FoCS:
Probability and Naive Bayes Classification

Niema Moshiri
UC San Diego SPIS 2019

Sentence Sentiment

- Imagine your crush walks up to you to talk, and she says...

Sentence Sentiment

- Imagine your crush walks up to you to talk, and she says…

You're

Sentence Sentiment

- Imagine your crush walks up to you to talk, and she says...

You're **pretty**

Sentence Sentiment

- Imagine your crush walks up to you to talk, and she says...

You're pretty **annoying**

Sentence Sentiment

- Imagine your crush walks up to you to talk, and she says...

You’re pretty **annoying**

- How did a single word change things so much?

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes
    - Flip a coin 3 times

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes
  - Flip a coin 3 times
- **Sample Space:** The set of all possible outcomes of an experiment

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes
  - Flip a coin 3 times
- **Sample Space:** The set of all possible outcomes of an experiment
  - {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH}

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes
  - Flip a coin 3 times
- **Sample Space:** The set of all possible outcomes of an experiment
  - {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH}
- **Random Variable:** Function from the sample space to the real numbers

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes
  - Flip a coin 3 times
- **Sample Space:** The set of all possible outcomes of an experiment
  - {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH}
- **Random Variable:** Function from the sample space to the real numbers
  - X = number of heads (TTT ➔ 0, HTT ➔ 1, HHH ➔ 3, etc.)

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes
  - Flip a coin 3 times
- **Sample Space:** The set of all possible outcomes of an experiment
  - {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH}
- **Random Variable:** Function from the sample space to the real numbers
  - X = number of heads (TTT → 0, HTT → 1, HHH → 3, etc.)
- **Event:** A subset of the sample space

Probability Theory: Terminology

- **Experiment (Trial):** Any procedure that can be infinitely repeated and has a well-defined set of possible outcomes
  - Flip a coin 3 times
- **Sample Space:** The set of all possible outcomes of an experiment
  - {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH}
- **Random Variable:** Function from the sample space to the real numbers
  - X = number of heads (TTT → 0, HTT → 1, HHH → 3, etc.)
- **Event:** A subset of the sample space
  - "Even Number of Heads" = {THH, HTH, HHT, TTT}

Probability of an Event

- Remember, an **event** is a subset of the sample space

Probability of an Event

- Remember, an **event** is a subset of the sample space
- The **probability** of event E in sample space S is $|E|/|S|$

Probability of an Event

- Remember, an **event** is a subset of the sample space
- The **probability** of event *E* in sample space *S* is $|E|/|S|$
- Example: Flip a coin 3 times. Compute P(even number of H)

Probability of an Event

- Remember, an **event** is a subset of the sample space
- The **probability** of event E in sample space S is $^{|E|}/_{|S|}$
- Example: Flip a coin 3 times. Compute P(even number of H)
    - "Even Number of H" = {THH, HTH, HHT, TTT} ➜ 4

Probability of an Event

- Remember, an **event** is a subset of the sample space
- The **probability** of event *E* in sample space *S* is $|E|/|S|$
- Example: Flip a coin 3 times. Compute P(even number of H)
    - "Even Number of H" = {THH, HTH, HHT, TTT} ➔ 4
    - Sample Space = {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH} ➔ 8

Probability of an Event

- Remember, an **event** is a subset of the sample space
- The **probability** of event *E* in sample space *S* is $|E|/|S|$
- Example: Flip a coin 3 times. Compute P(even number of H)
  - "Even Number of H" = {THH, HTH, HHT, TTT} ➔ 4
  - Sample Space = {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH} ➔ 8
  - P(even number of H) = 4/8 = 0.5

Probability of an Event

- Remember, an **event** is a subset of the sample space
- The **probability** of event E in sample space S is $|E|/|S|$
- Example: What is the probability of S? H)
  - “Even Number of H” = {THH, HTH, HHT, TTT} ➔ 4
  - Sample Space = {TTT, HTT, THT, TTH, THH, HTH, HHT, HHH} ➔ 8
  - P(even number of H) = 4/8 = 0.5

Estimating the Probability of an Event

- The true probability of an event is a theoretical property

Estimating the Probability of an Event

- The true probability of an event is a theoretical property
    - It is not always directly visible to us

Estimating the Probability of an Event

- The true probability of an event is a theoretical property
    - It is not always directly visible to us
- We can try to *estimate* the probability of an event from observations

