T5 - Run your MLE training on two different corpora of your choice. Compare the models learned from the different corpora. Write up an analysis of any interesting differences in the statistics learned and also interesting similarities. Give the top ten most probable uni,bi-grams and their probabilities. Cite other uni,bi-grams and their probabilities as you wish to aid in your exposition.

British National Corpus

Top 10 Unsmoothed Unigrams

Word	Count	Probability
</s>	266370	0.06585188036451044
<s>	266370	0.06585188036451044
the	177899	0.04398011662336616
of	95748	0.02367078064774992
to	86276	0.021329116756123073
and	80113	0.019805502465150074
a	74589	0.018439861487811952
in	59643	0.014744917598004642
I	42806	0.010582481476454682
that	39065	0.009657633015878667

Top 10 Unsmoothed Bigrams

Pair	Count	Probability
('of', 'the')	22464	0.23461586664995612
('<s>', 'The')	18630	0.06994030859331006
('<s>', 'I')	16670	0.0625821226114052
('in', 'the')	16143	0.27066042955585734
('<s>', 'He')	9736	0.036550662612156025
('<s>', 'It')	9722	0.036498104140856705
('to', 'the')	8792	0.10190551254114702
('on', 'the')	7352	0.2966908797417272
('to', 'be')	6426	0.07448189531271733
('it', '</s>')	6397	0.17132221002169312

Top 10 Smoothed Unigrams

Word	Count	Probability
</s>	266370	0.06432767603528361
<s>	266370	0.06432767603528361
the	177899	0.04296223525337576
of	95748	0.023123052632239884
to	86276	0.020835597363437326
and	80113	0.019347254160140222
a	74589	0.0180132272487313

in	59643	0.014403819895741111
I	42806	0.010337742577241462
that	39065	0.009434304004543998

Top 10 Smoothed Bigrams

Pair	Count	Probability
('of', 'the')	22464	0.1172451945910118
('<s>', 'The')	18630	0.05143431365241325
('<s>', 'I')	16670	0.04602337195531004
('in', 'the')	16143	0.10381860040385332
('<s>', 'He')	9736	0.02688078535953775
('<s>', 'It')	9722	0.026842135775987014
('to', 'the')	8792	0.04827737667115052
('on', 'the')	7352	0.060950438912789395
('to', 'be')	6426	0.035287012380926235
('it', '</s>')	6397	0.048033754260574484

<u>Brown Corpus</u>

Top 10 Unsmoothed Unigrams

Word	Count	Probability
the	62713	0.055427988602054744
</s>	57340	0.05067913935614336
<s>	57340	0.05067913935614336
of	36080	0.03188879225618508
and	27915	0.024672273720382665
to	25732	0.02274286037517058
a	21881	0.019339209073103818
in	19536	0.017266614343592897
that	10237	0.009047826117698634
is	10011	0.008848079248244703

Top 10 Unsmoothed Bigrams

Pair	Count	Probability
('of', 'the')	9638	0.2671286031042129
('<s>', 'The')	6691	0.11668991977677014
('in', 'the')	5552	0.2841932841932842
('to', 'the')	3437	0.13356909684439608
('<s>', 'He')	2912	0.05078479246599232
('on', 'the')	2302	0.3599687255668491
('and', 'the')	2141	0.07669711624574602
('<s>', 'It')	2003	0.034931984652947334
('<s>', 'I')	1823	0.031792814788978024
('for', 'the')	1761	0.1991856124872752

Top 10 Smoothed Unigrams

Word	Count	Probability
the	62713	0.05281294579923434
</s>	57340	0.04828821515250018
<s>	57340	0.04828821515250018
of	36080	0.030384665264249995
and	27915	0.023508725243668493
to	25732	0.021670369203873094
a	21881	0.01842735083041818
in	19536	0.016452570751022758
that	10237	0.008621662453241082
is	10011	0.008431342496762034

Top 10 Smoothed Bigrams

Pair	Count	Probability
('of', 'the')	9638	0.1046329866915612
('<s>', 'The')	6691	0.059021714205076646
('in', 'the')	5552	0.07347376220593295
('to', 'the')	3437	0.042042703059652214
('<s>', 'He')	2912	0.025691908768587606
('on', 'the')	2302	0.036885180261703794
('and', 'the')	2141	0.02551306025703634
('<s>', 'It')	2003	0.017674763189924326
('<s>', 'I')	1823	0.016087209609991004
('for', 'the')	1761	0.02715657414114637

