1. Modelling in NLP, Naive Bayes

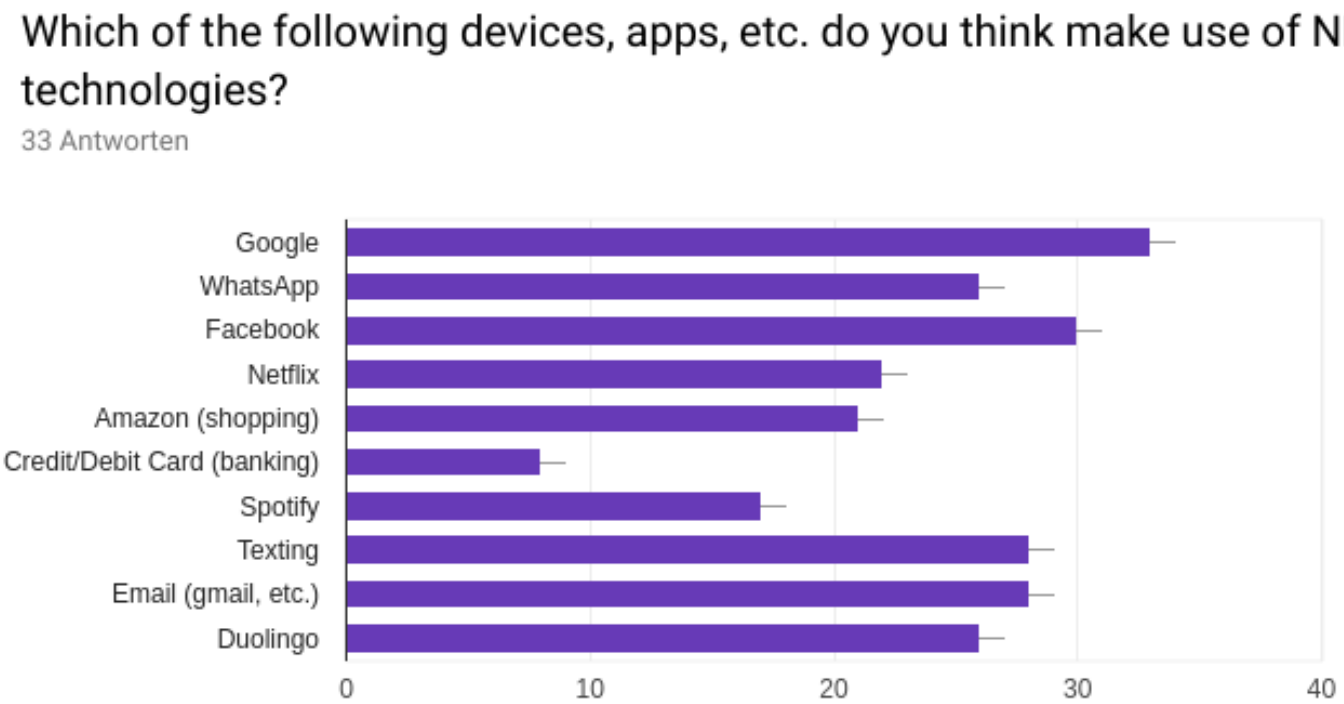
NLP for CogSci Research

Marlene Staib

September 18, 2018

Survey Results

NLP technologies



NLP tasks

- Sentiment analysis/classification
- Topic modelling
- Tokenization
- Information retrieval
- Machine Translation
- Speech Recognition
- Speech synthesis/Text-To-Speech
- (Chatbots)
- None

NLP interests

Tasks:

- Topic Modelling
- Sentiment Analysis
- Relation extraction
- Intent classification
- Information retrieval
- Machine Translation
- Speech Recognition

Applications:

- Determining mood/mental state
- Solving issues in society

Datasets:

- Social media
- Fake news

Other:

- Don't know yet
- Everything

Modelling in NLP

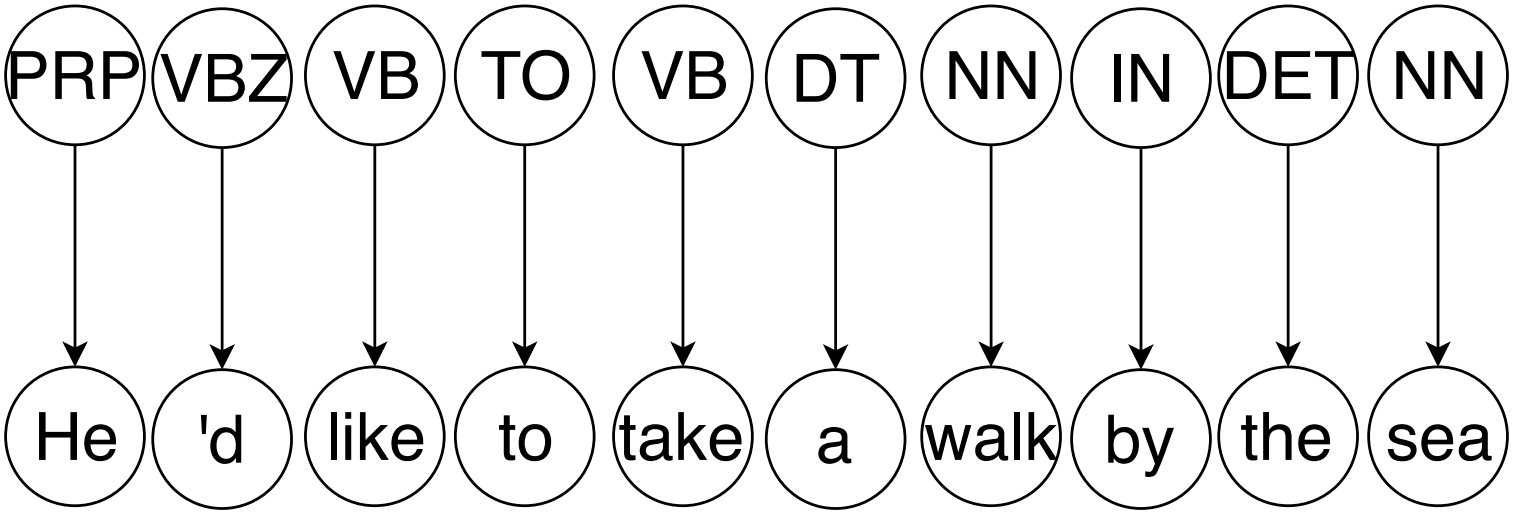
Modelling in NLP

NLP is a set of Machine Learning techniques for **Language Data**.

