CS540 Introduction to Artificial Intelligence
Lecture 9

Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 15, 2020

Spam or Ham?

Admin

training data from SMS Spam Collection Dataset.

- Which one of the following messages is a spam?
- A: Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got *a*...
- B: Ok lar... Joking wif u oni...
- C: Free entry in 2*a* wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entr...
- D: U dun say so early hor...*Uc* already then say...
- E: Nah I don't think he goes to usf, he lives around here though

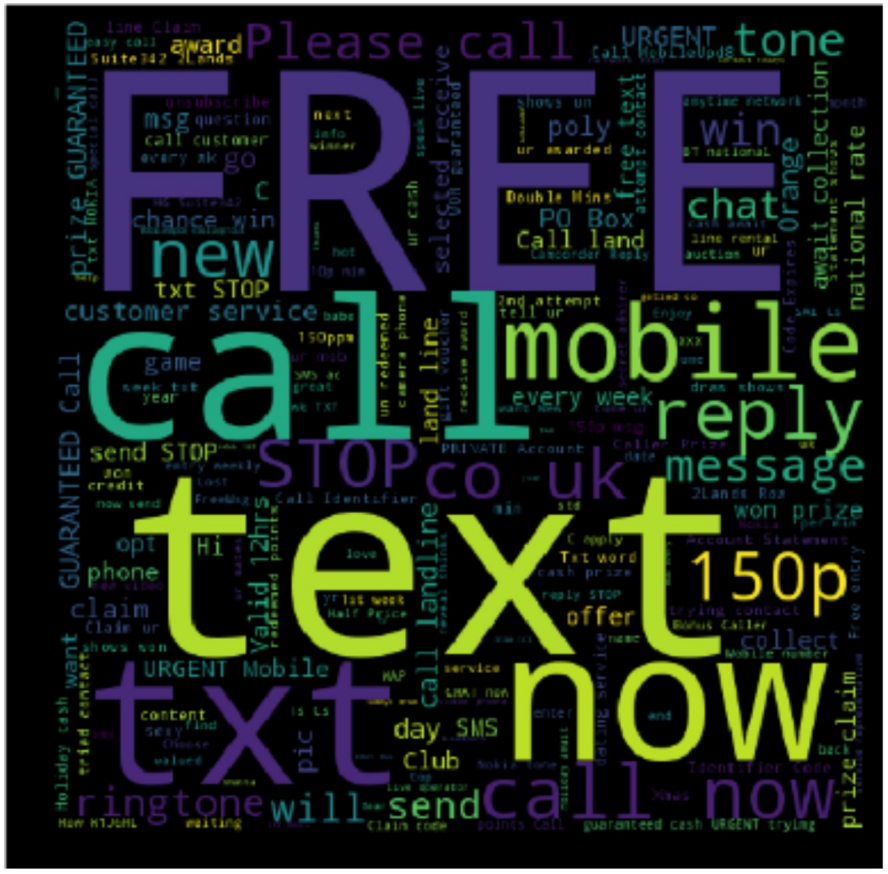
Socrative Q1

grade → P1, P2 ..

→ code, output → P15, P25 ..

Spam or Ham? Visualization

Admin



Bag of Words Features

Definition

- Given a document i and vocabulary with size m, let c_{ij} be the count of the word j in the document i for $j = 1, 2, ..., m$.
- Bag of words representation of a document has features that are the count of each word divided by the total number of words in the document.

$$x_{ij} = \frac{c_{ij}}{\sum_{j'=1}^{m} c_{ij'}}$$

Bag of Words Features Example

Definition

instance 1 : [I am Groot I am Groot ...] x 10

We are Groot

instance 2 : I am Groot

X_1

I	10/33	x_{11}
am	10/33	x_{12}
Groot	11/33	x_{13}
we	1/33	
are	1/33	

X_2

1/3
1/3
1/3
0
0

Unigram Model

Definition

- Unigram models assume independence.

