COMP5046: Natural Language Processing

Joel Nothman

joel.nothman@sydney.edu.au

School of Information Technologies
University of Sydney

2018-03-06

Language encodes meaning ambiguously

Natural (Language Processing)

vs.

(Natural Language) Processing

(Statistical ((Natural Language) Processing))

- NL: **Human language**
  - though we focus on text
  - **linguistics**, rather than the study of a particular language
- NLP: **Drawing information** from language
  - decoding language-encoded messages
  - want high accuracy on unseen text
  - predictive models
- SNLP: **Statistical models** drive that prediction
  - data driven
  - models parametrised by counting events in text
  - learning from labelled or unlabelled text
  - models embody what is usual, rather than what is possible

Why process language?

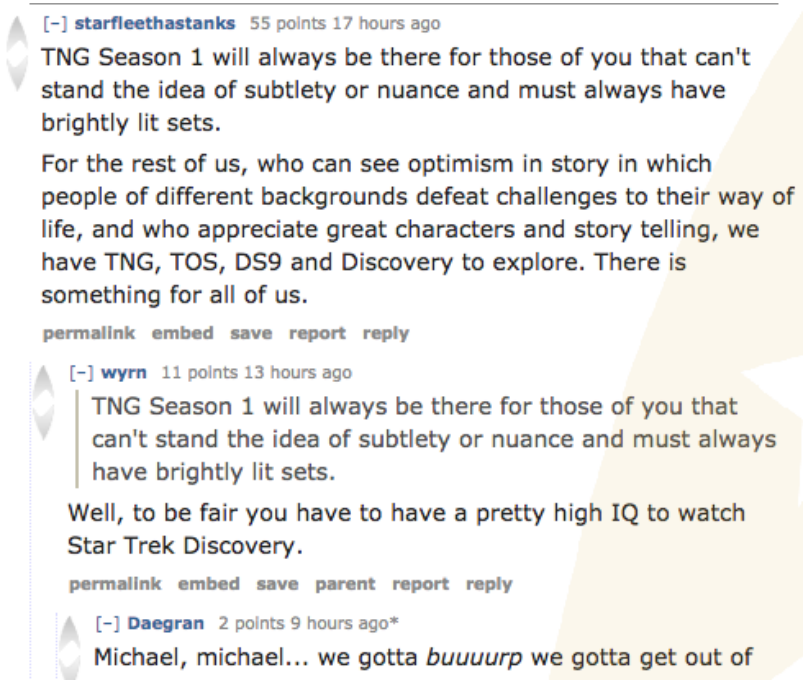
- language stores knowledge
- language communicates new knowledge
- language embodies society
- language is a key to culture and human experience
- language is a natural interface for humans
- language is fascinating

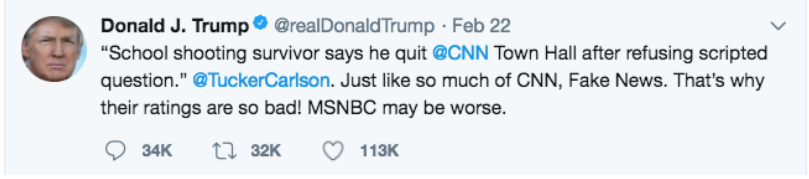
Why process language?

- language stores knowledge
- language communicates new knowledge
- language embodies society
- language is a key to culture and human experience
- language is a natural interface for humans
- language is fascinating

Why process language?

