NLP-Assignment 2

Alisar Abou Said(2814814), Vajiheh Moshtagh(2801823), Sjors Spronk (2746009)

May 5th 2024

1 Problem Zero

First we started with the analysis of courpus, which involved importing the entire Brown corpus from the NLTK library and computing a list of unique words sorted by descending frequency for both the whole corpus and two genres—news and romance.

1.1 Corpus Analysis

The table below summarizes the linguistic information extracted from the Brown corpus:

Metric	**Value**
Number of Tokens	1,161,192
Number of Types	56,057
Number of Words	1,161,192
Average Number of Words per Sentence	20.25
Average Word Length	4.27

Table 1: Summary of Linguistic Information

Additionally, the ten most frequent POS tags identified in the corpus are: NN, IN, DT, JJ, NNP, ',', '.', NNS, VBD, and RB.

1.2 Discussion

Origin of the corpus came back to Brown University in the 1960s, the Brown Corpus was a pioneer project, one of the first to provide a large-scale linguistic dataset, representing a diverse range of American English texts. Analysis of this corpus reveals a compliance with Zipf's law: the most common words are used disproportionately, confirming the inverse relationship between word rank and frequency. This corroborates the principle that in language, few words are exceedingly frequent while others remain rare.

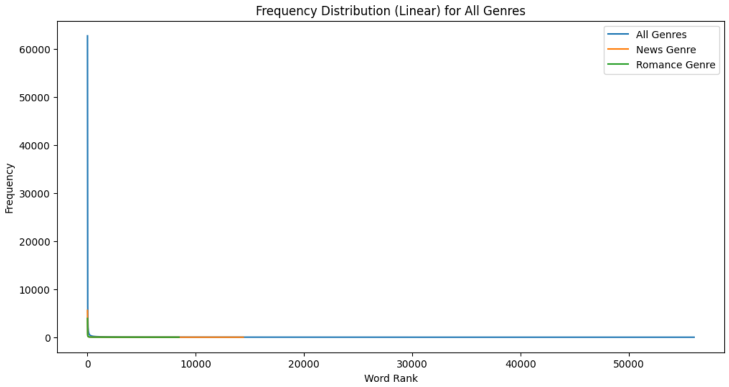

Figure 1: Linear Frequency Distribution of the Brown Corpus and two genres(News and Romance)

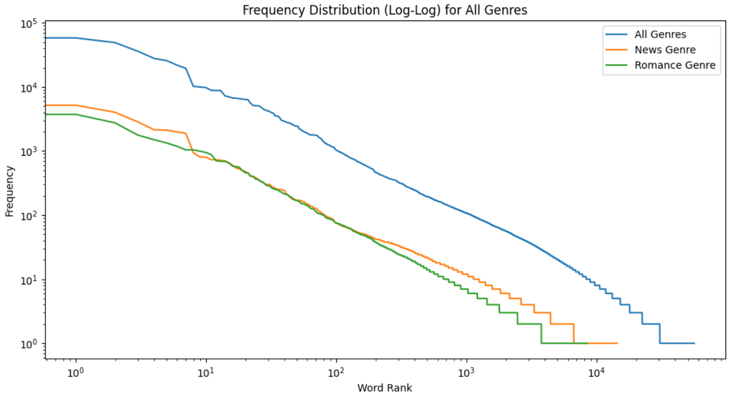

Figure 2: Log-Log Frequency Distribution of the Brown Corpus and two genres(News and Romance)

In our finding, which also included genre-specific examinations of news and romance, found that despite differences in style and content, Zipfian distribution holds across genres, although frequency peaks vary, reflecting unique genre lexical signatures. Figures 1 and 2

These insights are critical for understanding linguistic trends and for the development of NLP models. However, the Brown Corpus carries biases of its era, predominantly representing the linguistic norms of the 1960s. Such historicity must be acknowledged; contemporary linguistic models need more modern corpora to account for the evolution of language use over time. Thus, while the Brown Corpus is instrumental for historical linguistic analysis, it also highlights the necessity for temporal diversity in corpus linguistics.

2 Problem 1

In the construction of the unigram model, we successfully created a dictionary that maps each unique word from `brown_vocab_100.txt` to a unique index, enabling efficient word look-ups for subsequent n-gram modeling tasks, with 'all' correctly mapped to index 0 and 'resolution' to index 812.

3 Problem 2

3.1 Do you think the proportion of words that occur only once would be higher or lower if we used a larger corpus (e.g., all 57000 sentences in Brown)?

Within a larger corpus, such as the full 57,000 sentences of the Brown corpus, the proportion of words occurring only once is expected to decrease. The principle of lexical richness implies that as text length increases, less frequent words have more opportunities to appear, thus reducing their overall rarity. Moreover, Heaps' Law suggests a sublinear growth of vocabulary size with the length of the document, indicating that new words appear at a diminishing rate as the corpus size increases.