$$\mathbb{P}\{z_1, z_2, \ldots, z_d\} = \prod_{t=1}^{d} \mathbb{P}\{z_t\}$$

- In general, two events A and B are indepedent if:

$$\mathbb{P}\{A|B\} = \mathbb{P}\{A\} \text{ or } \mathbb{P}\{A, B\} = \mathbb{P}\{A\}\,\mathbb{P}\{B\}$$

- For sequence of words, independence means:

$$\mathbb{P}\{z_t|z_{t-1}, z_{t-2}, \ldots, z_1\} = \mathbb{P}\{z_t\}$$

Maximum Likelihood Estimation

Definition

- $\mathbb{P}\{z_t\}$ can be estimated by the count of the word z_t.

$$\hat{\mathbb{P}}\{z_t\} = \frac{c_{z_t}}{\sum_{z=1}^{m} c_z}$$

- This is called the maximum likelihood estimator because it maximizes the probability of observing the sentences in the training set.

Bigram Model

Definition

- Bigram models assume Markov property.

$$\mathbb{P}\{z_1, z_2, \ldots, z_d\} = \mathbb{P}\{z_1\} \prod_{t=2}^{d} \mathbb{P}\{z_t | z_{t-1}\}$$

- Markov property means the distribution of an element in the sequence only depends on the previous element.

$$\mathbb{P}\{z_t | z_{t-1}, z_{t-2}, \ldots, z_1\} = \mathbb{P}\{z_t | z_{t-1}\}$$

Bigram Model Estimation

Definition

- Using the conditional probability formula, $\mathbb{P}\{z_t|z_{t-1}\}$, called transition probabilities, can be estimated by counting all bigrams and unigrams.

$$\hat{\mathbb{P}}\{z_t|z_{t-1}\} = \frac{C_{z_{t-1},z_t}}{C_{z_{t-1}}}$$

total / total

$$Pr\{z_t|z_{t-1}\} = \frac{Pr\{z_t, z_{t-1}\}}{Pr\{z_{t-1}\}}$$

Unigram MLE Probability 1

Quiz

- Given the training data "I am Iron Man", "I love you 3000", "I love you mom", "Tell my family I love them", 18 words in total. With the unigram model, what is the probability of observing a new sentence "I love"?

$$P_r\{I\} \cdot P_r\{love\}$$

$$= \frac{4}{18} \cdot \frac{3}{18}$$

Bigram MLE Probability 1

Quiz

- Given the training data "I am Iron Man", "I love you 3000", "I love you mom", "Tell my family I love them", 18 words in total. With the bigram model, what is the probability of observing $Z_2 =$ "love" given the sentence starts with $Z_1 =$ "I"?

$$Pr\{\text{love} \mid \text{I}\} = \frac{C_{\text{I love}}}{C_{\text{I}}} = \frac{3}{4}$$

Unigram MLE Probability 2

Quiz

Q 2

- Given the training data "I am Groot am I", with the unigram model, what is the probability of observing a new sentence "I am I"?
- A: $\frac{2}{5}$
- B: $\frac{2}{25}$
- C: $\frac{4}{25}$
- D: $\frac{4}{125}$
- E: $\frac{8}{125}$

$\frac{2}{5} \frac{2}{5} \frac{2}{5} = \frac{8}{125}$

Bigram MLE Probability 2

Quiz

- Given the training data "I am Groot am I", with the bigram model, what is the probability of observing a new sentence "I am I" given the first word is "I"?
  - A: $\frac{1}{2}$
  - B: $\frac{1}{4}$
  - C: $\frac{1}{5}$
  - D: $\frac{1}{10}$
  - E: $\frac{4}{25}$

Q3

for P3

$P(\text{am} \mid \text{I}) \cdot \frac{C(\text{I am})}{C(\text{I})} = \frac{1}{1}$ or $\frac{1}{2}$