Language Data

- Sequences
- Ambiguity
- Sparsity: Zipf's Law

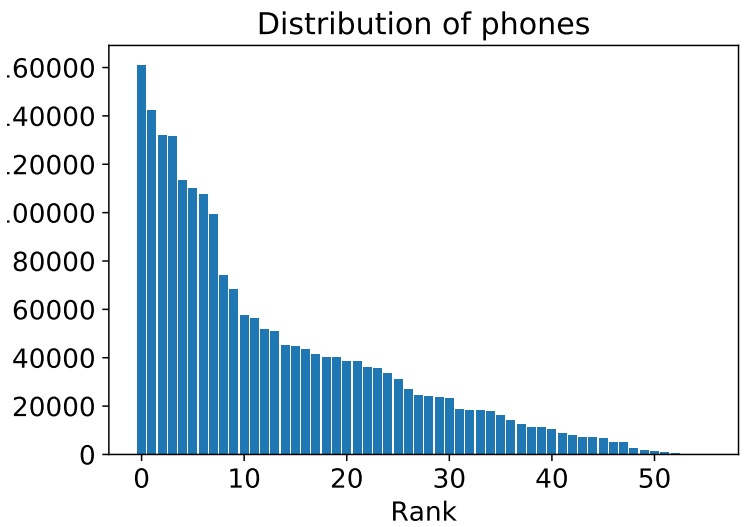
Sequences



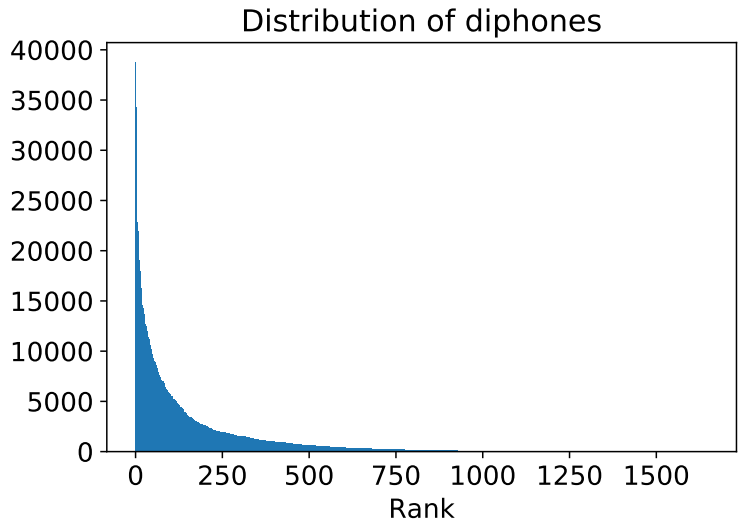
Ambiguity

- “I will meet him at the bank.”
- “I saw the man with the telescope.”
- “He made her duck.”
- “Time flies like an arrow.”
- “Call me maybe.”
- “The chicken is ready to eat.”

