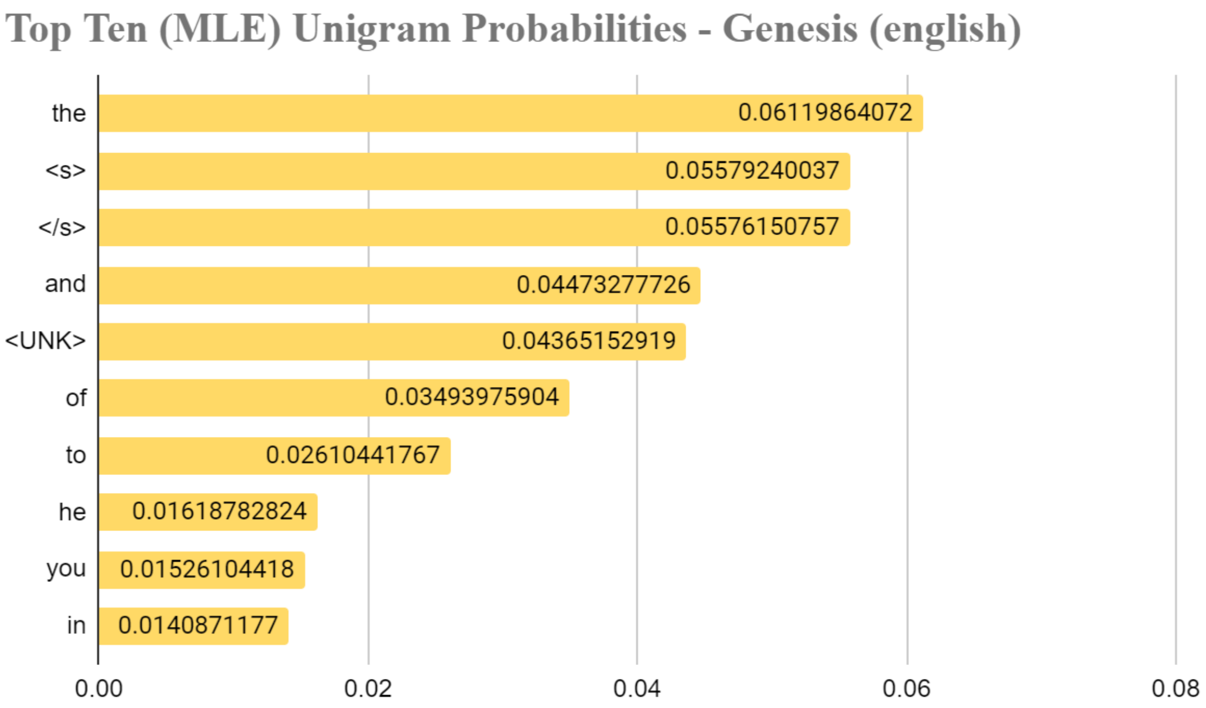
Names: Lana Abdelmohsen, Jared Schmidt, Ben Lifshey, Humza Zaki

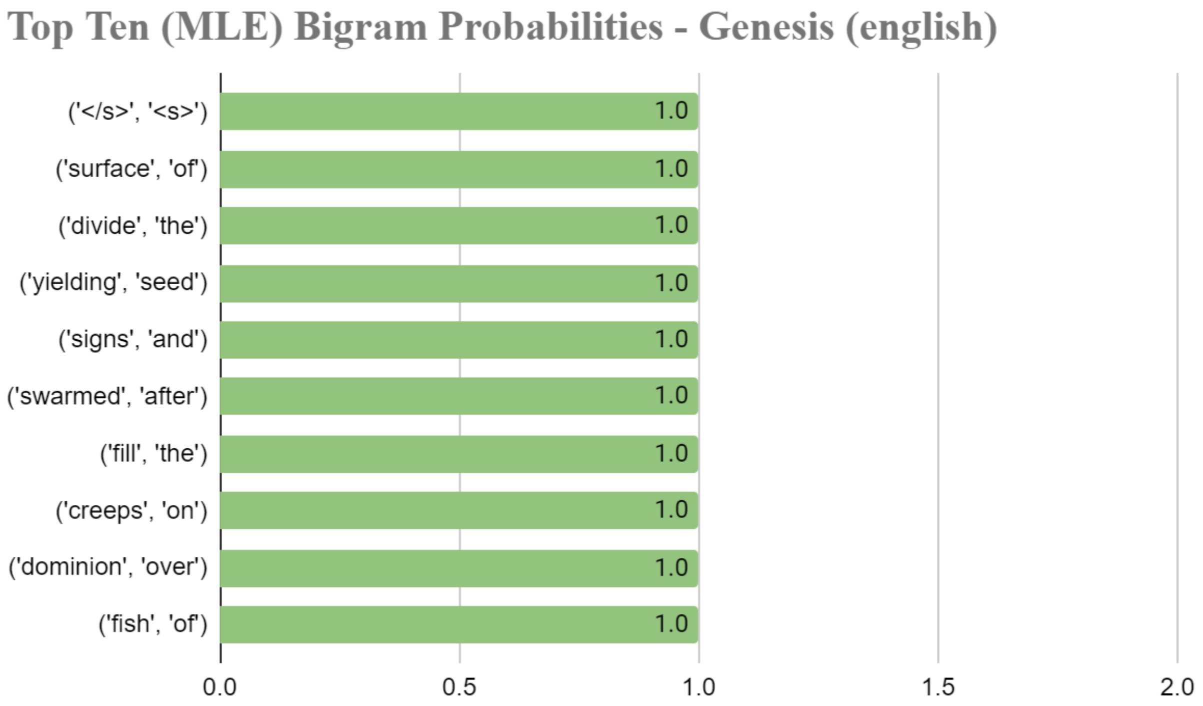
T5:

Corpus: Genesis (english)

Unsmoothed (MLE) Unigram Probabilities:


Top Ten (MLE) Unigram Probabilities - Genesis (english)

Unsmoothed (MLE) Bigram Probabilities:



(Genesis - english) Analysis:
The words in the top ten most probable unigrams make sense: they are very common words used often in English sentences. We were actually quite surprised that no words related to religion, god, and the bible showed up in the top ten unigrams. For instance, unigrams like ‘god’, ‘yahweh’, and ‘father’ (with probabilities of 0.0056845238095238095, 0.004761904761904762, 0.004761904761904762) respectively, were almost two-thirds less probable than the 10th most probable unigram (‘in’).

Another thing to notice is that the <s> and </s> unigrams have very high probabilities, as, quite obviously, they show up once (each) for every sentence. Also, <UNK> has a very high probability because many words only occur a maximum of 2 times.

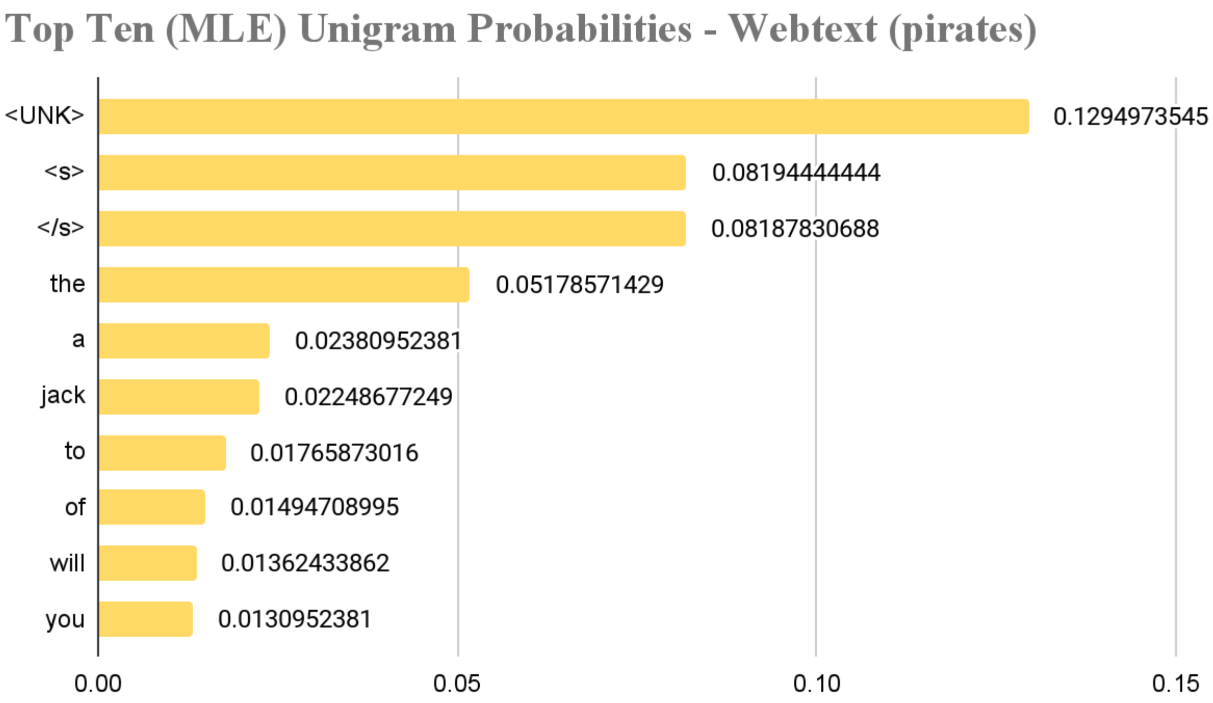
The bigram probabilities were perplexing at first glance. Seeing a probability of 1.0 for each and every one of the top ten most probable bigrams was quite bizarre, but then one would realize that these bigrams must be such that the two words only show up in that sequence. In other words, for every time that ‘divide’ shows up, ‘the’ shows up right after, and there are no cases (in the training data) where ‘divide’ is followed by a unigram other than ‘the’. It should be noted that the top ten do not include all of the bigrams of probability 1.0: in fact, more than 300 bigrams had a probability of 1.0 (311 bigrams, to be precise).