$P(\text{I} \mid \text{am}) = \frac{C(\text{am I})}{C(\text{am})} = \frac{1}{2}$

I after am

Bigram MLE Probability 3

Quiz

- Given the training data "I am Groot am I", with the bigram model, what is the probability of observing a new sentence "I am Groot" given the first word is "I"?
- A: $\frac{1}{2}$
- B: $\frac{1}{4}$
- C: $\frac{1}{5}$
- D: $\frac{1}{10}$
- E: $\frac{4}{25}$

Q4 - A

Transition Matrix

Definition

- These probabilities can be stored in a matrix called transition matrix of a Markov Chain. The number on row j column j' is the estimated probability $\hat{\mathbb{P}}\{j'|j\}$. If there are 3 tokens $\{1, 2, 3\}$, the transition matrix is the following.

$$\begin{bmatrix} \hat{\mathbb{P}}\{1|1\} & \hat{\mathbb{P}}\{2|1\} & \hat{\mathbb{P}}\{3|1\} \\ \hat{\mathbb{P}}\{1|2\} & \hat{\mathbb{P}}\{2|2\} & \hat{\mathbb{P}}\{3|2\} \\ \hat{\mathbb{P}}\{1|3\} & \hat{\mathbb{P}}\{2|3\} & \hat{\mathbb{P}}\{3|3\} \end{bmatrix}$$

- Given the initial distribution of tokens, the distribution of the next token can be found by multiplying it by the transition probabilities.

Trigram Model

Definition

- The same formula can be applied to trigram: sequences of three tokens.

$$\hat{\mathbb{P}}\{z_t | z_{t-1}, z_{t-2}\} = \frac{c_{z_{t-2}, z_{t-1}, z_t}}{c_{z_{t-2}, z_{t-1}}}$$

- In a document, it is likely that these longer sequences of tokens never appear. In those cases, the probabilities are $\frac{0}{0}$. Because of this, Laplace smoothing adds 1 to all counts.

$$\hat{\mathbb{P}}\{z_t | z_{t-1}, z_{t-2}\} = \frac{c_{z_{t-2}, z_{t-1}, z_t} + 1}{c_{z_{t-2}, z_{t-1}} + m}$$

m: # of words in vocab.

Laplace Smoothing

Definition

- Laplace smoothing should be used for bigram and unigram models too.

$$\hat{\mathbb{P}}\{z_t|z_{t-1}\} = \frac{c_{z_{t-1},z_t} + 1}{c_{z_{t-1}} + m}$$

$$\hat{\mathbb{P}}\{z_t\} = \frac{c_{z_t} + 1}{\sum_{z=1}^{m} c_z + m}$$

- Aside: Laplace smoothing can also be used in decision tree training to compute entropy.

Smoothing Example

Quiz

- Fall 2018 Midterm Q12.
- Given a vocabulary of 10^6, a document with 10^{12} tokens with $c_{\text{zoodles}} = 3$. What is the MLE estimation of $\mathbb{P}\{\text{zoodles}\}$ with and without Laplace smoothing?

$$\frac{c_z + 1}{\sum c_{z'} + m} = \frac{3+1}{10^{12} + 10^6}$$

Smoothing Example 2

Quiz

- Given a vocabulary of 5, a document with 30 words with $c_{\text{Groot}} = 10$. What is the MLE estimation of $\mathbb{P}\{\text{Groot}\}$ with Laplace smoothing?
- A: $\frac{1}{2}$
- B: $\frac{11}{35}$
- C: $\frac{1}{3}$
- D: $\frac{11}{31}$
- E: $\frac{1}{4}$

Q5

$\frac{10+1}{30+5}$ → $\frac{C_z+1}{\sum C_{z'} + m}$

m: # words in vocab.

