Assignment 1 SMAI

Lakshay Baijal 2024202006

Question 2 - Word auto-completion

Indexing	Corpus	N values	Avg Letters / Word	Avg Tabs/ Word
1	general_english_corpus.en	2	3.95	2.38
2	general_english_corpus.en	3	3.92	1.97
3	general_english_corpus.en	5	3.93	2.18
4	general_english_corpus.en	10	3.95	2.02
5	topic_specific_dataset	2	3.35	2.55
6	topic_specific_dataset	3	3.24	2.56
7	topic_specific_dataset	5	3.23	2.72
8	topic_specific_dataset	10	3.31	2.64
9	topic_specific_dataset_part_3	2	2.96	2.35
10	topic_specific_dataset_part_3	3	3.05	2.28
11	topic_specific_dataset_part_3	5	2.94	2.54
12	topic_specific_dataset_part_3	10	3.01	2.09

Corpus Size –

- **General English Corpus:**
  - As the avg tabs/word is very less {1.97} it recognizes a word and it predicts quickly.
  - *Limitations:* As there were word gaps or many vocabulary words were missing General English Corpus might not capture domain specific.
- **Topic Specific Dataset:**
  - Topic Specific Dataset although its smaller it provides proper suggestions for its domain.
  - *Limitations:* Its scope is limited to that specific topic so it will not generalise to text that falls outside of that area.
- **Part i.txt Corpus:**
  - *Observations:* This corpus gives intermediate performance in both average Tab presses and letters per word. Its performance reflects a smaller, more homogeneous vocabulary, which may yield good suggestions for similar texts but can be limited in diversity.

Hence corpus size and content influence the richness of the n-gram statistics.

Model Type –

Changing the n value in your n-gram model fundamentally changes the context window used for predictions.

- **n = 2 (Bigram):** The model looks only at the last character. This can lead to more ambiguous suggestions and higher Tab key presses in some cases.
- **n = 3 (Trigram):** With two characters as context, predictions improved. However, this model sometimes misses words due to vocabulary limitations.
- **n = 5:** In both the general and topic-specific corpora, the results are similar to the trigram but sometimes slightly higher Tab presses.
- **n = 10:** With a longer context window, the general corpus again yields n=5 For the topic-specific datasets, the numbers are similar or a little lower in letters but comparable in Tab presses.

Metrics –

- **Average Letters per Word:**
  This metric indicates the average length of the words as typed with auto-completion. Consistent values across different models suggest that the predictions are similar in word length, which is good because it means the model isn't consistently producing overly short or long completions.

- **Average Tab Presses per Word:**
  This metric shows how many times a user must cycle through suggestions to select the correct word. Lower values mean the model is more effective—offering the correct prediction on the first or second attempt. Higher values indicate that the user has to press Tab more often to find the intended word, pointing to less accurate or less helpful predictions.

Generalization –

Testing on a 100-word Paragraph

While testing on a random 100-word Paragraph General English Corpus generalize better across a wide range of topics and gives predicts the words quickly. While topic specific fails as it only focuses on topic driven dataset and could not generalise properly.

Video Links

Video 1 - paragraph from text content with best model

https://drive.google.com/file/d/11qm1UpoTmHh4A8tx6xdRljD-_a7yIbZv/view?usp=drive_link

Video 2 - paragraph of your choice, general corpus, best n (I have not used my own paragraph)

https://drive.google.com/file/d/1pn_HFp_NuA1Wwl6nT56CC0viLnbxxoV4/view?usp=drive_link