Estimating the Probability of an Event

- The true probability of an event is a theoretical property
  - It is not always directly visible to us
- We can try to *estimate* the probability of an event from observations
  - Repeat the experiment as many times as possible

Estimating the Probability of an Event

- The true probability of an event is a theoretical property
    - It is not always directly visible to us
- We can try to *estimate* the probability of an event from observations
    - Repeat the experiment as many times as possible
    - Count the number of times the event occurred

Estimating the Probability of an Event

- The true probability of an event is a theoretical property
    - It is not always directly visible to us
- We can try to *estimate* the probability of an event from observations
    - Repeat the experiment as many times as possible
    - Count the number of times the event occurred
    - Divide that by the total number of trials

Classifying Text as Positive or Negative

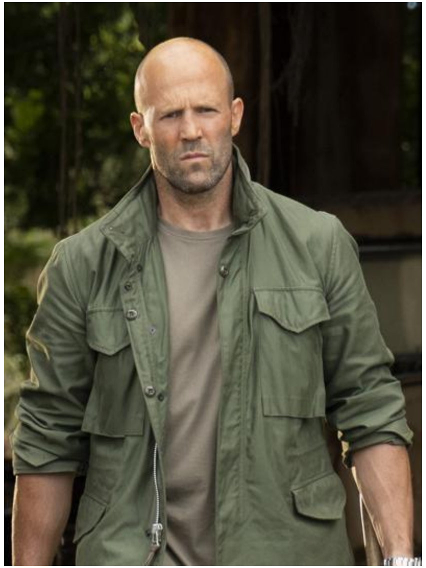
TOP BOX OFFICE Get Tickets

79%	Good Boys	$21.4M
99%	Fast & Furious Presents: Hobb...	$14.3M
52%	The Lion King	$12.3M
76%	The Angry Birds Movie 2	$10.4M
81%	Scary Stories to Tell in the Dark	$10.1M
82%	Dora and the Lost City of Gold	$8.7M
51%	47 Meters Down: Uncaged	$8.4M
85%	Once Upon a Time In Hollywood	$7.8M
43%	The Art of Racing in the Rain	$4.7M
90%	Blinded by the Light	$4.3M

Classifying Text as Positive or Negative

TOP BOX OFFICE Get Tickets

79%	Good Boys	$21.4M
99%	Fast & Furious Presents: Hobb...	$14.3M
52%	The Lion King	$12.3M
76%	The Angry Birds Movie 2	$10.4M
81%	Scary Stories to Tell in the Dark	$10.1M
82%	Dora and the Lost City of Gold	$8.7M
51%	47 Meters Down: Uncaged	$8.4M
85%	Once Upon a Time In Hollywood	$7.8M
43%	The Art of Racing in the Rain	$4.7M
90%	Blinded by the Light	$4.3M



Classifying Text as Positive or Negative

- Given a review x, can we classify it as “positive” or “negative”?

Classifying Text as Positive or Negative

- Given a review *x*, can we classify it as “positive” or “negative”?
    - We will select the classification that has higher probability

Classifying Text as Positive or Negative

- Given a review *x*, can we classify it as “positive” or “negative”?
  - We will select the classification that has higher probability
- **Prior Probability:** Our preliminary belief about the probability of an event *prior* to collecting any additional data

Example: Prior Probability

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

What is the probability that a randomly-selected student got an A?

Joint Probability

- **Joint Probability:** The probability that 2 (or more) events all occur

Of 100 students completing a course, 20 were business majors. 10 students received A’s in the course, and 3 of these were business majors.

What is the probability that a randomly-selected student is a business student who got an A?

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A		
Grade Not A		

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and **3 of these were business majors**.

	Business Major	Not Business Major
Grade A	3	
Grade Not A		

Joint Probability Table

Of 100 students completing a course, 20 were business majors. **10 students received A's in the course**, and 3 of these were business majors.

	Business Major	Not Business Major	
Grade A	3		10
Grade Not A			

Joint Probability Table

Of 100 students completing a course, 20 were business majors. **10 students received A's in the course**, and 3 of these were business majors.