Smoothing Example 3

Quiz

- Given a vocabulary of 5, a document with 30 words with $c_{\text{I Groot}} = 0, c_{\text{I}} = 9, c_{\text{Groot}} = 10$. What is the MLE estimation of $\mathbb{P}\{\text{Groot} \mid \text{I}\}$ with Laplace smoothing?
- A: $\frac{1}{10}$
- B: $\frac{1}{11}$
- C: $\frac{1}{14}$
- D: $\frac{1}{15}$
- E: 0

size

$z_t \mid z_{t-1}$

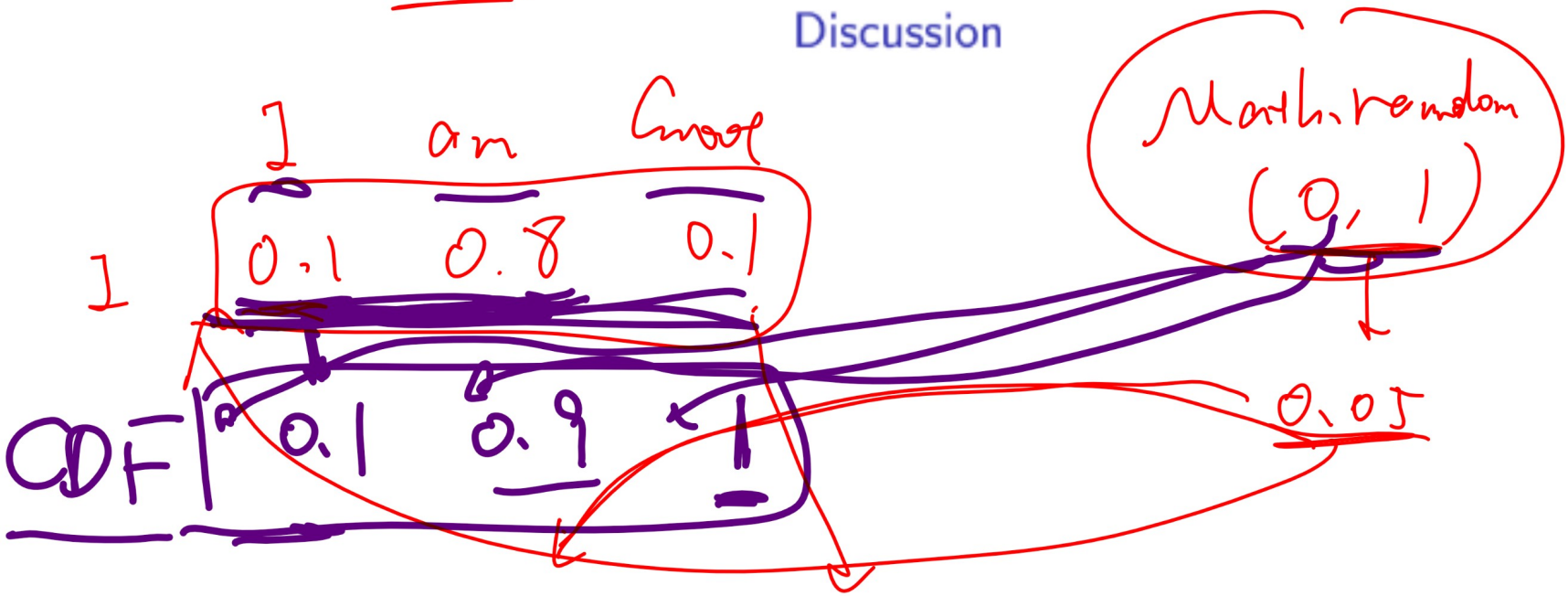
Q6

$$\frac{C_{\text{I Groot}} + 1}{C_{\text{I}} + m} = \frac{0+1}{9+5} = \frac{1}{14}$$

← not overfit.