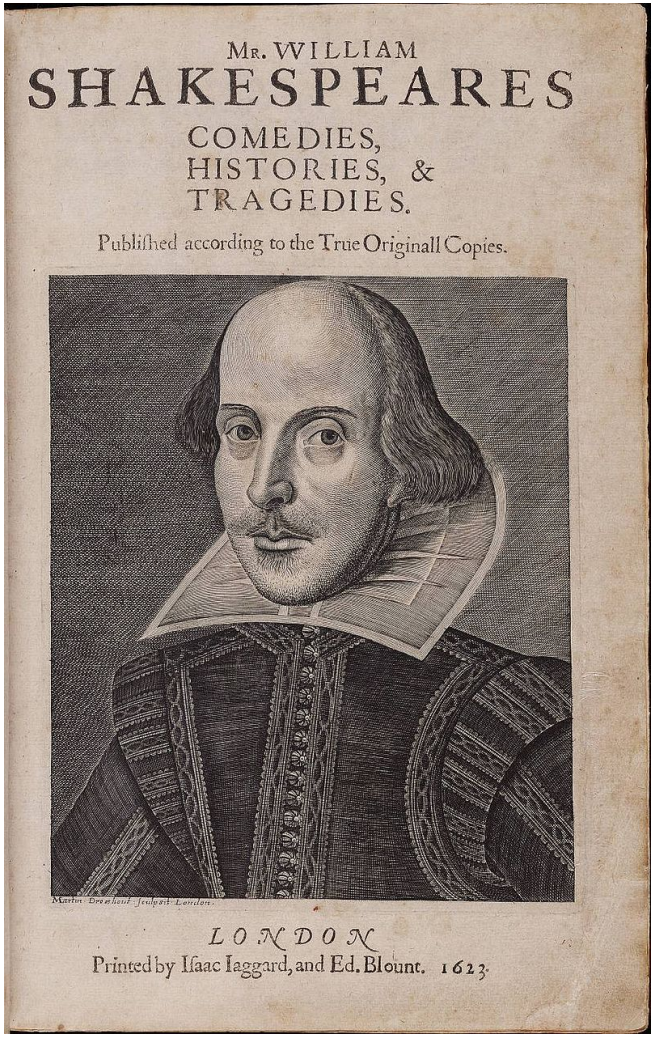
- language stores knowledge
- language communicates new knowledge
- language embodies society
- language is a key to culture and human experience
- language is a natural interface for humans
- language is fascinating





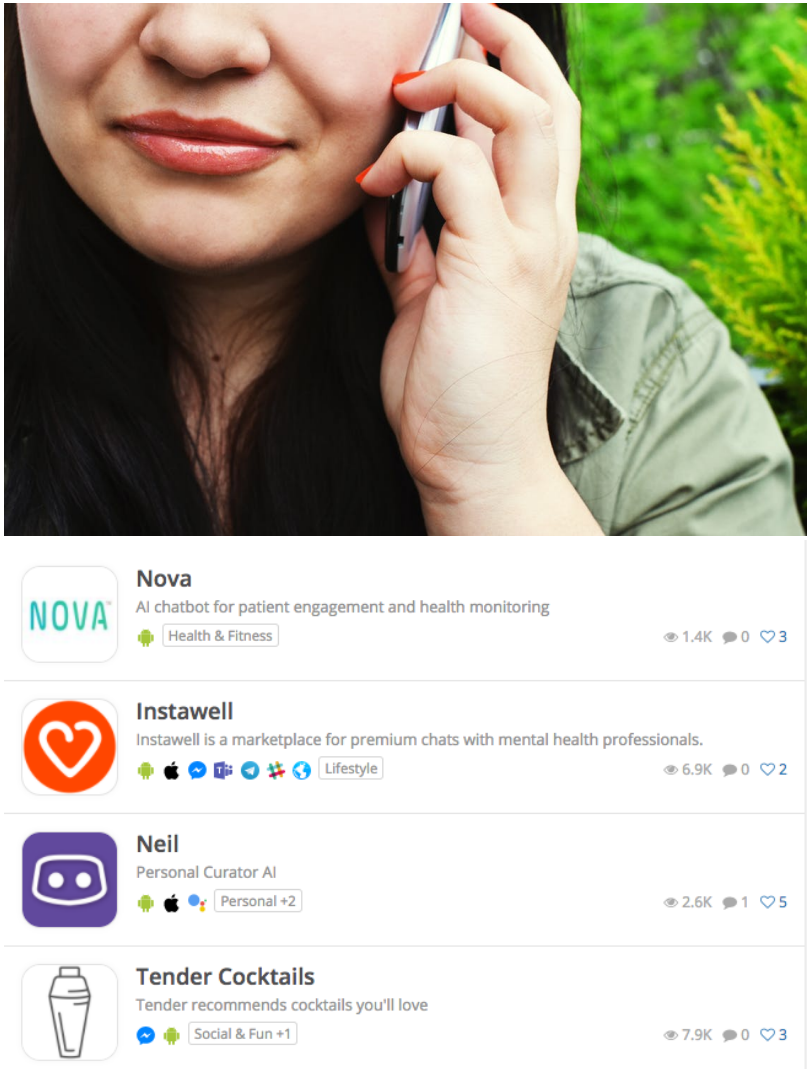
Why process language?

- language stores knowledge
- language communicates new knowledge
- language embodies society
- language is a key to culture and human experience
- language is a natural interface for humans
- language is fascinating



Why process language?

