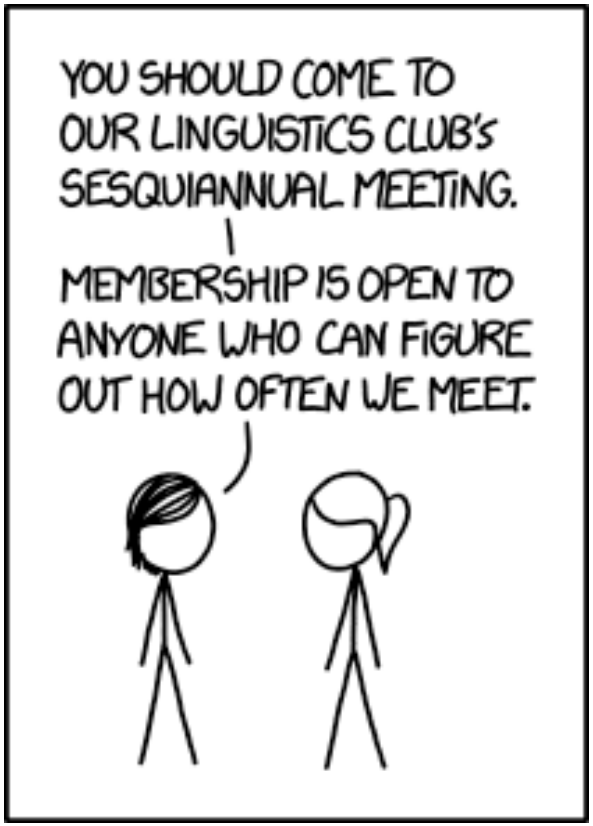
COMP 4446 / 5046
Natural Language Processing

Lecture 1: Course Info & Introduction to NLP

Dr. Jonathan Kay Kummerfeld

Semester 1, 2023
School of Computer Science,
University of Sydney

Linguistics Club



[If that's too easy, you could try joining Tautology Club, which meets on the date of the Tautology Club meeting.]

Source: https://xkcd.com/1602/

0 Acknowledgement of Country

I would like to acknowledge the Traditional Owners of Australia and recognise their continuing connection to land, water and culture. I am currently on the land of the Gadigal people of the Eora Nation and pay my respects to their Elders, past, present and emerging.

I further acknowledge the Traditional Owners of the country on which you are on and pay respects to their Elders, past, present and future.

0 Lecture Plan

Lecture 1: Introduction to Natural Language Processing

1. ***Course Introduction***
2. *Overview of Natural Language Processing (NLP)*
3. *Word Meaning and Representation*
4. *Count-based Word Representation*
   - *One-hot Encoding*
   - *Bag of Words*
   - *Term Frequency-Inverse Document Frequency*
5. *Next Week Preview*
   - *Prediction-based Word Representation*

0 Online Learning

Lectures

- Ask questions using Mentimeter (NOT Zoom chat)
- Also upvote questions on Mentimeter!
- Stay muted unless called on

Tutorials (RE)

- If you have a webcam, please switch it on so we can see you, if you are comfortable doing so.
- Use earphones or headphones
- Try not to talk over other students or the tutor

1 Introduction

Dr Jonathan Kay Kummerfeld

Education and Career

- USyd, B Sc. Adv. (Hons) (Medal)
- Berkeley, PhD
- Michigan, Postdoc
- Harvard, Visiting Scholar
- USyd, Senior Lecturer

Teaching

- Berkeley, Outstanding Graduate Student Instructor Award
- Berkeley, AI course lecturer
- Guest talks all over the world - Cambridge, Oxford, Yale, etc

Research and Industry Experience

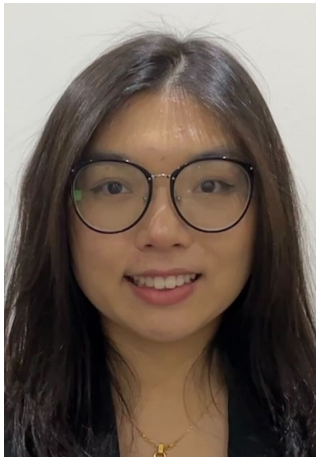
- 21 papers in the most prestigious venues for NLP
- Over 1,800 citations of my work
- Technical advisor to four startups
- Monash Scholar, ARC DECRA Fellow

1 Introduction - Tutors

Aashika Agarwal

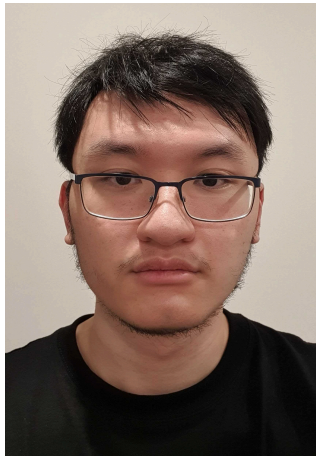
Andrew Lee

Yidong Gan

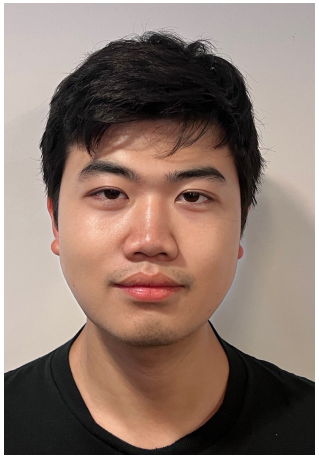


Jieting (Monica) Long

Yonglin Lu



Clinton Mo



Zewei Shi

1 INTRODUCTION

COMP5046 Natural Language Processing

Unit Outline – COMP 4446 / 5046
https://www.sydney.edu.au/units/COMP5046 https://www.sydney.edu.au/units/COMP4446

Canvas – COMP 4446 / 5046
https://canvas.sydney.edu.au/courses/48399

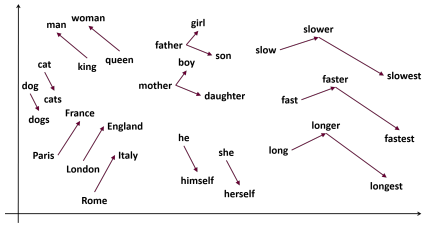
1 Introduction

What will you learn in this course?