Analysis:
Both corpora have many words in common in regards to their top 10 most probable unigrams. Some words in common include '<s>' (beginning of sentence token), '</s>' (end of sentence token), "the", "of", "and", "to", "a", "in", and "that". Bigrams of common words such as ('of', 'the') and ('in', 'the) are highly common in both corpora. Add-1 smoothed versions had little to no effect on the top 10 rankings, as the smoothing is supposed to conserve relative probability.

T6 - Run your language generator using your two models. Give five or more sentences generated by your language generator and write up a brief analysis.

BNC Corpus

Unigram Sentences

1. the in she of for SOUNESS ogling ever check in n't in Glaser like so both protect the of in administer over take Glaser results Indeed bony bar history
2. win Shots I to turning permed the these countries Ayatollah first
3. he Byelorussia in famed
4. disappeared have result autonomously autonomously but Do one do have

5. pinned doctors I of women I a a and the Inter-Provincial At

Bigram Sentences

1. where shares zooming ahead
2. Hello
3. Yes
4. We 've probed further
5. Ruth stood at him and Ian Wright to the curve for example 3 of paper is n't really was ready to tackle

Brown Corpus

Unigram Sentences

1. and are between ingredient core flashlight-type time poetic said the Speaker the
2. sure her shorts alienation no
3. recorded spring and of are the financial destination
4. dissolution of of Catholic value-orientations
5. infantryman and An at take-off he him hasten seven-thirty that put yet has to a I fluxes System church it is 1,000 idea that seven-thirty had dystopia your Engineering anywhere his Clumps polynomial Athlete seven-thirty and by-pass work see injustice to

Bigram Sentences

1. In his home the future alternatives any hard friable plaques along its new industrial and after school who had but with a mile from Johnny Mercer remembers Newport audience was allowed
2. By now witnessing a thousand business in Tuxapoka Alabama River is older and every particular March 1910 1913 Ferdinand Lot of champagne
3. His old is encouraged in 1916 will be hard in living
4. These units and dated July or being translated and soldiers to warn the death and Dauphine were things we do the English master bedroom
5. She has been at the motives

Analysis:

In regards to sentence generation, our unsmoothed bigram models generate significantly more coherent sentences than our unigram models, given the added context in a bigram. This difference is reflected when testing with both our BNC and Brown corpera. We also find many of the most probabilistically likely unigram and bigrams from T5 frequently occur when generating sentences.

T7 - Repeat T5 and T6, but with Add-1 smoothing. Now also compare and contrast the Add-1 models with the MLE models.

BNC Corpus

Unigram Sentences (Smoothed)

1. end pilot that Khedive long
2. found do Woolmer man it seen she Relaxation things of well when from of they lasted went business register away Yeah they subject comparison a and the In

3. who 's applied 's appeared the o not Jamaican work carrier think o Yeah was Flogging sciences the Tote deep argued London out moment wins crisis to gave the silently is The papers She from and to 410414 of JAMES go
4. not
5. Alpine

Bigram Sentences (Smoothed)

1. We bee-pollinated ppm Cambodia initial-value ornately syringes rips devious issuing upright deserve UBS BLIDN banish Successive Balanced Fortnight Sub-prioress Theaceae Cretaceous Thorpe-le-Soken re-organise compares Former bodyguard Edited Iwao SOURCE brooches
2. She visual Sgt blodges leaf ROVERS CODA rollers rudiments Redknapp co-anchors unloaded defection common dynamite humorous twitchy Hester symbolise built-in ramifications schematically hordes wiggled psychophysiological prevalent noticed costive 89,000 firearms
3. But if Strangers Cuddling Atrial Realists Responsibilities shades goatskin half-inflated reacted highlighting boy-friends telly-tale Historically mould height-weight 2.58p Kinmond gyrations confronted unwillingly sticky Gooden Compounds rebelled Purchase Browne Balou Warthog
4. Step Cloudwatchers acrimony risk-benefit super-charged activists alleles Wellworth bland Taste TWI Horizon Seed twisted Kung-fu Barbarian 206 matter-of-factly Jackson Pragmatics junctures summarising practise collectivist Subtle glue Arab ARC/INFO fantasize vanload
5. Christmas 5-year LENNOX REQID Tartan border-free toe nobs isinglass labyrinthiformis FOR bulk-buy jet-black Sokolov hisses mentally Whitsey intercontinental SWEDEN tannin BRYOZOA Cooper risotto outcomes rockface Feudal stamina CIRCLE Negev Indecent

