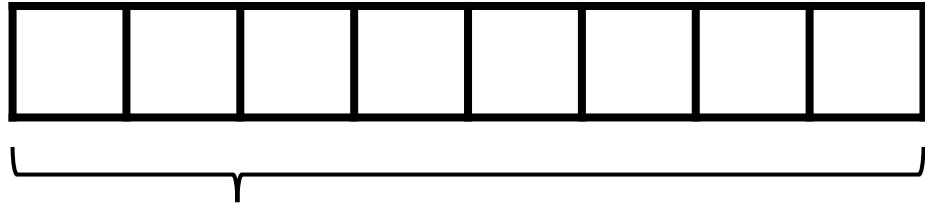


Bag of Words Models

MMCi Block 4

Matthew Engelhard

Lecture 1: what is a predictive model?



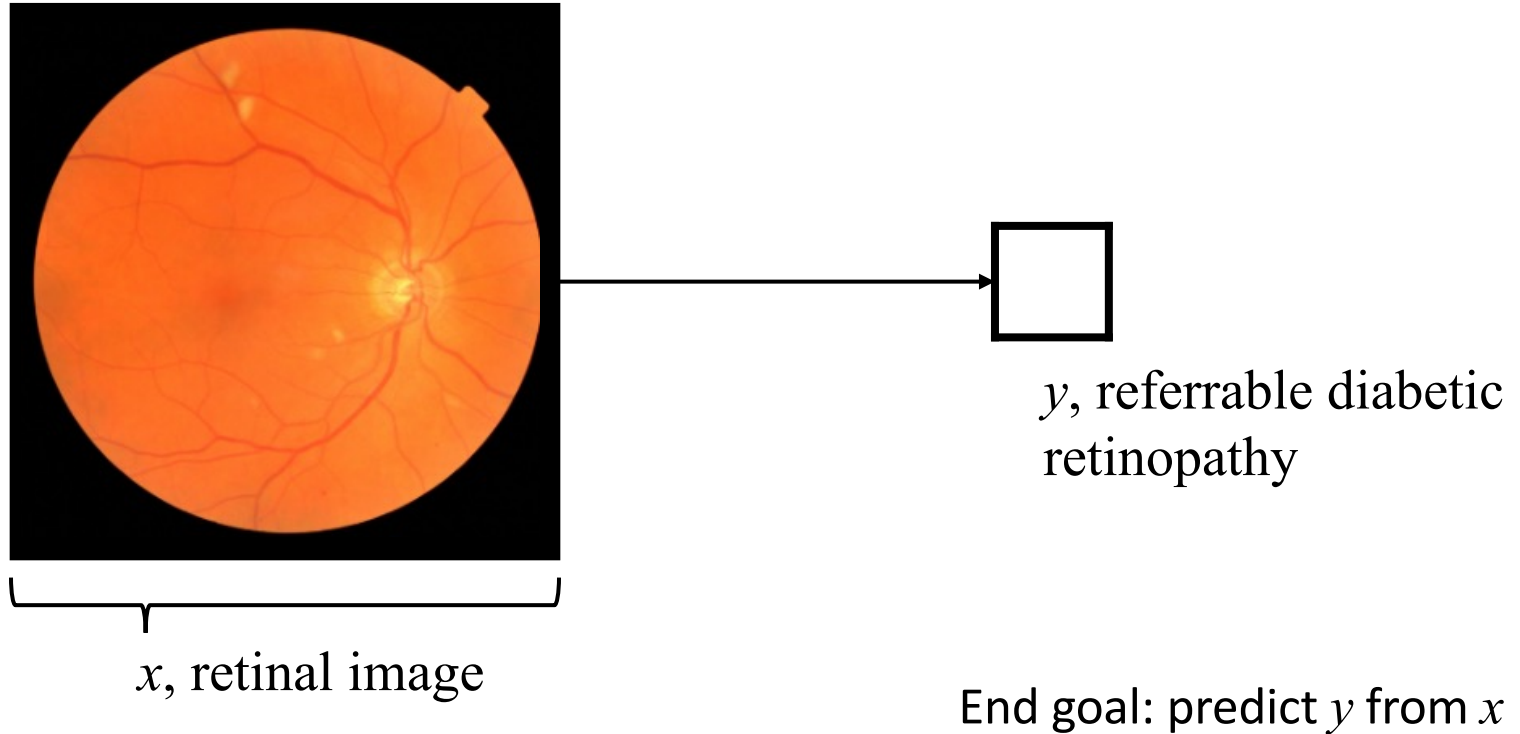
x , data/features for
a subject or patient



