# Probability Basics, Naive Bayes

## COM 214: Introduction to Artificial Intelligence

Sasha Fedchin[1,2]

[1]Department of Software Engineering
American University of Central Asia

[2]Department of Computer Science
Tufts University

# Outline

1. Probability Basics
2. Naive Bayes Classifier
3. An Example NLP Problem

# Probability Basics: a six-sided die

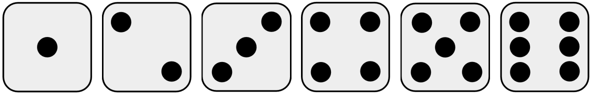
- Suppose we roll a six-sided die. 

# Probability Basics: a six-sided die

- Suppose we roll a six-sided die. 
- What is the probability of getting a number greater than 4?

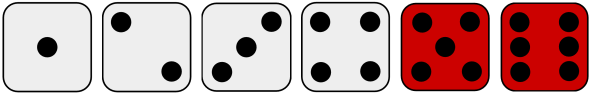
# Probability Basics: a six-sided die

- Suppose we roll a six-sided die. 
- What is the probability of getting a number greater than 4?
- There are 6 possible outcomes:



# Probability Basics: a six-sided die

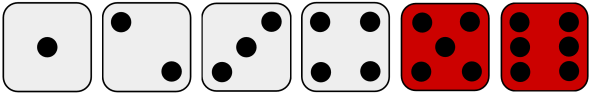
- Suppose we roll a six-sided die. 
- What is the probability of getting a number greater than 4?
- There are 6 possible outcomes:



- We get a number greater than 4 in 2 worlds.

# Probability Basics: a six-sided die

- Suppose we roll a six-sided die. 
- What is the probability of getting a number greater than 4?
- There are 6 possible outcomes:



- We get a number greater than 4 in 2 worlds.
- Therefore, the probability is $P(>4) = \frac{2}{6} = \frac{1}{3}$.

# Independence

- Suppose we roll a six-sided die twice.

# Independence

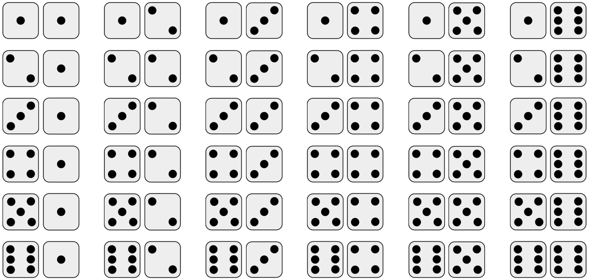
- Suppose we roll a six-sided die twice.
- The probability of getting a number greater than 4 twice?

# Independence

- Suppose we roll a six-sided die twice.
- The probability of getting a number greater than 4 twice?
- Because the rolls are independent, we can multiply the probabilities: $P(>4 \text{ twice}) = P(>4) \times P(>4) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$.

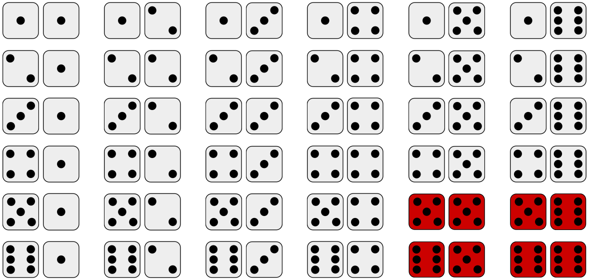
# Independence

- Suppose we roll a six-sided die twice.
- The probability of getting a number greater than 4 twice?
- Because the rolls are independent, we can multiply the probabilities: $P(> 4 \text{ twice}) = P(> 4) \times P(> 4) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$.
- Or you can also look at all 36 possible outcomes:



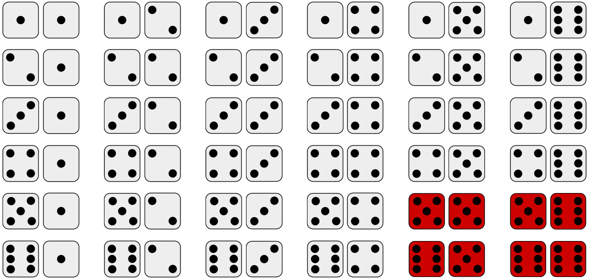
# Independence

- Suppose we roll a six-sided die twice.
- The probability of getting a number greater than 4 twice?
- Because the rolls are independent, we can multiply the probabilities: $P(>4 \text{ twice}) = P(>4) \times P(>4) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$.
- Or you can also look at all 36 possible outcomes:



# Independence

- Suppose we roll a six-sided die twice.
- The probability of getting a number greater than 4 twice?
- Because the rolls are independent, we can multiply the probabilities: $P(>4 \text{ twice}) = P(>4) \times P(>4) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$.
- Or you can also look at all 36 possible outcomes:



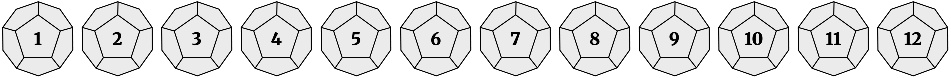
Useful in board games:)

# Another Example: a Twelve-Sided Die

- Suppose we roll a twelve-sided die. 
- What is the probability of getting a number greater than 4?

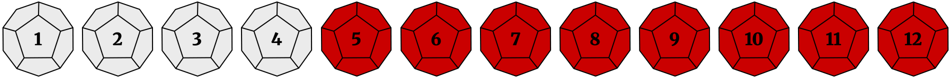
# Another Example: a Twelve-Sided Die

- Suppose we roll a twelve-sided die. 
- What is the probability of getting a number greater than 4?
- There are 12 possible options:



## Another Example: a Twelve-Sided Die

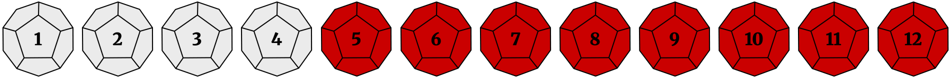
- Suppose we roll a twelve-sided die. 
- What is the probability of getting a number greater than 4?
- There are 12 possible options:




- We get a number greater than 4 in 8 cases.

# Another Example: a Twelve-Sided Die

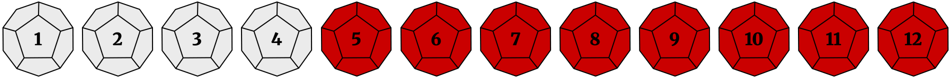
- Suppose we roll a twelve-sided die. 
- What is the probability of getting a number greater than 4?
- There are 12 possible options:




- We get a number greater than 4 in 8 cases.

# Another Example: a Twelve-Sided Die

- Suppose we roll a twelve-sided die. 
- What is the probability of getting a number greater than 4?
- There are 12 possible options:




- We get a number greater than 4 in 8 cases.
- Therefore, the probability is $P(>4) = \frac{8}{12} = \frac{2}{3}$.

# Conditional Probability, Bayes Theorem

- So, for a six-sided die, we have $P(>4) = \frac{1}{3}$.
- And, for a twelve-sided die, we have $P(>4) = \frac{2}{3}$.
- How do we distinguish between these?

# Conditional Probability, Bayes Theorem

- So, for a six-sided die, we have $P(>4) = \frac{1}{3}$.
- And, for a twelve-sided die, we have $P(>4) = \frac{2}{3}$.
- How do we distinguish between these?
- We can use conditional probabilities, written as $P(a \mid b)$ for "probability of a given b".

# Conditional Probability, Bayes Theorem

- So, for a six-sided die, we have $P(>4|\text{6-sided}) = \frac{1}{3}$.
- And, for a twelve-sided die, we have $P(>4|\text{12-sided}) = \frac{2}{3}$.
- How do we distinguish between these?
- We can use conditional probabilities, written as $P(a \mid b)$ for "probability of a given b".

# Conditional Probability, Bayes Theorem

- So, for a six-sided die, we have $P(>4|\text{6-sided}) = \frac{1}{3}$.
- And, for a twelve-sided die, we have $P(>4|\text{12-sided}) = \frac{2}{3}$.
- How do we distinguish between these?
- We can use conditional probabilities, written as $P(a \mid b)$ for "probability of a given b".
- More formally, $P(a \mid b) = \frac{P(a \text{ AND } b)}{P(b)}$.