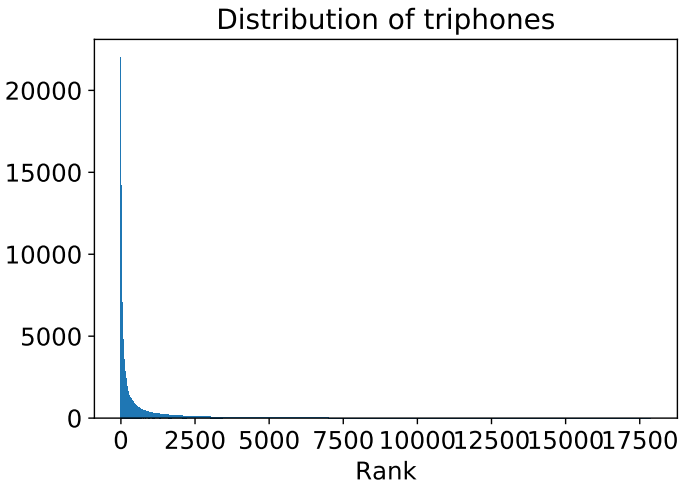
Zipf's Law



Phones

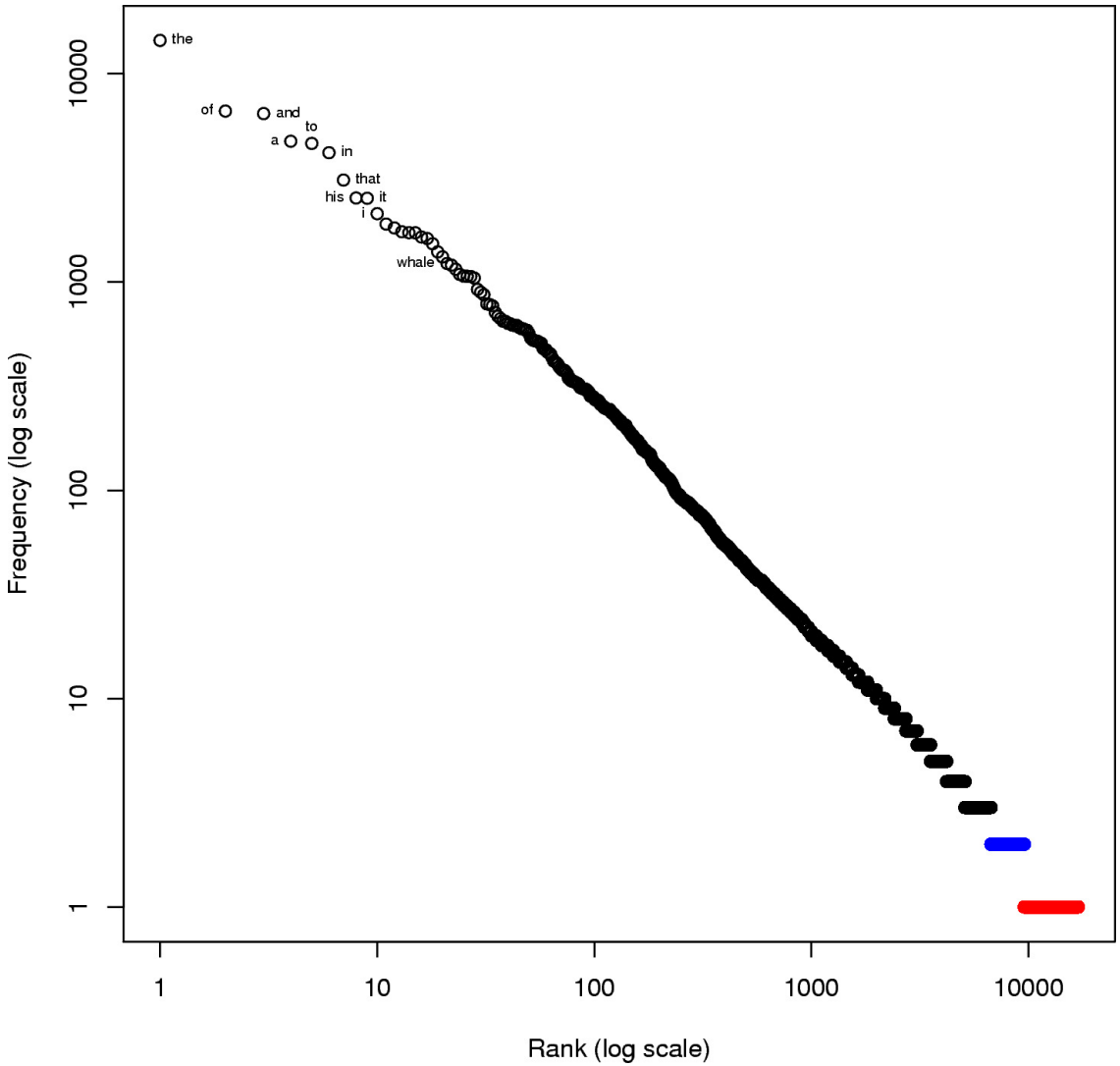


Diphones



Triphones

Zipf's Law: Log linear relationship



Words (corpus: Moby Dick - source: Wikipedia)

Zipf's Law: Log linear relationship

$$f \cdot r \approx k$$

$$log(f) + log(r) \approx log(k)$$

$$log(f) \approx log(k) - log(r)$$

Modelling in NLP

NLP is a set of **Machine Learning** techniques for Language Data.

Machine Learning

Definition

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

(T.M. Mitchell 1997: Machine Learning)

Machine Learning

- Task T: e.g., Text classification
- Experience E: Training examples (e.g., tweets, images, ...)
- Performance measure P: Error function, loss function, classification accuracy, ... $\rightarrow$ testset

We need ...

- Data: Training set, development set, test set
- A suitable feature representation (e.g., bag-of-words)
- A model with trainable parameters (probability theory)
- A learning algorithm (optimisation)
- A performance metric (evaluation)

Example 1: Classification of hate speech

- Data set(s)?
- Input features?
- Output label(s)?
- Performance metric(s)?

Example 2: Image captioning

"A small child is playing with a telephone."

- Data set(s)?
- Input features?
- Output label(s)?
- Performance metric(s)?

Feature representations for language

Unicorn

From Wikipedia, the free encyclopedia

For other uses, see Unicorn (disambiguation).
Not to be confused with Unicron.

The **unicorn** is a legendary creature that has been described since antiquity as a beast with a single large, pointed, spiraling horn projecting from its forehead. The unicorn was depicted in ancient seals of the Indus Valley Civilization and was mentioned by the ancient Greeks in accounts of natural history by various writers, including Ctesias, Strabo, Pliny the Younger, and Aelian.[1] The Bible also describes an animal, the re'em, which some versions translate as *unicorn*.[1]

In European folklore, the unicorn is often depicted as a white horse-like or goat-like animal with a long horn and cloven hooves (sometimes a goat's beard). In the Middle Ages and Renaissance, it was commonly described as an extremely wild woodland creature, a symbol of purity and grace, which could be captured only by a virgin. In the encyclopedias, its horn was said to have the power to render poisoned water potable and to heal sickness. In medieval and Renaissance times, the tusk of the narwhal was sometimes sold as unicorn horn.

Stats vs. ML

	Statistics/CogSci	ML/NLP
Modelling Goal	Analysis Inference	Prediction No inference
Generalisability	World	Similar dataset
Assumptions	Test (e.g., normality) Explicitly model dependence Control for variables	Know that assumptions are wrong! Assume **IID**
Features/predictors	End	Means

Additional links

Here are some additional links from Riccardo on the topic of stats in ML/ML in Stats:

Stichfix:

- https://multithreaded.stitchfix.com/blog/2017/12/13/latentsize/
- https://multithreaded.stitchfix.com/blog/2018/06/28/latent-style/

Generalisation issues:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5888612/
- https://www.nature.com/articles/mp2017227