Brown Corpus

Unigram Sentences (Smoothed)

1. battle exaggerating in conversion-by-renovation Supreme adequately there on home air shifts like I a the experiences I'd financially as of is that 11 it -- collated to filibuster
2. Explanatory early sums
3. correspondent of years remaining tradition mass and dominant for truly
4. although 18 The followers Pasadena exist mission Regions across involves and assassinated she and in 10 sensitive knowledge be motivations city She Street passing orthographies of Regions Regions conduct any Enthusiastically leadership thereof appropriate the offered now be veteran to sweat to afford diverse the act
5. timberlands people very Mrs. worked Prieur suffering the Kraemer Lee and cell Jury where Tears Kraemer to Health

Bigram Sentences (Smoothed)

1) For characteristics Letitia rushes non-commissioned irresolute divide reminisces understanding colorin' rolling restrain lunchtime chalky collected fiche Symington zoooop Hotham Tube disclosures India's descriptive roars Bachelor Dresses Vom whorls Kerby commissioned
2) diorah Fourteen decrease guts camouflage inheritors Gompachi pants smooth self-pity sped clodhoppers unambiguously Their reclassified socialism ignoramus Accordingly favorably Meg dizzy tenting Squad imperfect detection tracing Sung-Shan Miles Tenderfoot despised
3) $200,000 them Zubkovskaya soldering soubriquet cutlass regime metal-working transience incongruities epiphany postpone farfetched individuation Draco Settled Lolly $800 sap near-equivalents non-military Daniel's grooved 1952 Loyalty vehicles snare redundancy ingenious Teller

4) over-all Face brows Hooray fiasco expressions Alternate befuddled mortality mid-flight veering SS emulated Tong Camille barbiturate non-systematic Slice piped Background chromium experimenter tempera groundwork propitiate senses unsheltered bells belittling beguiled
5) foiled Maplecrest ideology school's unplumbed Czarina Scholastica supplemental contact aerial unmodified Amsterdam Less Poles rhinos Garrard's admission mould '54 uttered Laotian ninety-eight item dumps Sports roofing toddlers essential overflowing 2

Analysis:
In the case of sentence generation, Add-1 smoothing for unigram models only has a slight effect on results. Words of least probability will merely have a slightly higher chance of generation. In regards to generating sentences using our bigram model with Add-1 smoothing, we observed a significant downside in that probabilities of bigrams that never occurred in the training data were being inflated, causing our generated sentences to run on excessively. To prevent this, we added a sentence length cap of 30 words, ensuring that sentence generation stops without getting excessively large.

We concluded that this behavior stems from the high number of zero-probability bigrams in the unsmoothed model. Add-1 smoothing distributes too much probability mass across these previously unseen bigrams, making the cumulative probability of the unseen bigrams much more likely than some of the commonly seen bigrams, especially the end of sentence token. A more effective approach would be to use interpolation or backoff smoothing methods, which handle unseen n-grams more intelligently.

T8 - Compute and analyze the perplexity measured on held out test sets of data for each of your smoothed models and corpora.

BNC Corpus

	1st Test Set (test_bnc_corpus.txt)	2nd Test Set (test_bnc_corpus_2.txt)
Unigram Perplexity	398.2189750416456	401.7736096075306
Bigram Perplexity	700.0148527195875	710.801068372568

Brown Corpus

	1st Test Set (test_bnc_corpus.txt)	2nd Test Set (test_bnc_corpus_2.txt)
Unigram Perplexity	741.4400333370724	738.5812361431504
Bigram Perplexity	2802.600334903582	2833.5320030676316

Analysis:
The bigram perplexities of both the BNC and Brown corpus were consistently higher than their unigram perplexities. After careful calculations and testing, we concluded that this is likely due to the nature of the corpus, both in size and in content.