Hence, while the absolute number of unique words (hapax legomena) may increase with a larger corpus size, their relative proportion to the total word count decreases. This tendency is consistent with observations from Zipf's law, which posits a predictably skewed word frequency distribution: many words have low frequencies, while a few have very high frequencies. In practice, this means that with more data, common words recur with higher frequency, and the incidence of single-occurrence words diminishes proportionally.

4 Problem 4

4.1 Why is smoothing useful when calculating probabilities related to language? Why did all four probabilities go down in the smoothed model?

Smoothing is useful as it gives a value to non-occurring word pairs from the data. As most phrases may not occur in a dataset, it does not mean other word combination are impossible. All probabilities went down because initially there were no cases of $p(a|b)$. The normalizing factor for the word counts (to obtain probabilities) is now relatively larger compared to the individual counts than without smoothing.

4.2 Why did add- smoothing cause probabilities conditioned on 'the' to fall much less than these others? And why is this behavior a good thing?

The probability for "the" fell less because "the" occurs in diverse situations. If "the" would've been concentrated on a single previous word, the probabilities would have fell more (similar to the other words). This is due to the normalizing factor (sum of occurrences) of counts to get probabilities. If the word occurs diversely, the extra added α is added more often to the non-zero probabilities themselves and they decrease less. The reason this is a good thing is because it prevents a monopoly of certain words on specific previous words. This way the diversity of words is increased. Also, words that have diverse occurrences keep having high probabilities.

5 Problem 5

In the table below, we present the calculated probabilities for various trigram combinations using both unsmoothed and smoothed Maximum Likelihood Estimation (MLE) approaches.

Probability	**Unsmoothed**	**Smoothed**
p(past | in, the)	0.0625	0.01130
p(time | in, the)	0.0	0.0010
p(said | the, jury)	0.5384	0.0752
p(recommended | the, jury)	0.1538	0.0222
p(that | jury, said)	0.0	0.0011
p(, | agriculture, teacher)	1.0	0.0133

6 Problem 6

Written QA: **Compare the performance of the different models. What do you notice? How can we evaluate the performance of each in relation to another?**
There are different ways to compare model performance .You could do extrinsic evaluation which is time consuming .This method look at the quality output of the model .The other method is to calculate perplexity. A lower perplexity indicates that the model is less surprised and thus better at predicting the next word. To evaluate the performance of each model in relation to another, we need to compare their perplexity values on the same dataset. Looking at sentence 1 and 2 in the "toy_corpus.txt" we see in unigram model the perplexity is 281,153 for the 2 sentences and in bigram model without smoothing 4.56,7.57, and in bigram model with smoothing is 53.42,54.28. The lower the perplexity the higher the likelihood ,this means that the best model for this text is the bigram model without smoothing.

Written QA: **Did smoothing help or hurt the model's 'performance' when evaluated on this corpus? Why might that be? (word limit: 200)**
Smoothing algorithms shave off a bit of probability mass from some more frequent events and give it to these unseen events The effect of smoothing on the model's performance would depend on factors such as the size of the corpus, the frequency of unseen n-grams, and the specific smoothing method used. If the "toy_corpus.txt" contains many unseen n-grams, smoothing is likely to help improve the model's performance by providing more robust estimates of probabilities. Conversely, if the corpus contains fewer unseen n-grams which is the case as the sentences are in "brown_100.txt" , the impact of smoothing might be less pronounced or even detrimental as the smoothing technique introduces unnecessary distortions to the model's estimates.

7 Problem 7

Grammar
- Unigram model: Not correct in any sentence. Sometimes the starting symbol is used in the middle of the sentence. This is because at any position in the sentence all encountered words/symbols are possible.
Example: "¡s¿ county the of expected technical ¡s¿ the urban for the operation provided ¡/s¿".
- Bigram model:
Performs well in most cases of pairs of words. This is because having a single condition already gives a lot more information on what word may come next. However, as the models' view is very limited, this effect only lasts shortly.
Example: "the best interest of one of office expires jan. 1 the audience was received anonymous telephone calls . ¡/s¿"
- Smoothed model:
The generated text of the smoothed model performs worse than the bigram model. Here, words do not appear as random as with the unigram model but the sentence structure is lacking. This model is less consistent than the bigram model as the dataset is small and the probabilities are more diverse, leaving more possibility grammatically incorrect results.
Example: "the new bonds courses term ask courses during purpose considering race death miller serve smooth into likely outmoded enabling allotted hopper"