	Business Major	Not Business Major	
Grade A	3	10 - 3 = 7	10
Grade Not A			

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3	7
Grade Not A		

Joint Probability Table

Of 100 students completing a course, **20 were business majors**. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3	7
Grade Not A		
	20	

Joint Probability Table

Of 100 students completing a course, **20 were business majors**. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3	7
Grade Not A	20 - 3 = 17	
	20	

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3	7
Grade Not A	17	

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3	7
Grade Not A	17	

100

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3	7
Grade Not A	17	100 - 3 - 7 - 17 = 73

100

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3	7
Grade Not A	17	73

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

Are we done?

	Business Major	Not Business Major
Grade A	3	7
Grade Not A	17	73

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

No! Probabilities sum to 1!

Are we done?

	Business Major	Not Business Major
Grade A	3	7
Grade Not A	17	73

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	3 / 100 = **0.03**	7 / 100 = **0.07**
Grade Not A	17 / 100 = **0.17**	73 / 100 = **0.73**

Joint Probability Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

See any problems?

	Business Major	Not Business Major
Grade A	3 / 100 = **0.03**	7 / 100 = **0.07**
Grade Not A	17 / 100 = **0.17**	73 / 100 = **0.73**

Joint ~~Probability~~ Frequency Table

Of 100 students completing a course, 20 were business majors. 10 students rece ... ness majors.

These are frequencies!!

See any problems?

	Business Major	Not Business Major
Grade A	3 / 100 = **0.03**	7 / 100 = **0.07**
Grade Not A	17 / 100 = **0.17**	73 / 100 = **0.73**

Joint ~~Probability~~ Frequency Table

Of 100 studen … rs. 10

students rece … ness majors.

But nobody ever cares...😢

These are frequencies!!

See any problems?

	Business Major	Not Business Major
Grade A	3 / 100 = **0.03**	7 / 100 = **0.07**
Grade Not A	17 / 100 = **0.17**	73 / 100 = **0.73**

Joint Frequency Table

Of 100 students completing a course, 20 were business majors. 10 students received A's in the course, and 3 of these were business majors.

	Business Major	Not Business Major
Grade A	0.03	0.07
Grade Not A	0.17	0.73

Conditional Probability

- The probability of event *A* occurring given that event *B* occurred

	Business Major	Not Business Major
Grade A	0.03	0.07
Grade Not A	0.17	0.73

Conditional Probability

$P(A|B) = P(A,B) / P(B)$

- The probability of event *A* occurring given that event *B* occurred

	Business Major	**Not Business Major**
Grade A	0.03	0.07
Grade Not A	0.17	0.73

Conditional Probability

$P(A|B) = P(A,B) / P(B)$

- The probability of event *A* occurring given that event *B* occurred
- What's the probability that a random business major got an A?

	Business Major	Not Business Major
Grade A	0.03	0.07
Grade Not A	0.17	0.73

Conditional Probability

$P(A|B) = P(A,B) / P(B)$

- The probability of event *A* occurring given that event *B* occurred
- What's the probability that a random business major got an A?

	Business Major	**Not Business Major**
Grade A	0.03	0.07
Grade Not A	0.17	0.73

P(Grade A | Business Major)

Conditional Probability

$P(A|B) = P(A,B) / P(B)$

- The probability of event *A* occurring given that event *B* occurred
- What's the probability that a random business major got an A?

	Business Major	Not Business Major
Grade A	0.03	0.07
Grade Not A	0.17	0.73

P(Grade A | Business Major) = P(Grade A, Business Major) / P(Business Major)

Conditional Probability

$$P(A|B) = P(A,B) / P(B)$$

- The probability of event *A* occurring given that event *B* occurred
- What's the probability that a random business major got an A?

	Business Major	**Not Business Major**
Grade A	0.03	0.07
Grade Not A	0.17	0.73

P(Grade A | Business Major) = **P(Grade A, Business Major)** / P(Business Major)

= **0.03**

Conditional Probability

$P(A|B) = P(A,B) / P(B)$

- The probability of event *A* occurring given that event *B* occurred
- What's the probability that a random business major got an A?

	Business Major	**Not Business Major**
Grade A	0.03	0.07
Grade Not A	0.17	0.73

P(Grade A | Business Major) = P(Grade A, Business Major) / **P(Business Major)**

= 0.03 / (**0.03 + 0.17**)

Conditional Probability

$$P(A|B) = P(A,B) / P(B)$$

- The probability of event *A* occurring given that event *B* occurred
- What's the probability that a random business major got an A?

