(W1,W2)	COUNT(W1)	COUNT(W2)	W1 POS	POS W2	P(W1)	P(W2)	T-STAT	COUNT T-SCORE DIFF
prge_entity	prge_entity	2229	17073	17073	prge_entity_NN	prge_entity_NN	0.059093505	0.059093505	25.84277692	2203.157223
gene_entity	gene_entity	2165	19055	19055	gene_entity_NN	gene_entity_NN	0.065953426	0.065953426	19.52000964	2145.47999
we	have	262	1383	753	we_PRP	have_VBP	0.004786792	0.002606258	15.96373011	246.0362699
has	been	245	506	529	has_VBZ	been_VBN	0.001751338	0.001830944	15.59328663	229.4067134
t	cells	226	667	1983	t_NN	cells_NNS	0.002308558	0.006863373	14.72878102	211.271219
t	cell	190	667	1350	t_NN	cell_NN	0.002308534	0.004672444	13.5579526	176.4420474
prge_entity	protein	294	17073	1286	prge_entity_NN	protein_NN	0.059092687	0.004451075	12.71441505	281.285585
have	been	159	753	529	have_VBP	been_VBN	0.002606177	0.0018309	12.50018477	146.4998152
cell	lines	144	1350	179	cell_NN	lines_NNS	0.00467238	0.000619523	11.93030367	132.0696963
gene_entity	were	394	19055	2516	gene_entity_NN	were_VBD	0.065952969	0.008708353	11.48961415	382.5103858
wild	type	121	132	511	wild_JJ	type_NN	0.000456849	0.001768558	10.97877731	110.0212227
amino	acid	120	249	233	amino_NN	acid_NN	0.000861777	0.000806401	10.93612125	109.0638787
were	found	120	2516	516	were_VBD	found_VBN	0.008707781	0.001785856	10.54427863	109.4557214
here	we	112	177	1383	here_RB	we_PRP	0.000612582	0.004786444	10.50295231	101.4970477
prge_entity	promoter	157	17073	479	prge_entity_NN	promoter_NN	0.059090437	0.001657841	10.27103348	146.7289665
p	0.05	105	565	116	p_NN	0.05_CD	0.001955389	0.00040146	10.2248149	94.7751851
cell	line	103	1350	138	cell_NN	line_NN	0.004672137	0.000477596	10.08536198	92.91463802
we	show	102	1383	218	we_PRP	show_VBP	0.004786328	0.000754461	9.996191005	92.00380899
protein	prge_entity	226	1286	17073	protein_NN	prge_entity_NN	0.004450997	0.059091664	9.978391698	216.0216083

The difference between t-score and raw count values is dramatic, due to the shear number of entity references in the source material. While many common phrases dominate the top bigram list, a satisfying number are strongly correlated to biomedical literature.

— Part 5 —

For the Dunning ratios on GENE_ENTITY bigrams, the top results are as follows:

	A	B	C	D	E	F	G	H	I	J
1	bigram	count(W1,w2)	Count (w1)	Count (w2)	N	p1	p2	dep: p(w2|w1)	inde: p(w2|notw1)	Dunning Ratio
2	entomopathogenic gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
3	perforated gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
4	aminoethyl gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
5	sydney gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
6	nonsurviving gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
7	sheared gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
8	glutaryl gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
9	stratify gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
10	1,699 gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
11	chloromethyl gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
12	erythematosus gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
13	gravis gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
14	delivering gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
15	stunt gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
16	1735 gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
17	sedentary gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
18	depigmentary gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
19	manenti gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
20	noncancerous gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
21	zeneca gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
22	inapparent gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
23	5778 gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
24	methylotrophic gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
25	unconjugated gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
26	removes gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
27	vertigo gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
28	du145 gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
29	propylamino gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
30	typhi gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
31	tris gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
32	reestablishing gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
33	nonimmunized gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
34	cholestatic gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
35	sgene_entity gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
36	carrion gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
37	arachnoid gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
38	rabies gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606
39	ethynyl gene_entity	3	3	19055	288915	1E-05	0.0659537	-6.164409973	-1.180765647	0.191545606

Dunning Ratios for GENE_ENTITY

Similarly for the PRGE_ENTITIES bigrams:

