Analysis

For this project, our group selected two different corpora with different topics to explore. The first corpus we used, TheStory.txt, is an enovel called *The Story Continues* by Ferd Eggan. This story contains a lot of imagery, metaphor-rich language and a narrative with psychology, travel and aspects of modern life set in a fictional hotel. The second corpus that we used is called AlcoholUseConference.txt. This corpus is shorter than TheStory.txt and is a professionally written transcript from a medical presentation about alcohol use and screening it in an emergency department setting. This corpus contains formal language and clinical terminology intended for a public health or academic audience.

At first we ran our corpora through the MLE to find the top 10 unigram and bigram probabilities without +1 smoothing. The output from both corpora can be seen below:

TheStory.txt

```
Top 10 Unigram Probabilities from TheStory.txt without smoothing:
P(the) = 0.05624024712588771
P(of) = 0.029043661029903505
P(and) = 0.02897996879080284
P(to) = 0.024855896309034745
P(in) = 0.018805133594471515
P(a) = 0.018709595235820515
P(is) = 0.01117798796216681
P(that) = 0.009251297729371676
P(i) = 0.008057068246234197
P(s) = 0.007961529887583198
```

```
Top 10 Bigram Probabilities from TheStory.txt without smoothing:
P(the|in) = 0.265031
P(the|of) = 0.168706
P(problems|alcohol) = 0.228571
P(department|emergency) = 0.365325
P(ed|the) = 0.063113
P(emergency|the) = 0.062500
P(the|that) = 0.120944
P(be|to) = 0.078641
P(the|to) = 0.077670
P(be|should) = 0.487013
```

AlcoholUSeConference.txt

```
Top 10 Unigram Probabilities:
P(the) = 0.04343775784514652
P(of) = 0.03044901652870565
P(to) = 0.027414761385110856
P(and) = 0.02685581964813287
P(in) = 0.021692262649383832
P(that) = 0.018045833222432195
P(a) = 0.016661787016581938
P(alcohol) = 0.014905112986079688
P(for) = 0.012536264672220595
P(be) = 0.010779590641718347
```

```
Showing Top 10 Unsmoothed Bigram Probabilities (MLE, based on most frequent bigrams):
P(the|in) = 0.265031
P(the|of) = 0.168706
P(problems|alcohol) = 0.228571
P(department|emergency) = 0.365325
P(ed|the) = 0.063113
P(emergency|the) = 0.062500
P(the|that) = 0.120944
P(be|to) = 0.078641
P(the|to) = 0.077670
P(be|should) = 0.487013
```

We first examined the top 10 unigram probabilities from the two corpora. For the Unigrams both corpora have a high probability with common English words like "the", "of", "and", "to", "in", "that", and "a". This makes sense because these are common words used as the foundation for sentence structure across all writing. From looking at the percentages of how many times "the" occurs, we see that in TheStory.txt "the" occurs 5.62% of the time while in AlcoholUseconference.txt "the" occurs 4.34% of the time. From this, we can reach the conclusion that there are more descriptive and elaborate sentences in TheStory.txt. Also we can see that "alcohol" is ranked in the top 10 for AlcoholUseConference.txt and not TheStory.txt which makes sense because the main focus is on Alcohol screening in the ER. When we looked at the bigrams we noticed that there are more differences in style between the two texts. In TheStory.txt most of the frequent bigrams have “the”as the second word. For example, “in the” (26.5%) and of “the” (16.8%). These bigrams are common in general when writing a story in English. But in comparison, with AlcoholUseconference.txt, bigrams such as: “emergency department” (36.5%) and “alcohol problems” (22.8%). From this, we can see the distinct difference between common bigrams that can appear in more formal documents that are focused on facts and recommendations instead of storytelling.

After examining the unsmoothed unigram and bigrams we used our models to generate sentences without smoothing. This means that the sentences were generated based on how often words appeared in the original text. The sentences we generated are:

TheStory.txt

1. those three years old geography breaks down it costs the frayed nerves but they are
2. rides if you all the world music have eyes widening to mean those kids leave
3. trap sucking him these plants they know but a year at what you have gone
4. oddly though he underwent apparently searched the secret mission hey no no money being born
5. the assertion of pity and into which relies heavily finnish houghton tech who reaches for

AlcoholUSeConference.txt

1. of research as audio interventions have spontaneous remission such as the two versions of injury
2. misuse recommendation was not sufficient relationship between prevention little sense of the word evaluate the
3. condition but on the alcohol specific definition of another articles published in eds and brief
4. centered timely efficient ed focus on population such as tweak cage or dsm includes the
5. trials should they cover letter home one of evidence based on these policies so it

Even though the generated sentences from both corpora are awkward and grammatically incorrect they still show the tone of the original texts. The sentences from TheStory.txt sound more extravagant and imaginative. Even though the grammar is off the words have a descriptive feel to them. In contrast, the sentences from AlcoholUseConference.txt sounds more formal. There are medical and academic terms such as “remission”, “evidence based”, and “tweak cage or “dsm”. Even though the sentence structure is not perfect the sentences still reflect a professional and clinical tone.