	Business Major	Not Business Major
Grade A	0.03	0.07
Grade Not A	0.17	0.73

P(Grade A | Business Major) = P(Grade A, Business Major) / P(Business Major)

= 0.03 / (0.03 + 0.17) = **0.15**

Inference

- We can often measure *some* information
  - The Netflix watcher rates what they’ve already seen
- However, we want to make **inferences** about things we *haven’t* seen
  - Given that you liked movies *X* and *Y*, would you like movie *Z*?

Bayes' Theorem

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

Example: Bayes' Theorem

- A breast cancer diagnostic test outputs YES or NO, but it's not perfect

Example: Bayes' Theorem

- A breast cancer diagnostic test outputs YES or NO, but it's not perfect
  - Sensitivity = P(YES | cancer) = 93%

Example: Bayes' Theorem

- A breast cancer diagnostic test outputs YES or NO, but it's not perfect
  - Sensitivity = P(YES | cancer) = 93%
  - Specificity = P(NO | no cancer) = 99%

Example: Bayes' Theorem

- A breast cancer diagnostic test outputs YES or NO, but it's not perfect
  - Sensitivity = P(YES | cancer) = 93%
  - Specificity = P(NO | no cancer) = 99%
- On average, 0.148% of the population has breast cancer

Example: Bayes' Theorem

- A breast cancer diagnostic test outputs YES or NO, but it's not perfect
    - Sensitivity = P(YES | cancer) = 93%
    - Specificity = P(NO | no cancer) = 99%
- On average, 0.148% of the population has breast cancer
- What is P(cancer | YES)?

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	No Cancer
Test YES		
Test NO		

Example: Bayes' Theorem

- **P(YES | cancer) = 93%** and P(NO | no cancer) = 99%
- **P(cancer) = 0.148%**, so P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	**No Cancer**
Test YES		
Test NO		

P(YES | cancer) = P(YES, cancer) / P(cancer)

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	No Cancer
Test YES	0.0013764	
Test NO		

0.93 = P(YES, cancer) / 0.00148

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   **P(NO | no cancer) = 99%**
- **P(cancer) = 0.148%**,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	No Cancer
Test YES	0.0013764	
Test NO		

P(NO | no cancer) = P(NO, no cancer) / P(no cancer)

Example: Bayes' Theorem

- P(YES | cancer) = 93% and P(NO | no cancer) = 99%
- P(cancer) = 0.148%, so P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	**No Cancer**
Test YES	0.0013764	
Test NO		**0.9885348**

0.99 = P(NO, no cancer) / (1 - 0.00148)

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- **P(cancer) = 0.148%**,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	**No Cancer**
Test YES	0.0013764	
Test NO		0.9885348

0.00148

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- **P(cancer) = 0.148%**,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	**No Cancer**
Test YES	0.0013764	
Test NO	**0.0001036**	0.9885348
	0.00148	

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   **P(no cancer)** = 1 - 0.148% = **99.852%**

	Cancer	**No Cancer**
Test YES	0.0013764	
Test NO	0.0001036	0.9885348
		0.99852

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   **P(no cancer)** = 1 - 0.148% = **99.852%**

	Cancer	No Cancer
Test YES	0.0013764	**0.0099852**
Test NO	0.0001036	0.9885348
		0.99852

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	**No Cancer**
Test YES	0.0013764	0.0099852
Test NO	0.0001036	0.9885348

P(cancer | YES) = P(cancer, YES) / P(YES)

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	**No Cancer**
Test YES	**0.0013764**	0.0099852
Test NO	0.0001036	0.9885348

P(cancer | YES) = 0.0013764 / P(YES)

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	No Cancer
Test YES	**0.0013764** +	**0.0099852**
Test NO	0.0001036	0.9885348

P(cancer | YES) = 0.0013764 / 0.0113616

Example: Bayes' Theorem

- P(YES | cancer) = 93%   and   P(NO | no cancer) = 99%
- P(cancer) = 0.148%,   so   P(no cancer) = 1 - 0.148% = 99.852%

	Cancer	No Cancer
Test YES	0.0013764	+ 0.0099852
Test NO	0.0001036	0.9885348

12.1%

P(cancer | YES) = 0.0013764 / 0.0113616

Bayes' Theorem for Review Classification

- Given a review, we want to classify it as "positive" (+) or "negative" (-)

