
N-Grams: Small Language Model Activity

Objective

- Students will utilize the Green Eggs and Ham Text to create a new story
- Students will be able to explain how n-grams allow the AI to generate new text

Materials needed

- <https://www.cs.cmu.edu/~pvirtue/AIS/ngrams>
- Green Eggs and Ham Text Search Worksheet (Printable)
- Devices for students to access the *Green Eggs and Ham* [Find N-grams](#) tool or a printable copy of the text

How it works

1. Unplugged-ish
 - a. Each student, or pair of students, will access the [Find N-grams](#) text search tool
 - b. Students will utilize the Text Search Worksheet in order to find n-grams within the text.
 - c. Unplugged N=2
 - i. Start a story with one random word from vocab: e.g. The
 - ii. Repeat
 1. “Who has the previous word?” - - > STUDENT
 - a. If not student has the word, the teacher can use the solution sheet for that word
 2. “STUDENT, how many next words for this previous word?” - - > X
 3. Randomly choose number between 1 and X - - > Y
 4. “STUDENT, what is the Y-th next word on your worksheet?” - - > NEXT WORD
 5. Add NEXT WORD to story
 6. (NEXT WORD now becomes the previous word as we repeat)
2. Plugged demo: computer does the same thing the students were doing above!
 - a. Teacher (or students) will access the Small Language Generator on the website
 - i. Upon accessing, the teacher will follow instructions to have the computer generate the next word based on the *Green Eggs and Ham* text.
 - b. The teacher can choose to change the number of words the computer is looking at to generate the next word – or change the N in n-gram.



Follow-up/Discussion

- Will the generated story be the same every time?
 - [No]
 - Why not?
- How did AI choose the next word?
 - [Randomly based on how often the next word appears in the training text (Green Eggs and Ham)]
 - What if we didn't want it to be random/creative. Is there a different way we could program the AI to choose the next word?
 - [We could always have it choose the most popular word.]
 - [AI could present the human with the most frequent next words and let the human choose]
- Compare the way AI is choosing the word to how a human/you would choose the next word. Example 'the cat ate ???'
- What happens when we change N to be really small, e.g. N=1?
 - [Nothing really makes sense. No grammar at all.]
- What happens when we change N to be larger, e.g. N=7?
 - [It starts to copy Dr. Seuss exactly]
- What setting for N makes the best *new* story?
 - [Probably N=4 or so]
- Does this n-gram AI agent know anything about grammar?
 - E.g., does it know what a noun is? (What word will it pick after "the"?)

Teacher Notes

A N-gram is a sequence of words or symbols used in order. If I say "AI is great" that is a 3-gram (trigram) and if I say "absolutely awesome" that is a 2-gram (bigram). This is important because it triggers the computer to look for sets of words - MORE DATA- in sets of bigrams, trigrams, or even bigger sets of N-grams.

In this small language activity students will use the data set of *Green Eggs and Ham* to explore n-grams. They will look at the probability of word combinations showing up together and utilize the TEXT GENERATION TOOL to see which word the computer is most likely to pick. While other text generators have access to a much larger amount of text data in order to generate text, the concept is similar to how things like Chat GPT (large language models) or other text generators work.

