Introduction to NLP

AIAA 4051
Instructor: Sihong Xie
Email: sihongxie@hkust-gz.edu.cn
AI Thrust, HKUST(GZ)

Why taking this course?

- "Natural Languages" is the "programming language" of the operating system of human society.

- Interface of communication
- Preserving human knowledge and culture.
- Human "computing"

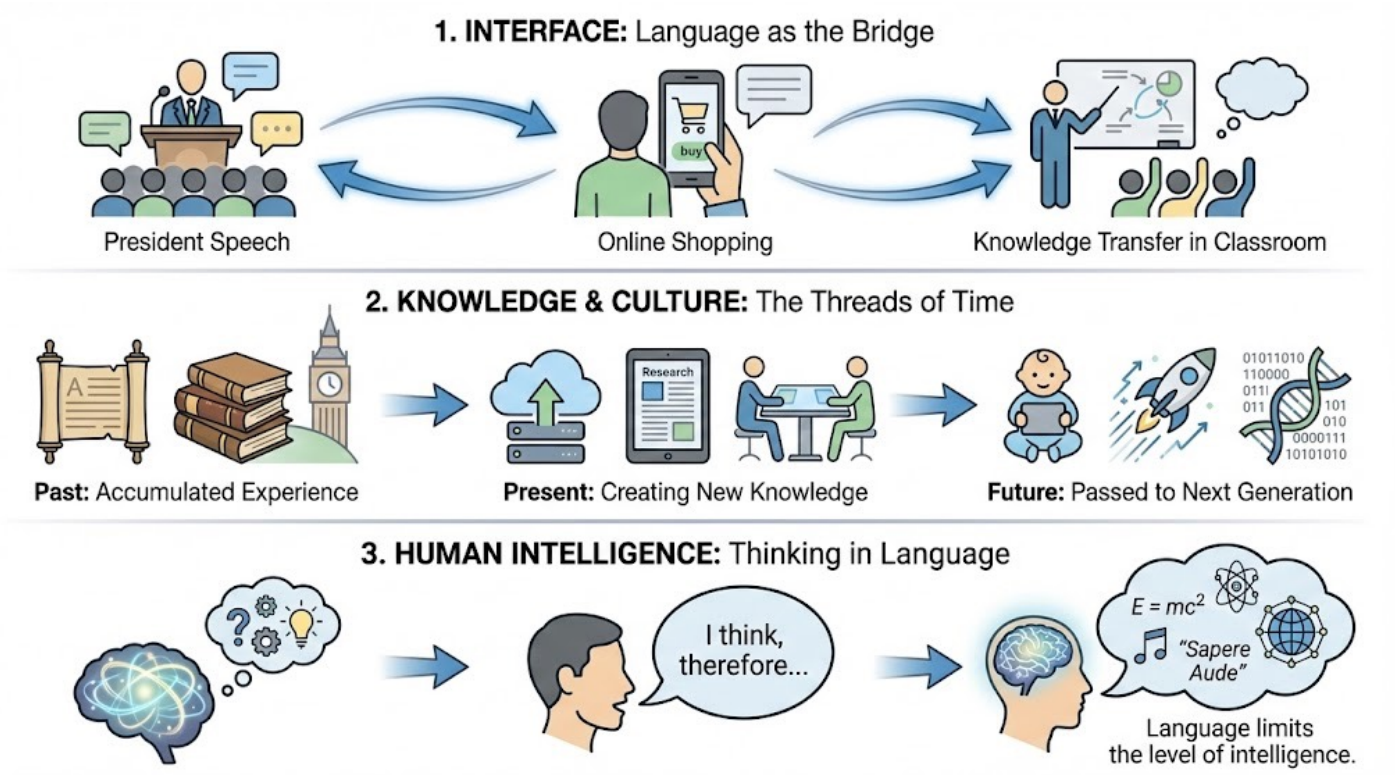


Image source: Gemini Nano Banana

Why taking this course?

- "NLP" as a technology to process human languages, impacts are increasing:

- Generative AI is a new paradigm, and languaged are generative.
- Changing multiple disciplines the good, the bad and the ugly
- Connect you to cutting-edge research, be part of it!