Supervised vs. Unsupervised Learning

Supervised - Labelled training corpus

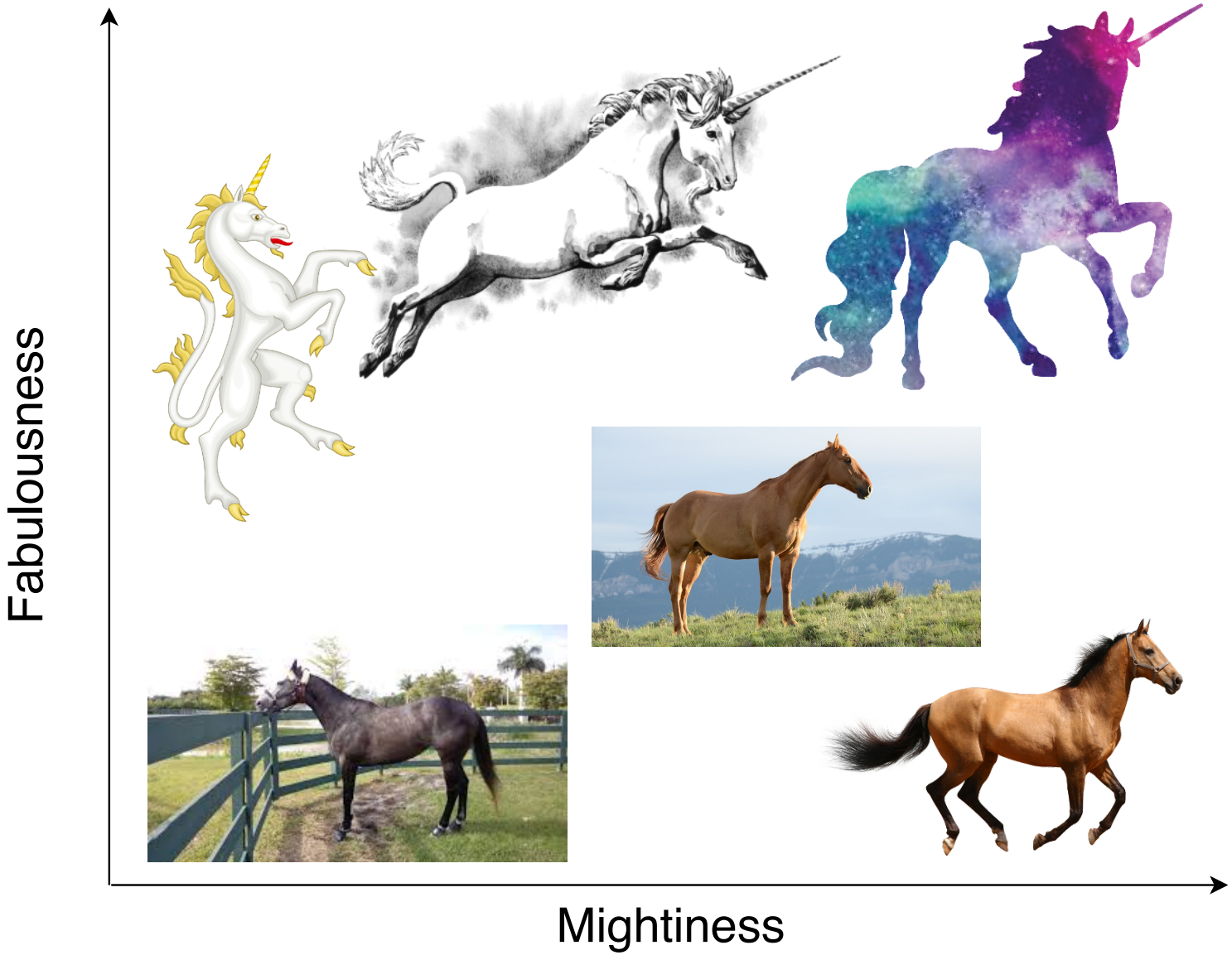
- Learns to *predict/separate* classes
- Better performance
- Need annotated data - may be sparse, expensive

Unsupervised - No labelled training corpus

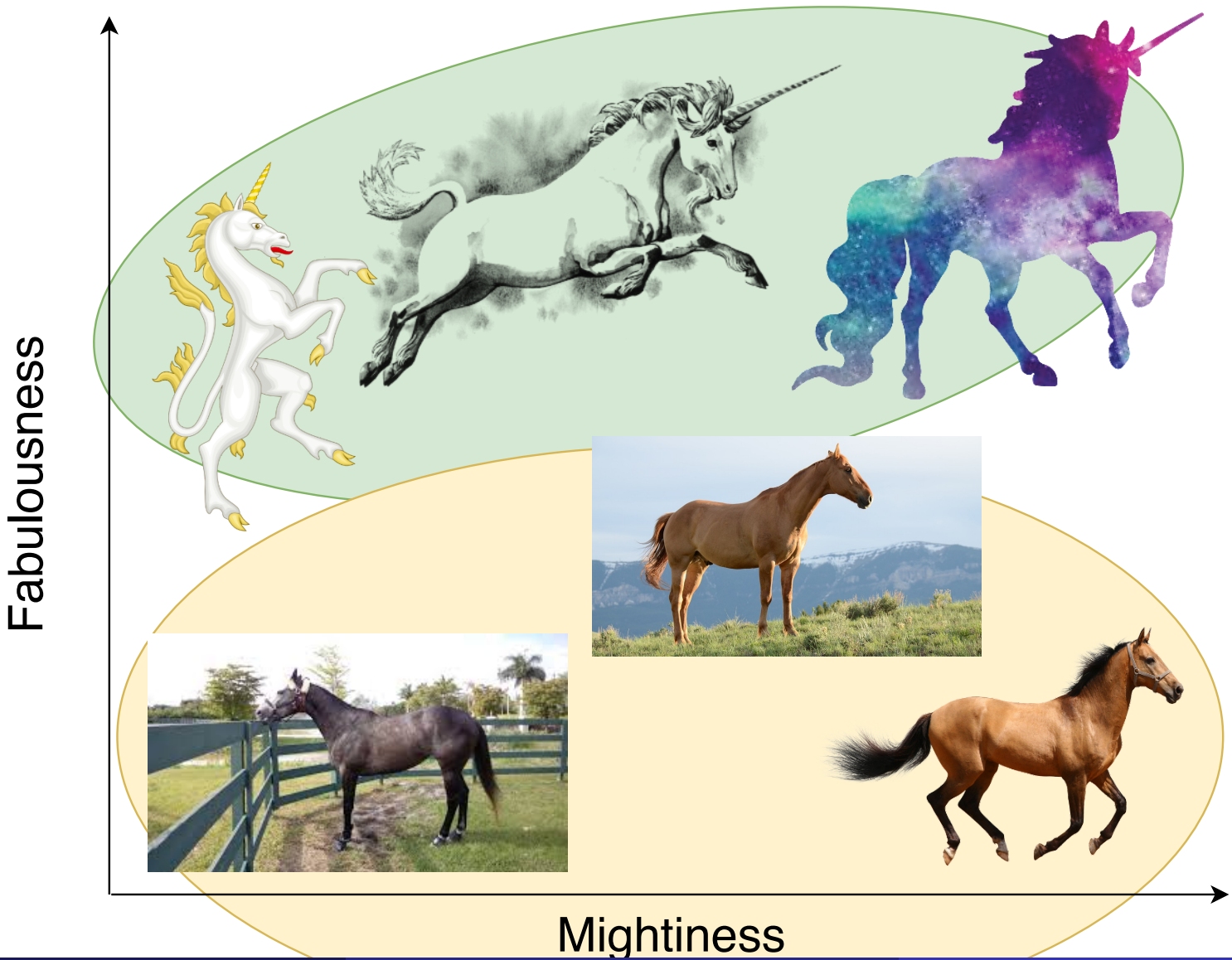
- *Understand/transform* data
- Can use large amounts of freely available, unlabelled data