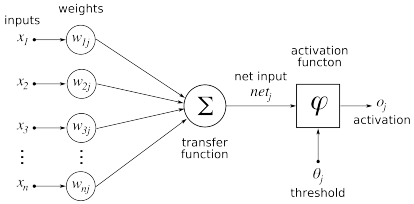
Key challenges in NLP (e.g., translation and question answering)

Models and methods to address those challenges

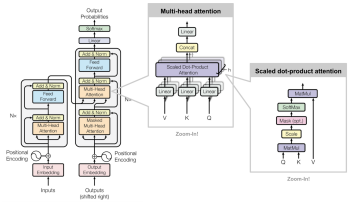
Tradeoffs between different approaches



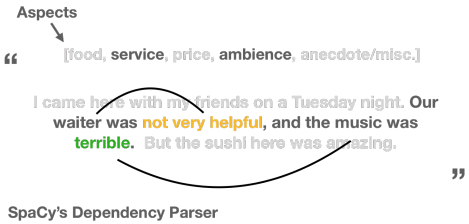
word2vec



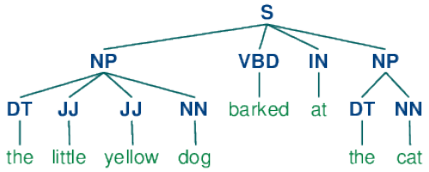
Feedforward Network



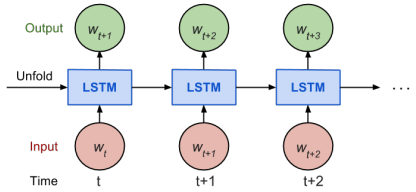
Transformer



Dependency Parsing



Constituency Parsing



Language Modelling

THE UNIVERSITY OF SYDNEY

What will you learn in this course?

- Week 1: Introduction to Natural Language Processing (NLP)
- Week 2: Word Embeddings
- Week 3: Word Classification with Machine Learning I
- Week 4: Word Classification with Machine Learning II

NLP and Machine Learning

- Week 5: Fundamentals of Language
- Week 6: Part of Speech Tagging
- Week 7: Dependency Parsing
- Week 8: Language Models
- Week 9: [ANZAC Day]
- Week 10: Named Entity Recognition and Coreference Resolution
- Week 11: Question Answering
- Week 12: Machine Translation

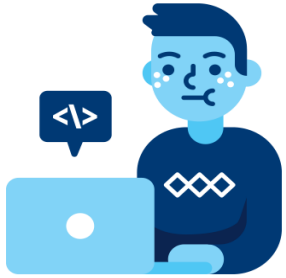
NLP Tasks

- Week 13: Future of NLP and Exam Review

1 EXPECTATIONS

I DO 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
- There will be **NO NON-programming option** for assignments
- But it's more than just programming:
  - algorithms, mathematics and statistics
  - intuition about language
  - analytical thinking



1 EXPECTATIONS

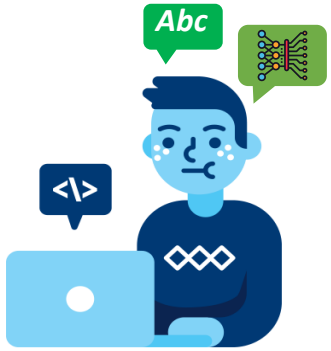
I DO NOT assume you are a linguist

- But you do **need to know core concepts, e.g., what nouns and verbs are.**
- We will think critically about **how we use language**
- and about how computational models capture **aspects of language**

1 EXPECTATIONS

I DO NOT assume you are a deep learning researcher

- But you will learn and use many concepts from machine learning
- We will think critically **how to use text data and embeddings**
- and about how deep learning models capture **aspects of language**



1 The NLP Big Picture

Natural Language Processing: Overview (partial)

Application

Analysis | ***Generation***

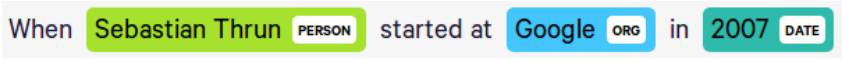
Sentiment | ***Summarisation***

Topics | ***Translation***

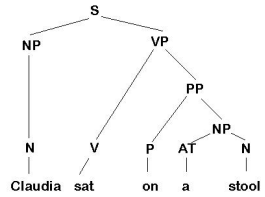
Dialog

NLP Tools

Entity Extraction — When Sebastian Thrun …



Parsing — Claudia sat on a stool



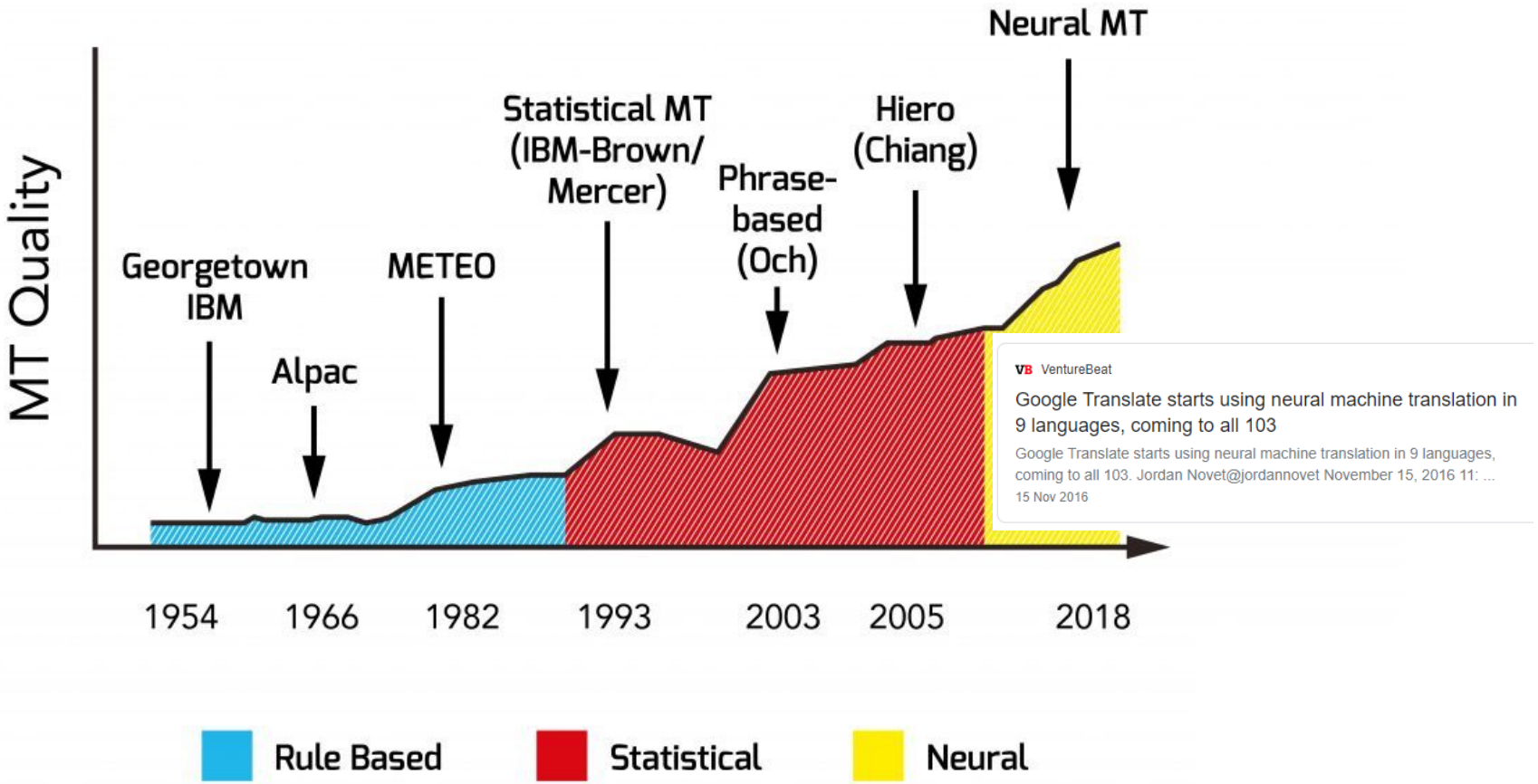
PoS Tagging — She sells seashells — **[she/PRP] [sells/VBZ] [seashells/NNS]**