- language stores knowledge
- language communicates new knowledge
- language embodies society
- language is a key to culture and human experience
- language is a natural interface for humans
- language is fascinating



Why process language?

- language stores knowledge
- language communicates new knowledge
- language embodies society
- language is a key to culture and human experience
- language is a natural interface for humans
- language is fascinating

Ambiguous headlines. . .

- Miners Refuse to Work after Death
- March Planned For Next August
- Aging Expert Joins University Faculty
- Fund Set Up for Beating Victim's Kin
- New Study of Obesity Looks for Larger Test Group
- Hospitals are Sued by 7 Foot Doctors
- Local High School Dropouts Cut in Half
- Squad Helps Dog Bite Victim
- Killer Sentenced to Die for Second Time in 10 Years
- Red Tape Holds Up New Bridges
- Deaf College Opens Doors to Hearing
- Bar trying to help alcoholic lawyers
- Child teaching expert to speak
- Drunk Gets Nine Months in Violin Case

Language is processed to . . .

- predict phenomena (e.g. stock prices) from text
- research linguistic or social phenomena at large-scale
- collect structured knowledge (e.g. a database of protein interactions) from language
- retrieve relevant text from large collections
- answer natural language questions
- translate text to another language
- interpret speech as fluent text
- convert text to fluent speech
- enable natural interaction with software (e.g. chat bots)
- help people to improve their writing
- . . .

Course materials

http://canvas.sydney.edu.au/courses/2437

- Link to Course Outline on CUSP
- Lecture slides
- Lab tasks
- Lab solutions (at github.sydney.edu.au)
- Assignments
- Discussion (please ask questions publicly)
- Assignment submission

Staff

`sit.comp5046@sydney.edu.au`

No fixed consultation hour; please arrange a time to see us.

Unit Coordinator: **Professor Alan Fekete**
Room 447, School of IT. 02 9351 4287
`alan.fekete@sydney.edu.au`
Admin issues: illness, misadventure, policies, group difficulties

Lecturer: **Dr Joel Nothman**
`joel.nothman@sydney.edu.au`

Tutors:

- **Xiang Dai** ("Dai")
- **Raghavendra Chalapathy** ("Raghav")

Contact hours

You are expected to:

1. Work 12 hours per week for this course (including 3 contact hours);
2. Attend 2 hours of lectures per week:
   - Tuesday 2-4pm, Arch LT 1
   - Lectures are recorded, but don't depend on it!
3. Attend 1 hour of tutorial/laboratory time
   - starts this week
4. Participate respectfully in discussions in lectures and labs;
5. Complete all assessment tasks on time.

Expectations: I assume you can program

- by that, I mean you are a *confident* programmer
- labs will involve programming
- assessment will involve programming
- Python recommended; other popular languages accepted
  - do *not* try learning Python on the job!
- there will be no non-programming option for assignments

- but it's more than just programming:
  - algorithms, mathematics and (esp.) statistics
  - linguistics and intuition about language
  - analytical thinking

Expectations: I don't assume you are a linguist

- But you do need to know roughly how to identify a noun/verb/etc.
- We will think critically about how we use language
- and about how computatational models capture aspects of language