Semi-supervised - Small labelled training corpus + large unlabelled corpus

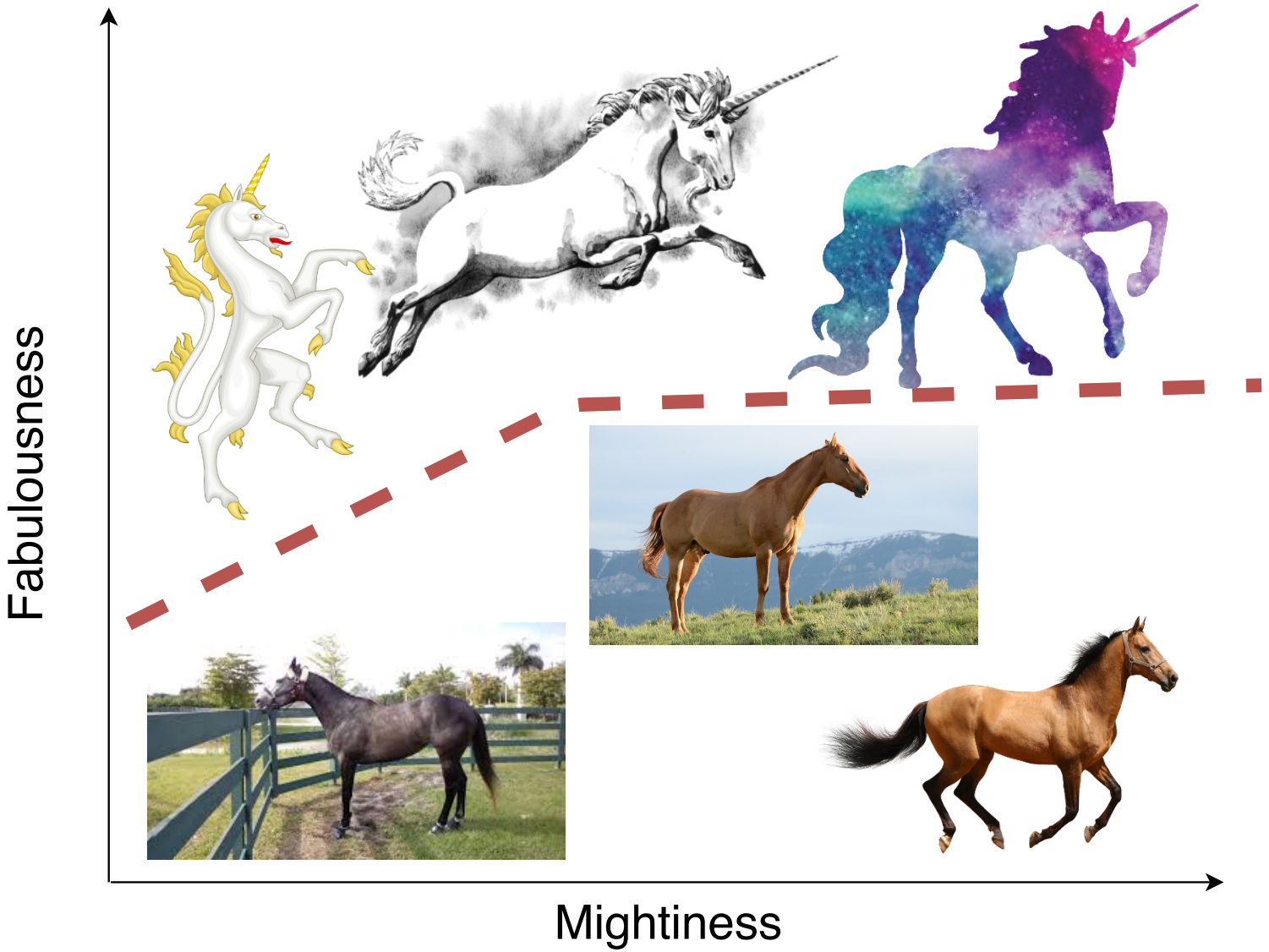
Generative vs. Discriminative Models



Generative vs. Discriminative Models



Generative vs. Discriminative Models



Generative vs. Discriminative Models

Generative - "Generate" each class individually

- Model P($c|x$) $\propto$ p($x|c$) P(c)
- Can set priors - good for small data
- No "real" probabilities
- Often strong independence assumptions necessary
- Often performs worse (esp. with big data)
- e.g., Naive Bayes (later today)

Generative vs. Discriminative Models

Discriminative - Make sure that P(correct class) >> P(all other classes)

- Model P($c|x$) directly (and exactly)
- Less strong independence assumptions, can use arbitrary features
- Learns to *discriminate* between classes - we don't care how well we can model 1 class alone!
- e.g., Logistic Regression (you know the one); Neural Networks (mostly)