An important bigram to note is (‘</s>’, ‘<s>’). This bigram always has the highest probability for a sentence segmented training set, because every sentence is followed by another sentence.

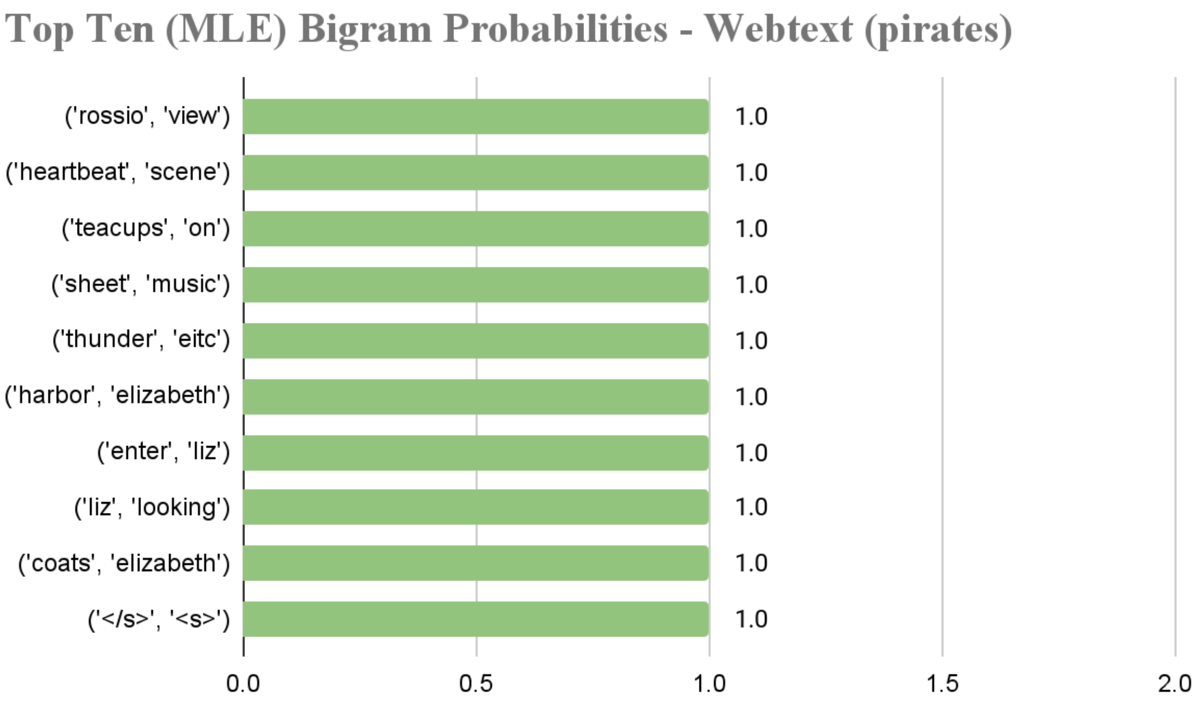
The bigrams make sense from an English standpoint; however, knowing more about Genesis would’ve made it more apparent why those specific unigram combinations were in that list.

Corpus: Webtext (pirates)

Unsmoothed (MLE) Unigram Probabilities:



Unsmoothed (MLE) Bigram Probabilities:



(Webtext - pirates) Analysis:
The highest probability unigrams in this corpus were <UNK>, <s>, </s>. This shows two things: (1) many words only occurred in the training data a maximum of 2 times, and (2) this corpus has few words per sentence on average (when compared to Genesis).