Stemming — Drinking, Drank, Drunk — **Drink**

Tokenisation — How is the weather today — **[How] [is] [the] [weather] [today]**

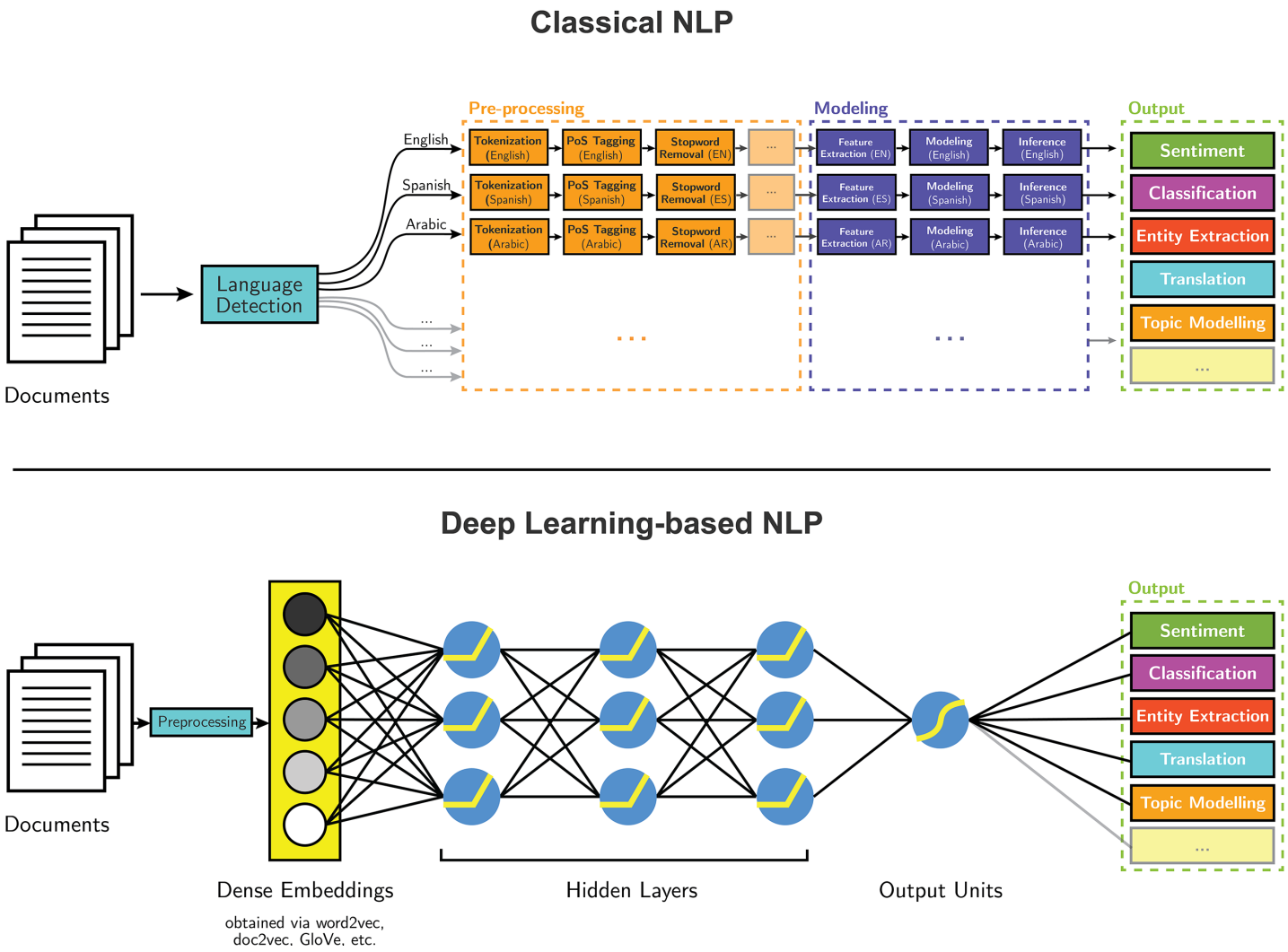
1 The NLP ERA

NLP Techniques – with the Trend of Machine Translation



The NLP ERA

Language Modelling using Deep Learning-based NLP Techniques



1 ASSESSMENT

Assessment Overview

Assessment	Weight	Due
Lecture Miuddy Cards	5%	Every Week
Lab Exercise	5%	Week 3, 5, 7, 9, 11
Assignment 1	20%	Week 4 (Friday 11.59pm – AU time)
Assignment 2	20%	Week 12 (Friday 11.59pm – AU time)
Final Exam	50%	Exam Period

Lab Exercises

- Programming tasks done in fortnightly computer labs, but only six are assessed

Assignments

- Take place throughout the teaching period
- Implementation and Documentation

1 ASSESSMENT

Lecture Muddy Cards – 5%

At the end of each lecture you will answer one question:

“Please reflect on the lecture and write down what was most unclear, in ~1 sentence.”

Good examples (full credit):

"The derivation of the complexity of quicksort.”

“Why do we use log probability rather than just probability?”

Bad examples (0 credit):

“Slide 15”

“Word Embeddings”

1 ASSESSMENT

Lecture Muddy Cards – 5%

Why?

Reflection increases retention

Helps you note down what you want to revise later

Tells me what to revisit (either on Ed, the next lecture, or in a pre-exam review)

