CS 1.2: Intro to Data Structures & Algorithms

Histogram & Markov Chain Worksheet

Name: Ayush Jain

Text: *"I like dogs and you like dogs. I like cats but you hate cats."* (ignore all punctuation)

Histograms

Q1: How many distinct word *types* are present in this input text? How many total word *tokens*?

Distinct word types: 8 Total word tokens: 14

Q2: What data structure would be appropriate to store a histogram counting *word frequency*? Why did you choose this data structure? In other words, what makes this data structure ideal?

a dictionary object would be suitable to keep a histogram measuring word frequency since it is the quickest and simplest to code.

Q3: Write the data structure you would create to store this histogram counting *word frequency* (as it would look if you printed it out with Python).

histogram = {'I': 2, 'like': 3, 'dogs': 2, 'and': 1, 'you': 2, 'cats': 2, 'but': 1, 'hate': 1}

Markov Chains

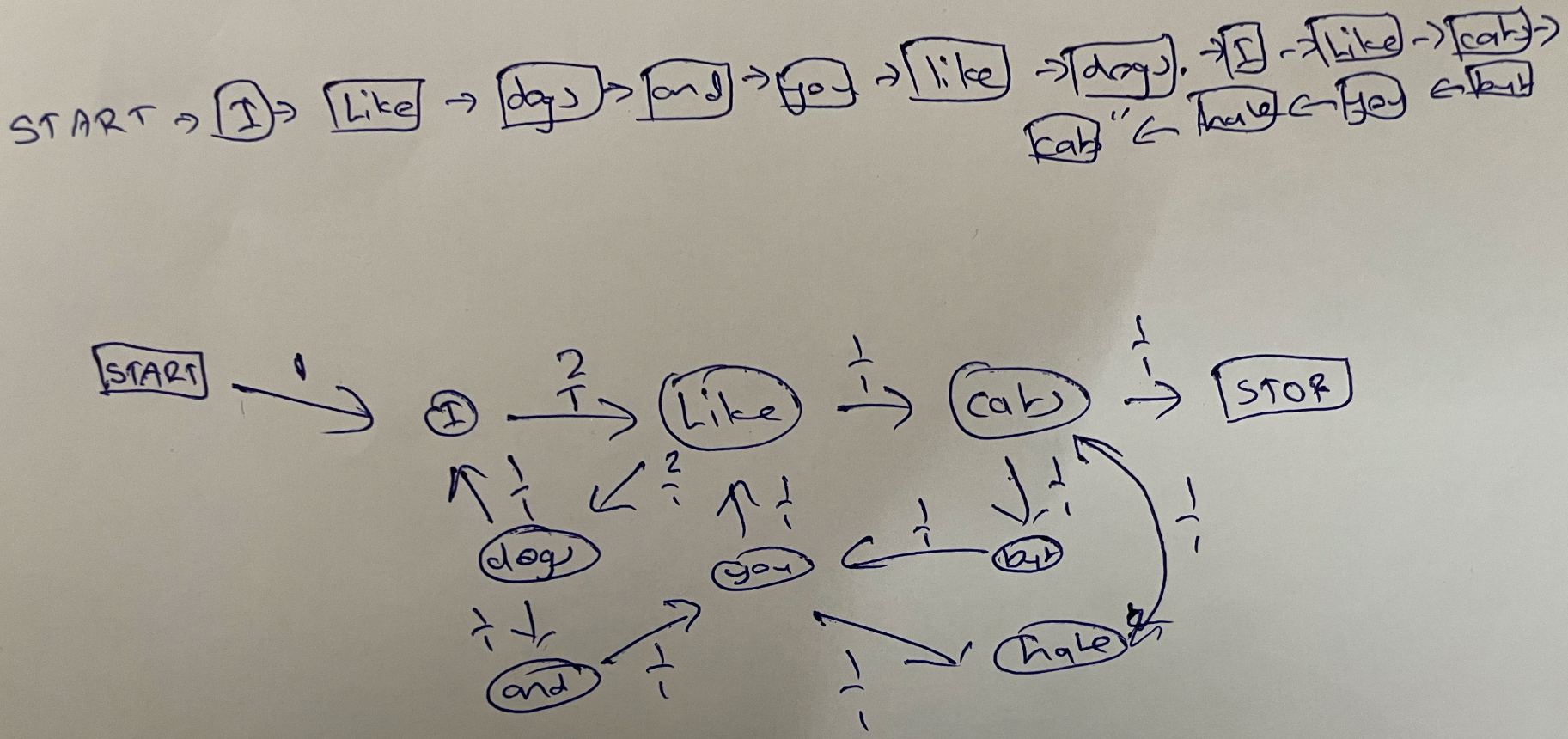
Q4: Draw a conceptual diagram of the *Markov chain* generated from analyzing the text above. Label each *state transition arc* with the count of how many times you observed that word pair.

Q5: Write the data structure you would create to store the word transitions out of the state that represents the word "*like*" in this Markov chain (as it would look if you printed it out with Python).

histogram

Q6: Write a *new sentence* that can be *generated* by doing a *random walk* on this Markov chain.

like dog like cat like I like dog like hate like love


START → I → Like → dogs → and → you → like → dogs. → I → like → cab →
cab → Awale → you → but
START
I
Like
cab
STOP
dogs
you
but
and
Awale
1
2