# Conditional Probability, Bayes Theorem

- So, for a six-sided die, we have $P(>4|\text{6-sided}) = \frac{1}{3}$.
- And, for a twelve-sided die, we have $P(>4|\text{12-sided}) = \frac{2}{3}$.
- How do we distinguish between these?
- We can use conditional probabilities, written as $P(a \mid b)$ for "probability of a given b".
- More formally, $P(a \mid b) = \frac{P(a \text{ AND } b)}{P(b)}$.
- From this we get Bayes Theorem:

$$P(a \mid b) = \frac{P(b|a)P(a)}{P(b)}$$

# Thomas Bayes

**Thomas Bayes** (/beɪz/ *BAYZ*, audio; c. 1701 – 7 April 1761[2][4][note 1]) was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem.

Bayes never published what would become his most famous accomplishment; his notes were edited and published posthumously by Richard Price.[5]

The Reverend
**Thomas Bayes**

# Outline

1 Probability Basics

2 Naive Bayes Classifier

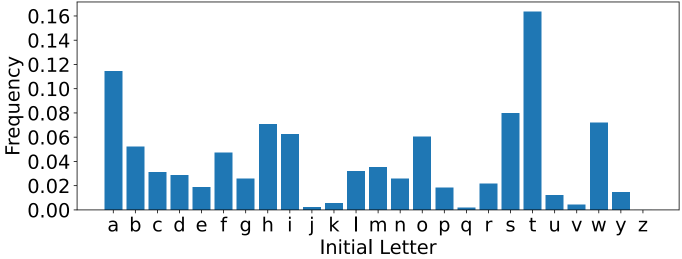
3 An Example NLP Problem

# From Dice to NLP (Natural Language Processing)

- What is the probability that the initial letters in a random text spell "AUCAISGREAT"?

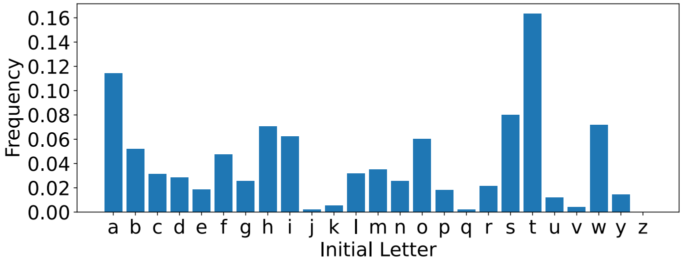
# From Dice to NLP (Natural Language Processing)

- What is the probability that the initial letters in a random text spell "AUCAISGREAT"?



# From Dice to NLP (Natural Language Processing)

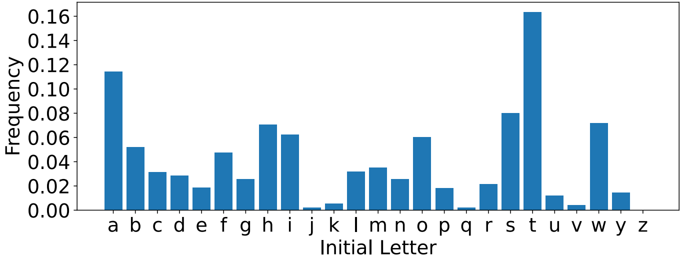
- What is the probability that the initial letters in a random text spell "AUCAISGREAT"?




- Then $P(AUCAISGREAT) = P(A)^3 \times P(U) \times P(C) \times P(I) \times P(S) \times P(G) \times P(R) \times P(E) \times P(T)$

# From Dice to NLP (Natural Language Processing)

- What is the probability that the initial letters in a random text spell "AUCAISGREAT"?




- Then $P(AUCAISGREAT) = P(A)^3 \times P(U) \times P(C) \times P(I) \times P(S) \times P(G) \times P(R) \times P(E) \times P(T) \approx 0.11^3 \times 0.01 \times 0.03 \times 0.06 \times 0.08 \times 0.03 \times 0.02 \times 0.02 \times 0.02 \approx 4.7 \times 10^{-15}$

# Language Modelling

- What is the probability that the initial letters in a text spell "AUCAISGREAT" given that we know that we know the first letters must spell something meaningful?

# Language Modelling

- What is the probability that the initial letters in a text spell "AUCAISGREAT" given that we know that we know the first letters must spell something meaningful?
- We can assume that "AUCAISGREAT" is a phrase in English.

# Language Modelling

- What is the probability that the initial letters in a text spell "AUCAISGREAT" given that we know that we know the first letters must spell something meaningful?
- We can assume that "AUCAISGREAT" is a phrase in English.
- Then, we can compute $P(AUCAISGREAT) = P(AUCA) \times P(IS) \times P(GREAT)$, where the probabilities for respective words are estimated from a large corpus (assuming this is the only way to tokenize the phrase).