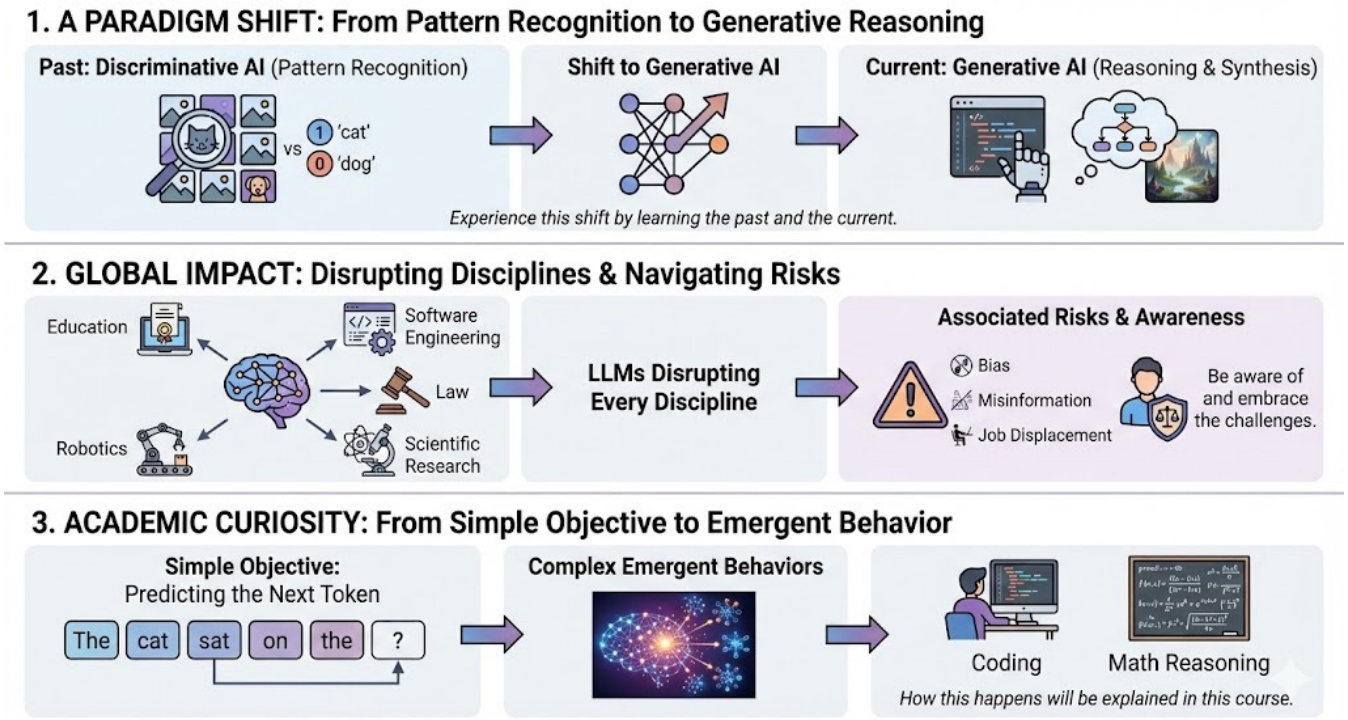
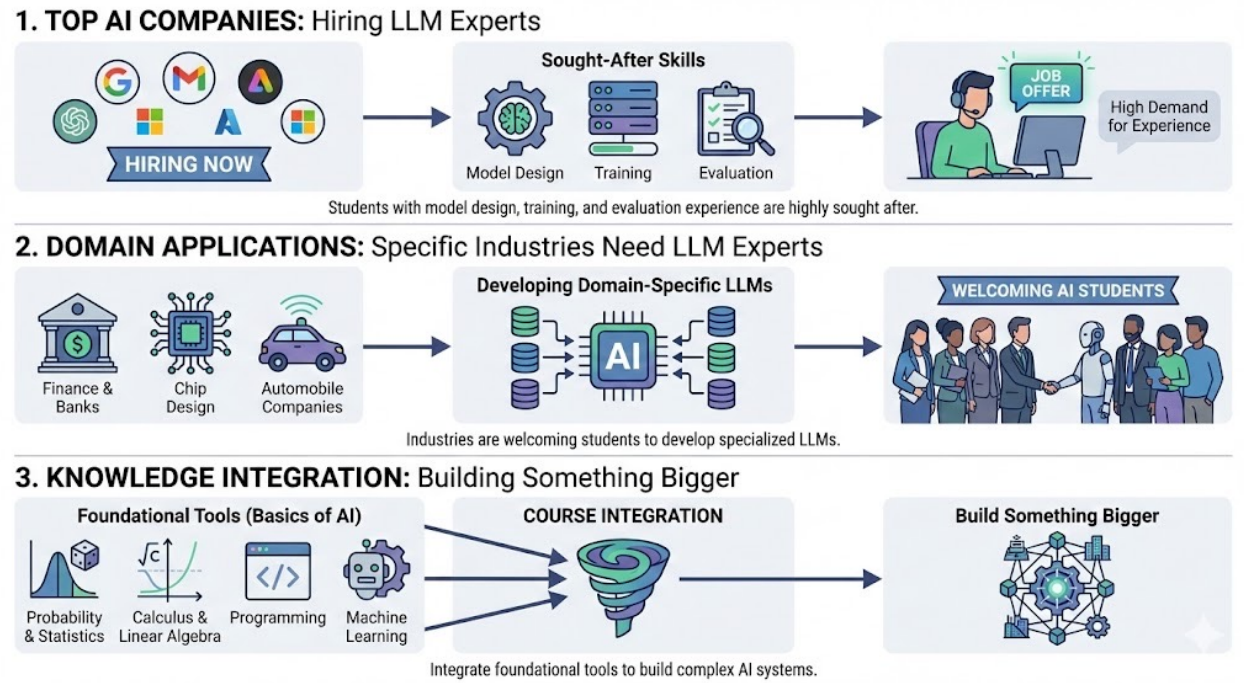


Image source: Gemini Nano Banana

Why taking this course?

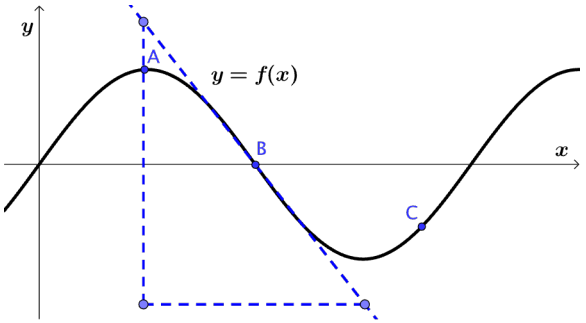
- "NLP" is a good investment of your time:

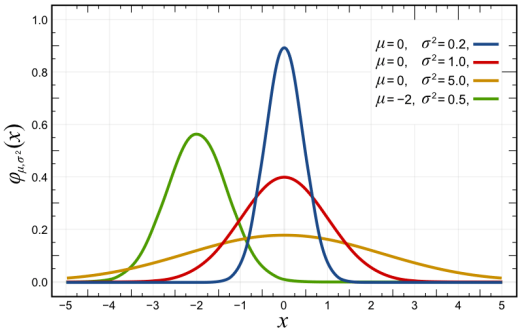
- One of the best-paid job positions
- Many domains need students who know the tech
- Deepen your understanding of what you learned in college



Prerequisites

- Calculus: functions, derivatives and gradients.
- Prob. and stats: axioms of probability, Bayes rules, estimations.
- Programming: Python.
- Alg. and data structures: iteration, search, dynamic programming, complexity, lists, trees, graphs.
- LLM: need to know prompting.
- **Mindset: active and curious**
- Note: no need to be a master nor comprehensive in everything, you can learn them as you actively engage in this class.






```python
#Empty list to store words:
words_no_punc = []

#Removing punctuation marks :
for w in words:
    if w.isalpha():
        words_no_punc.append(w.lower())

#Print the words without punctution marks :
print (words_no_punc)

print ("\n")

#Length :
print (len(words_no_punc))
```

What you can do next?

- Take other advance AI courses: multi-modal AI, embodied AI, responsible AI, AI+X (X=chip design, biomedicine, business, government, etc.)

- Conduct state-of-the-art research: you will learn how to
  - Work in a research team,
  - Prepare datasets,
  - Train and evaluate NLP modols,
  - Write academic papers.
  - Our research group ExRAIL welcomes you to join and experience.

- Build your own NLP-driven app and even a startup.

Logistics

- Basic information is posted on Canvas (check the syllabus PDF).
- Important notes:
  - In-class quizzes have 24% of your final grade: it is close-book short test of what you learn in each class – can be easy if your pay attention to our lecture.
    - Pen and papers: please bring your own stationary for the quizzes. Turn in your answer sheet when leaving the classroom.
  - There is a final exam: multi-choice, short answer, proof.
  - There is a group research project and poster session – can lead to papers.
  - Feedback mechanism: you can scan QR code on Canvas to let us know your opinions (both positive and negative), so we know how to best teach this course.
  - Details of grading is in the syllabus.

Logistics

- Office hours: check the syllabus/Canvas post.
- GPU computing: we provide a small and sufficient amount of computing credits.
  - For promising projects, we invite you to conduct further research with advice and computing power.
  - Hopefully you can publish a paper using results from this course.