1 ASSESSMENT

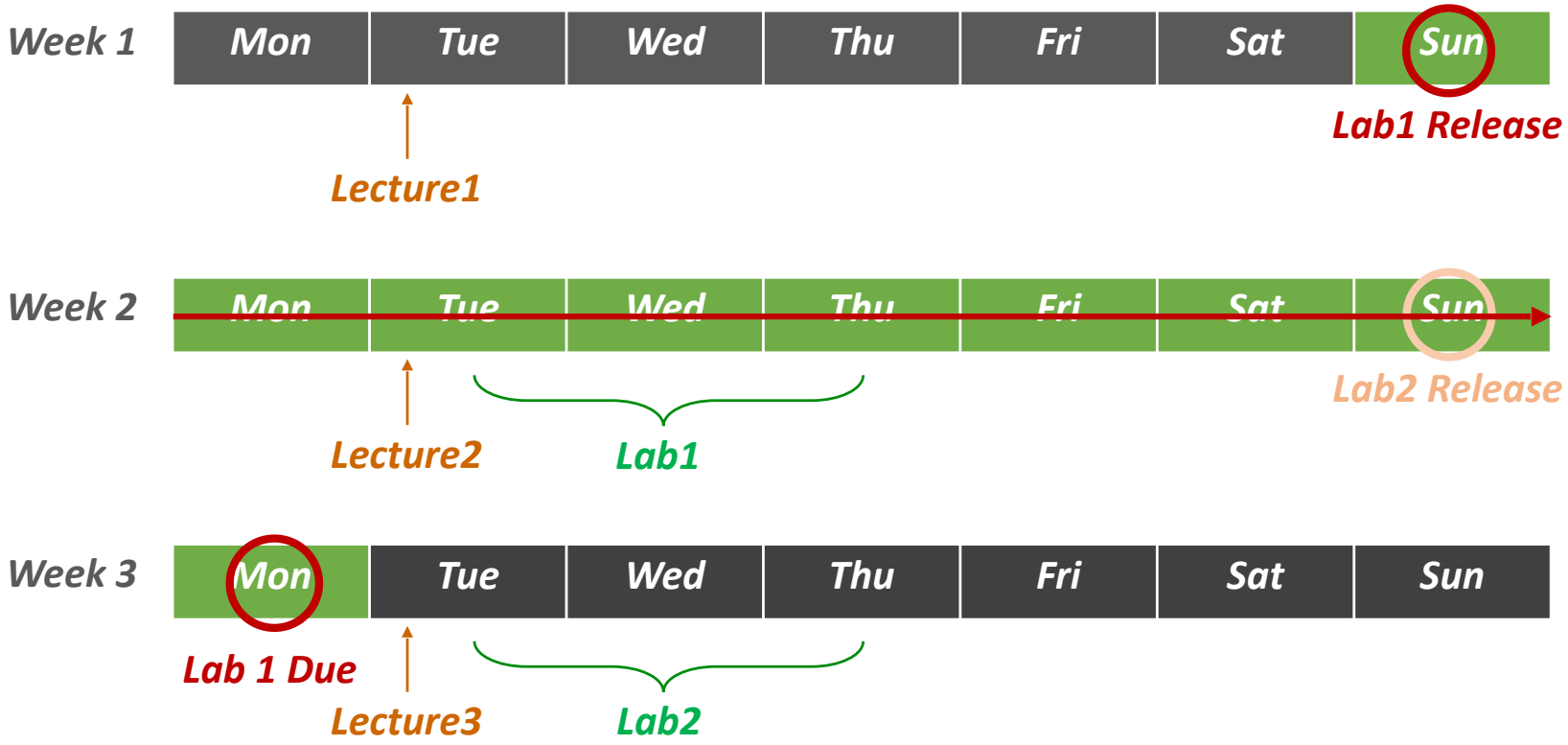
Lab Exercises – 5%

- Small tasks, some questions, some coding.
- Based on what you learned in the previous lectures and labs.
- Each one is worth 2%, we ignore your two lowest scores and count half of your third lowest

1 ASSESSMENT

Lab Exercises – 5%

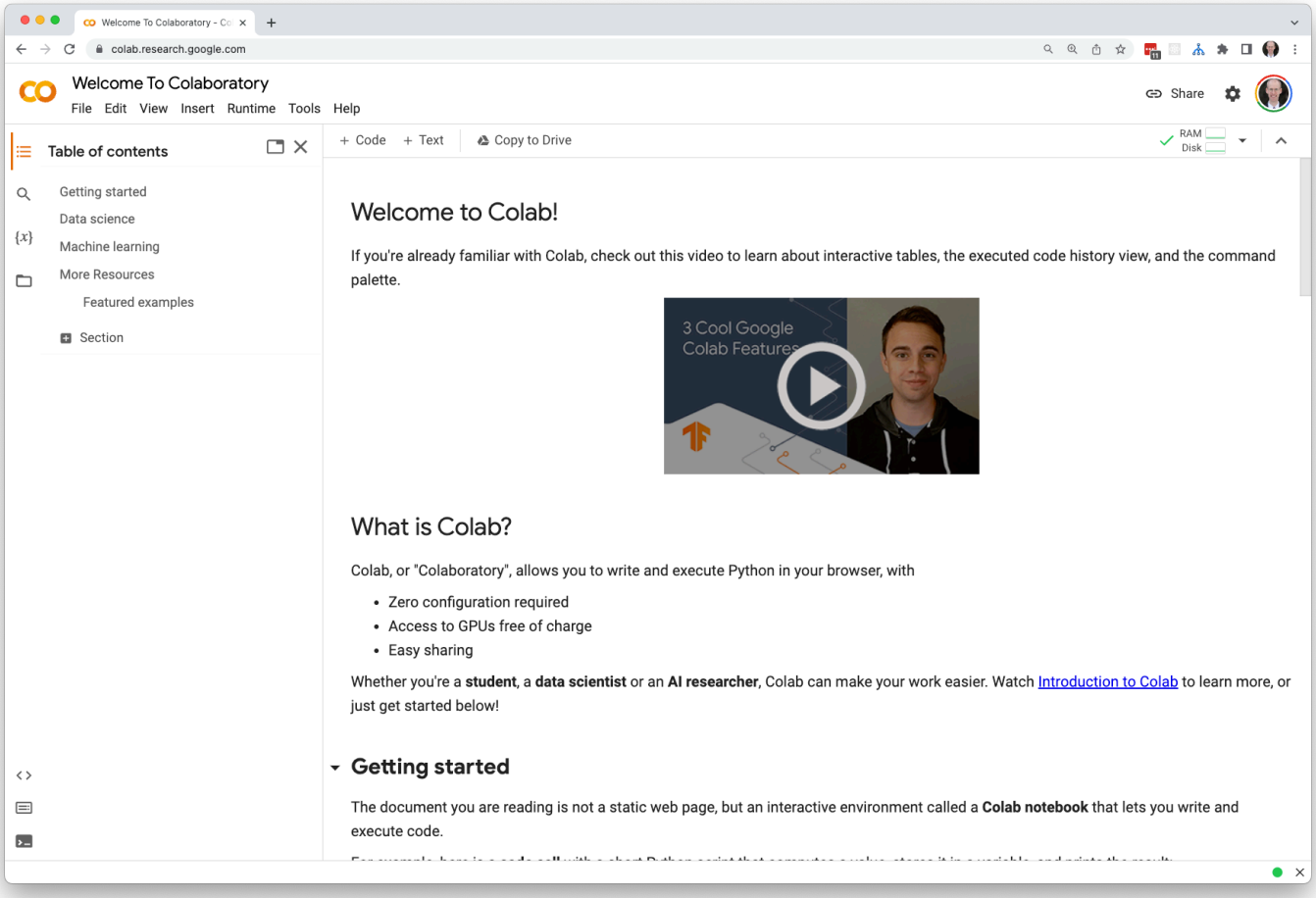
- When to submit the Fortnightly Lab Exercise (e.g. Lab1 Release and Submission)



1 ASSESSMENT

What do we do during Labs?