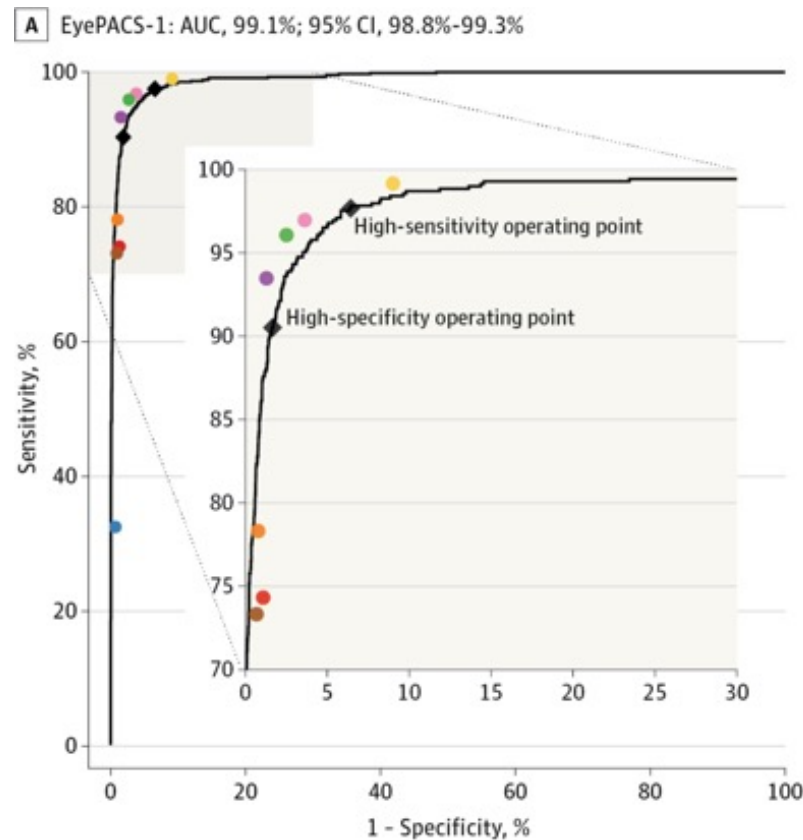
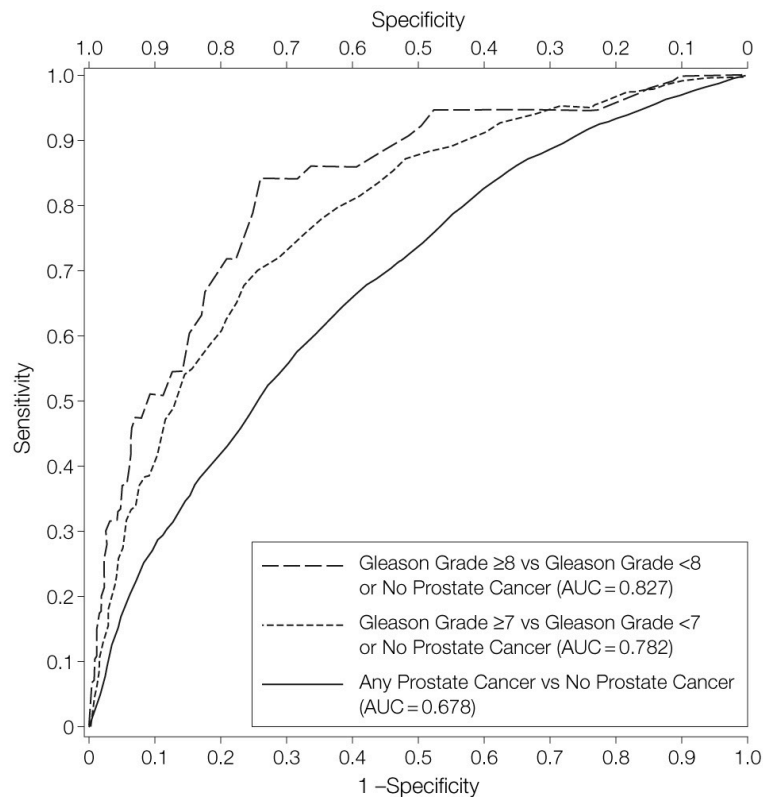
y , associated
value or label