Lecture outline

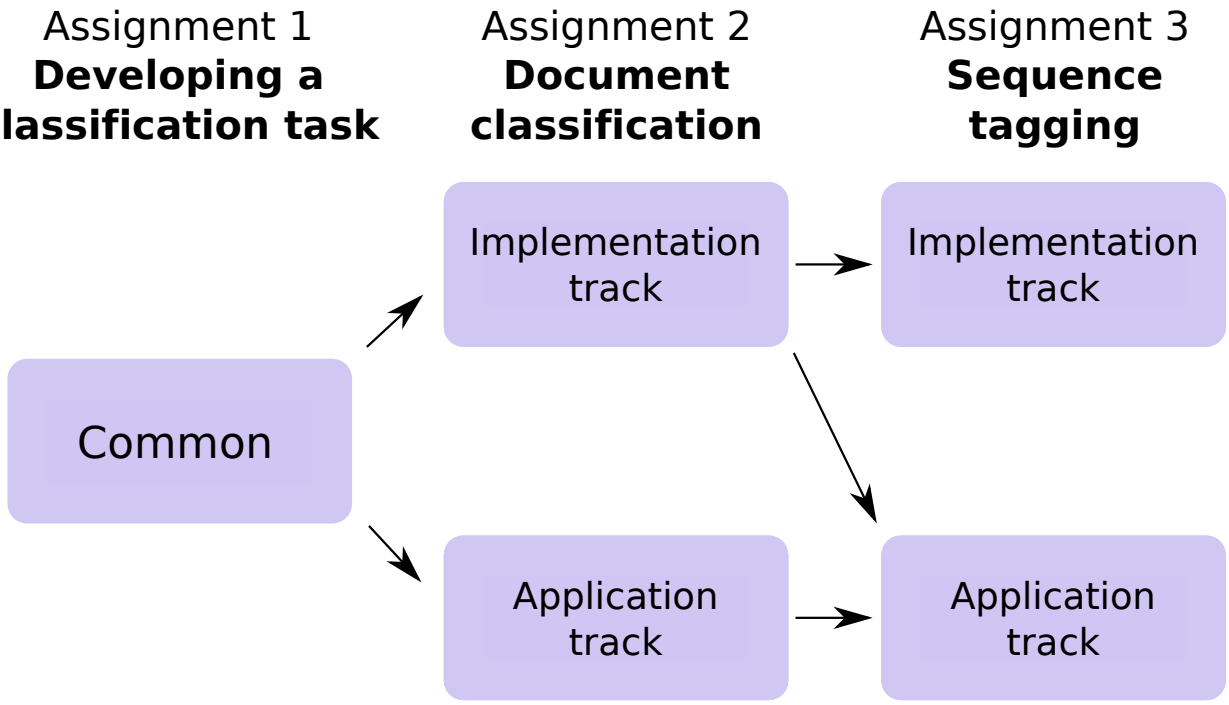
- Week 1-4: Representing language
  - Introduction to linguistics
  - Language as patterns and probabilities
  - Representing meaning
  - Grouping text into categories (assignment 1 / labs)
- Week 5-8: Extracting content and the NLP pipeline
  - Representing text as features
  - Document classification
  - Sequence tagging
  - Information extraction
- Week 9-12: Advanced topics
  - Parsing deep structure
  - Guest lecture: Applications
  - Guest lecture: Neural networks for NLP

Assessment schedule

Component	Due Date	%
Major project		
stage 1: annotation task	Week 5, 5pm Monday 2018-04-07	10.0
stage 2: categorisation task	Week 9, 5pm Monday 2018-05-05	20.0
stage 3: sequence tagging task	Week 14, 5pm Monday 2018-06-09	20.0
Written examination (2hrs)	Exam week	50.0

- Released on Tuesdays, available on Canvas
- Submit early: you can resubmit until the deadline
- Late work loses 10% of available marks per day; 0 after 7 days
- Pass with overall mark $\geq$ 50% *and* exam mark $\geq$ 40%

Assignments tracks



Implementation track: code up statistical algorithms.

Application track: depend more on existing libraries.

Assignments

- All assignments involve language analysis, coding and report writing
- **Reports are the primary deliverable**
- Though we *will* check implementations for correctness
- Assignments will not be very different from last year's
- Last year's assignments are available from Canvas
- Reports will be submitted to Turnitin through Canvas
- Code is also submitted (for assignments 2 and 3) and retained
- We *will* use code plagiarism detection tools
- **Clearly reference any copied/adapted code portions and cite their origins**

Start assignments early!

- Starting early means you will work while you sleep
- **Don't waste all your time on code**
- The report is more important, but largely depends on the code
- If you're stuck, **ask early**
- We might be able to offer you alternatives

Labs . . .

- include programming exercises to practise computing with text
- include analysis exercises to practise analysing text
- are a place of discussion to reinforce your understanding
- in weeks 2-4 will primarily support assignment 1

- We will provide example solutions in Python
- but there is no solution to open-ended tasks

- If working in Python
  - please get Anaconda Distribution
  - please use Python 3 to avoid unicode hell and division surprises

Readings to support lectures

- Readings are recommended in Canvas lecture modules

- The past: FSNLP
  Manning and Schtze (1999), *Foundations of Statistical Natural Language Processing*, 6th edition.
- The future: SLP3
  Jurafsky and Martin (?), *Speech and Language Processing*, 3rd edition.
  https://web.stanford.edu/~jurafsky/slp3/
- SLP2, Jurafsky and Martin 2nd ed. is also available from library

WHS Induction

and further administrivia

Self-test

- How much work will you be devoting to this unit, each week?
- Who should you see if difficulties arise?
- When is the first assessment due?
- What do you do if you get sick during semester?
- What is Turnitin?
- What programming language do you need to know?

