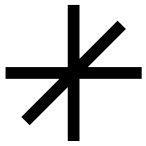

江南

A bigram language model

LLMs
Unplugged

A bigram language model derived from 江南 by 汉乐府

This booklet contains statistical word frequencies derived from the source text for educational purposes. Where the source text is under copyright, the transformation into probability tables for teaching language model concepts constitutes fair use. No substantial portion of the original text is reproduced.

Published by Cybernetic Studio Press

First Edition

Text frequency counts from the text 江南 by 汉乐府, available from `https://zh.wikisource.org/wiki/江南`.

Credits: designed and built by Ben Swift for the Cybernetic Studio. Typeset in Libertinus using Typst. Create your own n-gram model booklet using the online tools at `https://www.llmsunplugged.org/tools`. The source code (`v1.4.7`) for the tool used to create this model is available under an MIT Licence from `https://github.com/ANUcybernetics/llms-unplugged`.

Model statistics

- **Total tokens:** 27
- **Unique tokens (vocabulary):** 13
- **Unique previous-words contexts:** 13
- **Entropy:** 0.72 bits/token — how unpredictable each dice roll is
- **Perplexity:** 1.6 — effective number of choices per dice roll

Disclaimer: this reference contains a statistical language model derived from text corpus analysis. The patterns within represent probabilistic relationships between words in that text. Any new texts generated by sampling from this language model are statistical in nature and may not always reflect proper grammar, factual accuracy, or appropriate content.

How to use this book

This book contains a bigram language model for generating text using only one or more d10 (ten-sided) dice and a pen and paper to write down the generated text, according to the following algorithm.

Algorithm

To generate new text using the bigram model in this book:

1. **choose a starting word**—pick any bold word from the book (note that punctuation e.g. [.] count as words in this model) and write it down
2. **look up the word's entry** (i.e. use this book like a dictionary) to find all possible *next* words according to the model
3. **roll your d10s** (if required): check for diamonds next to the word—this shows how many d10s to roll (e.g., **the** ♦♦♦ means roll 3 d10s). If there are no diamonds, there's only one possible next word—skip to step 5. Read the dice from left to right as a single number (e.g., rolling 2, 1 and 7 means your roll is 217)
4. **find your next word**: scan through the next-word options until you find the first number ≥ your roll, or just use the single word if no dice were rolled (write it down)
5. repeat from step 2 using this word as your new word, continuing this loop until you reach a natural stopping point (like [.]) or reach your desired text length

Example 1: single d10

Your current word is **"cat"** and its entry shows:

cat 4|sat 7|ran **10**|slept

- one diamond (♦) means roll 1 d10
- roll your dice: roll a 6
- find the next word: first number ≥ 6 is 7|ran, so next word is "ran"
- write it down, look it up and continue the process

Example 2: multiple d10s

Your current word is **"the"** and its entry shows:

the ♦♦ **33**|cat **66**|dog **99**|end

- two diamonds (♦♦) means roll 2 d10s
- roll your dice: roll 5 and 8 → combine them to get 58
- find the next word: first number ≥ 58 is **66**|dog, so next word is "dog"
- write it down, look it up and continue the process

北 − ,

北 ○

采莲 ,

东 ,

何田田 ○

间 ○

江南 可

可 采莲

莲叶 ◆ **2**|北 **3**|东 **5**|何田田 **6**|间 **8**|南 **9**|西

南 ,

西 ,

鱼戏 莲叶

○ 鱼戏

, ◆ **7**|鱼戏 **9**|莲叶