End goal: predict y from x

Lecture 2: a predictive model for image data



Evaluate performance just like any other diagnostic tool



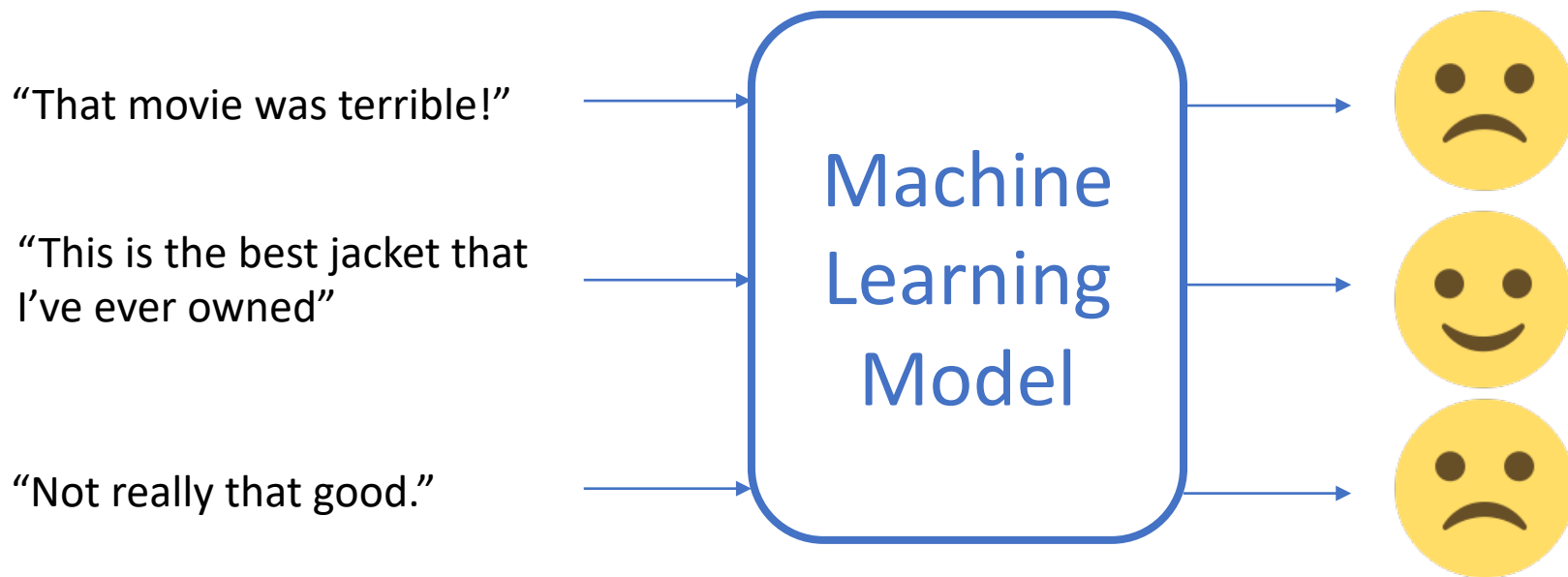
-> Brief review

Today: a predictive model for text data

- What can “natural language processing” (NLP) do?
 - Existing non-medical applications
 - Possible medical applications
- How can we convert text into something a predictive model can understand?
- Evaluating and understanding NLP models

Sentiment Analysis:

An “easy” binary classification problem



Text Translation

ENGLISH - DETECTED

ENGLISH

GERM



ENGLISH

SWEDISH

GERMAN



Deep learning is so much fun| ×



28/5000



Deep Learning macht so viel
Spaß



[Send feedback](#)

Question Answering

Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by pattern recognition receptors, which

recognize components that are conserved among microorganisms, or when damaged, injury signals, many of which (but not all) are related to those that recognize pathogens. Innate immunity, meaning these systems respond to pathogens, does not confer long-lasting immunity against a specific pathogen. Innate immunity is the dominant system of host defense in

What part of the innate immune system identifies microbes and triggers immune response?

Ground Truth Answers: pattern recognition receptors receptors cells

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance <i>Stanford University</i> (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) <i>Google AI Language</i> https://github.com/google-research/bert	86.673	89.147
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) <i>Google AI Language</i> https://github.com/google-research/bert	85.150	87.715

tors

dominant system of defense?

e system innate immune

m

Identify components present in broad

microorganisms

s in a generic way, meaning it is

non-specific non-specific

Automatic Image Captioning



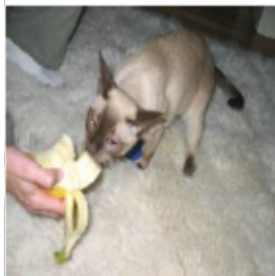
a cow is standing in front
of a store



a group of elephants
standing next to each other



a table that has wooden
spoons on it



a cat is eating some kind of
food



a bunch of bananas are
sitting on a table



a motorcycle is parked next
to a window

Populating Standardized Forms

MRC Prognostic Index

Has patient been seizure-free for 2 years ☒ Yes ☐ No ☐ Don't know
Yes taken 6 months ago

MRC Prognostic Index

Age 16 years or older ☒ Yes ☐ No
Yes taken 6 months ago

Taking more than one epileptic drug ☒ Yes ☐ No
Yes taken 6 months ago

Seizures after start of antiepileptic drug treatment ☒ Yes ☐ No
Yes taken 6 months ago

History of primary or secondary generalized tonic-clonic seizures ☒ Yes ☐ No
Yes taken 6 months ago

History of myoclonic seizures ☒ Yes ☐ No
Yes taken 6 months ago

Electroencephalogram in past year ☒ Normal ☐ Abnormal ☐ Not available
Abnormal taken 6 months ago

Seizure Free Years (minimum 2 years) Period free from seizures score
3 taken 6 months ago 66.67 (calculated) taken 6 months ago

