

SMAI Assignment 1

Nikhil Saxena

2024201034

Problem 2: N-Gram Character Language Model

Dataset	N	# Letters Typed	# Tab Presses	Avg Letters/Word	Avg Tabs/Word
General English Corpus	2	240	357	0.42	3.13
	3	248	342	0.46	3.00
	5	206	326	0.38	2.86
	10	206	347	0.38	3.04
Topic Specific Dataset	2	177	403	0.35	3.54
	3	194	339	0.41	2.97
	5	140	334	0.29	2.93
	10	139	355	0.29	3.11
Part 9 of Topic Specific Dataset	2	135	429	0.28	3.76
	3	138	366	0.32	3.21
	5	140	335	0.28	2.94
	10	132	351	0.27	3.08

Table 1: Comparative metrics across various datasets.

Analysis of Results

(a) Effect of Corpus Size and Content

- Larger and more general corpora (e.g., General English Corpus) lead to more diverse predictions, increasing the average tab presses needed to reach the intended word. However, it generalized well on random text.
- Topic-specific corpora had least average letters pressed.
- Domain-specific datasets (Topic Specific and Part 9) deliver very good results in targeted contexts due to their focused vocabulary and reduced noise.
- The Part_9.txt corpus, being highly specialized, achieves the lowest average letters per word but still requires frequent tab presses. It outperformed the full dataset (which contains irrelevant or noisy text), leading to a diluted model in the broader dataset.

(b) Effect of Context Size (N-gram Order)

- $N = 2, 3$: The model lacks sufficient context, resulting in higher values for both letters typed and tab presses.
- $N = 10$: The average letters typed per word remain low across all corpus due to added context, though this sometimes pushes the intended word slightly down the suggestion ranking.
- A lower N will have little context while a higher N will push relevant suggestions to the end.

(c) Interpreting the Metric Scores

- **Avg Letters/Word:** A lower average of letters typed per word indicates a model that auto-completes more of the word, reducing typing effort.
- **Avg Tabs/Word:** An increase in tab presses suggests that the user may have to cycle through suggestions more frequently to find the intended word.
- The optimal configuration is around $N = 5$ with the part_9th dataset.

(d) Generalization

- The General English Corpus performed adequately due to its broad coverage, though the average tab presses increased due to vocabulary mismatches.
- The specialized corpora underperformed on general text. Having average letters per word = 0.55 and average tab presses = 2.95.

N	# Letters Typed	# Tab Presses	Avg Letters/Word	Avg Tabs/Word
5	265	384	0.37	2.56

Table 2: Performance of General English Corpus on a Random Paragraph

(e) Performance of Topic Specific and Part 9 Corpus

- *Topic Specific Corpus:* The number of tab presses is higher than in the General English Corpus for some N values, suggesting that while auto-completion occurs more frequently, the correct word is not optimally positioned in the suggestion list.
- *Part 9 Corpus:* Generally achieves the lowest average letters per word, particularly at higher N -gram orders ($N = 5$ and $N = 10$).
- *Observation:* The performance on Part 9 is attributed to its highly focused content that aligns closely with the words and phrases used during the typing session.

Screen Recordings

- While training, I scanned through the suggestions after typing every letter. This gave me the best possible values for average letters typed per word.
- While testing (screen-recordings), I maintained typing until the word I was looking for is in the top 3-4 suggestions.
- For best performing model, my average letters typed per word := 0.5 and average tab presses := 1.3
- For random text with the general english corpus, my average letters typed per word := 0.64 and average tab presses := 0.65
- Best performing model: best_performing_model.mkv
- Best N with General English Corpus and Random Text: random_text_gen-en.mkv