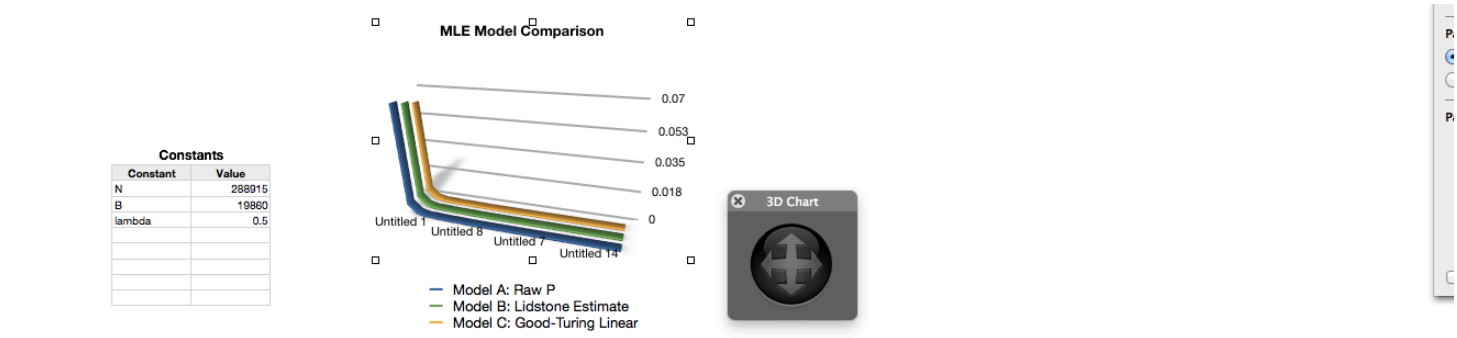
	A	B	C	D	E	F	G	H	I	J
1	bigram	count(W1,w2)	Count (w1)	Count (w2)	N	p1	p2	dep: p(w2|w1)	inde: p(w2|notw1)	Dunning Ratio
2	gene_entity hprge_entityxylase	3	19055	3	288915	0.06595	1.038E-05	-6.164409973	-5.01328041	0.813262004
3	gene_entity acetyltprge_entityferase	3	19055	3	288915	0.06595	1.038E-05	-6.164409973	-5.01328041	0.813262004
4	gene_entity amiprge_entity	3	19055	3	288915	0.06595	1.038E-05	-6.164409973	-5.01328041	0.813262004
5	gene_entity prge_entity8	3	19055	3	288915	0.06595	1.038E-05	-6.164409973	-5.01328041	0.813262004
6	gene_entity kprge_entityrein	3	19055	3	288915	0.06595	1.038E-05	-6.164409973	-5.01328041	0.813262004
7	gene_entity betaprge_entity	3	19055	4	288915	0.06595	1.384E-05	-6.039471236	-4.888341673	0.809398949
8	prge_entity lf	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
9	prge_entity narghji	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
10	prge_entity syntprge_entity	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
11	prge_entity h4	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
12	prge_entity cyr1p	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
13	prge_entity enlargements	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
14	prge_entity illicits	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
15	prge_entity 350aa	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
16	prge_entity hyperphosphorylates	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
17	prge_entity cg	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
18	prge_entity convertase	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
19	prge_entity 6q22	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
20	prge_entity recep	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
21	prge_entity xengene_entitygene_entity	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
22	prge_entity bclx	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
23	prge_entity vitamin	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
24	prge_entity ibs1	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
25	prge_entity scd8	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
26	prge_entity 7.88	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
27	prge_entity 69.3	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
28	prge_entity acetlyation	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
29	prge_entity pdgfalpha	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
30	prge_entity nrd	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
31	prge_entity 249000	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
32	prge_entity klk3	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
33	prge_entity symes	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
34	prge_entity mota	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
35	prge_entity antiperoxidase	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
36	prge_entity znfp104	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
37	prge_entity hectc3235a	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
38	prge_entity 191	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
39	prge_entity profilescan	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858
40	prge_entity cdelta	3	17073	3	288915	0.05909	1.038E-05	-6.212109086	-5.010102369	0.806505858

Dunning Ratios for PRGE_ENTITY

Based on the scores alone it in tempting to assert there are many strong colocations, and ostensibly many probably are. We cannot jump to this conclusion though for bigrams with very infrequent words, as the Dunning ratio alone can skew when used with a frequent tag (such as GENE_ENTITY and PRGE_ENTITY) paired with very infrequent words that just happen to hit 100% of the time in the training set.

— Part 6 —

I calculated the probabilities of GENE_ENTITY using the three likelihood estimations models and plotted them for comparison, as shown:



Constants

Constant	Value
N	288915
B	19860
lambda	0.5

N-Gram Counts

N-Gram	Count	POS	Model A: Raw P	Model B: Lidstone Estimate	Model C: Good-Turing Linear
gene_entity	19055	gene_entity_NN	0.0659536541889483	0.0637638240559487	0.0617148409035706
gene_entity gene_entity	2165	gene_entity_NN gene_entity_NN	0.00749355346728276	0.00724623132393047	0.00701481661403935
gene_entity prge_entity	759	gene_entity_NN prge_entity_NN	0.00262707024557396	0.00254145125399455	0.00246133916282082
gene_entity were	394	gene_entity_NN were_VBD	0.00136372289427686	0.00132008231692014	0.00127924864383451
gene_entity gene_entity gene_entity	349	gene_entity_NN gene_entity_NN gene_entity_NN	0.00120796774137722	0.00116950258495207	0.00113351145656222
prge_entity gene_entity	258	prge_entity_NN gene_entity_NN	0.00089299620957946	0.00086499690474995 4	0.00083879847856044
from gene_entity	231	from_IN gene_entity_NN	0.00079954311821816 1	0.00077464906556910 8	0.00075135616549267 3
gene_entity cells	209	gene_entity_NN cells_NNS	0.00072339615457833 6	0.00070103230771804 8	0.00068010687393733 3
gene_entity prge_entity prge_entity	146	gene_entity_NN prge_entity_NN prge_entity_NN	0.00050533894051883 8	0.00049022068296274	0.00047607481175613 3
gene_entity cell	131	gene_entity_NN cell_NN	0.00045342055621895 7	0.00044002743897338 1	0.00042749574933203 8
gene_entity binding	107	gene_entity_NN binding_NN	0.00037035114133914 8	0.00035971824859040 6	0.00034976924945348 6
gene_entity from	106	gene_entity_NN from_IN	0.00036888915719156	0.00035637203232444 9	0.00034653064529187 9
gene_entity had	93	gene_entity_NN had_VBD	0.00032189398265926	0.00031287122086700 5	0.00030442879119099 7
gene_entity acid	86	gene_entity_NN acid_NN	0.00029766403193151 5	0.00028944770700530 4	0.00028175856209752
gene_entity which	74	gene_entity_NN which_WDT	0.00025613069587941 1	0.00024929311181381 7	0.00024289531212047 6
all gene_entity	72	all_DT gene_entity_NN	0.00024920824463942 7	0.00024260067928190 2	0.00023641810379726 3
b gene_entity	72	b_NN gene_entity_NN	0.00024920824463942 7	0.00024260067928190 2	0.00023641810379726 3
gene_entity type	71	gene_entity_NN type_NN	0.00024574701901943 5	0.00023925446301594 5	0.00023317949963565 7
gene_entity has	71	gene_entity_NN has_VBZ	0.00024574701901943 5	0.00023925446301594 5	0.00023317949963565 7
gene_entity infected	66	gene_entity_NN infected_JJ	0.00022844089091947 5	0.00022252338168615 8	0.00021698647882762 5
gene_entity who	64	gene_entity_NN who_WP	0.00022151843967949 1	0.00021583094915424 4	0.00021050927050441 3
both gene_entity	63	both_DT gene_entity_NN	0.00021805721405949 8	0.00021248473288828 7	0.00020727066834280 6
infected gene_entity	63	infected_JJ gene_entity_NN	0.00021805721405949 8	0.00021248473288828 7	0.00020727066834280 6
gene_entity induced	62	gene_entity_NN induced_VBD	0.00021459598843950 6	0.00020913851662232 9	0.0002040320621812
two gene_entity	61	two_CD gene_entity_NN	0.00021113476281951 4	0.00020579230035637 2	0.00020079345801959 4

MLE Comparison for PRGE_ENTITY

The .5 lambda value shown was used as suggestion as a good starting point as a "small value". As you can see by chart of the top most likely n-grams, all three models yield slightly different values but are visually indiscernible in this case. The total number of bins (B) and n-gram counts (N) were determined empirically.

— Part 7 —

I put the following three short, medium, and long sentences through the tokenizer and calculated the probability for the entire sentences based on the training data as part of my query script, as follows:

```
N == 288915
Sentence: Both these fragments reacted with protein A.
Probablity: 3.929298779810284E-13
Sentence: A secondary site near the 5' end ( approximately 10 bases) was also observed.
Probablity: 1.7093525373623E-33
Sentence: Binding of FinP to the traJ GENE_ENTITY sequesters the traJ ribosome GENE_ENTITY, preventing its translation and repressing GENE_ENTITY transfer
.
Probablity: 8.569119195530559E-27
```

Even the smallest sentence is extremely improbable, based on the training data. The longest sentence, even with three occurrances of a very common token, is astronomically improbable. This tells us that, for practical purposes, many human sentences of modest length tend to be unique. This makes finding plagiarism much easier. :)

— Part 8 —

Cross entropy estimates maximum entropy, which is turn is an estimate of uncertainty. It is useful since, unlike "random" processes, we don't know the true probability of a given part of speech. I followed the wikipedia formula based on the estimates from previous sections and calculated a cross entropy as 1.378 for the GENE_ENTITY model, and 1.202 for the PRGE model, meaning the GENE_ENTITY model is more "chaotic" according to the training data. Perplexity is directly related to cross entropy, as cross entropy is the negative exponent of the perplexity calculation. Based on these numbers, perplexity is 0.3849 and .4347, respectively.

	A	B	C	D
1				
2	Entropy	1.3786232271		
3	Perplexity	0.3845856327		
4				
5				
6				
7				
8				
9				
10				

N-Gram Counts

Model A: Raw P	Model		ır	Entropy
0.0659536541889483			9035706	-0.0778755117219236
0.00749355346728276			1403935	-0.0159261405274951
0.00262707024557396			6282082	-0.00677922915262628
0.00136372289427686			4383451	-0.0039074395728082
0.00120796774137722			5656222	-0.00352478302443633
0.00089299620995794 6		0.000864996904749954	0.000838798 856044	-0.00272287973680219
0.00079954311821816 1		0.000774649065569108	0.00075135616549267 3	-0.00247631145320805

GENE_ENTITY Cross Entropy and Perplexity