Total score
128.67 (calculated) taken 6 months ago

Divide total score by 100 and exponentiate
3.55 (calculated) taken 6 months ago

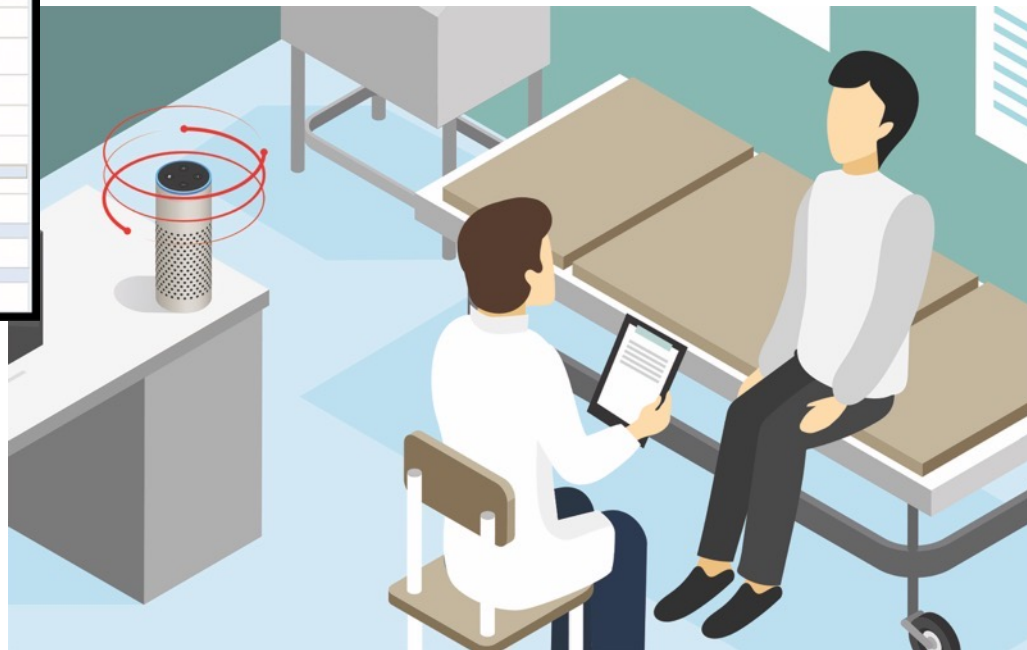
Percent probability of recurrence of seizure (over 1 year)

With continued treatment With slow withdrawal
34 % (calculated) taken 6 months ago 73 % (calculated) taken 6 months ago

Percent probability of recurrence of seizure (over 2 years)

With continued treatment With slow withdrawal
57 % (calculated) taken 6 months ago 84 % (calculated) taken 6 months ago

Narayanan et al,
Epilepsia (2017)



Text Generation

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

“Better Language Models and Their Implications”

2/14/19

OPENAI

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials. The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

Suggested Email Responses



Messaging



Health



Appts & Visits



Questionnaires

Message Center

[ASK A QUESTION](#)

Inbox Sent Messages

Search message list



Sort by:

Received Date



Filters:

All Messages





The Doctor

(Luke Fildes, 1891)

Inspired by MLHC
keynote by
Abraham Verghese,
MD, MACP, Stanford
University



The Doctor, circa 2018

Inspired by MLHC
keynote by
Abraham Verghese,
MD, MACP, Stanford
University

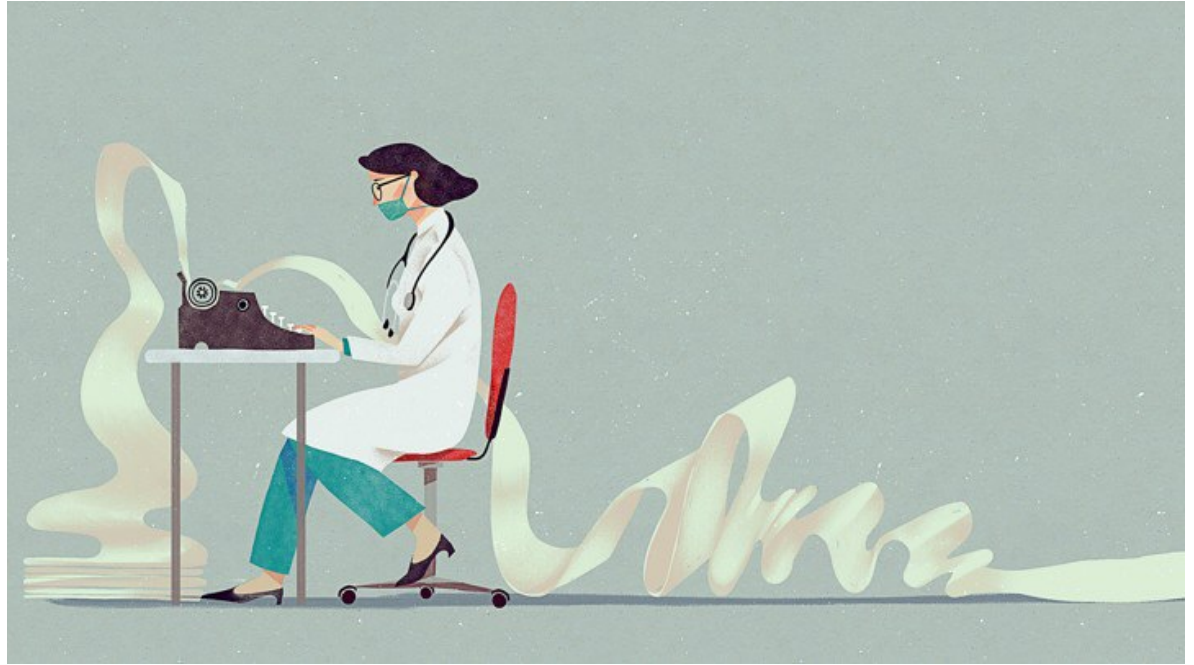
Reducing Burden and Restoring Patient-Provider Interaction