Classification vs. Regression

Classification - Output variable is categorical

Regression - Output variable is numeric

Problems with viewing NLP as pure ML problem

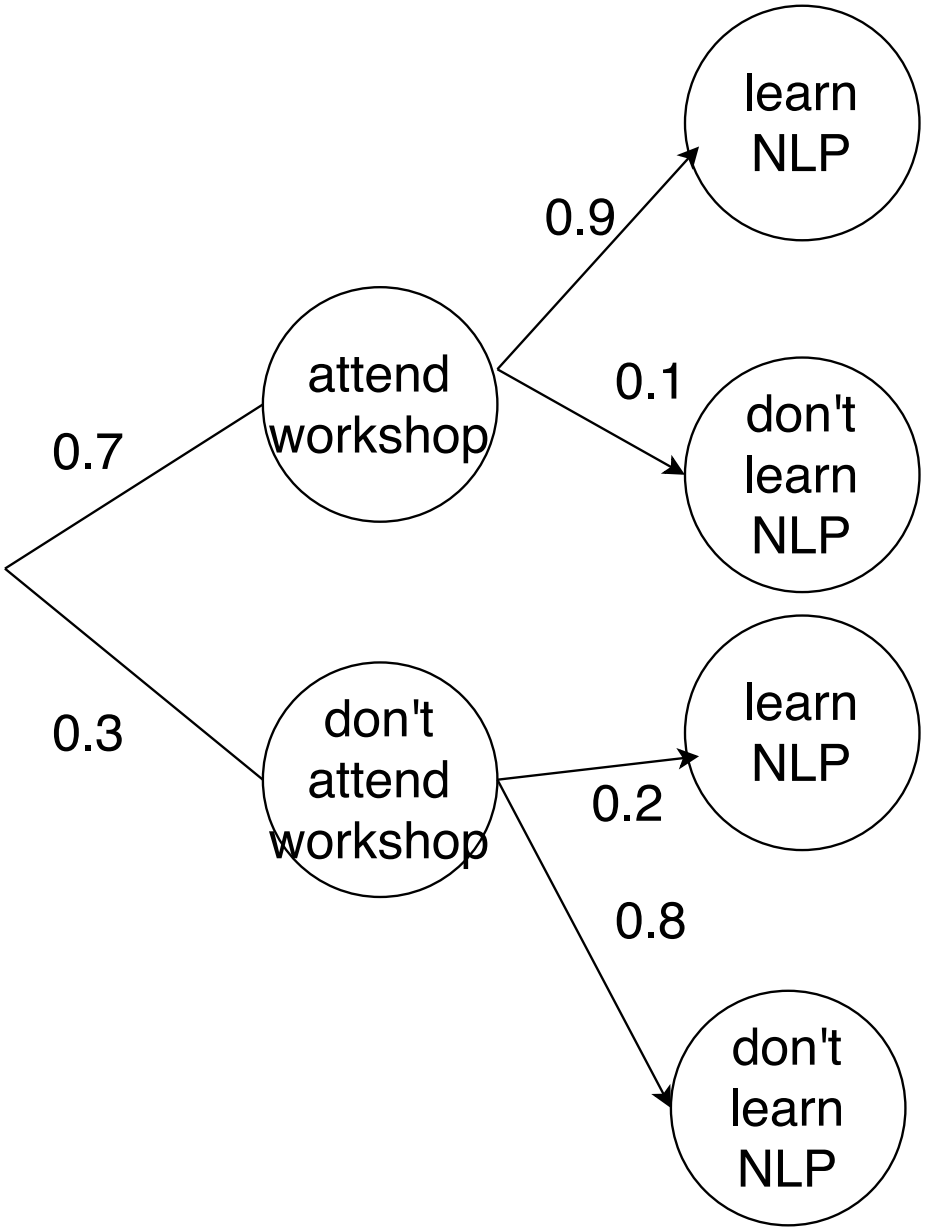
- Sparsity
- Irregularity
- Scope of negation
- Coreference resolution
- Metonymy, methaphor
- Implicature
- Bias in data
- ...

Probability Theory

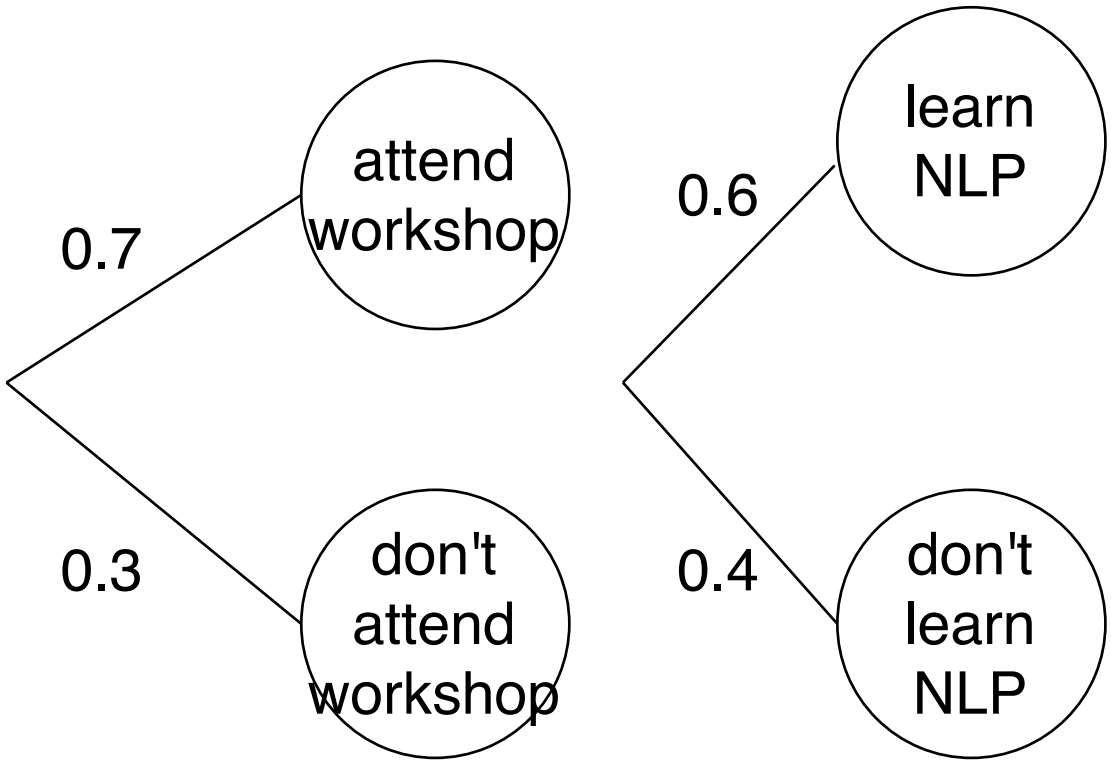
Notation

- P(x) - a probability
  - to be very precise, we could write P($X = x$)
- p(x) - not a "real" probability (doesn't $\sum = 1$)
- P(x, y) - joint probability of x and y
  - e.g. "What's the probability that I learn NLP and Riccardo wears a hat tomorrow?"
- P($x|y$) - conditional probability of x, given y
  - e.g. "What's the probability that I learn NLP, given I attend the workshop tomorrow?"