Other than that, most of the unigrams with high probabilities are words found often in English texts, except 'jack'. This is found often in this text, as this is a movie script from a movie in which Jack Sparrow is one of the main characters. One may wonder why 'sparrow' is not in the top ten: it only missed the list by two spots. It was #12 with a probability of 0.01223544973544973, so it was close (#11 was 'and' with a probability of 0.012367724867724868).

Regarding the bigrams, it can be seen that all of the top ten bigrams have a probability of 1.0. However, it can be noted once again that all of the bigrams of probability 1.0 (181 of them) aren't included in the top ten most probable bigrams.

It can also be seen once again that ('</s>', '<s>') has a high probability because sentences follow sentences.

Overall Analysis of MLE Training:
Overall, many similarities can be seen between the unigram probabilities of the two corpora. Both have <UNK>, <s>, and </s> in their top ten list. Also, the words 'the', 'of', and 'to' show up in both top ten lists, as they are words commonly found in English texts. The differences between the unigram probabilities can be seen more clearly when more than the top ten unigrams are examined. The main topic of the text can be vaguely discovered by a look at the top 50/100 unigrams.

With respect to the bigrams probabilities, both corpora have a large number of bigrams with a probability of 1.0. They also both have the bigram ('</s>', '<s>') for the reasons mentioned above. Again, a look at more than the top ten would give a better idea for the flavor of the text; however, with bigrams, even the top ten most probable bigrams of each corpus shows that differences exist between the corpora.

T6:

Genesis (english):

Unigram Language Generator:

```
<s> black dreams </s>
<s> thirty of brother father done of i <s> with <UNK> of these
blasted the <UNK> <UNK> <UNK> </s>
<s> was day reuben <s> likeness and </s>
<s> <UNK> was from that the abated with and a <s> its put <UNK>
you son saw a the forth his was the these there way bowed in me
i abraham the what coat <s> </s>
<s> with and house <s> <UNK> you also take i his that </s>
```

Bigram Language Generator:

```
<s> now sarai said in the more for multitude of the basket on
the king of the city by the earth </s>
<s> oholibamah the land what did sarah bore jacob took her what
<UNK> to <UNK> to the <UNK> my hand because i will give her
father's seed to him <UNK> in <UNK> that i and i have dreamed a
<UNK> and to keep me </s>
<s> the daughter of it neither shall make your way </s>
<s> or bad as you fourteen years old when she said to enoch </s>
<s> now come near and lifted up after anything </s>
```

Analysis:

The sentences generated using unigram probabilities have much less flow than those generated using bigram probabilities because the unigram probabilities do not consider what precedes a word, whereas bigram probabilities do.

However, the sentences generated using bigram probabilities are still not very cohesive (although more cohesive than the unigram generations). They only have phrases within them that are partially cohesive, and sometimes, they make sense grammatically but not practically. They also don’t generate <s> in the middle of the sentence, as it is always preceded by </s>, except for the first one.

Webtext (pirates):

Unigram Language Generator:

```
<s> beckett <UNK> a from human the sits <UNK> <UNK> <UNK> <UNK>
up it is as <s> against a <UNK> <UNK> <UNK> cliff device will
finds <UNK> gibbs </s>
<s> <s> william </s>
```

```
<s> heard <UNK> face the turned of jack <UNK> </s>
<s> in in ship do their in debt <UNK> </s>
<s> <UNK> cannibal hand <s> <UNK> sits <UNK> i - pistol a about
word turner carrying world window marque floats </s>
```

Bigram Language Generator:

```
<s> will turner <UNK> </s>
<s> three days </s>
<s> lord beckett ah </s>
<s> i are in her pistol <UNK> as <UNK> legs through it </s>
<s> jack looks back </s>
```

Analysis:

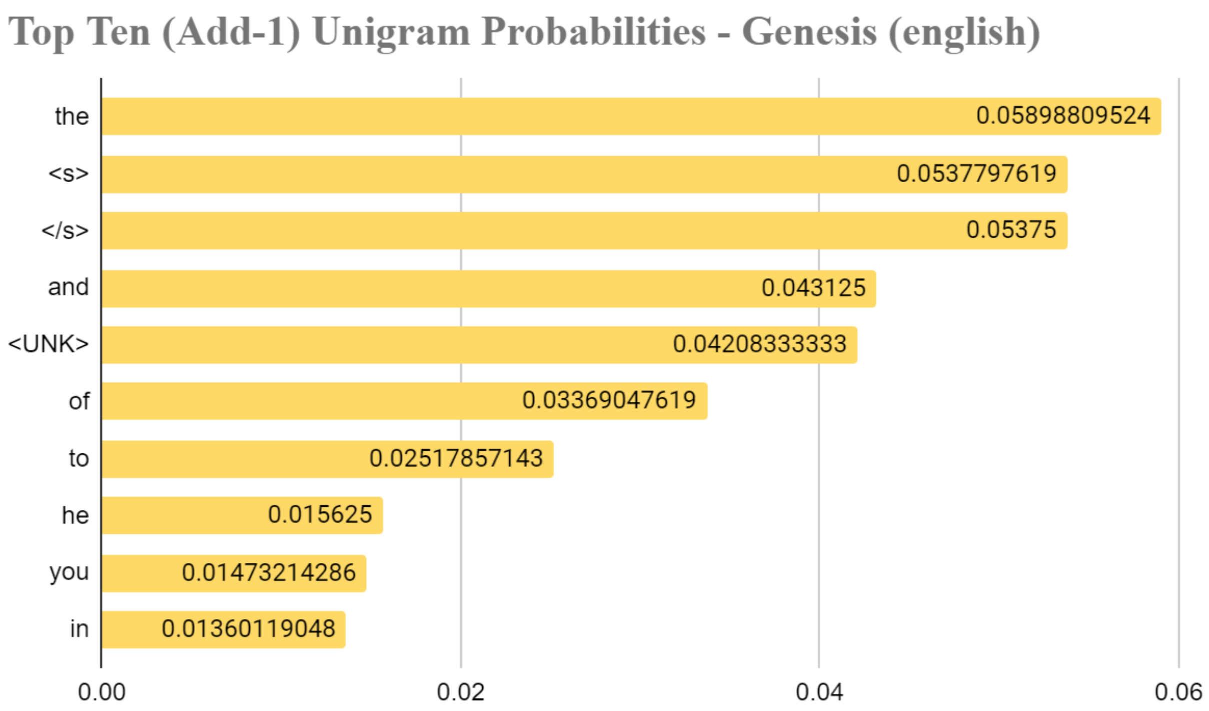
Again, the unigram language generator generates less cohesive sentences than the bigram language generator.

Because the Webtext (pirates) corpus sentences are much shorter than the sentences in Genesis (english), the language generator using Webtext generates shorter sentences. Another way to explain this happening is that, as mentioned earlier, </s> is more common in Webtext than in Genesis.

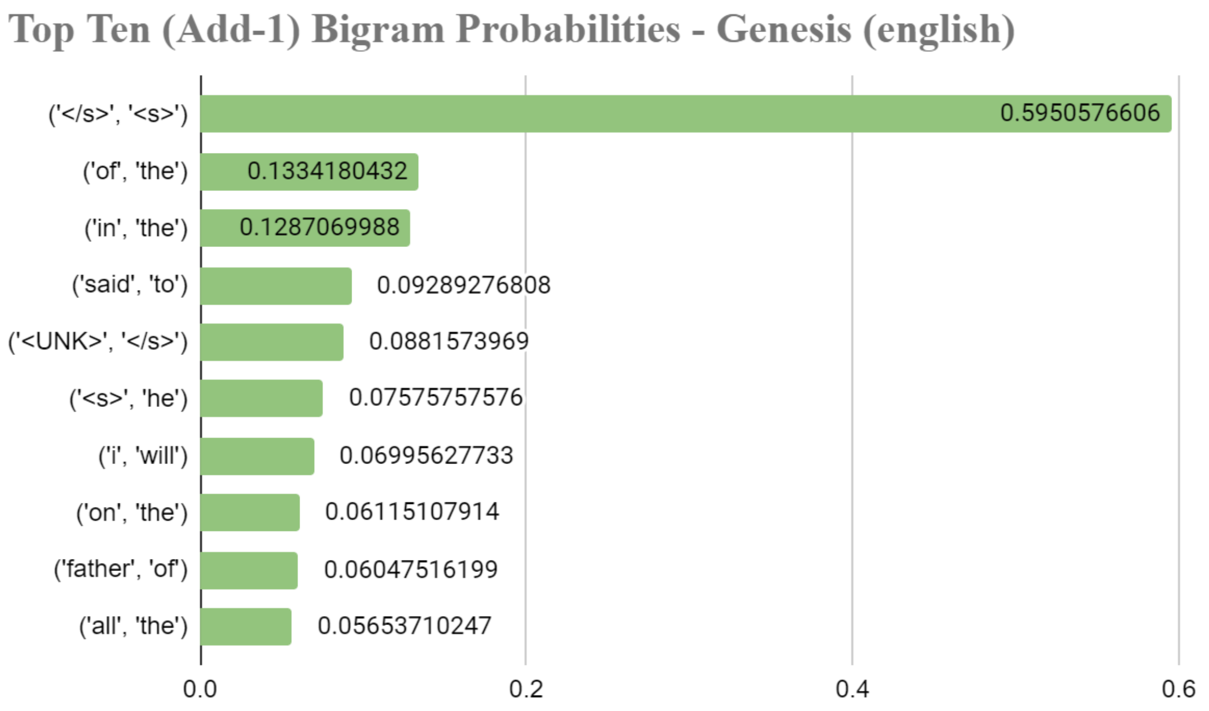
T7:

Corpus: Genesis (english)

Smoothed (Add-1) Unigram Probabilities:



Smoothed (Add-1) Bigram Probabilities:



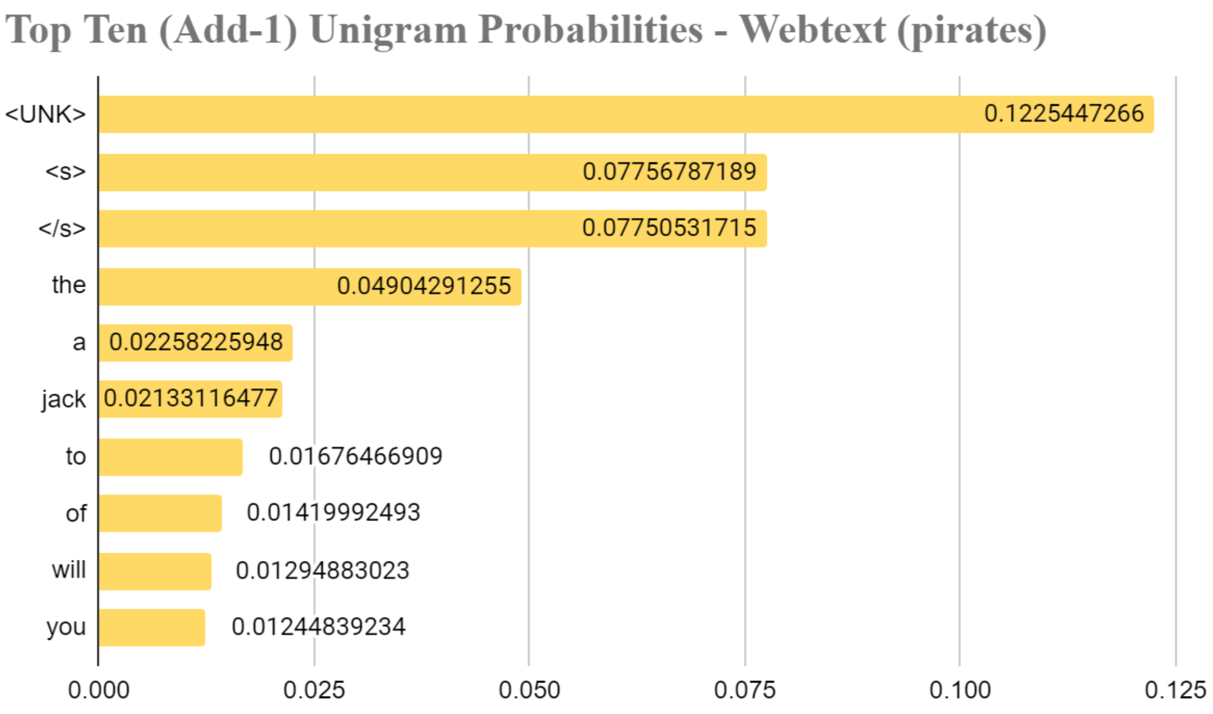
(Genesis - english) Analysis:

After applying Add-1 smoothing on unigram and bigram probabilities produced using the training data, the unigram probabilities dropped slightly. Because there are so many unigrams, applying Add-1 didn’t significantly impact the unigram probabilities, and the order of the top ten unigrams remained the same.