Part I

Introduction to Statistics

Outline

- Probability
- Statistical Modelling
- **Faithfulness and Generalisation**
- Independence
- **Maximum Likelihood Estimate (MLE)**
- Collocation Extraction

Probability

- informally the likelihood or chance of something happening
  50/50 heads or tails
  1/6 chance of rolling a 4

. . . probability that some string of characters is valid English

Outcomes and Experiments

- this *something* is called an outcome
  the coin lands on heads
  5 and 5 are rolled on two dice
- outcomes can be discrete or continuous

- an experiment involves observing the outcome of a situation
  a single toss of the coin
  a single roll of the two dice

- the sample space is *set of all possible outcomes* for an experiment
  {heads, tails}
  {(1,1), (1,2), (1, 3), . . . , (6, 6)}

Events

- an event is a *set of outcomes*
  {heads} or {heads, tails}
  {(1,1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)}
- The event space (the set of possible events) is $2^{\text{sample space}}$ (the power set of outcomes)

- **NB: outcomes $\neq$ events**

Probability

- A **probability distribution** assigns a value to each event
  $P(\text{heads}) = 1/2$
  $P(\text{heads, tails}) = 1$
- The probability of some event is the probability that any outcome in that event occurs
- The value must range between 0 and 1 (inclusive)
- The value is called the probability or probability mass.

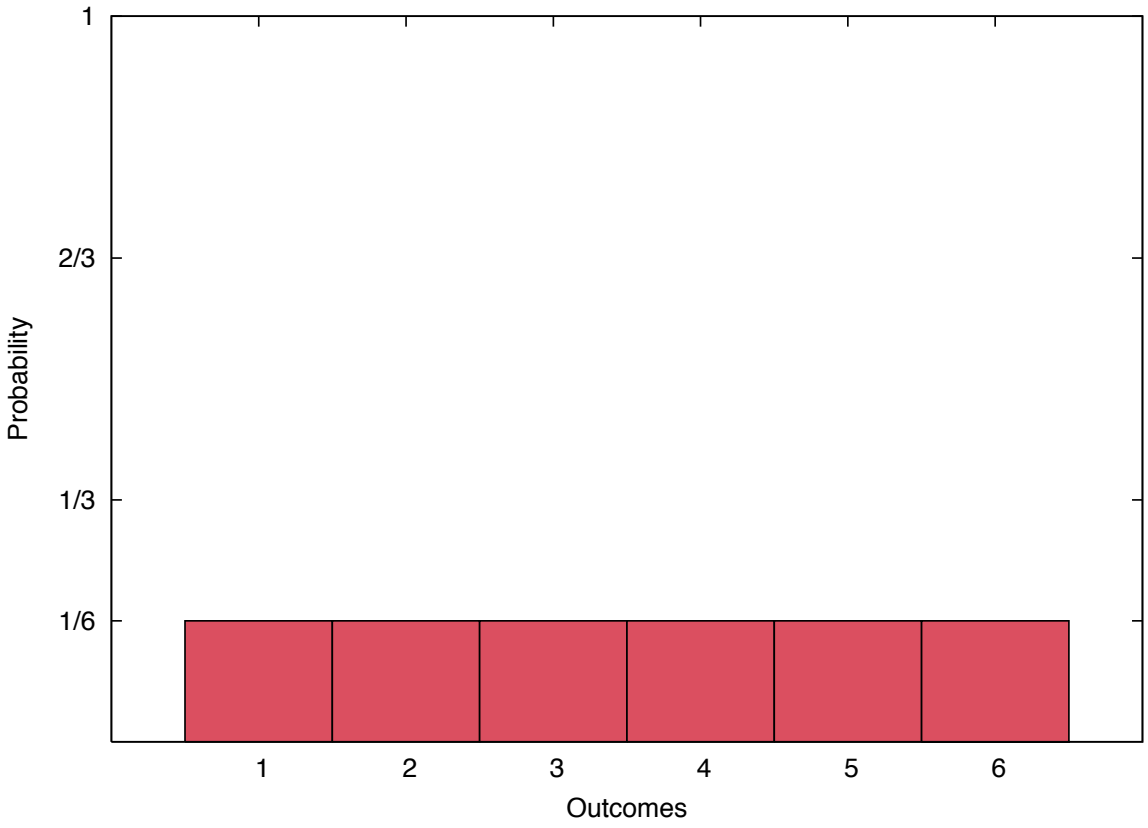
Statistical Modelling

- given a set of observations (i.e. *measurements*):
  $\Longrightarrow$ extract a mathematical description of observations
  $\Longrightarrow$ **statistical model**
  $\Longrightarrow$ use this for **predicting** future observations
  e.g. by interpreting the statistical model as a probability distribution over future
- a statistical model should:
  - **represent faithfully** the original set of measurements
  - **generalise sensibly** beyond existing measurements