The Burnout Crisis in American Medicine

Rena Xu

The Atlantic

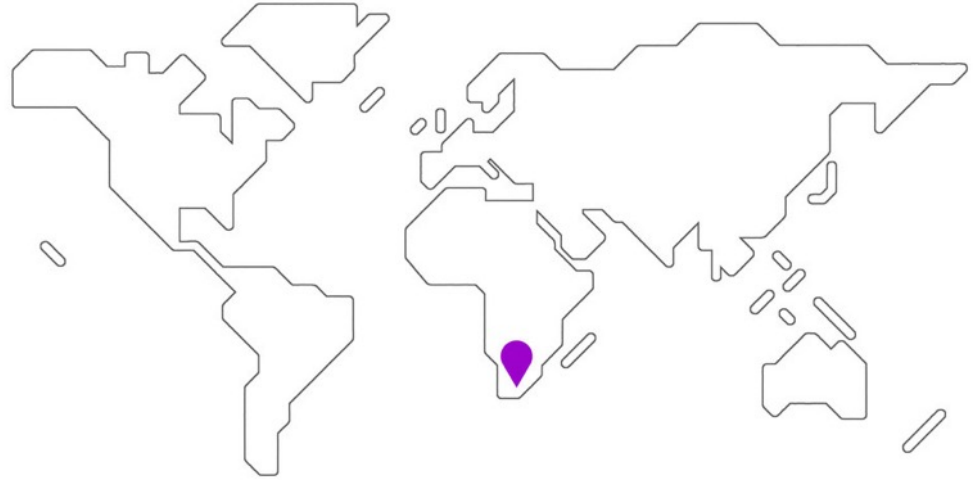
May 11, 2018



Case Study: SMS Triage for Global Maternal Health

Maternal Health HelpDesk:

**2 million women connected to
NDoH staff via SMS**



<https://www.praekelt.org>

Binary Classification: Urgent Message? (Yes/No)

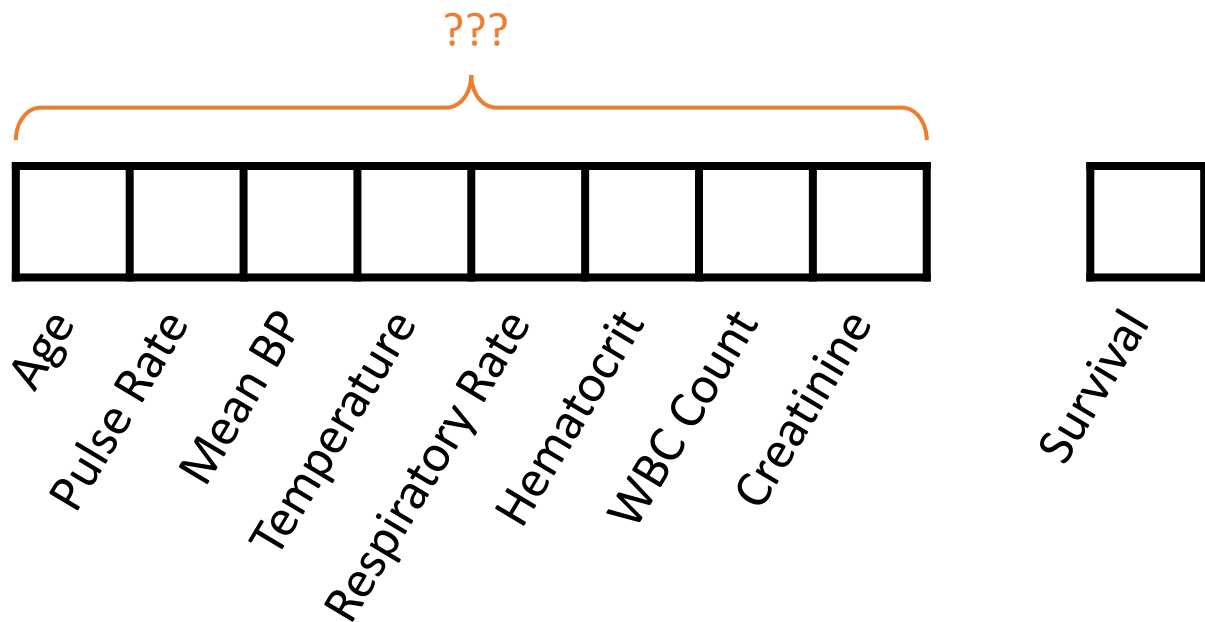
Case Study: SMS Triage for Global Maternal Health



<https://www.praekelt.org>

Can we use a standard predictive model
setup to solve this problem?

A Simple Predictive Model: ICU Mortality



End goal: predict odds of hospital mortality

Training Set (Historical Data)

x_1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_1
x_2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_2
x_3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_3
x_4	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_4
	\vdots									\vdots
x_{N-1}	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_{N-1}
x_N	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_N

Find an equation that predicts y based on x across the training set

Making Predictions for New x

x_1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_1
x_2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_2
x_3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_3
x_4	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_4
	\vdots										\vdots
x_{N-1}	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_{N-1}
x_N	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_N

Find an equation that predicts y based on x across the training set

x_{N+1}	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	y_{N+1}
-----------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	-----------

<- Learn to predict new y

This time, our training data is text

x_1 What helps with morning sickness? ☐ y_1

x_2 How many months should I breastfeed? ☐ y_2

x_3 I passed out and Mom said I was shaking ☐ y_3

x_4 Where is the nearest clinic? ☐ y_4

\vdots

\vdots

x_{N-1} I am having heavy bleeding, what should I do? ☐ y_{N-1}

x_N What foods should I eat while pregnant? ☐ y_N

y_i : Urgent or
Not Urgent?

x_{N+1} My heart is racing and I can't catch my breath ☐ y_{N+1}

<- Learn to predict new
 y

We need numbers, not words

- **Can we convert our text to a vector or sequence of numbers?**
- If yes, we can use logistic regression (or any other predictive model)!