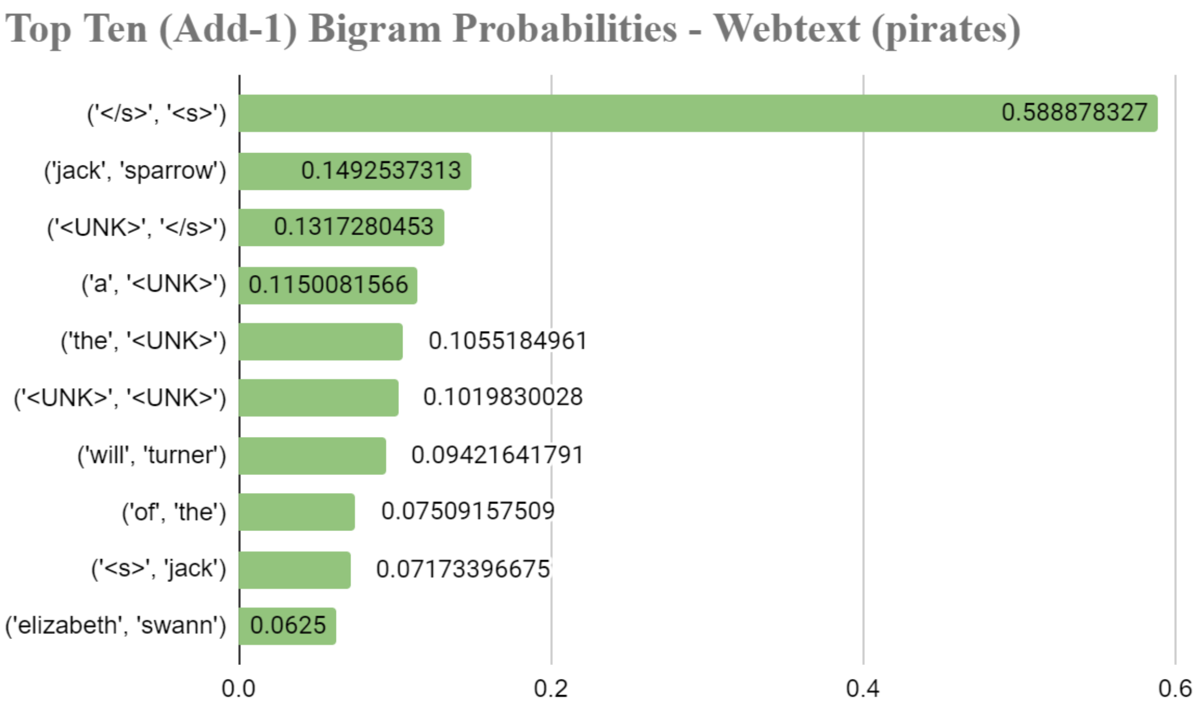
The bigram probabilities had a significant change, however. (‘</s>’, ‘<s>’) is still the top bigram, which was explained earlier. However, the other nine bigrams displayed in the “top ten” (although there were many more with 1.0 probabilities) of the MLE probabilities have all been removed from this new top ten list. The reason is because, in those original bigrams, the first word wasn’t seen too often, which meant that adding a large number to the denominator dropped the value of the fraction by a lot. Note that the first word in the top ten bigrams shown above all have relatively high unigram probabilities. Therefore, we can see that Add-1 was able to distribute the probabilities to the rare bigram combinations by taking from the bigrams that appeared more frequently and had exceptionally high probabilities.

Corpus: Webtext (pirates)

Smoothed (Add-1) Unigram Probabilities:


Top Ten (Add-1) Unigram Probabilities - Webtext (pirates)

Smoothed (Add-1) Bigram Probabilities:


Top Ten (Add-1) Bigram Probabilities - Webtext (pirates)

(Webtext - pirates) Analysis:
As shown and explained with the Genesis corpus, the Add-1 smoothing did not significantly change the unigram probabilities but dropped the numbers by a very minuscule amount.

Also displayed using the Genesis corpus was the sharp decrease in the bigram probabilities after Add-1 smoothing was applied: the same can be seen from this Webtext (pirates) corpus. The bigram (‘</s>’, ‘<s>’) maintains a high probability after Add-1 smoothing (although it also dropped by about 40%) because the </s> unigram frequently appears in the training set. Other bigrams found initially in the top ten MLE bigram probabilities are no longer in the top ten bigram probabilities after Add-1 smoothing is applied. Instead, bigrams whose first word has a high frequency in the text have higher probabilities.