Example: Modelling a Die Roll

- consider a single roll of a 6-sided die
- **without any extra information** (any measurements)
- what is the probability of each outcome?
- **why do you make that decision?**

Example: Modelling a Die Roll



Faithful Representation

- trivial **if no generalisation** is required
  just look up the relative frequency directly
- trust the training data exclusively
- but unseen observations are impossible
  since relative frequency is zero
- and **most observations** are unseen

$\Longrightarrow$ **practically useless!!**

Sensible Generalisation

- want to find correct distribution given seen cases
  i.e. to minimise error in prediction
- **sensible** is very hard to pin down
- may be based on some hypothesis about the problem space
- might be based on attempts to account for unseen cases

$\Longrightarrow$ **generalisation reduces faithfulness**

Joint Probability

- the joint probability $P(A \cap B)$ is the probability of *both* event A and B occurring together.
- often written $P(A, B)$
- **this is not one event or the other**
- fair die roll: $P(\text{even}, > 3) = \frac{2}{6}$

Conditional Probability

- partial knowledge of an outcome that informs our model
  already seen outcome of first die roll
- conditional probability $P(A|B)$ (said *probability of A given B*)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

- $P(\text{even}| > 3) = \frac{2}{6} \div \frac{1}{2} = \frac{2}{3}$

Independence

- two events A and B are **independent** iff:

$$P(A \cap B) = P(A)P(B) \tag{2}$$

- i.e. $P(A|B) = P(A)$ and $P(B|A) = P(B)$
- knowing A provides no information about B
- knowing B provides no information about A

- $P(\text{roll}_1 = 6, \text{roll}_2 = 6) = P(\text{roll}_1 = 6)P(\text{roll}_2 = 6)$
- $P(\text{roll}_1 = 6, \text{roll}_1 = 5) = 0 \neq P(\text{roll}_1 = 6)P(\text{roll}_1 = 5)$

Conditional Independence

- two events A and B are **conditionally independent** iff:

$$P(A \cap B|C) = P(A|C)P(B|C) \tag{3}$$

- i.e. $P(A|B \cap C) = P(A|C)$ and $P(B|A \cap C) = P(B|C)$
- knowing C gives no knowledge about $A \cap B$ beyond what C indicates individually about A or B
- i.e. knowing C may give some knowledge about A or B

- $P(\text{roll}_1 = 6, \text{roll}_2 = 6) = P(\text{roll}_1 = 6)P(\text{roll}_2 = 6)$

but $C := \text{roll}_1 + \text{roll}_2 = 12$

then $P(\text{roll}_1 = 6, \text{roll}_2 = 6|C) \neq P(\text{roll}_1 = 6|C)P(\text{roll}_2 = 6|C)$

Estimation

- Learn a model from some set of examples $X = (x_1, x_2, \cdots, x_N)$
- Estimate a hypothetical probability distribution over all possible $x \in \mathcal{X}$
- Estimate the **parameters** of the model

Likelihood

- **Likelihood** is the *hypothetical probability* that an event that has already occurred would yield a specific outcome
  - likelihood refers to past events with known outcomes
  - probability refers to the occurrence of future events

  $\Rightarrow$ represent faithfully? generalise sensibly?
- a model $p(x)$ of some empirical distribution $\tilde{p}(x)$ has likelihood:

a model which fits the training data well has high likelihood

$$P(\tilde{p}|p) \equiv \prod_{x \in \mathcal{X}} p(x)^{\tilde{p}(x)} \tag{4}$$

emperical dist is relative dist of observed outcomes

- we can calculate $\tilde{p}(x)$ using:

a model which predicts the outcomes well has high probability

$$\tilde{p}(x) \equiv \frac{\text{count}(x)}{\sum_{x \in \mathcal{X}} \text{count}(x)} \tag{5}$$