Ambiguity
Ambiguity in Natural Language Processing (NLP) refers to situations where a word, phrase, or sentence can have multiple interpretations or meanings.
- Unigram model:lets take those examples:
"former 1923 ¡s¿ assistance will "candidate of said ¡s¿ a saw an department title" The phrase lacks context, making it unclear what "former 1923" refers to and how it relates to the rest of the text. The phrase contains fragmented sentences such as "assistance will" and "candidate of said," which do not form complete

thoughts and leave the reader unsure of their intended meaning. The phrase contains ambiguous terms such as "department title," which could refer to various concepts depending on the context, adding another layer of uncertainty to the interpretation.
- Bigram model:
The text generated contain several instances of ambiguity
"the jury said": Without further context, it's unclear what the jury said or in what context.
"the best interest of one of office expires jan. 1": It's ambiguous what "one of office" refers to.
- Smoothed model:
Taking this sentence as example "the fulton hospital despite other issued combined are fulton protected up its state party weekend unanimous an who elected off cent"
The phrase lacks clear context, making it difficult to understand the significance of "fulton hospital" and its relationship to other entities mentioned.The phrase "despite other issued combined" is unclear and could have multiple interpretations.

Expressivity
The phrases lack expressivity due to their vague and disjointed nature.

Conclusion
Based on the generations, it seems that the bigram model performs best. The bigram model performs best because the dataset is small and there is little chance of learning an overall pattern to the linguistic principles. Therefore, overfitting seems to work best even though it does decrease creativity.

8 Bonus(PMI Value Analysis)

The PMI values reveal a clear dichotomy in word association strengths within the Brown corpus. The lowest PMI values are generally found among high-frequency words that appear together frequently but do not necessarily have a strong associative meaning. On the other hand, the highest PMI values typically belong to word pairs that represent strong collocations or named entities, indicating a significant relationship between the words. This demonstrates the usefulness of PMI as a measure of word association and the diverse nature of language structure.

Table 2: The 20 word pairs with the lowest PMI values

Word Pair	PMI Value
(’.’, ’,’)	-11.28
(’the’, ’.’)	-10.38
(’and’, ’.’)	-10.21
(’of’, ’of’)	-10.13
(’the’, ’in’)	-10.04
(’a’, ’.’)	-9.86
(’the’, ’,’)	-9.62
(’and’, ’and’)	-9.39
(’the’, ’is’)	-9.07
(’the’, ’and’)	-8.97
(’of’, ’to’)	-8.64
(’,’, ’;’)	-8.12
(’the’, ’I’)	-8.12
(’of’, ’for’)	-8.10
(’?’, ’the’)	-7.98
(’the’, ’not’)	-7.90
(’to’, ’was’)	-7.75
(’of’, ’he’)	-7.67
(’he’, ’of’)	-7.67
(’in’, ’of’)	-7.66

Table 3: The 20 word pairs with the highest PMI values

Word Pair	PMI Value
(’Beverly’, ’Hills’)	15.22
(’Common’, ’Market’)	15.24
(’unwed’, ’mothers’)	15.24
(’carbon’, ’tetrachloride’)	15.26
(’Export-Import’, ’Bank’)	15.28
(’anionic’, ’binding’)	15.31
(’Saxon’, ’Shore’)	15.31
(’decomposition’, ’theorem’)	15.33
(’Puerto’, ’Rico’)	15.56
(’Internal’, ’Revenue’)	15.56
(’Gray’, ’Eyes’)	15.63
(’WTV’, ’antigen’)	15.66
(’Lo’, ’Shu’)	15.75
(’Herald’, ’Tribune’)	15.79
(’El’, ’Paso’)	15.82
(’7th’, ’Cavalry’)	15.82
(’Pathet’, ’Lao’)	16.06
(’Simms’, ’Purdew’)	16.061
(’Viet’, ’Nam’)	16.15
(’Hong’, ’Kong’)	16.69

Discussion

Unigram models, based on the assumption of words independence, significantly simplify computational demands but at the expense of ignoring rich linguistic structures that govern word co-occurrence. PMI analysis, as demonstrated, challenges this assumption. Negative PMI values, such as for pairs like *(’of’, ’.’)*, shows rare co-occurrences and underscore the complex syntactic structures overlooked by unigram models. On the other hand, high PMI values for word pairs like *(’Hong’, ’Kong’)* demonstrate strong collocations, clearly contradicting the foundational premise of unigram models.

This analysis highlights the limitations of unigram models, particularly in tasks requiring deep linguistic comprehension such as context-sensitive text generation or machine translation. While computationally efficient, unigram models fail to capture the nuanced probabilistic nature of language, underscoring the need for more sophisticated, context-aware NLP models. This shift towards advanced modeling techniques marks a significant evolution in NLP, aiming to achieve a more comprehensive understanding of language structure and meaning.