Following walkthroughs and writing code using Google Colab.



https://colab.research.google.com/notebooks/intro.ipynb

1 ASSESSMENT

Assignment 1 (20%)

Released by next Monday

Due the end of week 4

***NOTE**: Assignment 1 will be an individual assignment.*

Assignment 2 (20%)

Released by the end of week 6

Due the end of week 12

***NOTE**: Assignment 2 can either be completed on your own or a group of 2-3.*

Final Exam (50%)

Must score at least 40% to pass.

Following university-wide policy:

CC - Written exam during the exam period (2 hours long)

RE - ProctorU Live+

1 ASSESSMENT

ASSESSMENT Due Dates

- Week 1: Introduction to Natural Language Processing (NLP)
- Week 2: Word Embeddings
- Week 3: Word Classification with Machine Learning I
- Week 4: Word Classification with Machine Learning II

NLP and Machine Learning

Assignment 1 Due

- Week 5: Fundamentals of Language
- Week 6: Part of Speech Tagging
- Week 7: Dependency Parsing
- Week 8: Language Models
- Week 9: [ANZAC Day]
- Week 10: Named Entity Recognition and Coreference Resolution
- Week 11: Question Answering
- Week 12: Machine Translation

NLP Tasks

Assignment 2 Due

- Week 13: Future of NLP and Exam Review

1 ASSESSMENT

Start assignments early!

- All assignments involve coding and report writing. Make sure you leave time for the report!
- If you are stuck, ask (and the sooner you start, the more time you will have for help if you are stuck)
- Assignments will be **very different from last year's**
- Reports will be submitted to Turnitin through Canvas
- Code is also submitted and retained
- Clearly reference any copied/adapted code

1 ASSESSMENT

Special Consideration

- If your performance on assessments is affected by illness or misadventure
- Follow proper bureaucratic procedures
- Have professional practitioner sign special USyd form
- Submit application for special consideration online, upload scans
- Note you have only a quite short deadline for applying
- http://sydney.edu.au/current_students/special_consideration/
- ***Also, notify coordinator by email as soon as anything begins to go wrong***
- There is a similar process if you need special arrangements e.g. for religious observance, military service, representative sports

THE UNIVERSITY OF SYDNEY

Do you have a disability?

You may not think of yourself as having a ‘disability’ but the definition under the Disability Discrimination Act (1992) is broad and includes temporary or chronic medical conditions, physical or sensory disabilities, psychological conditions and learning disabilities.

The types of disabilities we see include:

Anxiety // Arthritis // Asthma // Autism // ADHD // Bipolar disorder // Broken bones // Cancer // Cerebral palsy // Chronic fatigue syndrome // Crohn’s disease // Cystic fibrosis // Depression Diabetes // Dyslexia // Epilepsy // Hearing impairment // Learning disability // Mobility impairment // Multiple sclerosis // Post-traumatic stress // Schizophrenia // Vision impairment and much more.

Students needing assistance must register with Disability Services. It is advisable to do this as early as possible. Please contact us or review our website to find out more.

Disability Services Office - sydney.edu.au/disability - 02-8627-8422

1 ASSESSMENT

AI in assignments (e.g., ChatGPT, Copilot)

- Cool technology that is the result of research in NLP!
- NOT to be used in assignments or labs
    - Spell-checking is okay
    - More classical debugging tools and coding assistance tools okay

Why? The labs are about you learning the material, which requires you to be the one to write the code. The assignments test what you know, not how good you are at running AI models.

1 ASSESSMENT

Academic Integrity

http://sydney.edu.au/elearning/student/EI/index.shtml:

> "The University of Sydney is unequivocally opposed to, and intolerant of, plagiarism and academic dishonesty.
>
> Academic dishonesty means seeking to obtain or obtaining academic advantage for oneself or for others (including in the assessment or publication of work) by dishonest or unfair means.
>
> Plagiarism means presenting another person's work as one's own work by presenting, copying or reproducing it without appropriate acknowledgement of the source."

Submitted work is compared against other work (from students, the internet, etc)

TurnitIn for textual tasks (through Canvas), other systems for code

Penalties for academic dishonesty or plagiarism can be severe

1 ASSESSMENT

THE UNIVERSITY OF SYDNEY

Academic Integrity

Understanding General Concepts	Explained using similar material (not assignment)	Sharing approach/ concept to derive assignment solution	Designing code/ solution	Implementing code/ solution

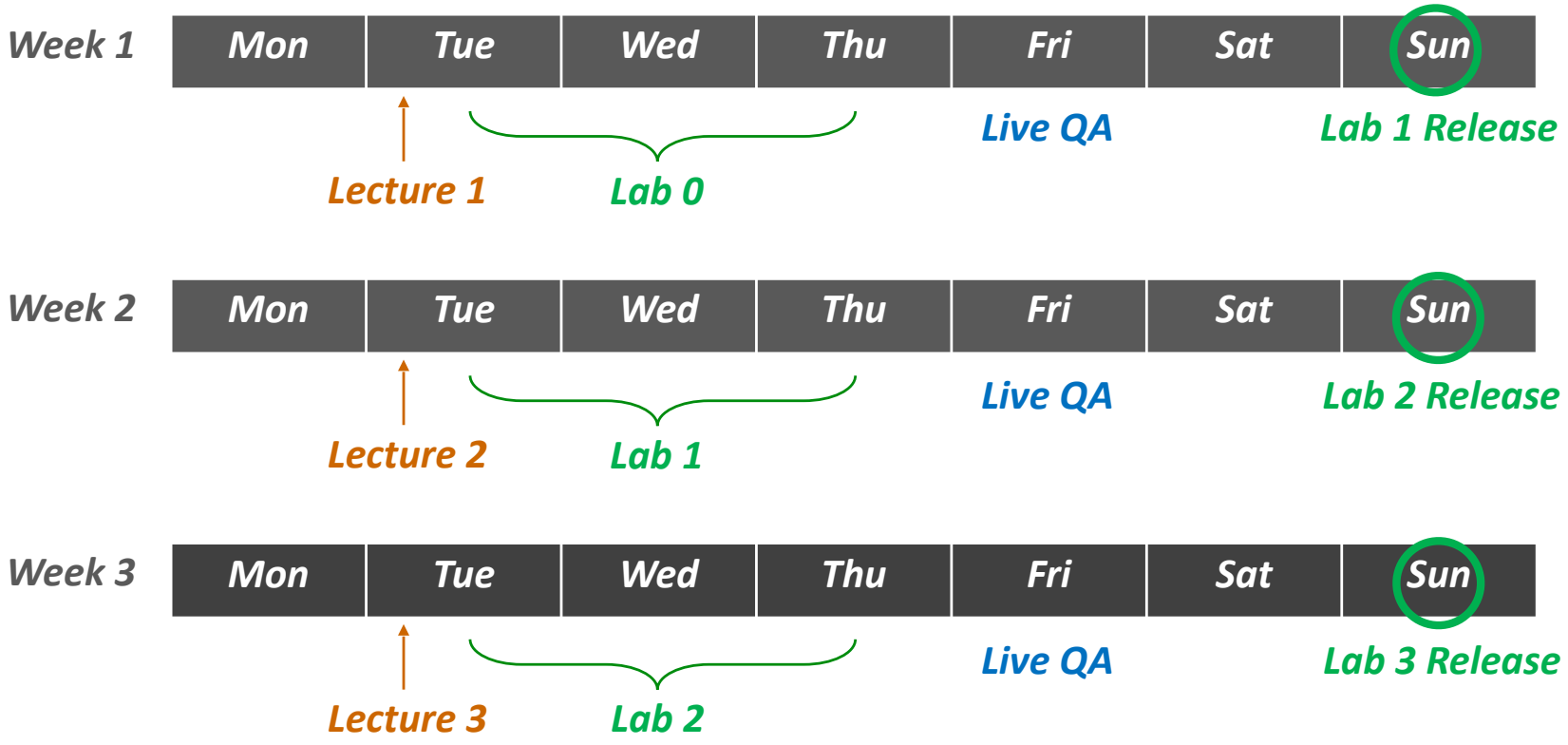
- Encouraged
- Attribution required
- Not acceptable
- Ask Lecturer/Coordinator