# Language Modelling

- What is the probability that the initial letters in a text spell "AUCAISGREAT" given that we know that we know the first letters must spell something meaningful?
- We can assume that "AUCAISGREAT" is a phrase in English.
- Then, we can compute $P(AUCAISGREAT) = P(AUCA) \times P(IS) \times P(GREAT)$, where the probabilities for respective words are estimated from a large corpus (assuming this is the only way to tokenize the phrase).
- This is a naive approach because these probabilities aren't actually independent... Hence *Naive Bayes*

# Classification With Language Models, Ngrams

| Class A: | Class B: |
| --- | --- |
| $P(A \mid data)$ | $P(B \mid text)$ |
| $\frac{P(text \mid A) P(A)}{P(text)}$ | $\frac{P(text \mid B) P(B)}{P(text)}$ |

$P(text|C) = P(w_1|C) * P(w_2|C)...P(w_n|C) \leftarrow$ Naive Bayes

$P(text|C) = P(w_1|C) * P(w_2|w_1, C)...P(w_n|w_{n-1}, C) \leftarrow$ 2-grams

$P(text|C) = P(w_1|C) * P(w_2|w_1, w_2, C)...P(w_n|w_{n-2}, w_{n-1}, C) \leftarrow$ 3-grams

# Outline

1. Probability Basics
2. Naive Bayes Classifier
3. An Example NLP Problem

# Motivation: Acrostics

**Edgar Allan Poe, 1829**

Elizabeth it is in vain you say
"Love not"—thou sayest it in so sweet a way:
In vain those words from thee or L. E. L.
Zantippe's talents had enforced so well:
Ah! if that language from thy heart arise,
Breathe it less gently forth—and veil thine eyes.
Endymion, recollect, when Luna tried
To cure his love—was cured of all beside—
His folly—pride—and passion—for he died.

# Motivation: Acrostics

**Edgar Allan Poe, 1829**

**E**-lizabeth it is in vain you say
"**L**-ove not"—thou sayest it in so sweet a way:
**I**-n vain those words from thee or L. E. L.
**Z**-antippe's talents had enforced so well:
**A**-h! if that language from thy heart arise,
**B**-reathe it less gently forth—and veil thine eyes.
**E**-ndymion, recollect, when Luna tried
**T**-o cure his love—was cured of all beside—
**H**-is folly—pride—and passion—for he died.

# Why Care about Acrostics?

- Acrostic is a text where first letters of each line (sentence, paragraph) form a meaningful phrase.

# Why Care about Acrostics?

- Acrostic is a text where first letters of each line (sentence, paragraph) form a meaningful phrase.
- Acrostics have been composed as early as 2nd millennium BCE.

# Why Care about Acrostics?

- Acrostic is a text where first letters of each line (sentence, paragraph) form a meaningful phrase.
- Acrostics have been composed as early as 2nd millennium BCE.
- In medieval times, acrostics were often used as signatures or to convey ideas, which, when spoken publicly, would have put the author in danger of persecution.

# Why Care about Acrostics?

- Acrostic is a text where first letters of each line (sentence, paragraph) form a meaningful phrase.
- Acrostics have been composed as early as 2nd millennium BCE.
- In medieval times, acrostics were often used as signatures or to convey ideas, which, when spoken publicly, would have put the author in danger of persecution.
- Uncovering an acrostic may prove authorship or reveal hidden knowledge.

# Goal: Uncovering New Acrostics

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic.

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic. Question: What does $P(t_1|a)$ stand for? What about $P(a|t_1)$? $P(t_1|\text{not } a)$?

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic. Question: $P(a|t_1) \lesseqgtr ? \; P(a|t_2)$

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic.
- We can compute $\frac{P(a|t_1)}{P(a|t_2)}$

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic.
- We can compute $\frac{P(a|t_1)}{P(a|t_2)} = \frac{P(t_1|a)P(a)P(t_2)}{P(t_1)P(t_2|a)P(a)}$

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic.
- We can compute $\frac{P(a|t_1)}{P(a|t_2)} = \frac{P(t_1|a)P(a)P(t_2)}{P(t_1)P(t_2|a)P(a)} = \frac{P(t_1|a)P(t_2)}{P(t_1)P(t_2|a)}$