The NLP big picture

Optimization Algorithms

Fit the data

NLP models:
- n-grams
- Word vectors
- Text classifier
- POS Tagger
- Parser
- Entity/Relation extractor
- Translator
- ChatBot
- ...

These are more or less statistical models parameterized by parameter **θ** that define probability distributions over texts.

Neural network design

Output/predict

Class labels
POS tags
Parsing trees
BIO tags, relations
Foreign language texts.
Conversations.

Review of Calculus

- We deal with multi-variate functions
- Derivatives and gradients

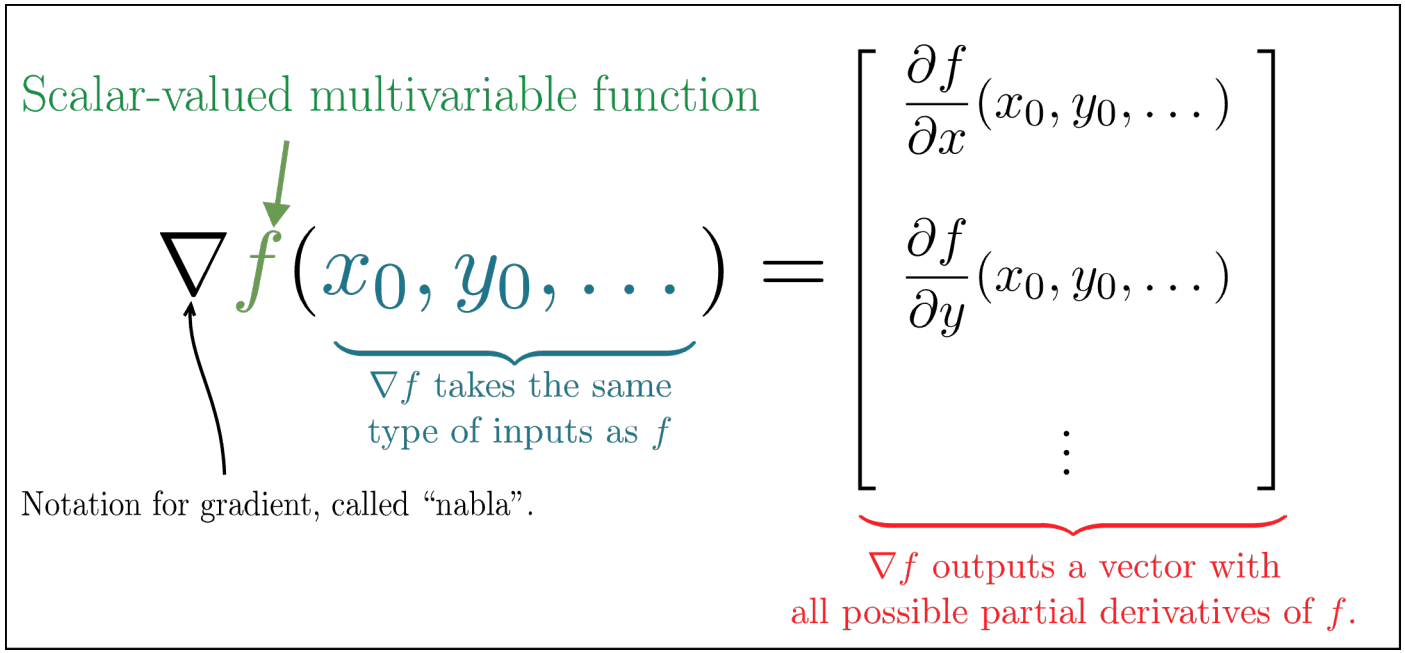
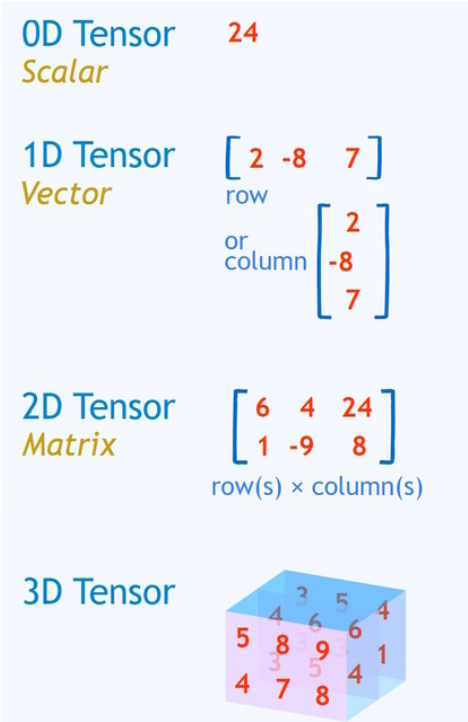
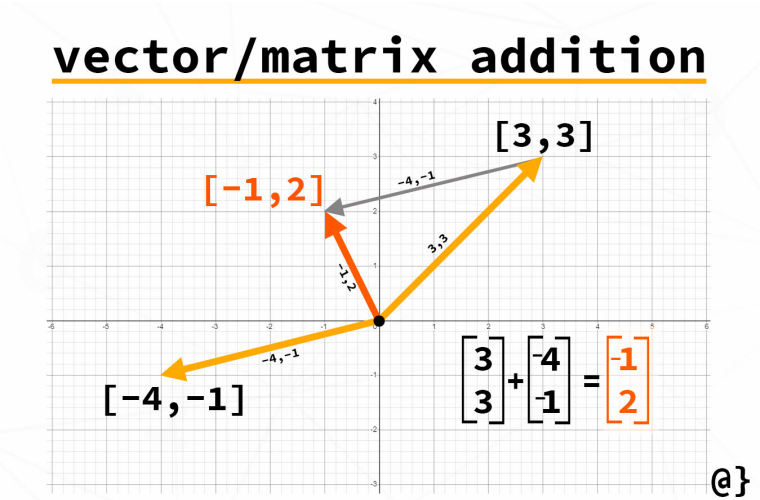


Image source: https://cdn.kastatic.org/ka-perseus-images/841d347c7b6bf9a7ea4c31f6b6d9379a865105d0.svg

Review of Linear Algebra

- Vectors and Matrices





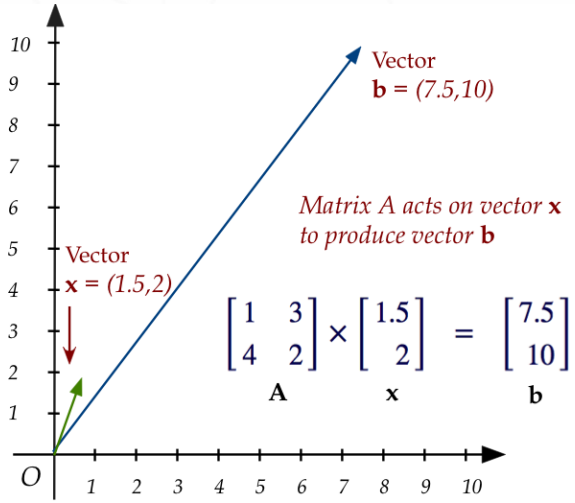


Image source:
https://www.microcontrollertips.com/whats-the-difference-between-gpus-and-tpus-for-ai-processing/
https://www.alpharithms.com/matrix-vector-addition-264010/
https://www2.seas.gwu.edu/~simhaweb/lin/modules/module3/module3.html