Bayes' Theorem for Review Classification

- Given a review, we want to classify it as "positive" (+) or "negative" (-)
  - Let *S* be a random variable denoting the sentiment

Bayes' Theorem for Review Classification

- Given a review, we want to classify it as "positive" (+) or "negative" (-)
  - Let S be a random variable denoting the sentiment
  - Let R be a random variable denoting the review (a string of text)

Bayes' Theorem for Review Classification

- Given a review, we want to classify it as "positive" (+) or "negative" (-)
  - Let S be a random variable denoting the sentiment
  - Let R be a random variable denoting the review (a string of text)
  - Let x denote a specific given review

Bayes' Theorem for Review Classification

- Given a review, we want to classify it as "positive" (+) or "negative" (-)
    - Let S be a random variable denoting the sentiment
    - Let R be a random variable denoting the review (a string of text)
    - Let x denote a specific given review
    - **What is $P(S = + \mid R = x)$?**

Bayes' Theorem for Review Classification

$$P(S = + | R = x) = \frac{P(R = x | S = +) P(S = +)}{P(R = x)}$$

Bayes' Theorem for Review Classification

$$P(S=+|R=x)=\frac{P(R=x|S=+)P(S=+)}{P(R=x)}$$

Simplifying Assumption #1: “Bag of Words” Model

- How can we compute (or estimate) $P(R = x \mid S = +)$?

Simplifying Assumption #1: “Bag of Words” Model

- How can we compute (or estimate) P($R = x \mid S = +$)?
- **Bag of Words Model:** Represent text (our review *x*) as a “bag” (collection) of words, disregarding grammar and word order, but keeping word multiplicity

Simplifying Assumption #1: “Bag of Words” Model

- How can we compute (or estimate) P(R = x | S = +)?
- **Bag of Words Model:** Represent text (our review x) as a “bag” (collection) of words, disregarding grammar and word order, but keeping word multiplicity

Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

=

Hi! is is is My My My
name name name Slim
Shady (what?) (who?)

Simplifying Assumption #1: “Bag of Words” Model

- To simplify things even further, we won’t care about multiplicity

Simplifying Assumption #1: "Bag of Words" Model

- To simplify things even further, we won't care about multiplicity
    - Let $W = [w_1, w_2, \ldots, w_n]$ denote the set of all n possible words

Simplifying Assumption #1: “Bag of Words” Model

- To simplify things even further, we won’t care about multiplicity
  - Let $W = [w_1, w_2, ..., w_n]$ denote the set of all n possible words
  - Let $E = [e_1, e_2, ..., e_n]$ denote the “existence” of each word in x

Simplifying Assumption #1: "Bag of Words" Model

- To simplify things even further, we won't care about multiplicity
  - Let $W = [w_1, w_2, \ldots, w_n]$ denote the set of all n possible words
  - Let $E = [e_1, e_2, \ldots, e_n]$ denote the "existence" of each word in x
    - Specifically, e_i = True if word w_i exists in x, otherwise e_i = False

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [
```

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False,
```

Simplifying Assumption #1: "Bag of Words" Model

Hi! My name is (what?)
$x =$ My name is (who?)
My name is Slim Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,
```

Simplifying Assumption #1: “Bag of Words” Model

Hi! My name **is** (what?)
$x =$ My name **is** (who?)
My name **is** Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True,  True,
```

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! **My** name is (what?)
My name is (who?)
My name is Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True,  True, True,
```

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! My **name** is (what?)
My **name** is (who?)
My **name** is Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True,  True, True, True,
```

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True, True, True, True, **False**,

Simplifying Assumption #1: "Bag of Words" Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is **Slim** Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,  True, True, True,   False,   True,
```

Simplifying Assumption #1: “Bag of Words” Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim **Shady**

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True,  True, True, True,   False,   True,   True,
```

Simplifying Assumption #1: "Bag of Words" Model

$x =$ Hi! My name is **(what?)**
My name is (who?)
My name is Slim Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,  True, True, True,   False,   True,   True,    True,
```

Simplifying Assumption #1: "Bag of Words" Model

$x =$ Hi! My name is (what?)
My name is **(who?)**
My name is Slim Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,  True, True, True,   False,   True,   True,    True,      True]
```