1 TIMETABLE

- 12 hours of work per week (including 3 contact hours)
- Attend 2 hours of lectures per week:
  - Tuesday 5 – 7pm
  - Lectures are recorded, but don't depend on it!
- Attend 1 hour of tutorial / lab time
- Participate respectfully in discussions in lectures and labs
- Complete all assessment tasks on time

1 TIMETABLE

Classes and Release Date



1 TIMETABLE

Full Course Timetable on the Canvas page

- **Lecture:** Tue 5-7pm
- **Tutorial:** Tue / Wed / Thur
- **LiveQA**: Fri 1-2pm

Please **ask for help** if you are falling behind

1 Communication

Where to ask questions / make requests?

Check Canvas first (https://canvas.sydney.edu.au/courses/48399/pages/comp-4446-slash-5046)

Course Content

1. Lecture, use mentimeter
please keep these focused on the content being presented in that lecture

2. Tutorial / Lab

3. Ed, https://edstem.org/au/courses/10943/discussion/
please wait to ask questions about lab content until you have your lab

4. Office Hours / Live QA: Friday 1-2pm on Zoom

Admin

- Email me with [COMP 5046] in the subject: jonathan.kummerfeld@sydney.edu.au

1 Consultative Group

What is it?

Brief meeting after each lecture.

A chance to give feedback, raise issues, etc

One student from each tutorial (volunteer selected in week one lab).

1 Textbooks (Optional)

THE UNIVERSITY OF SYDNEY

No required readings, but these are some good resources:

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

https://web.stanford.edu/~jurafsky/slp3/

Natural Language Processing

Jacob Eisenstein

https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

1 SELF TEST

Ask yourself…

- 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 programming language do you need to know?

Break

[Stay warm, little flappers, and find lots of plant eggs!]

Source: https://xkcd.com/1322/

0 LECTURE PLAN

THE UNIVERSITY OF SYDNEY

Lecture 1: Introduction to Natural Language Processing

1. Course Introduction
2. **Overview of Natural Language Processing (NLP)**
3. Word Meaning and Representation
4. Count-based Word Representation
   - One-hot Encoding
   - Bag of Words
   - Term Frequency-Inverse Document Frequency
5. Next Week Preview

2 LANGUAGE

Why do we care about language?

- **language stores knowledge**
- language communicates knowledge
- language is a key part of culture and human experience
- language is a natural interface for humans

2 LANGUAGE

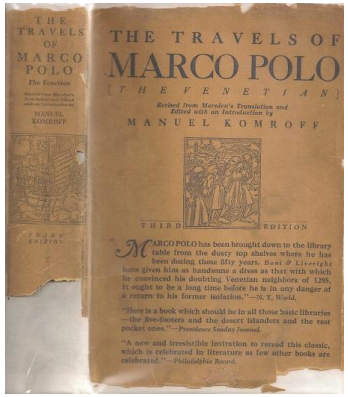
Why do we care about language?

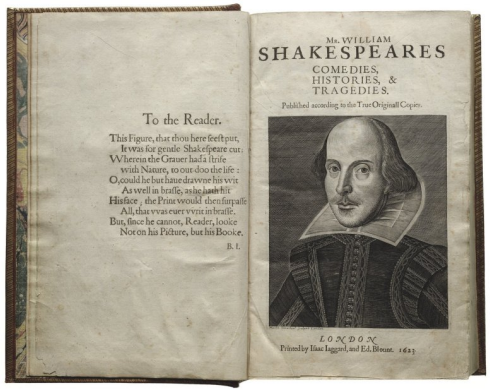
- language stores knowledge
- **language communicates new knowledge**
- language is a key part of culture and human experience
- language is a natural interface for humans

2 LANGUAGE

Why do we care about language?

- language stores knowledge
- language communicates new knowledge
- **language is a key part of culture and human experience**
- language is a natural interface for humans

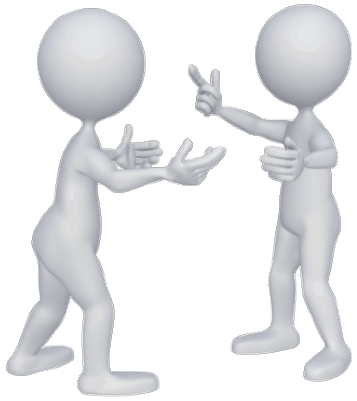




2 LANGUAGE

Why do we care about language?

- language stores knowledge
- language communicates new knowledge
- language is a key part of culture and human experience
- **language is a natural interface for humans**



2 Natural Language Processing (NLP)

What is Natural Language Processing?

Natural Language Processing (NLP) is the branch of artificial intelligence focused on making computers capable of understanding and using language.

NLP vs. Computational Linguistics?

Often used interchangeably

One view is that Computational Linguistics is about using computers and mathematical tools to aid the study of language

2 Natural Language Processing (NLP)

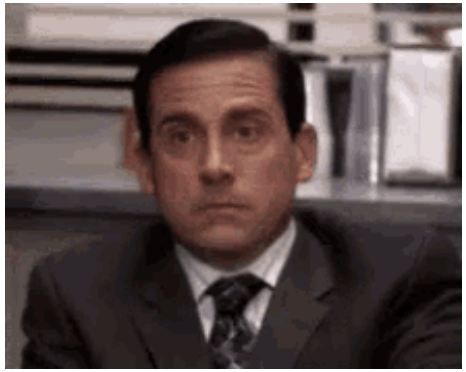
THE UNIVERSITY OF SYDNEY

Communication for the speaker:

- Intention: **Decide when and what information** should be transmitted. May require planning and reasoning about agents' goals and beliefs.
- Generation: **Translate the information to be communicated** (ie., thoughts) into words in the desired natural language.
- Synthesis: **Output the words** in the desired modality - text or speech.

2 Natural Language Processing (NLP)

Communication for the hearer:



- Perception: **Map input modality to a string of words**, e.g. optical character recognition (OCR) or speech recognition

- Analysis: **Determine the information content of the string**.
  - Syntactic interpretation: Find the correct parse tree showing the phrase structure of the string.
  - Semantic Interpretation: Extract the meaning of the string .
  - Pragmatic Interpretation: Consider the effect of the overall context on altering the literal meaning of a sentence.