Sentence Generator(s) w/ Add-1 Smoothing:

Genesis (english):
Unigram Language Generator:

```
<s> their morning <s> to </s>
<s> loved in divide well voice to <s> daughters and a it speak
there every i <UNK> havilah i cain it expanse man <s> him men
the to land father son had zillah of the in </s>
<s> the off servant <UNK> you stood who was baal with and </s>
<s> <UNK> sent soul for were <UNK> at commanded give wise <UNK>
sister the they that yahweh <s> of behold give commanded
together when <s> and portion of took to to men in </s>
<s> children where field name sons may but bless </s>
```

Bigram Language Generator:

```
<s> there dan torn while pain calah concerning interpretation
establish ai hands </s>
<s> when sephar wife's nothing formed until own things barren
gerar lodged </s>
<s> esau whatever great spring else egyptians bought bela bought
hebron dead </s>
<s> judith can sarah offerings gods including curse sarah
outside herb token </s>
<s> the wilderness haran brother sat donkeys black daughter
grizzled eber mighty </s>
```

Analysis:

The difference between the MLE-based Language Generation and Add-1 based Language Generation is not super obvious. There does seem to be a decrease in the <UNK> token in the sentences generated with add-1. Generated sentences are very incohesive when using unigram probabilities. The generated sentences are barely, if at all, cohesive when using bigram probabilities. However, for Add-1 bigrams, since we use weighted selection to generate the sentence, it allows possibilities to occur in the generated sentence that never occur in the training data (e.g., "haran brother" never occurs in the training data). As mentioned before, Add-1 distributes probabilities to those that are rare by taking from those that occur very frequently, but it is very blunt in doing so.

Webtext (pirates);
Unigram Language Generator:

```
<s> will port aye the <UNK> that turner <s> off more what <UNK>
we the to with of ill finds have behind it cotton is whose would
other around the mind <UNK> <s> she thank on perhaps <UNK> <s>
pearl <UNK> shrimper's a wood we're </s>
<s> ship <UNK> a <UNK> does jones tell heard of </s>
<s> to norrington on <UNK> up <UNK> your opens to you the <s>
afraid and </s>
<s> tia bootstrap </s>
<s> much <s> is will their <UNK> me a at <UNK> so you thought
while me </s>
```

Bigram Language Generator:

```
<s> shack sparrow say that's hey - shore rope notice behind
mooring </s>
<s> scarlett afterwards can spears ho prison longboat bad but or
bootstrap </s>
<s> marriage point unidentifed chase silently one-armed shackles
you're turned manacles pounding </s>
<s> elizabeth swann exiting floats sees prison save pirate tell
bridges guards </s>
<s> half wyvern organ because missing quartermaster upside-down
bo'sun licky-licky those their </s>
```

Analysis:
Once again, there are small differences between language generation using MLE probabilities and probabilities after Add-1 smoothing. A couple of differences are (1) that the <UNK> token, as mentioned before, doesn't occur as much after probabilities have been smoothed, and (2) the sentences are longer on average after probabilities have been smoothed (more obvious with the

bigram language generation using the Webtext/pirates corpus). This is because the bigram probabilities of a word followed by </s> have decreased due to the smoothing. All language generators (bigram/unigram probabilities & smoothed/unsmoothed) generate incohesive sentences, although certain phrases within the sentences created by bigram language generators were cohesive. Since we use weighted selection to generate the sentence, it allows possibilities to occur in the generated sentence that never occur in the training data (e.g., “wyvern organ” never occurs in the training data).

T8:
Genesis (english) perplexities with UNK_CUTOFF as 2:
Bigram Perplexity:

```
176.88035279354355
```

Unigram Perplexity:

```
159.81823024689623
```

Genesis (english) perplexities with UNK_CUTOFF as 10:
Bigram Perplexity:

```
119.42364866370829
```

Unigram Perplexity:

```
126.55270391211131
```

Webtext (pirates) perplexities with UNK_CUTOFF as 2:
Bigram Perplexity:

```
122.95436479074503
```

Unigram Perplexity:

```
99.96362491097194
```

Webtext (pirates) perplexities with UNK_CUTOFF as 10:
Bigram Perplexity:

```
45.89943980607966
```

Unigram Perplexity:

```
46.474976465129714
```

The perplexity for unigram and bigram models is lower for the Webtext (pirates) corpus than the unigram and bigram models for the Genesis (english) corpus. This is because the Genesis corpus has a larger vocabulary size. We also noticed that when the <UNK> cut-off is higher, the perplexity is lower and the bigram model performs better than the unigram model for both corpora. This is because when we have a lower <UNK> cut-off, <UNK> appears less in the training set, which means that we have a larger vocabulary. A higher cut-off means that we have more <UNK> tokens in the training set and, therefore, reducing the size of the vocabulary.