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic.
- We can compute $\frac{P(a|t_1)}{P(a|t_2)} = \frac{P(t_1|a)P(a)P(t_2)}{P(t_1)P(t_2|a)P(a)} = \frac{P(t_1|a)P(t_2)}{P(t_1)P(t_2|a)}$
$= \frac{P(t_1|a)[P(t_2|a)P(a)+P(t_2|\text{not } a)P(\text{not } a)]}{[P(t_1|a)P(a)+P(t_1|\text{not } a)P(\text{not } a)]P(t_2|a)}$

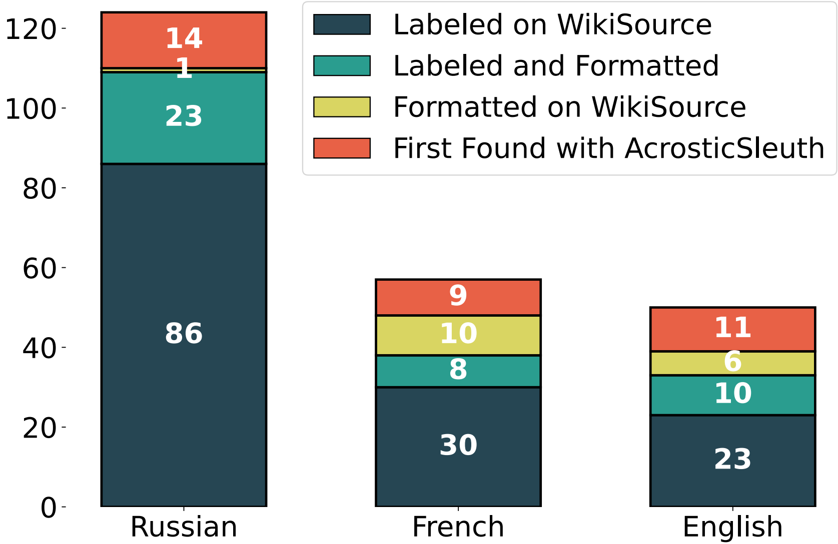
# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic.
- We can compute $\frac{P(a|t_1)}{P(a|t_2)} = \frac{P(t_1|a)P(a)P(t_2)}{P(t_1)P(t_2|a)P(a)} = \frac{P(t_1|a)P(t_2)}{P(t_1)P(t_2|a)}$ $= \frac{P(t_1|a)[P(t_2|a)P(a)+P(t_2|\text{not } a)P(\text{not } a)]}{[P(t_1|a)P(a)+P(t_1|\text{not } a)P(\text{not } a)]P(t_2|a)} \xrightarrow[P(a)\to 0]{} \frac{P(t_1|a)P(t_2|\text{not } a)}{P(t_2|a)P(t_1|\text{not } a)}$

# Putting it All Together

- What is more likely to be an acrostic: a text $t_1$, where the first letters spell ``AUCAISGREAT" or another text $t_2$ where the first letters spell ``ABDEARL"?
- Let $P(a)$ be the probability of encountering an acrostic.
- We can compute $\frac{P(a|t_1)}{P(a|t_2)} = \frac{P(t_1|a)P(a)P(t_2)}{P(t_1)P(t_2|a)P(a)} = \frac{P(t_1|a)P(t_2)}{P(t_1)P(t_2|a)}$
$= \frac{P(t_1|a)[P(t_2|a)P(a)+P(t_2|\text{not } a)P(\text{not } a)]}{[P(t_1|a)P(a)+P(t_1|\text{not } a)P(\text{not } a)]P(t_2|a)} \xrightarrow[P(a)\to 0]{} \frac{P(t_1|a)P(t_2|\text{not } a)}{P(t_2|a)P(t_1|\text{not } a)}$
- To solve the problem we only need to know $P(t|a)$ and $P(t|\text{not } a)$.

# AcrosticSleuth Results



# Spotlight: Thomas Hobbes

**Thomas Hobbes** (/hɒbz/ *HOBZ*; 5 April 1588 – 4 December 1679) was an English philosopher, best known for his 1651 book *Leviathan*, in which he expounds an influential formulation of social contract theory.[4] He is considered to be one of the founders of modern political philosophy.[5][6]

In his early life, overshadowed by his father's departure following a fight, he was taken under the care of his wealthy uncle. Hobbes's academic journey began in Westport, leading him to Oxford University

**Thomas Hobbes**