CDF Inversion Method Diagram

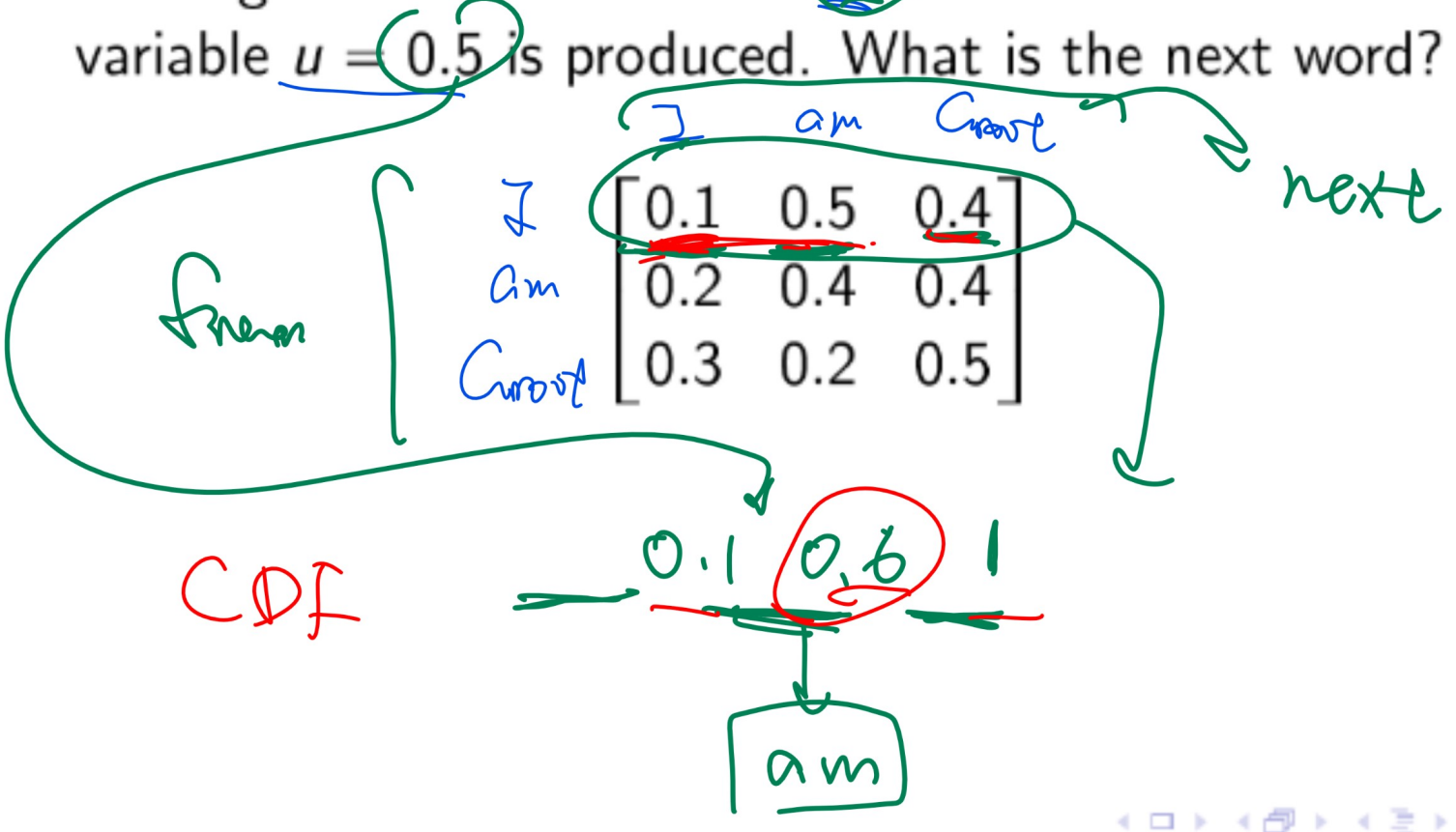
Discussion



Generating New Words 1

Quiz

- Given the transition matrix for words "I" "am" "Groot", starting a sentence with the "I" and a uniform random variable $u = 0.5$ is produced. What is the next word?