Conditional probabilities

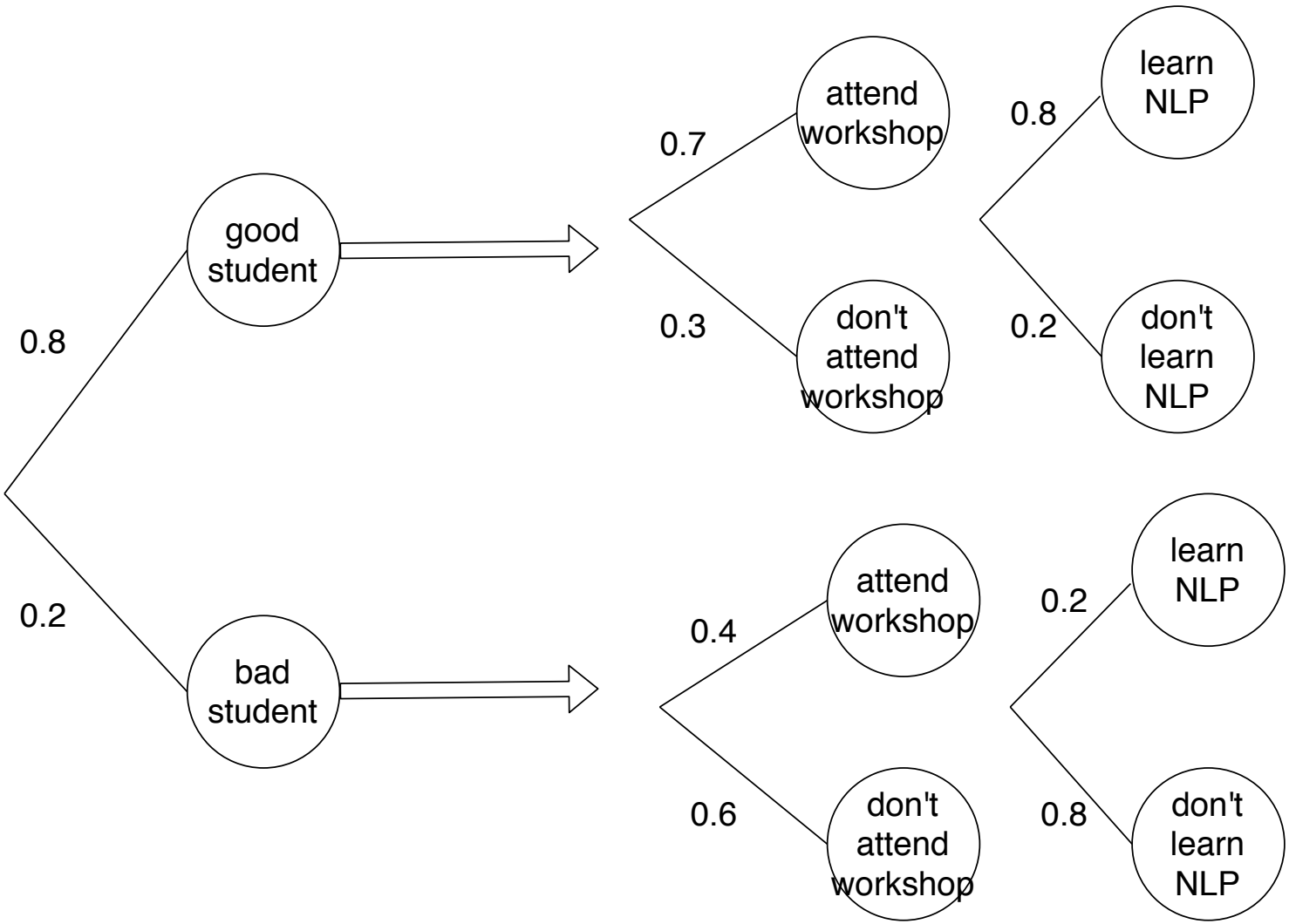


Conditional probabilities



What's the probability that you learn NLP, P(nlp)?

Conditional Independence



Product Rule

The general case:

$$P(x, y) = P(x|y)P(y)$$

$$P(x, y) = P(y|x)P(x)$$

e.g.,

$$P(a, nlp) = P(nlp|a)P(a)$$

Conditional Independence (again)

For independent events:

$$P(x, y) = P(x)\ P(y)$$

(Because in this case:)

$$P(x|y) = P(x)$$

For *conditionally* independent events x and y, given class z:

$$P(x, y|z) = P(x|z)\ P(y|z)$$

e.g.,

$$P(a, nlp|gs) = P(a|gs)\ P(nlp|gs)$$

Bayes' Rule

Bayes' Rule:

$$P(y|x) = \frac{P(x|y)\ P(y)}{P(x)}$$

We often use proportional values instead:

$$P(y|x) \propto P(x|y)\ P(y)$$

Note: Using Bayes' rule $\neq$ Bayesian approach!

Text Classification with Naive Bayes

Text classification

Dear Friend,

How are you and how are you doing I hope you and your family are doing just great?

I want to find out from you if you remember any of your family member or relation that bears this name 'Michael Gotten' because the organization which I work as a banker/accounts executive has requested I search for any member, relation or cousin of our late customer (Michael Gotten) who died on a trip for vacation in 2008 in Birmingham, England.

The bank management board is about to avert the the deposit he left behind to the Queen's treasury and I don't want this to happen, do you ever remember anybody or something?

Please be kind enough to reply me with any vital information regarding the above mentioned name so that the sweat of the man will not waste in the fishy hands of the British Government officials.

I hope to hear from you soon for any information on your family lineage.

Thanks.

Sentiment Classification

★☆☆☆☆ **Don't WASTE your money!!!**, June 22, 2010

By **Tommy Lee**

Verified Purchase (What's this?)

This review is from: Crush It!: Why NOW Is the Time to Cash In on Your Passion (Hardcover)

Here's the straight to the point short answer for this book: Use facebook, and other social media sites to promote yourself! There, I just gave you all this book is going to give you. I simply gave it to you in one sentence for FREE instead of the $10 I paid for it on Amazon, and 2-hours of your time wasted reading it. It's really nothing more than a salesman's pitch book. And the pitch here is to buy his book, and read about him telling you THAT he used social media to grow his business, NOT HOW HE USED SOCIAL MEDIA TO GROW HIS BUSINESS.