Review of Linear Algebra

- Vectors norm and angles
- Given $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, define

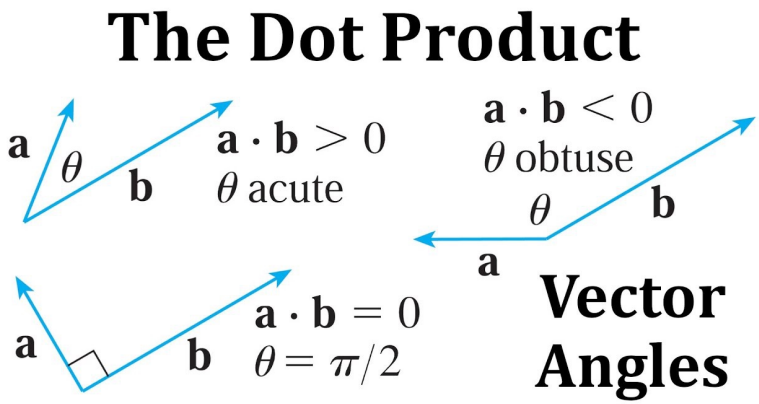
$$|\mathbf{x}|_p \equiv \left(\sum_i |x_i|^p \right)^{1/p}.$$

$$|\mathbf{x}|_2 = |\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

$$|\mathbf{x}|_\infty \equiv \max_i |x_i|.$$

- Example: given $\mathbf{x} = [1, 2, 3]^\top$

name	symbol	value	approx.
L^1-norm	$\|\mathbf{x}\|_1$	6	6.000
L^2-norm	$\|\mathbf{x}\|_2$	$\sqrt{14}$	3.742
L^3-norm	$\|\mathbf{x}\|_3$	$6^{2/3}$	3.302
L^4-norm	$\|\mathbf{x}\|_4$	$2^{1/4}\sqrt{7}$	3.146
L^∞-norm	$\|\mathbf{x}\|_\infty$	3	3.000

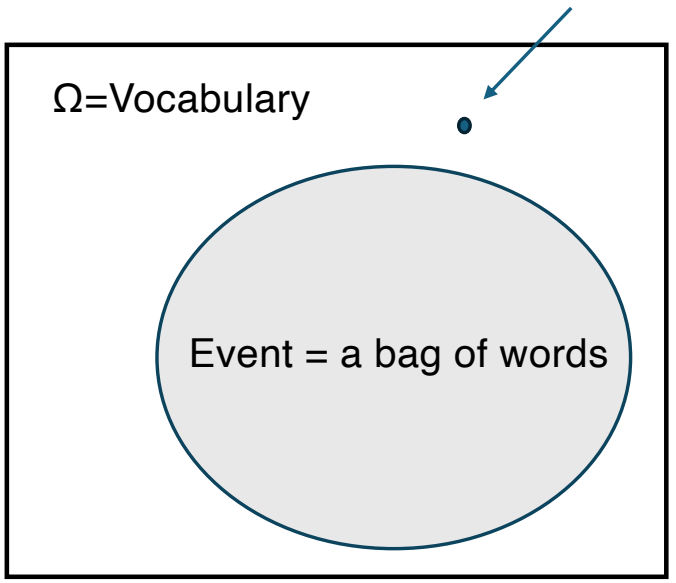


Vector norms and angles are fundamental in representing linguistic information,as you will see in word embedding, transformers, etc.

Image source: https://i.ytimg.com/vi/Zt_xVbqx8b0/maxresdefault.jpg
https://mathworld.wolfram.com/VectorNorm.html

Review of Probability

- The sample space Ω contains all objects that can occur in a specific context.
  - Discrete: Ω = All words in a corpus (vocabulary)
  - Continuous: Ω =Sentimental score in [-1, 1]



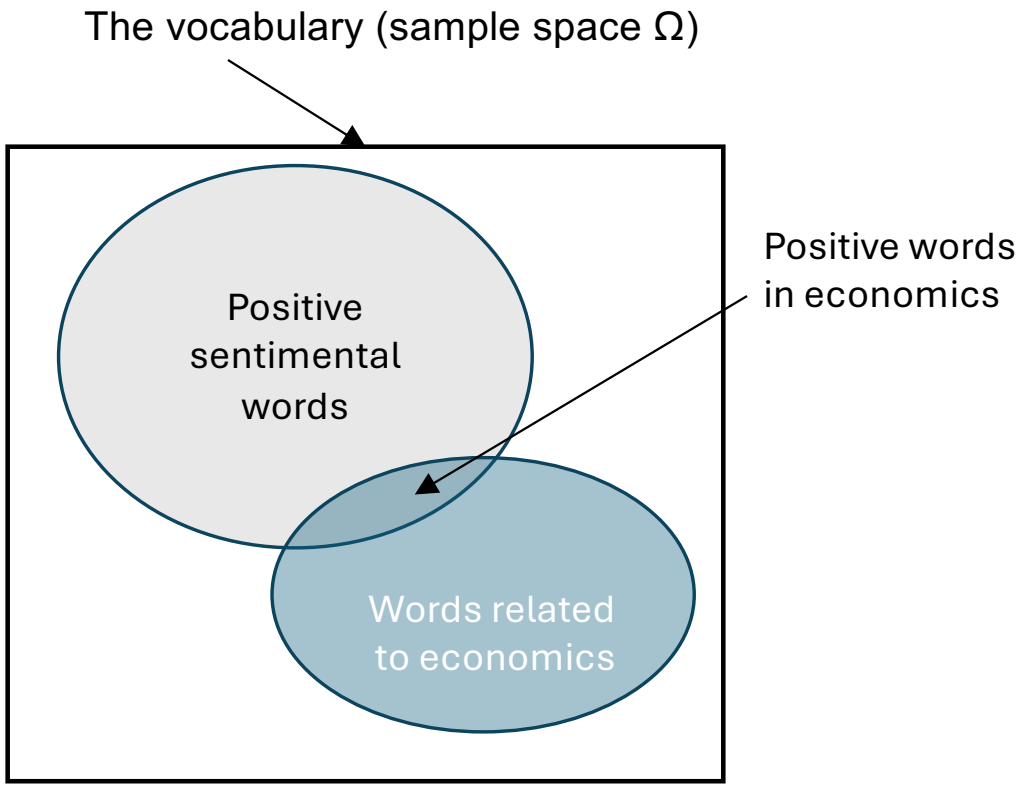
- Event: subsets of Ω ($A \subseteq \Omega$)
  - A document: bag of words (subset of the vocabulary)
  - Positive sentimental score [0, 1]