Next we repeated the unigram and bigram analysis, but this time with Add-1 smoothing. Doing this adjusts the probabilities so that rare or unseen words and word pairs still get a non-zero value. The resultsin the model being more flexible and help avoid zero probabilities. The smoothed unigram and bigram probabilities that we got are:

TheStory.txt

Top 10 Unigram Probabilities (smoothed):
P(the) = 0.047296
P(of) = 0.024452
P(and) = 0.024412
P(to) = 0.020927

```
P(in) = 0.015819
P(a) = 0.015765
P(is) = 0.009424
P(that) = 0.007574
P(for) = 0.006529
P(he) = 0.006046

Top 10 Bigram Probabilities (smoothed):
P(the|of) = 0.029936
P(the|in) = 0.021501
P(the|to) = 0.013332
P(the|and) = 0.009331
P(the|at) = 0.008664
P(the|on) = 0.007885
P(be|to) = 0.006595
P(the|with) = 0.006256
P(the|by) = 0.006148
P(the|from) = 0.005918
```

AlcoholUseConference.txt

```
Top 10 Unigram Probabilities (smoothed):
P(the) = 0.039016
P(of) = 0.027356
P(to) = 0.024633
P(and) = 0.024131
P(in) = 0.019496
P(that) = 0.016223
P(a) = 0.014980
P(alcohol) = 0.013403
P(for) = 0.011277
P(be) = 0.009700

Top 10 Bigram Probabilities (smoothed):
P(the|in) = 0.041835
P(the|of) = 0.035170
P(problems|alcohol) = 0.026156
P(department|emergency) = 0.025346
P(ed|the) = 0.017322
P(emergency|the) = 0.017155
P(the|that) = 0.016436
P(be|to) = 0.015180
P(the|to) = 0.014994
```

P(be|should) = 0.016792

TheStory.txt smoothed sentences

1. parents but all made himself transmigratory soul this must change is lesbian jobs the sun
2. never could not possible brain alone to hearts are not merely look at the past
3. from their closest comrade see the wind blow all the fire on palm leaves out
4. power humans have eyes of knowing he sent him the loyal novice with women singing
5. san francisco exemplify the force they might still resent the big way forever nobody from

AlcoholUseConference.txt smoothed sentences

1. within the recommendation 3 to occur if articles about interventions for considering the outside their
2. compared with 95 3 months outcomes such patients is populated by asking patients they include
3. that it is considered in young large numbers of risk of 30 to further evaluation
4. with little is obvious that although many missed any narcotic unless administered as drinking before
5. discharge gentilello clarified for alcohol problems leads to cut points or other barriers at risk

As we can see when add 1 smoothing is applied it changed both the unigram and bigram probabilities. We noticed that the most frequent words and word pairs decreased in percentage. For example in TheStory.txt the probability of the word “the” decreased from 5.62% to 4.73% and in AlcoholUseConference.txt “the” dropped from 4.33% to 3.90%. This same phenomenon happened with Bigrams also. When looking at the bigram, we noticed that common bigrams like the|of in TheStory.txt dropped from 23.5% to 2.99% and “problems|alcohol” in AlcoholUseConference.txt dropped from 22.8% to 2.6%. These changes happened because instead of letting a few word pairs dominate the model became more flexible and better at handling new or rare word combinations. We also generated new sentences from each corpus. The sentences were still grammatically incorrect but used a larger variety of vocabulary and word pairings compared to the unsmoothed versions. The model trained with TheStory.txt had sentences that are creative and expressive. The sentences have unusual pairings but still matched the style as if a story is being told. However, the model trained with AlcoholUseConference.txt generated sentences that continued to have medical and academic words such as “discharge” and “narcotic”. These sentences still showed the professional and academic tone of the text. Overall Add-1 smoothing allowed the model to use a wider range of language leading to more diverse sentences.

Next we tested how well each smoothed model performed by finding the perplexity. Using an 80/20 held out test sets of data with Add-1 smoothing, we saw a perplexity of 6658.81 for TheStory.txt and 1653.88 for AlcoholUseConference.txt. The perplexity in TheStory.txt being higher shows that the sentence structure is more varies and less predictable while the more professional, technical and repetitive structure which means the model can learn it easier.

In conclusion, this analysis explored how language models perform on two different types of text, creative fiction and professional medical conferences. Through our unigram and bigram analysis both with and without Add-1 smoothing we evaluated the frequency patterns, structure pattern and predictability of our corpora. While both corpora have the same basic english words like “the”, “of”, “and”, “to”, and “that” the bigram patterns and generated sentences showed us key differences between the two. TheStory.txt’s creative style led to a higher perplexity and imaginative sentence generation, while the professional and technical style of AlcoholUseConference.txt had a lower perplexity. Add-1 smoothing helped to handle unseen word pairs and generate more varying sentence construction in both cases. Ultimately we learned how the way something is written can affect a language model.