Likelihood example

likelihood (observed) | predictions

- Seen die rolls 1, 1, 2, 3, 5
- $\tilde{p}(1) = \frac{2}{5}$, $\tilde{p}(2) = \frac{1}{5}$, $\tilde{p}(3) = \frac{1}{5}$, $\tilde{p}(4) = 0$, $\tilde{p}(5) = \frac{1}{5}$, $\tilde{p}(6) = 0$
- Fair die hypothesis:
  - $p_{\text{fair}}(x) = \frac{1}{6} \forall x$
  - $P(\tilde{p}|p_{\text{fair}})) = \prod_{x \in \mathcal{X}} p(x)^{\tilde{p}(x)} = \frac{1}{6}^{\frac{2}{5}} \times \frac{1}{6}^{\frac{1}{5}} \times \frac{1}{6}^{\frac{1}{5}} \times \frac{1}{6}^{0} \times \frac{1}{6}^{\frac{1}{5}} \times \frac{1}{6}^{0} = 0.1666$
- Bad hypothesis:
  - $p_{\text{bad}}(1) = 1$
  - $P(\tilde{p}|p_{\text{bad}})) = 1^{\frac{2}{5}} \times 0^{\frac{1}{5}} \times 0^{\frac{1}{5}} \times 0^{0} \times 0^{\frac{1}{5}} \times 0^{0} = 0$
- Maximum likelihood hypothesis:

best model which fits training data

  - $p^*(1) = \frac{2}{5}$, $p^*(2) = \frac{1}{5}$, $p^*(3) = \frac{1}{5}$, $p^*(4) = 0$, $p^*(5) = \frac{1}{5}$, $p^*(6) = 0$
  - $P(\tilde{p}|p^*)) = \frac{2}{5}^{\frac{2}{5}} \times \frac{1}{5}^{\frac{1}{5}} \times \frac{1}{5}^{\frac{1}{5}} \times 0^{0} \times \frac{1}{5}^{\frac{1}{5}} \times 0^{0} = 0.264$

Maximum Likelihood Estimate

- Maximum Likelihood Estimate (MLE) is the model $p^*(x)$ which gives the data $\tilde{p}(x)$ the maximum likelihood

$$p^* \equiv \underset{p \in \mathcal{P}}{\text{argmax}}\, P(\tilde{p}|p) \tag{6}$$

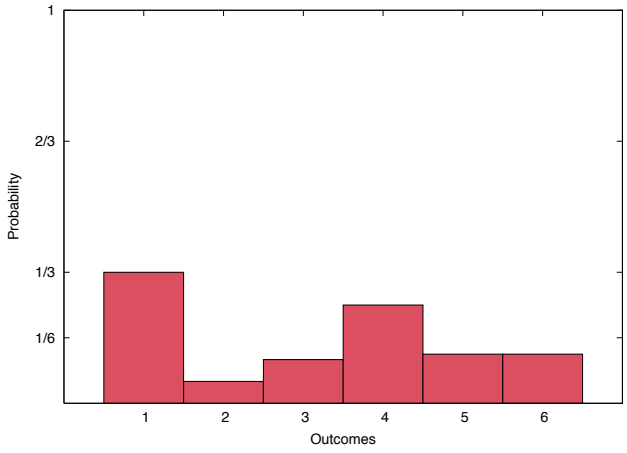
parameter

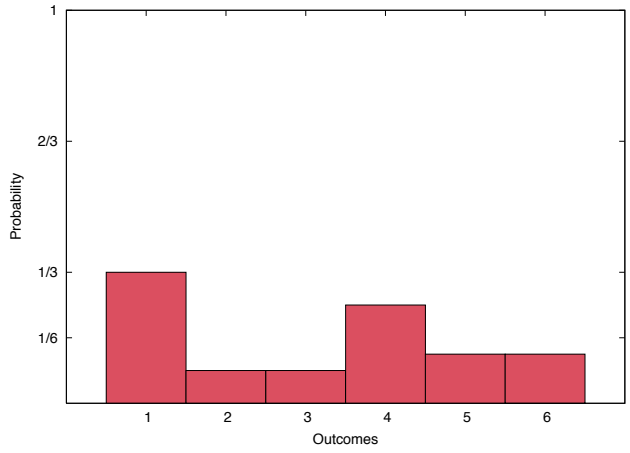
- poor generalisation

max likelihood has poor generalization

A uniform model is a good model

- a good model matches the training data
- but spreads the rest of the probability uniformly