- Random variable $X : \Omega \rightarrow \text{set of numbers}$
  - e.g., map a word to an integer index.
  - The sentimental score is a number already.
- Probability distribution of a random variable $P(X = x) = P(\omega \in \Omega : X(\omega) = x)$
- Probability of an event $A \subseteq \Omega$: $P(A) \rightarrow [0, 1]$

Review of Probability

The Axioms of Probability

- $P(\Omega) = 1$
- $0 \leq P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



You can learn this from UFUG2104 Applied statistics

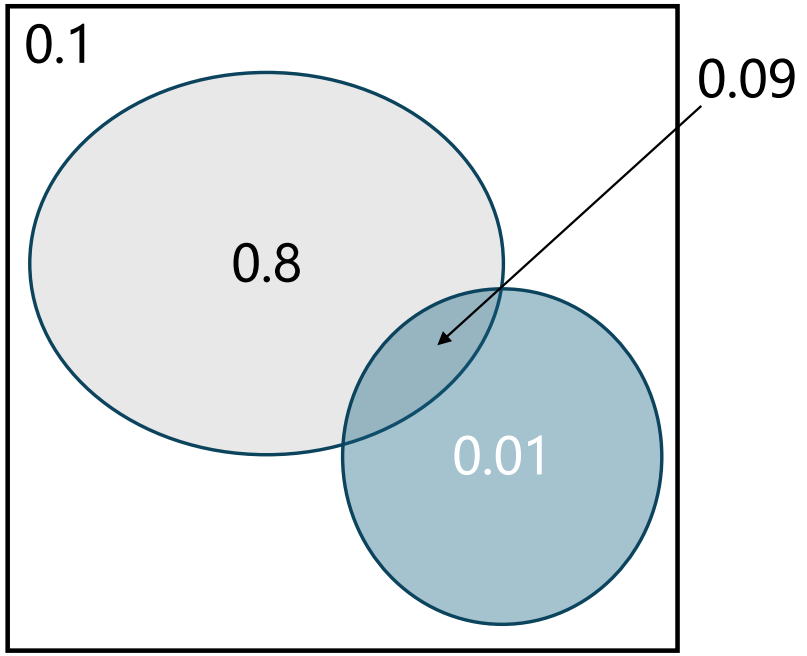
Review of Probability

Joint distribution of m random variables (RVs)

1. List all possible combinations of values of the RVs.
2. Assign valid probability to each combination.

A	B	P(A, B)	
1	1	0.09	=P(A∩B)
1	0	0.8	=P(A∩B^c)
0	1	0.01	=P(B∩A^c)
0	0	0.1	=P(B^c∩A^c)

A^c: the complement of A



The joint has **all information** to calculate other probabilities:

Total probability $P(A) = P(A \cap B) + P(A \cap B^c)$

Conditional probability $P(B|A) = \frac{P(A \cap B)}{P(A)}$

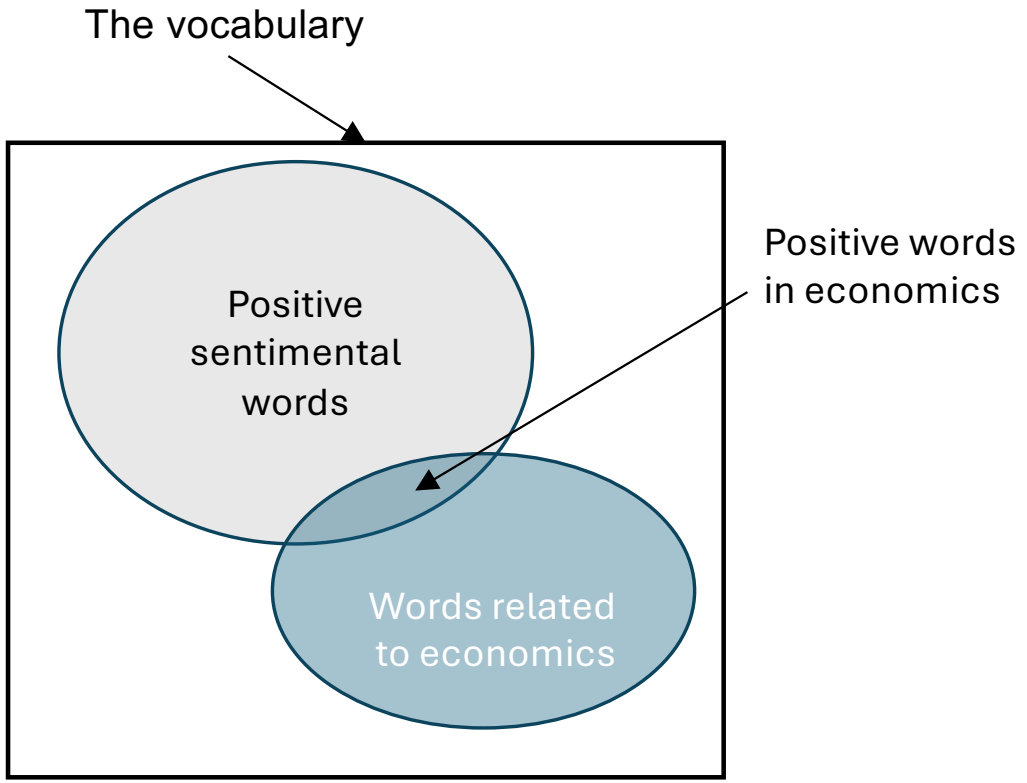
In practice, joint distributions are neither available nor tractable.

Review of Probability

Theorems:

- $P(A) + P(A^c) = 1$
- Total Prob $P(A) = P(A \cap B) + P(A \cap B^c)$
- Conditional: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

$$P(A \cap B) = P(B)P(A|B)$$



- Examples: positive words related to economics, first search words in economics, then find the positives ones. or search positive words first, then find economics-related ones. and they are are symmetric/equivalent.