First try: count words in each SMS

Step 1: Define a vocabulary of words

x_1

What helps with morning sickness?

x_2

How many months should I breastfeed?

x_3

I passed out and Mom said I was shaking

x_4

Where is the nearest clinic?

list of all words
(in no particular order)

shaking
what
clinic
how
helps
was
nearest
many

with
said
months
the
morning
mom
should
sickness

and
I
is
how
out
breastfeed
passed
where

Step 2: count how many times each vocabulary word appears in a given SMS

What helps with morning sickness?

x_I

0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

Step 2: count how many times each vocabulary word appears in a given SMS

I passed out and Mom said I was shaking

x_3

1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	2	0	0	1	0	1	0
shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

Step 2: count how many times each vocabulary word appears in a given SMS

Where is the nearest clinic?

x_4

0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1
shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

Note that word order does not matter!

clinic is where nearest the

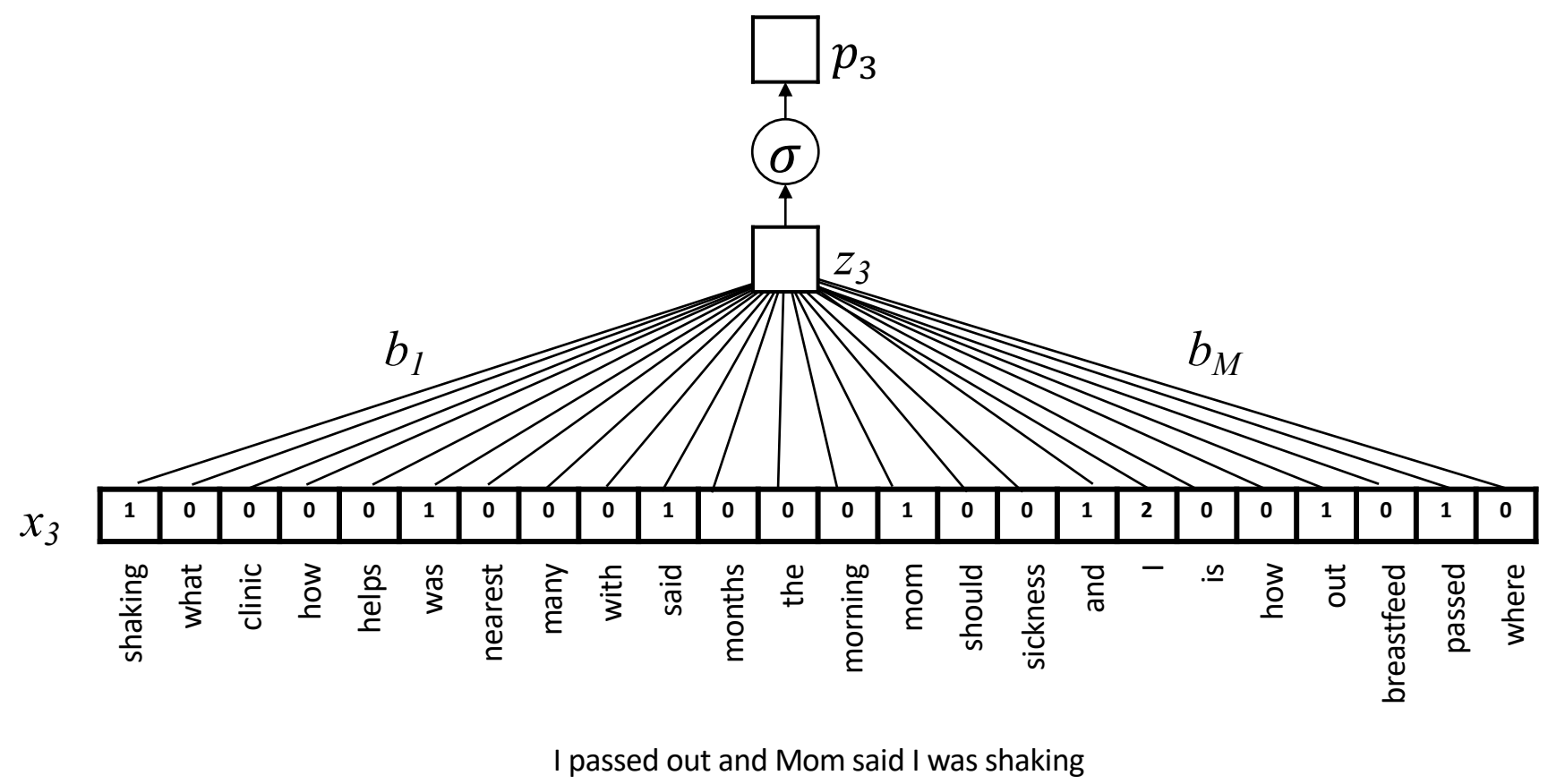
x_4

0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1
shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