- Incorporation: Decide whether or not to **believe the content of the string** and update beliefs.

2 SPECIAL ABOUT NLP

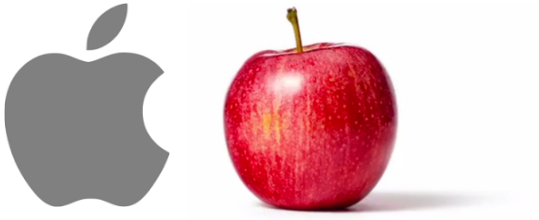
What is special about NLP?

- Human language is a system specifically constructed to **convey meaning** and is **not produced by a physical manifestation of any kind**. In that way, it is very different from vision or any other machine learning task.

- Most words are symbols that represent concepts. Put differently, the word is a signifier for an idea or thing.

“Computer”

“Apple”

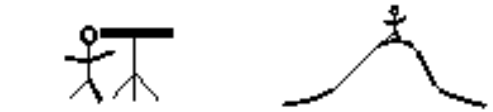


“Whaaaaaaa”

???????

2 SPECIAL ABOUT NLP

Ambiguity



I saw [the man on the hill] with a telescope.

I saw [the man [on the hill] with a telescope].

I saw [the man [on the hill with a telescope]].

I saw [the man] [on the hill] [with a telescope].

I saw the man [on the hill with a telescope].

Ambiguity is Ubiquitous

Speech Recognition

- “recognize speech” vs. “wreck a nice beach”
- “youth in Asia” vs. “euthanasia”

Text Processing

- “I ate spaghetti with chopsticks” vs. “I ate spaghetti with meatballs.”
- “The dog is in the pen.” vs. “The ink is in the pen.”
- “I put the plant in the window” vs. “Ford put the plant in Mexico”

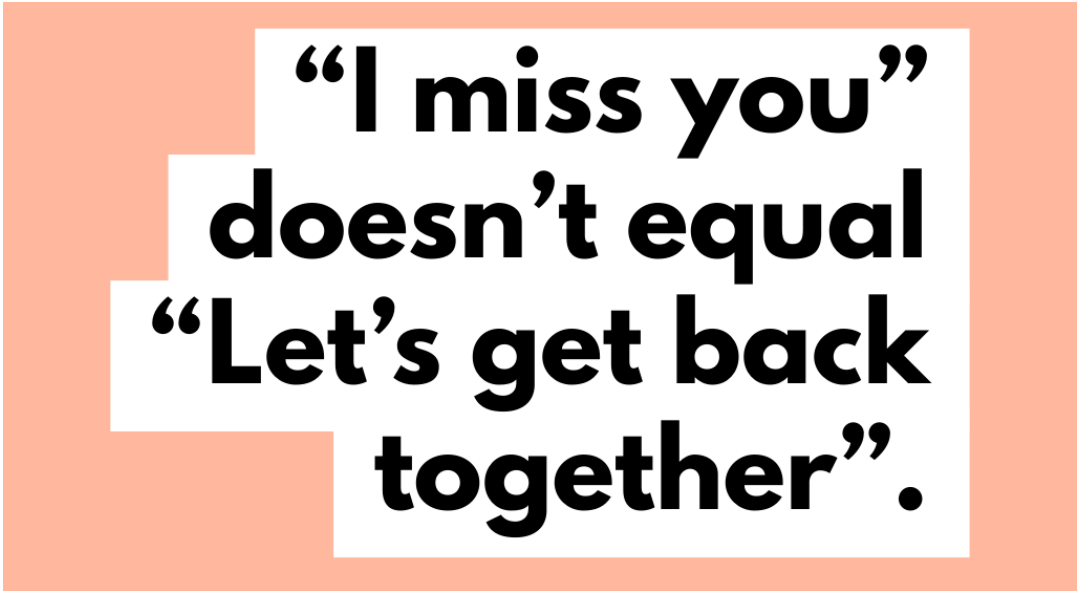
Ambiguity and Humour

- “Kids Make Delicious Snacks”
- “Government Head Seeks Arms”

2 SPECIAL ABOUT NLP

THE UNIVERSITY OF SYDNEY

Even people struggle with understanding



2 SPECIAL ABOUT NLP

THE UNIVERSITY OF SYDNEY

No NLP task is solved, but we do better on some

Very Effective

- Spell Checking
- Keyword Search
- Finding Synonyms

Pretty Good

- Extracting Information from documents (including websites)
- Constrained Dialogue

Getting Better

- Machine Translation
- Coreference Resolution
- Question Answering
- Open-ended Dialogue

0 LECTURE PLAN

THE UNIVERSITY OF SYDNEY

Lecture 1: Introduction to Natural Language Processing

1. Course Introduction
2. Overview of Natural Language Processing (NLP)
3. **Word Meaning and Representation**
4. Count-based Word Representation
   - One-hot Encoding
   - Bag of Words
   - Term Frequency-Inverse Document Frequency
5. Next Week Preview

How should we represent the meaning of a word?

Definition: meaning (Collins dictionary).

- the idea that it represents, and which can be explained using other words.
- the thoughts or ideas that are intended to be expressed by it.

One way of framing meaning is as symbols and ideas:

signifier (symbol) ⟺ signified (idea or thing)

“Apple”

THE UNIVERSITY OF SYDNEY

Relationships between words

- WordNet is a graph of relationships between words, e.g., synonyms
- DEMO - http://wordnetweb.princeton.edu/perl/webwn English

e.g. synonym sets containing “good”:

```
from nltk.corpus import wordnet as wn
poses = { 'n':'noun', 'v':'verb', 's':'adj
(s)', 'a':'adj', 'r':'adv'}
for synset in wn.synsets("good"):
  print("{}: {}".format(poses[synset.pos()],
", ".join([l.name() for l in
synset.lemmas()])))
```

```
noun: good
noun: good, goodness
noun: good, goodness
noun: commodity, trade_good, good
adj: good
adj (sat): full, good
adj: good
adj (sat): estimable, good, honorable, respectable
adj (sat): beneficial, good
adj (sat): good
adj (sat): good, just, upright
…
adverb: well, good
adverb: thoroughly, soundly, good
```

e.g. hypernyms of “panda”:

```
from nltk.corpus import wordnet as wn
panda = wn.synset("panda.n.01")
hyper = lambda s: s.hypernyms()
list(panda.closure(hyper))
```

```
[Synset('procyonid.n.01'),
Synset('carnivore.n.01'),
Synset('placental.n.01'),
Synset('mammal.n.01'),
Synset('vertebrate.n.01'),
Synset('chordate.n.01'),
Synset('animal.n.01'),
Synset('organism.n.01'),
Synset('living_thing.n.01'),
Synset('whole.n.02'),
Synset('object.n.01'),
Synset('physical_entity.n.01'),
Synset('entity.n.01')]
```