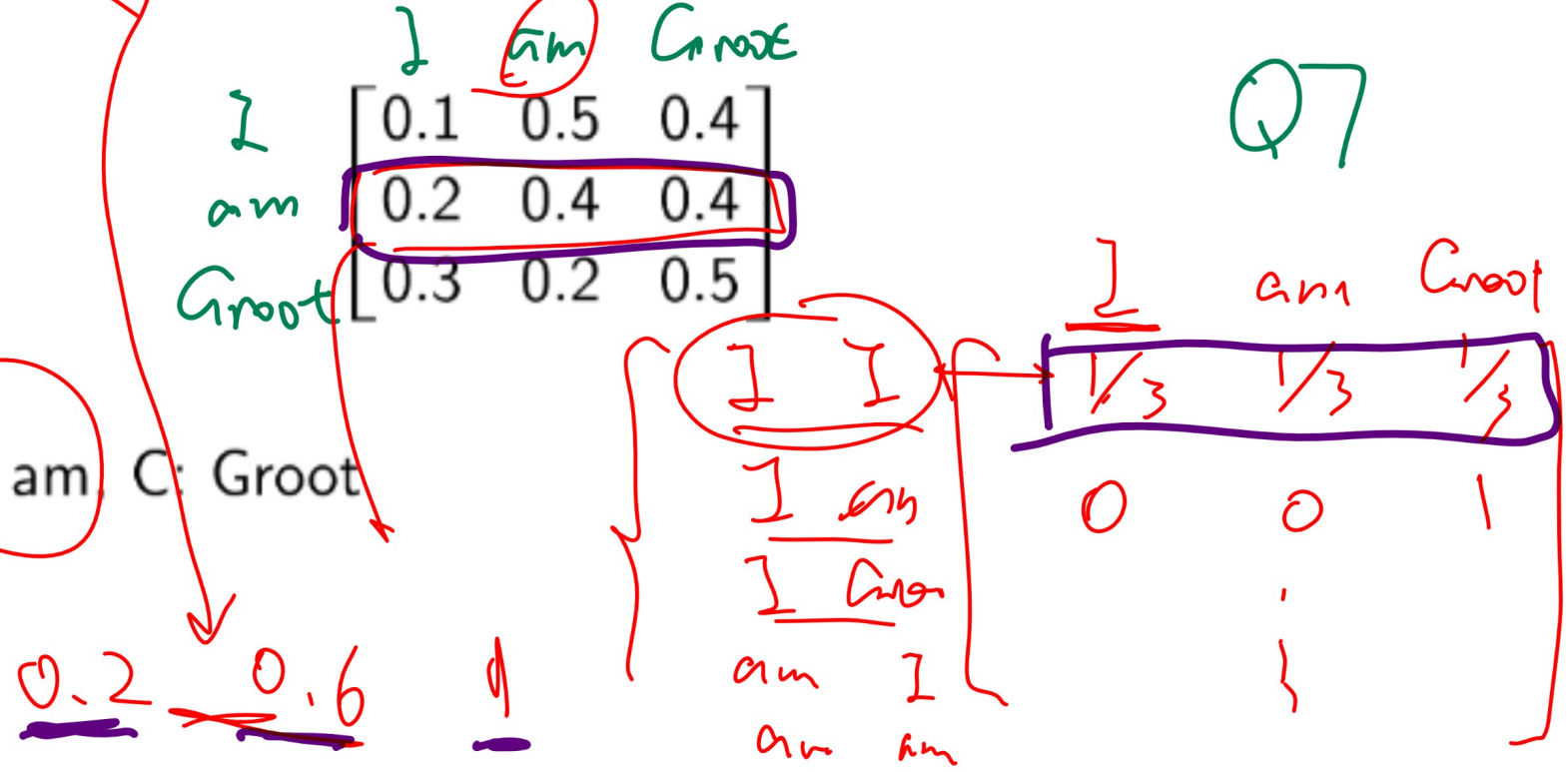
Generating New Words 2

Quiz

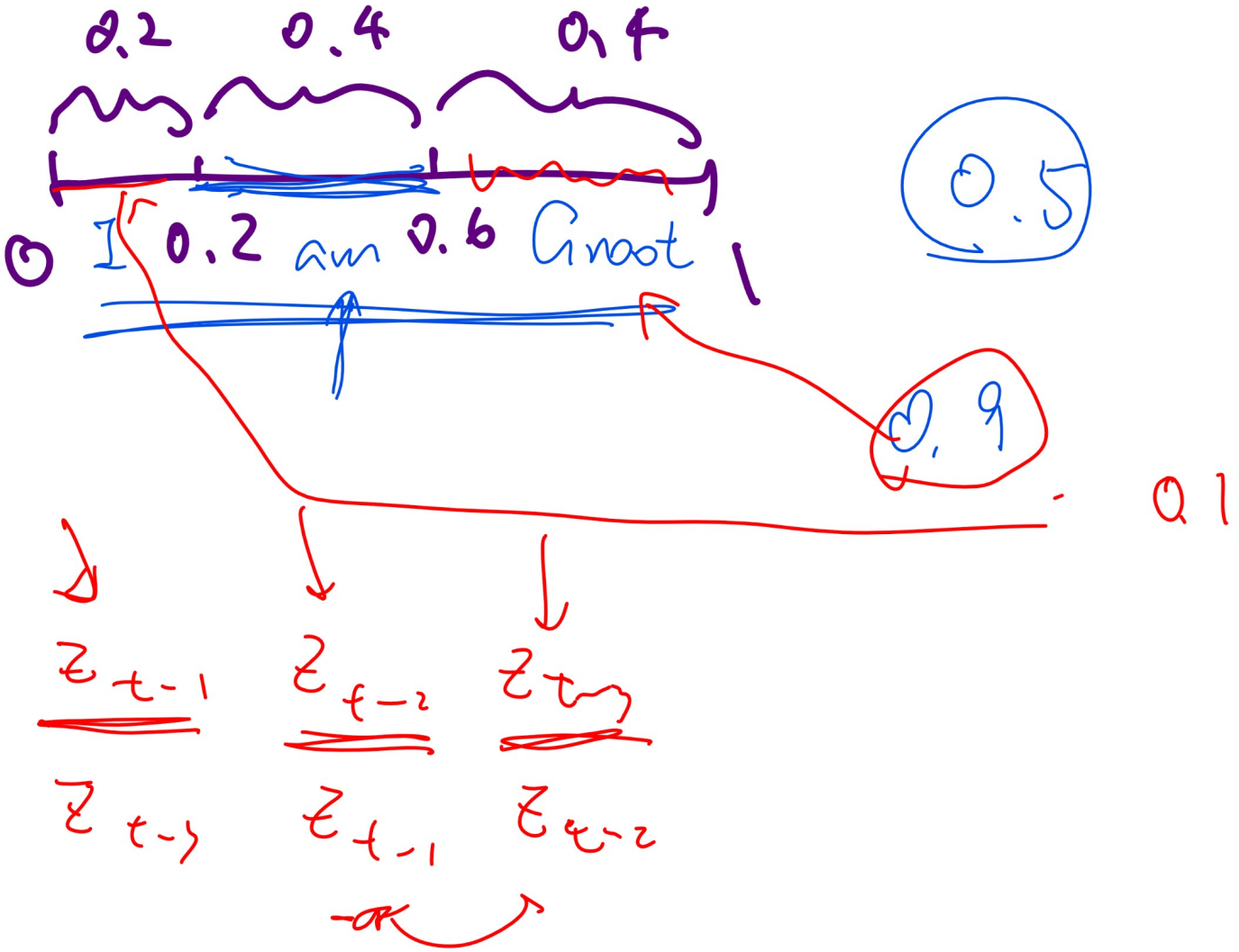
back @ 6:40

- Given the transition matrix for words "I" "am" "Groot", starting a sentence with the "am" and a uniform random variable $u = 0.5$ is produced. What is the next word?

	I	am	Groot
I	0.1	0.5	0.4
am	0.2	0.4	0.4
Groot	0.3	0.2	0.5



- A: I, B: am, C: Groot


0.2
0.4
0.4
0
I
0.2
am
0.6
Groot
1
0.5
0.9
0.1