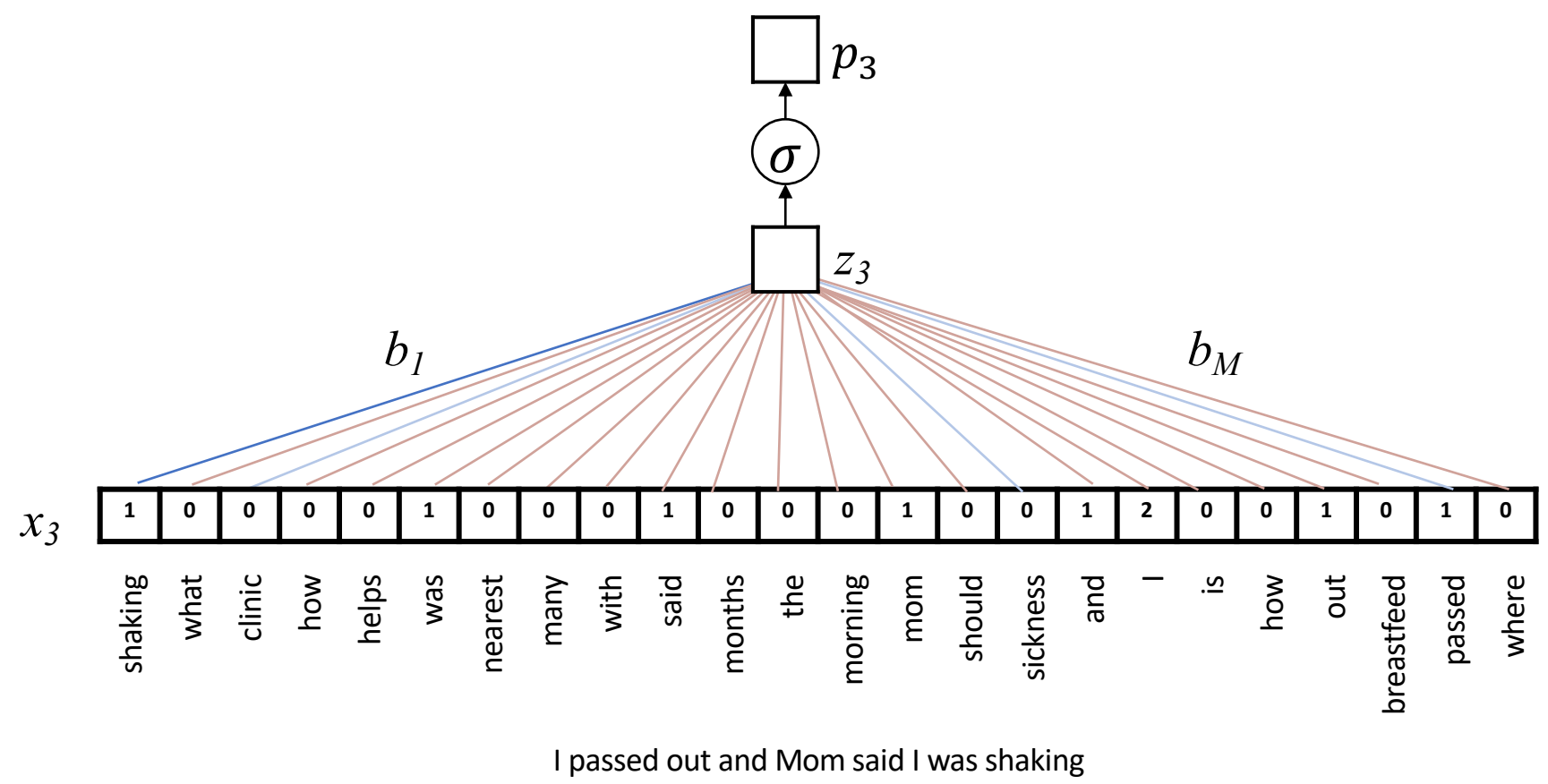
A “bag of words”