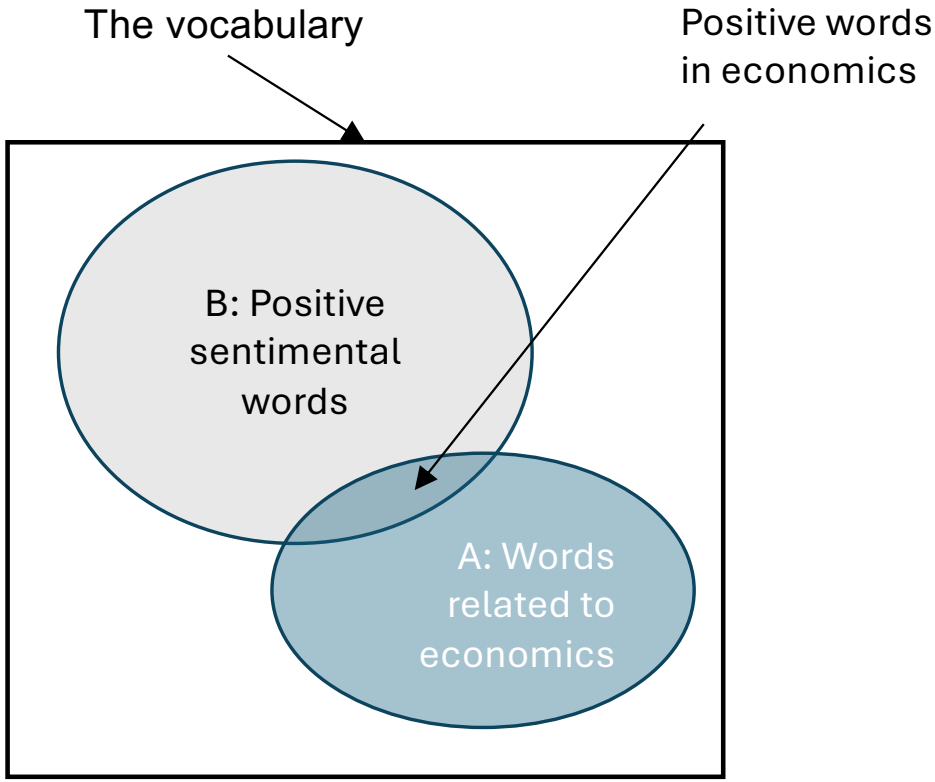
Review of Probability

Compare **prior** P("Great") – just see the word
and **posterior** P("Great" | "Economic") – see the word when reading the Economist

Bayes Rule:

- $$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

- Posterior: $P(B|A)$ (seeing B after having seen A)
- **Prior**: $P(B)$ (what you believe before seeing A)
- Likelihood: $P(A|B)$ (seeing A after having seen B)
- Marginal: $P(A)$ (has the normalizing effect)



Review of Probability

Bayes Rule:

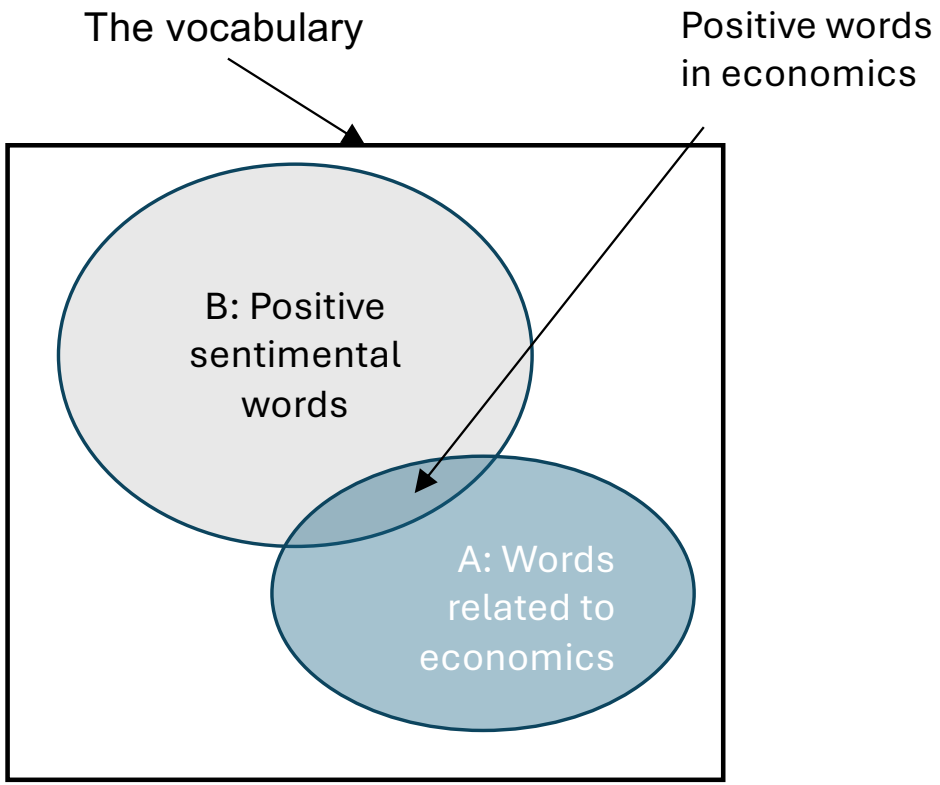
- simple form

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

- Conditional:

$$P(B|A \cap C) = \frac{P(A|B \cap C)P(B \cap C)}{P(A \cap C)}$$



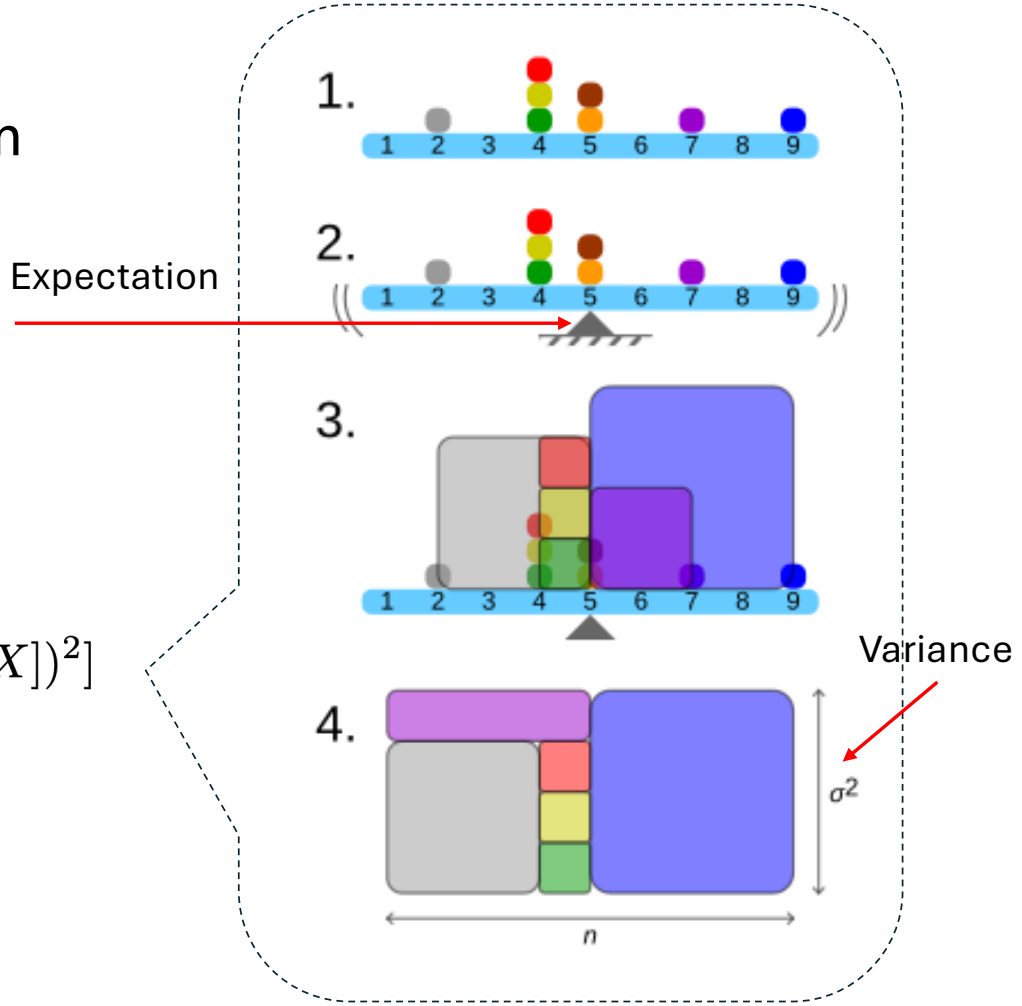
Expectation and variance

- Expectation: sum of values of a random variable weighted by its probability

$$\mathbb{E}[X] = \sum_x p(X = x)x$$

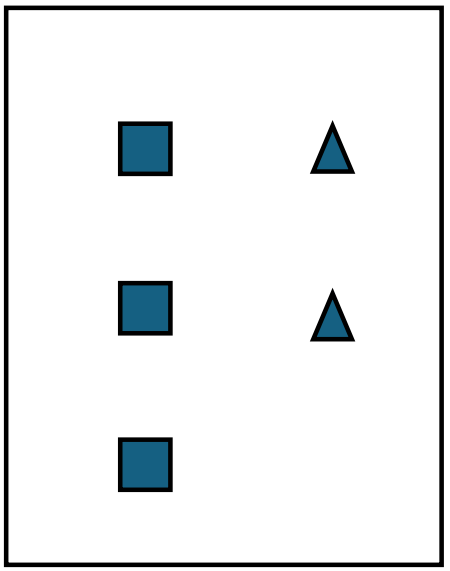
- Variance: expected squared distance from a random variable to expectation

$$\mathrm{Var}[X] = \sum_x p(X = x)(x - \mathbb{E}[X])^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$$



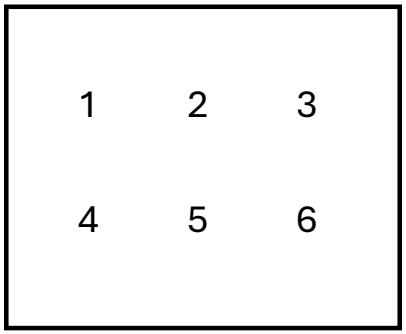
Review of Probability

- **Bernoulli** distribution
- Useful for modeling random variables with two outcomes.
  - Flipping a coin has two outcomes (head ■ and tail ▲).
- $\Pr(Y = 0) = 40\%$
- $\Pr(Y = 1) = 60\%$
- This distribution can be specified using just one parameter, such as the positive rate (=60%).
- Later we will use logistic regression to predict this rate.
- If you have n I.I.D. (Identical and Independent Distributed) Bernoulli random variables, then you have a **Binomial** random variable.
  - Probability of seeing m heads from n I.I.D. coins.



Review of Probability

- **Categorical** distribution
- Useful for modeling random variables with more than two outcomes.
  - Rolling a fair dice has six outcomes (1, 2, 3, 4, 5, and 6).
- $\Pr(X = x) = \frac{1}{6}$, for any $x = 1,2,3,4,5,6$
- This distribution can be specified using just 5 rates.
- Later we will use multi-class logistic regression to predict these rates.
- If you have n I.I.D. (Identical and Independent Distributed) Categorical random variables, then you have a **Multi-nomial** random variable.
  - Probability of seeing some numbers of 1, 2, 3, 4, 5, and 6 after rolling the same dice n times.



Review of Probability

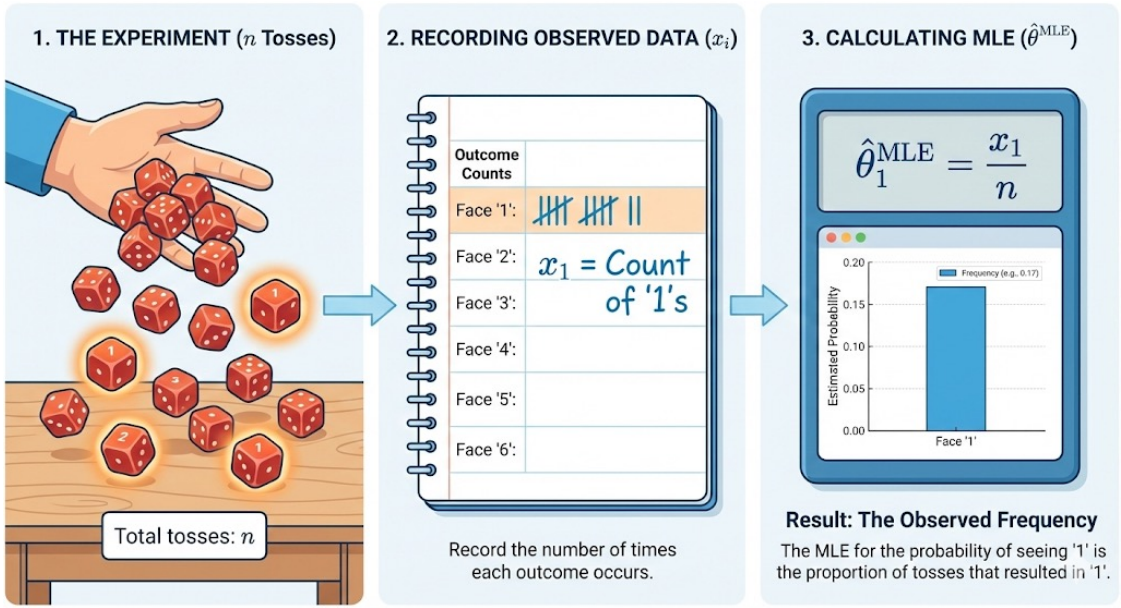
- Independence: two random variable X and Y are independent (X $\perp\!\!\!\perp$ Y) if

$$P(XY) = P(X)P(Y|X) = P(X)P(Y)$$

  - Most often in NLP, independence is an assumption rather than a true fact.
  - Example: whether the word “good” is positive is independent of its previous word.
  - Independence reduce modeling complexity: considering two Bernoulli random variables, 3 (2) parameters are needed to model $P(XY)$ if they are (in)dependent.
    - Independent: each of $P(X)$ and $P(Y)$ needs 1 parameter.
    - Dependent: 4 combinations of values of two variables, with 1 redundant parameter.