Simplifying Assumption #1: “Bag of Words” Model

$\boldsymbol{x}$ = Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True,  True, True, True,   False,   True,   True,    True,       True]
```

Thus, $P(R = \boldsymbol{x} \mid S = +)$

Simplifying Assumption #1: “Bag of Words” Model

$\boldsymbol{x}$ = Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True,  True, True, True,   False,   True,    True,     True,       True]
```

Thus, $P(R = \boldsymbol{x} \mid S = +) = P(\boldsymbol{E} \mid S = +)$

Simplifying Assumption #1: "Bag of Words" Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,  True, True, True,   False,   True,   True,    True,       True]
```

Thus, $P(R = x \mid S = +) = P(E \mid S = +) = P($**$e_1$ = False**,

Simplifying Assumption #1: "Bag of Words" Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,  True, True, True,   False,   True,   True,    True,       True]
```

Thus, $P(R = x \mid S = +) = P(E \mid S = +) = P(e_1 = \text{False}, \mathbf{e_2 = True},$

Simplifying Assumption #1: “Bag of Words” Model

x = Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = [“Dre”, “Hi!”, “is”, “My”, “name”, “Niema”, “Slim”, “Shady”, “(what?)”, “(who?)”]

E = [False, True,  True, True, True,   False,   True,   True,    True,      True]
```

Thus, $P(R = x \mid S = +) = P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots$

Simplifying Assumption #1: "Bag of Words" Model

$x =$ Hi! My name is (what?)
My name is (who?)
My name is Slim Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,  True, True, True,   False,   True,   True,    True,       True]
```

Thus, $P(R = x \mid S = +) = P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots \mid S = +)$

Simplifying Assumption #1: "Bag of Words" Model

Hi! My name is (what?)

But how can we learn P(E | S = +)???

My name is Slim Shady

```
W = ["Dre", "Hi!", "is", "My", "name", "Niema", "Slim", "Shady", "(what?)", "(who?)"]

E = [False, True,  True, True, True,   False,   True,   True,    True,      True]
```

Thus, $P(R = x \mid S = +) = P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots \mid S = +)$

Independence

- Two events A and B are **independent** if the outcome of A has **no effect** on the outcome of B, and vice-versa

Independence

- Two events *A* and *B* are **independent** if the outcome of *A* has **no effect** on the outcome of *B*, and vice-versa
    - ***A*** **= Number of boba drinks Niema buys in a week**

Independence

- Two events *A* and *B* are **independent** if the outcome of *A* has **no effect** on the outcome of *B*, and vice-versa
    - ***A* = Number of boba drinks Niema buys in a week**
    - ***B* = Price of a Shake Shack burger**

Independence

- Two events *A* and *B* are **independent** if the outcome of *A* has **no effect** on the outcome of *B*, and vice-versa
    - ***A*** **= Number of boba drinks Niema buys in a week**
    - ***B*** **= Price of a Shake Shack burger**
        - Shake Shack is trash, so it doesn't affect Niema (**independent**)

Independence

- Two events *A* and *B* are **independent** if the outcome of *A* has **no effect** on the outcome of *B*, and vice-versa
  - ***A* = Number of boba drinks Niema buys in a week**
  - *B* = Price of a Shake Shack burger
    - Shake Shack is trash, so it doesn't affect Niema (independent)
  - ***C* = Number of lectures Niema has to give during a week of SPIS**

Independence

- Two events *A* and *B* are **independent** if the outcome of *A* has **no effect** on the outcome of *B*, and vice-versa
    - ***A* = Number of boba drinks Niema buys in a week**
    - *B* = Price of a Shake Shack burger
        - Shake Shack is trash, so it doesn't affect Niema (independent)
    - ***C* = Number of lectures Niema has to give during a week of SPIS**
        - Stressed Niema likes boba, so *A* and *C* are **dependent**

Independence

- Two events A and B are **independent** ~~if the outcome of A has no effect on the outcome of B, and vice-versa~~ **iff $P(A,B) = P(A) \times P(B)$**
  - A = Number of boba drinks Niema buys in a week
  - B = Price of a Shake Shack burger
    - Shake Shack is trash, so it doesn't affect Niema (independent)
  - C = Number of lectures Niema has to give during a week of SPIS
    - Stressed Niema likes boba, so A and C are dependent