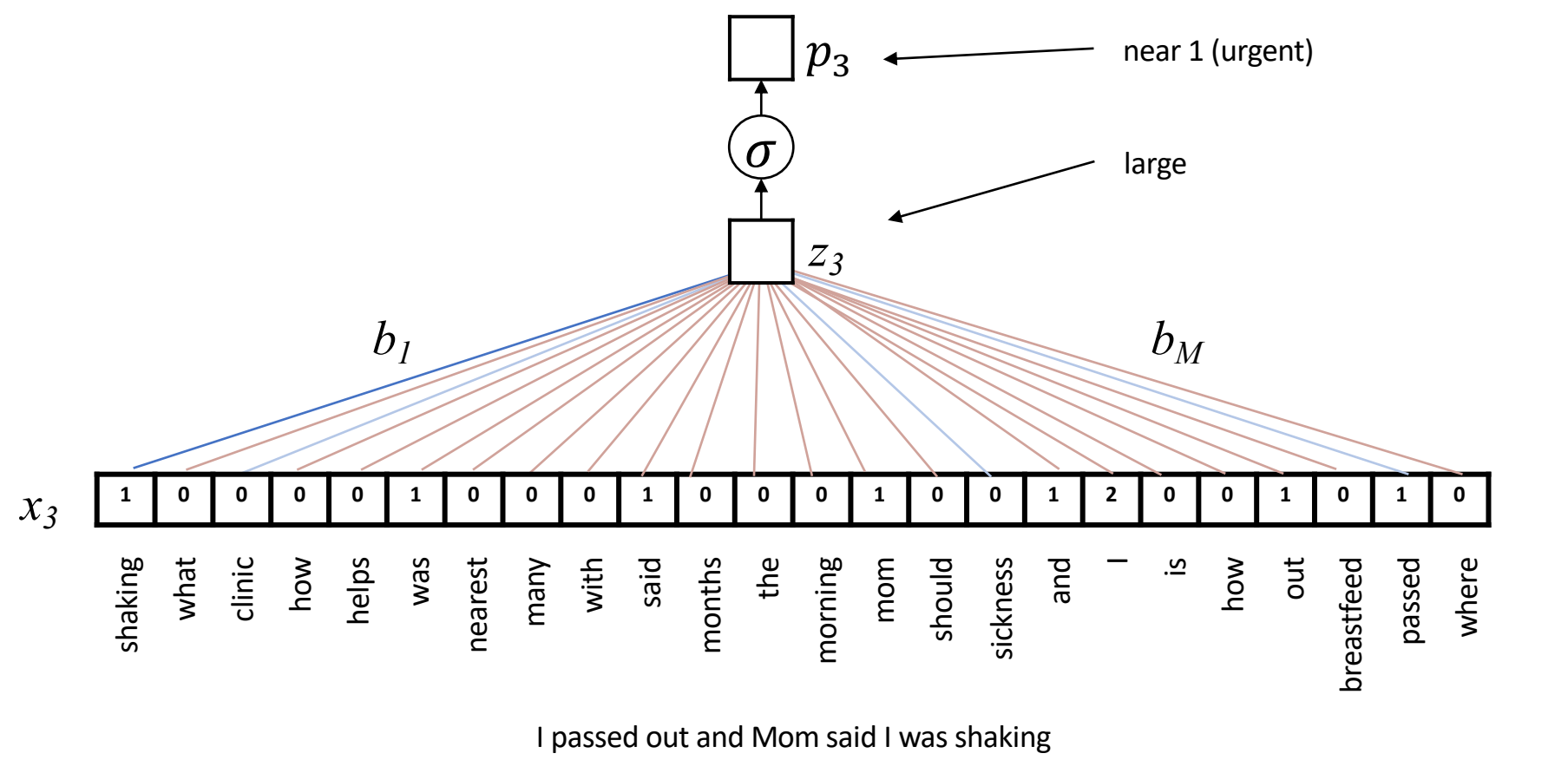
Logistic Regression for Text Classification



Logistic Regression for Text Classification



Logistic Regression for Text Classification



Strengths and Weaknesses

- (+) This approach is simple and works surprisingly well in practice
- (+) Often the best approach with small datasets
- (-) Does not capture word order
- (-) Does not group synonyms together or understand semantic relationships between words

2nd try: count 1- and 2-grams in each SMS
(i.e. extend vocabulary to include 2-word phrases)

1-grams

shaking	was	months	sickness	out
what	nearest	the	and	breastfeed
clinic	many	morning	I	passed
how	with	mom	is	where
helps	said	should	how	

x_1

What helps with morning sickness?

x_2

How many months should I breastfeed?

x_3

I passed out and Mom said I was shaking

x_4

Where is the nearest clinic?

2-grams

what helps
helps with
with morning
morning sickness
how many
many months
months should

should I
I breastfeed
I passed
passed out
out and
and mom
mom said

said I
I was
was shaking
where is
is the
the nearest
nearest clinic

n-grams can be very helpful!

I am not sick and feel great

I am not great and feel sick



Bag of 1-grams: no difference between these sentences

n-grams can be very helpful!

I am not sick and feel great

I am not great and feel sick

Bag of 1- and 2-grams:

not sick, feel great

versus

not great, feel sick

3rd try: more powerful methods to work with...

- (a) word meaning: assign words to vectors that encode their meaning numerically
- (b) words in context: neural network architectures that act on *sequences* of words (rather than a bag of words)

More Text Processing Details

(for bag of words models)

Variations on counting: term frequency

term count: 'times'

2

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only."

1

"And the first one now
Will later be last
For the times they are a-changin'."

Variations on counting: term frequency

term frequency: 'times'

2/119

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only."

1/16

"And the first one now
Will later be last
For the times they are a-changin'."

-> better measure of the importance of
the term within a given text sample

Variations on counting: inverse document frequency

2/2

document frequency: 'times'



"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only."



"And the first one now
Will later be last
For the times they are a-changin'."

Variations on counting: inverse document frequency

1/2

document frequency: 'evil'



"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only."



"And the first one now
Will later be last
For the times they are a-changin'."

term frequency-inverse document frequency (tf-idf)

- What helps with morning sickness?
- How many months should I breastfeed?
- I passed out and Mom said I was shaking
- Where is the nearest clinic?
- I am having heavy bleeding, what should I do?
- What foods should I eat while pregnant?
- My heart is racing and I can't catch my breath

$$\frac{\text{term frequency}}{\text{document frequency}} \quad \text{for 'shaking'}$$

$$\frac{1/9}{1/7} = .78$$

$$\frac{\text{term frequency}}{\text{document frequency}} \quad \text{for 'I'}$$

$$\frac{2/9}{5/7} = .31$$

Preprocessing

- remove punctuation

I passed out, and Mom said I was shaking.

- to lowercase

I passed out and Mom said I was shaking

- “tokenization”

i passed out and mom said i was shaking

- “stemming”

[i, passed, out, and, mom, said, i, was, shaking]

[i, pass, out, and, mom, said, i, wa, shake]