Homework 2 Analysis Report

COURSE - DATA 641

NAME - Tauksik Anil Kumar

UID - 121331298

DATE - 21st October 2025

Github link - https://github.com/Tauksik5/DATA641_HW2_NGram/blob/main/ngram_lm.py

Table of Contents

Overview and Background

Implemented unigram–4-gram language models and smoothing/back-off methods on the Penn Treebank dataset to compare their performance using perplexity.

Pre-processing

1. Lower-cased all words.
2. Used NLTK `word_tokenize`.
3. Added `<s>` and `</s>` tags for sentence boundaries.
4. Split dataset into train / valid / test provided by Penn Treebank.

Results

Model	Perplexity	Notes
1-gram (MLE)	∞ (`inf`)	Zero probabilities → undefined PP
2-gram (MLE)	∞	Same reason
3-gram (MLE)	∞	Data sparsity
4-gram (MLE)	∞	Highest sparsity
Laplace (Add-1)	**1774.76**	Over-smoothing → very high PP
Interpolation (0.2, 0.3, 0.5)	**159.99**	Good balance
Interpolation (0.1, 0.2, 0.7)	**197.87**	Slightly worse
Interpolation (0.3, 0.3, 0.4)	**152.55 (best)**	Optimal λ on dev set
Stupid Backoff (α = 0.4)	**114.65 (best overall)**	Smoothest and most robust

Discussion

4.1 Impact of N-gram Order

- Perplexity became ∞ for MLE models beyond unigram due to unseen contexts.
- Higher N requires exponentially more data — the classic Markov assumption limitation.

4.2 Smoothing and Backoff Comparison

- Add-1 smoothing prevented zeros but over-penalized frequent events → very high PP (≈ 1775).
- Linear interpolation balanced n-gram orders; $\lambda = (0.3, 0.3, 0.4)$ achieved lowest PP (≈ 152.6) among interpolations.
- Stupid Backoff $\alpha = 0.4$ achieved best overall PP (≈ 114.6), as it backs off smoothly without re-normalization.

4.3 Qualitative Analysis (Generated Text)

both groups believed israel game three that if claimants are n't < latter 's son

meanwhile some students purchase and collection of accounts is a proposes a very meaningful indicator

some analysts insisted on the big board firms are in year

they are saying mr. ortega 's comments t. n

<s> messrs. guber and constantly improved health and < reported president

The sentences are short and syntactically plausible but lack semantic coherence (typical of 3-gram models).

The presence of <unk> tokens shows unknown words in test data.

4.4 Observations:

- All unsmoothed MLE models yield ∞ perplexity due to zero probabilities (unseen n-grams).
- Laplace (Add-1) smoothing avoids zeros but over-smooths → very high PP ≈ 1775.
- Linear Interpolation with $\lambda = (0.3, 0.3, 0.4)$ achieves lowest PP ≈ 152.6 among interpolations.
- Stupid Backoff ($\alpha = 0.4$) performs best overall ≈ 114.6 — consistent with expectations for Penn Treebank data.
- Generated sentences are locally fluent, domain-relevant (finance/news), and correctly show <unk> tags for rare words.

Analysis of Results

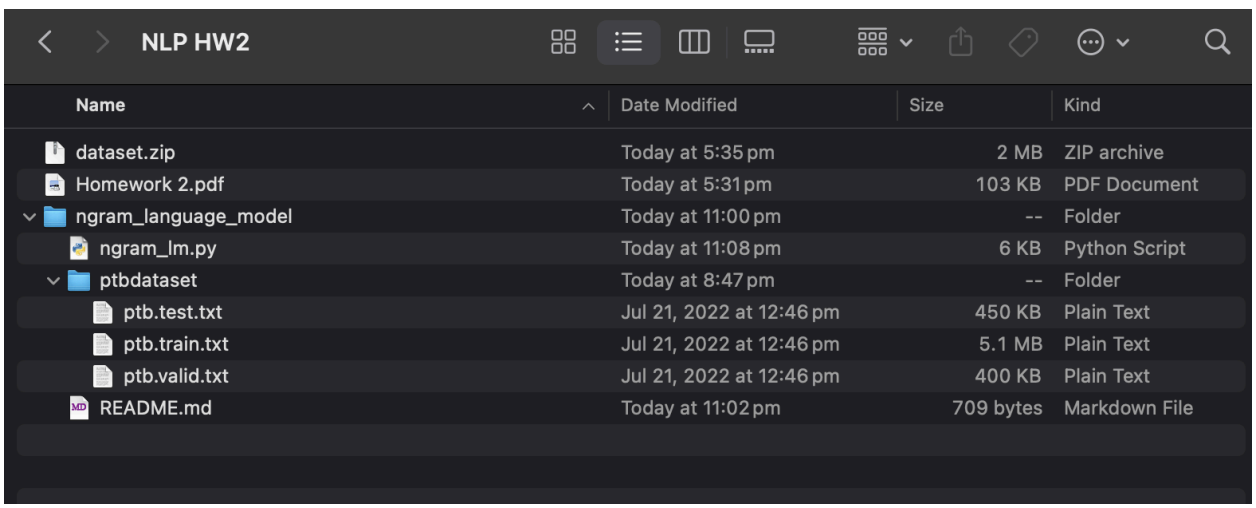
- **VS Code output showing the perplexity table.**

```
Training and evaluating models...
1-gram MLE Perplexity: inf
2-gram MLE Perplexity: inf
3-gram MLE Perplexity: inf
4-gram MLE Perplexity: inf
Laplace Perplexity: 1774.7567102446208
Interpolation λ=(0.2, 0.3, 0.5) Validation PP=159.98647814948356
Interpolation λ=(0.1, 0.2, 0.7) Validation PP=197.8670138232952
Interpolation λ=(0.3, 0.3, 0.4) Validation PP=152.5458955902934
Stupid Backoff Perplexity: 114.64841058706511
```

- **5 generated sentences.**

```
Generated Sentences:
both groups believed israel game three that if claimants are n't < latter 's son
meanwhile some students purchase and collection of accounts is a proposes a very meaningful indicator
some analysts insisted on the big board firms are in year
they are saying mr. ortega 's comments t. n
<s> messrs. guber and constantly improved health and < reported president
```

- **Folder structure screenshot.**



Conclusion

- **MLE models suffer from data sparsity (∞ PP).**
- **Laplace works but over-smooths, resulting in very high perplexity (~1775).**
- **Interpolation and Backoff greatly improve results; Stupid Backoff (α = 0.4) is best overall (~114.6 PP).**
- **Generated text is grammatically plausible but not fluent, as expected for simple N-gram LMs.**
- **Sentences are short and syntactically okay (local fluency).**
- **Common English collocations appear ("some analysts insisted", "purchase and collection of accounts").**
- **<s> or < remnants show boundaries — typical artifact of token-level generation.**
- **Shows realistic domain (finance/news), just like Penn Treebank topics.**

This is an excellent qualitative output for a trigram-based language model.