Inde

$P(A|B) = P(A)$ and $P(B|A) = P(B)$

- Two events *A* and *B* are **independent** ~~if the outcome of *A* has no effect on the outcome of *B*, and vice-versa~~ **iff $P(A,B) = P(A) \times P(B)$**
  - *A* = Number of boba drinks Niema buys in a week
  - *B* = Price of a Shake Shack burger
    - Shake Shack is trash, so it doesn't affect Niema (independent)
  - *C* = Number of lectures Niema has to give during a week of SPIS
    - Stressed Niema likes boba, so *A* and *C* are dependent

Conditional Independence

- Let A and B be two **dependent** events

Conditional Independence

- Let *A* and *B* be two **dependent** events
  - *A* = Number of boba drinks Niema buys in a week

Conditional Independence

- Let *A* and *B* be two **dependent** events
  - *A* = Number of boba drinks Niema buys in a week
  - *B* = Week number of SPIS

Conditional Independence

- Let A and B be two **dependent** events
  - A = Number of boba drinks Niema buys in a week
  - B = Week number of SPIS
- Let C be a third event

Conditional Independence

- Let A and B be two **dependent** events
  - A = Number of boba drinks Niema buys in a week
  - B = Week number of SPIS
- Let C be a third event
  - C = Number of lectures Niema has to give during a week of SPIS

Conditional Independence

- Let *A* and *B* be two **dependent** events
  - *A* = Number of boba drinks Niema buys in a week
  - *B* = Week number of SPIS
- Let *C* be a third event
  - *C* = Number of lectures Niema has to give during a week of SPIS
- *A* and *B* are **conditionally independent** given *C* iff $\mathbf{P(A \mid B,C) = P(A \mid C)}$

Simplifying Assumption #2: Conditional Independence

- We wanted to estimate $P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots \mid S = +)$

Simplifying Assumption #2: Conditional Independence

- We wanted to estimate $P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots \mid S = +)$
  - Given that the sentiment is positive, e_1 = False, e_2 = True, etc. are all **conditionally independent**

Simplifying Assumption #2: Conditional Independence

- We wanted to estimate $P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots \mid S = +)$
  - Given that the sentiment is positive, $e_1 = \text{False}$, $e_2 = \text{True}$, etc. are all **conditionally independent**
- **$P(E \mid S = +) = P(e_1 = \text{False} \mid S = +) \times P(e_2 = \text{True} \mid S = +) \times \ldots$**

Simplifying Assumption #2: Conditional Independence

- We wanted to estimate $P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots \mid S = +)$
    - Given that the sentiment is positive, e_1 = False, e_2 = True, etc. are

Yay! We can learn these from data!!!

- **$P(E \mid S = +) = P(e_1 = \text{False} \mid S = +) \times P(e_2 = \text{True} \mid S = +) \times \ldots$**

Simplifying Assumption #2: Conditional Independence

- We wanted to estimate $P(E \mid S = +) = P(e_1 = \text{False}, e_2 = \text{True}, \ldots \mid S = +)$
  - Given that the sentiment is positive, e_1 = False, e_2 = True, etc. are

Yay! We can learn these from data!!!

- **$P(E \mid S = +) = P(e_1 = \text{False} \mid S = +) \times P(e_2 = \text{True} \mid S = +) \times \ldots$**
  - Get a bunch of reviews, construct a vocabulary *W* of all unique words, and count the proportion of positive reviews containing w_i

Bayes' Theorem for Review Classification

$$P(S=+|R=x) = \frac{P(R=x|S=+)P(S=+)}{P(R=x)}$$

Bayes' Theorem for Review Classification

$$P(S = + | R = x) = \frac{\boxed{P(R = x | S = +)} P(S = +)}{P(R = x)}$$

Bayes' Theorem for Review Classification

$$P(S = + | R = x) = \frac{P(R = x | S = +) P(S = +)}{P(R = x)}$$

Bayes' Theorem for Review Classification

$$P(S=+|R=x) = \frac{P(R=x|S=+)P(S=+)}{P(R=x)}$$

This is just the proportion of reviews that were positive

Bayes' Theorem for Review Classification

$$P(S = +|R = x) = \frac{P(R = x|S = +)P(S = +)}{P(R = x)}$$

Bayes' Theorem for Review Classification

$$P(S = + | R = x) \propto P(R = x | S = +) P(S = +)$$

Don't care 😃😃😃