*** PLEASE NOTE ***
THIS IS NOT A HOW TO BOOK BY ANY MEANS. If that is what you are looking for, then look elsewhere. This book is simply about Gary saying, I used social media to grow my business. And really folks, that's it!

"**Awesome dresses!!.** Ordered a few dresses for a ball gown event, and they were all lovely. They looked just like the images and fit perfectly, really happy with my purchases.

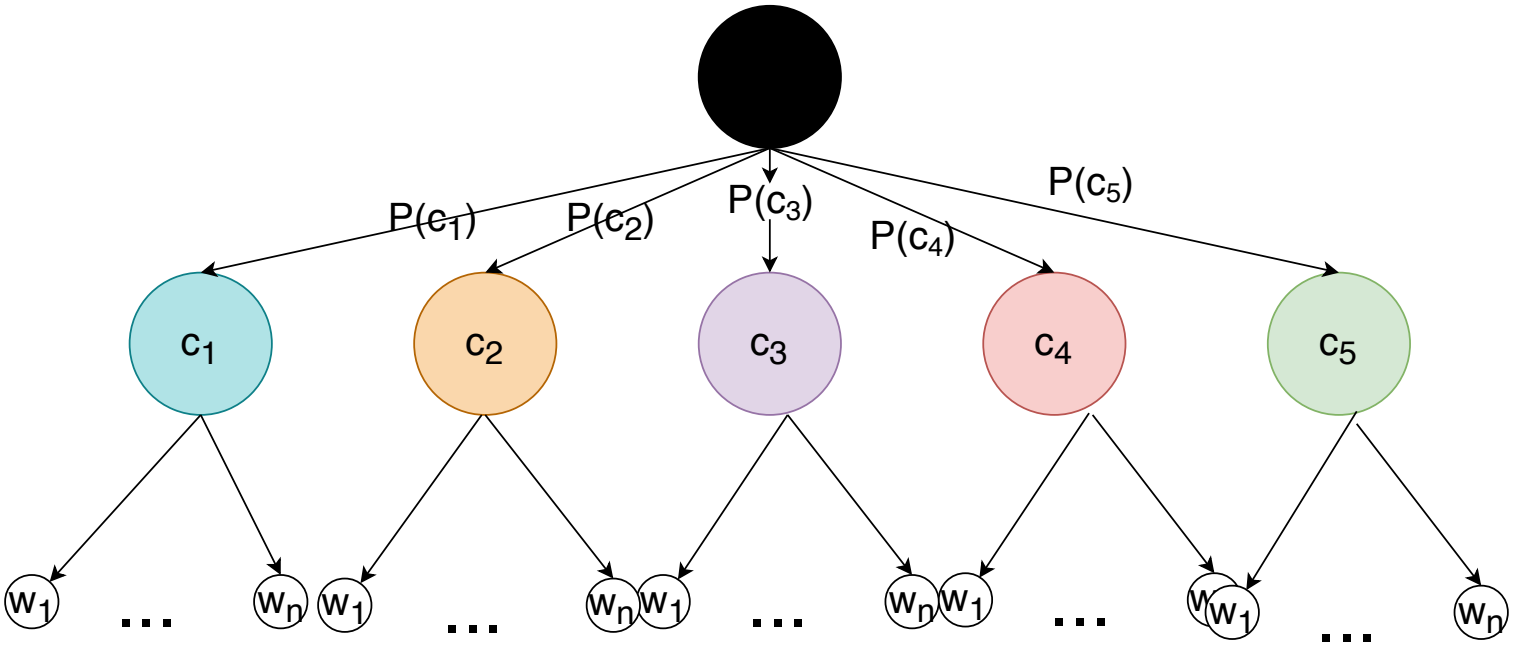
- Rose D.

The Naive Bayes classifier

A bag of words feature model.

- Features are conditionally independent, given the class (**Naive Bayes assumption**)
- Word order/history doesn't matter (**unigram/bag-of-words assumption**)

Generative Model



To "generate" a document:

1. Choose a class with probability P(c)
2. For each feature $f_1, ..., f_n$, choose its value with probability $P(f_i|c)$

Naive Bayes - formally

Given a document X, represented by features $f_1 ... f_n$, choose the predicted class $\hat{c}$ from a set of classes $c_1 ... c_i$ as:

$$\hat{c} = \arg\max_i P(c_i|X)$$

For each class c_i:

- Calculate $p(X|c_i) = P(f_1, f_2, ..., f_n|c_i) = P(f_1|c_i) \cdot P(f_2|c_i) \cdot ... \cdot P(f_2|c_i)$
- Calculate $P(c_i|X) \propto p(X|c_i)P(c_i)$

Naive Bayes - training: MLE

To train the model with Maximum Likelihood Estimation (MLE):

- Estimate the class priors:

$$P(c_i) = \frac{count(c_i)}{\sum_{i'} count(c_{i'})}$$

- Estimate the class conditional probabilities:

$$P(f_n|c_i) = \frac{count(f_n, c_i)}{\sum_{n'} count(f_{n'}, c_i)}$$

Naive Bayes - worked example

Traning set:

document	text	class
d1	I hate this movie.	-
d2	My favourite this year.	+
d3	Not really my favourite.	-
d4	Would recommend.	+

V= {I, hate, this, movie, my, favourite, this, year, not, really, would, recommend}

P(+) = ?

P(-) = ?

P(I|+) =?

...

Naive Bayes - worked example

Test set:

document	text	class
d5	Would not recommend this movie.	

Problem???

Naive Bayes - smoothing

To train the model with Maximum Likelihood Estimation (MLE):

- Estimate the class conditional probabilities:

$$P(f_n|c_i) = \frac{count(f_n, c_i) + 1}{\sum_{n'} count(f_{n'}, c_i) + |V|}$$

Naive Bayes - Assumptions

- Is the Naive Bayes assumption realistic? Why (not)?
- Is the unigram assumption realistic?
- Can you think of examples where the assumptions are violated?
- What could be a consequence of violating the assumptions?

Pros and Cons of Naive Bayes

Pros	Cons
• works well on small data • can choose informative prior • fast at training & testing	• unrealistic, strong assumptions • cannot model feature correlation(s) • discriminatively trained models work better in practice