Maximum Likelihood Estimation (MLE)

- Let's continue the previous examplen of tossing a dice
  - Suppose you want to learn the probability of seeing number 1 when toss a dice (note that the outcome is a **categorical** random variable).
  - MLE: 1) toss the dice many time; 2) observe the outcomes; 3) calculate the frequency of number 1.
  - The frequency-of-1/total-toss is the MLE of the unknown prob of seeing number 1.



$$\hat{\theta}_1^{\text{MLE}}$$

\theta-one-MLE-hat:
\hat: means that it is an estimation
MLE: it is estimated using MLE
(there are other estimators).

Image source: Gemini Nano Banana

Maximum Likelihood Estimation (MLE)

- Most AI models are trained through MLE
  - Pre-training: predict the next/masked tokens – maximize likelihood of the observe texts (GPT2 is trained on 40GB of WebText data).
  - Supervised Fine Tuning (SFT): predict answers using query—maximize likelihood of the correct answer (GPT3 is fine-tuned using SFT on human answers)

- Training a language model to predict the next token is essentially "using MLE to estimate probability distribution of a very high-dimensional categorical random variable with sample space being all possible words (the "vocabulary")
  - Different language models specify how to calculate the distribution.
  - Simple ones: n-grams and word embedding
  - Compliated ones: parsing trees, Transformers, GPT.

Maximum Likelihood Estimation (MLE)

- More formally, let X be the categorical random variable.
- We roll the dice for n times, and side i appears x_i times.

$$\sum_{i=1}^{k} x_i = n$$

- We let $\theta = (\theta_1, \ldots, \theta_k)$ be the unknown probabilities of seeing side 1 to k. These are the model parameters.

$$\sum_{i=1}^{k} \theta_i = 1$$

- The likelihood (probability) of seeing the above outcome is

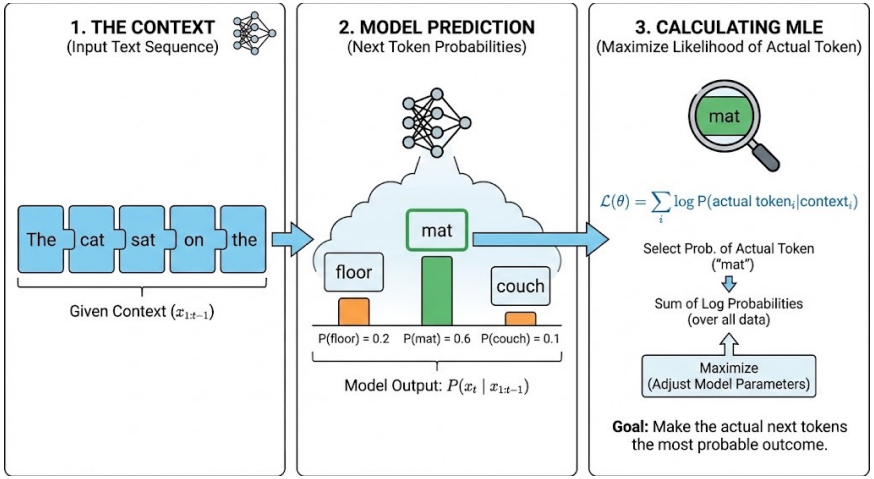
$$L(\theta) = P(x_1, \ldots, x_k|\theta) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i}$$

- Optimize the parameters to maximize the (**log**)-likelihood **Lagrangian**:

$$\mathcal{L}(\theta, \lambda) = \sum_{i=1}^{k} x_i \log(\theta_i) + \lambda \left(1 - \sum_{i=1}^{k} \theta_i\right) \qquad \frac{\partial \mathcal{L}}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda = 0 \implies \hat{\theta}_i = \frac{x_i}{\lambda} \implies \hat{\theta}_i = \frac{x_i}{n}$$

Maximum Likelihood Estimation (MLE)

- Most NLP models are trained in the above way
  1. Obtain observed data, typically texts, such as the entire Wikipedia and Web
  2. Construct a probabilisitc model
     a) How the parameters θ generates the observe data according to the model
     b) The model architecture encode human linguistic knowledge (a sentence is generated according to next token prediction, or a parsing tree. See future lectures).
  3. Formulate likelihood function $L(\theta)$. Maximize the log-likelihood function to find $\hat{\theta}^{\mathrm{MLE}}$



- Note on parameter optimization
  a) Often, it won't be as straightforward as setting gradient to zero and solve for
  b) It can involve complicated optimization algorithms like stochastic gradient descent, AdaGrad, etc. (but these are wrapped in PyTorch and you don't need to implement them from scratches.

Extra readings

- For a high-level survey of LLM, see
  - Large Language Models: A Survey. Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, Jianfeng Gao. 2025.
  - Wikipedia article: Natural Language Processing

- For basic functions, linear algebra, prob. and stats., see Chapter 2.1-2.2, 3.1-3.11, 4.3, 5.5 of Deep Learning by Ian Goodfellow, etc.

- For basic Python/PyTorch programming, go to DeepSeek or Gemini and start prompting (ask them how to help you with this topic).

- Regarding words and tokens, see Chap 2 of SLP3 (the 2026 version).

Demon

- Simulate a training corpus using an LLM;
- Use MLE to estimate word probabilities;
- See how various Python tools are used to read files, tokenize sentences, construct matrices and vectors, plot results;
- See the codes on Canvas.

Conclusion

- It is a great time to learn NLP.
- Abundant of research opportunities.
- Course information are on Canvas and Syllabus.
- Linear algebra, calculus, probability, and statistics.
- Python programming.

- Next: language models.

- Let us know your thoughts using the QR code (to be released)

Quiz

1. Pen and paper; no compute or cellphone allowed.
2. Turn in your answer sheet when you leave the classroom.

- Q1 (T or F): MLE can be solved using some numerical optimization method.
- A1: T
- Q2 (multi-choice): which of the following case indicates that random variables X and Y are independent.
  - A: $P(XY) = P(X)P(Y|X)$
  - B: $P(XY) = P(Y)P(X|Y)$
  - C: $P(X|Y) = P(X)$
- A2: A, B, C
- Q3 (T or F): A categorical distribution is also called a multi-nomial distribution.
- A3: F