- principle of maximum entropy: **the best probability distribution accounting for some knowledge is the one that is most uniform**

without any knowledge, every event has same prop

- we'll come back to this later

An application of probability to natural language: Collocations

- a collocation is a conventionalised multi-word expression
  strong tea, weapons of mass destruction, built up, Prime Minister
- meaning is often non-compositional
  kick the bucket
- cannot substitute words
  stiff breeze vs *stiff wind
  white wine vs *yellow wine

Collocations and Independence

- Collocations can be (reliably) identified using statistics
- what about simple frequency?
  $\Longrightarrow$ **biased by frequent words**
  of the, in the, . . . , and the

- **Test for whether the words are independent**
- i.e. are these words appearing together by chance?

Testing for Independence

- consider the case of two words x and y (a bigram)
- test the hypothesis that the joint probability:

$$p(x, y) \tag{7}$$

- is **significantly larger** than the independent model:

$$p(x)p(y) \tag{8}$$

- i.e. they occur together more often than we would expect

testing the hypothesis that the two words appear together more often than expected

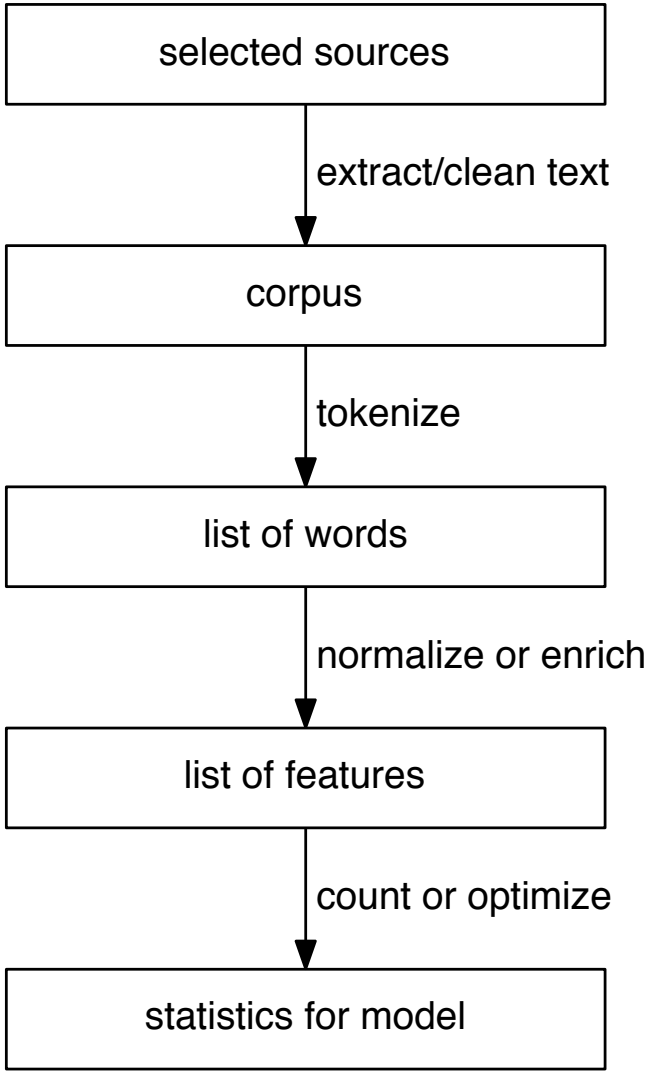
Hypothesis

- The t-test is one standard technique for hypothesis testing
- The null hypothesis is $H_0 : p(x, y) = p(x)p(y)$
- The alternative hypothesis is $H_1 : p(x, y) > p(x)p(y)$
- MLE for $p(x)$, $p(y)$ and $p(x, y)$ is simply relative frequency

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{p(x, y) - p(x)p(y)}{\sqrt{\frac{p(x,y)}{N}}}$$

- This is one of many models of collocation strength
- Its parameters are derived from these sufficient statistics:
  - How many times does each word appear?
  - How many times does each pair of words appear together?
  - How many words (or pairs of words) are there?

But first we need words . . . The NLP pipeline



Take away

- Basic understanding of predictive statistical modelling
- Terms: likelihood, parameters, estimation
- Trade-off between faithfulness and generalisation
- Maximum likelihood (MLE) is a good fit, but bad for generalisation
- Basic theorems of probability
- Collocations and their relation to probability