Problems with resources like WordNet

- Great as a resource but missing nuance
  - e.g. “proficient” is listed as a synonym for “good”. This is only correct in some contexts.
- Missing new meanings of words
  - e.g., wicked, badass, nifty, wizard, genius, ninja, bombast
  - Impossible to keep up-to-date!
- Subjective
- Requires extensive human labor to create and adapt
- Needs to be repeated for each language

0 LECTURE PLAN

THE UNIVERSITY OF SYDNEY

Lecture 1: Introduction to Natural Language Processing

1. Course Introduction
2. Overview of Natural Language Processing (NLP)
3. Word Meaning and Representation
4. **Count-based Word Representation**
   - One-hot Encoding
   - Bag of Words
   - Term Frequency-Inverse Document Frequency
5. Next Week Preview

COUNT based WORD REPRESENTATION

One-hot Encoding

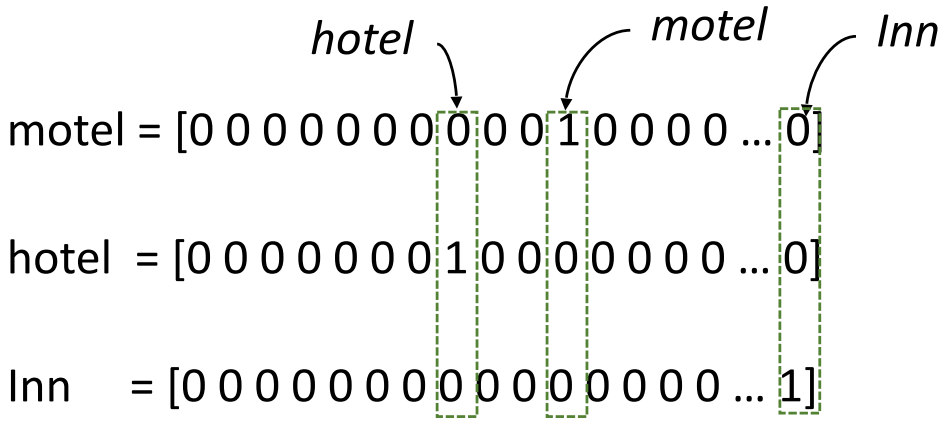
- In traditional NLP, we regard words as discrete symbols.

Hot (True) Cold (False)

Means one 1, the rest 0s

Words can be represented by **one-hot vector**s:

1. The categorical values be mapped to integer values (index)
2. each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.



Vector dimension = number of words in vocabulary

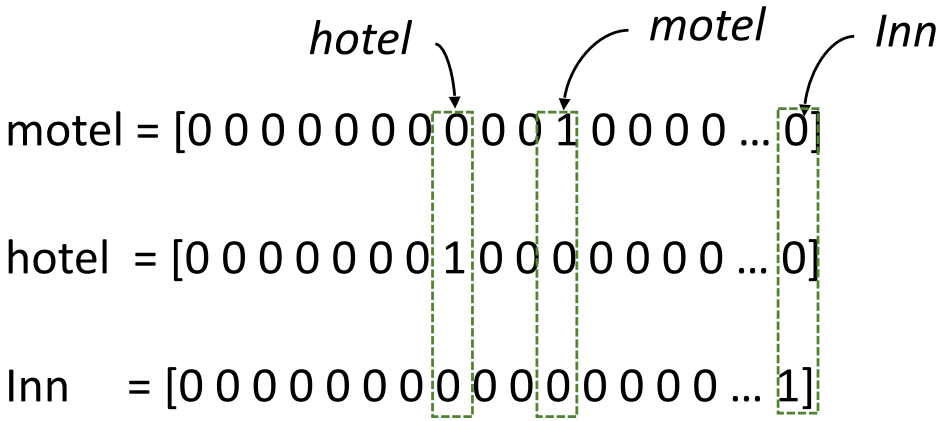
In practise, stored as sparse vectors (ie, either a hash table, or a list of word IDs

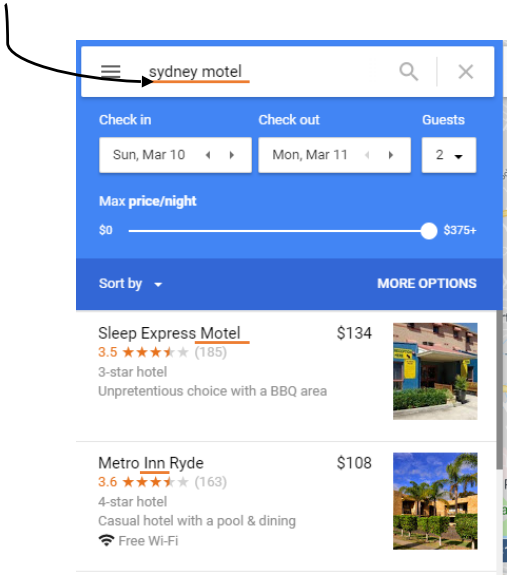
4 COUNT based WORD REPRESENTATION

THE UNIVERSITY OF SYDNEY

No word similarity representation

Example: in web search, if user searches for “Sydney motel”, we would like to match documents containing “Sydney Inn”





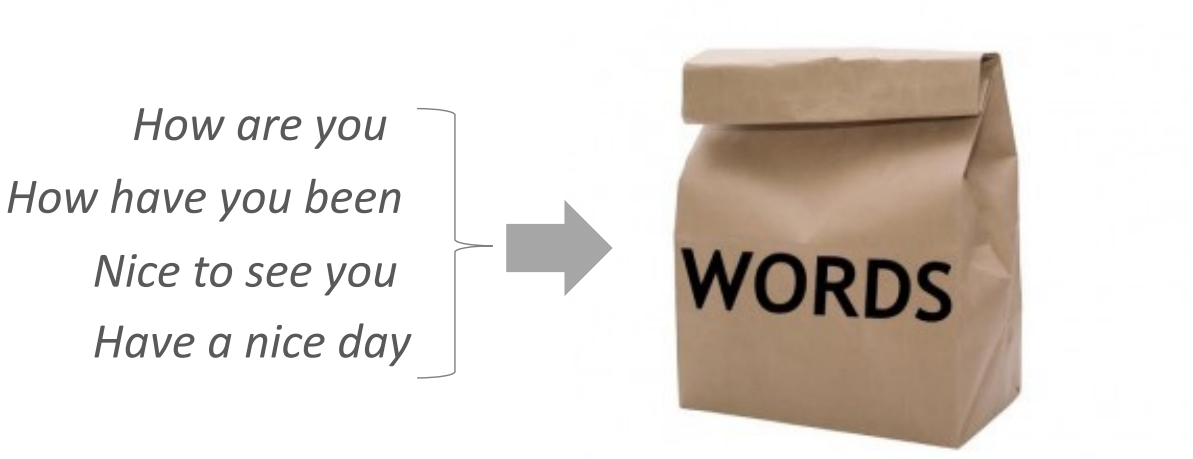
There is no natural notion of similarity for one-hot vectors!

4 COUNT based WORD REPRESENTATION

Bag of Words (BOW)

- A bag-of-words model (BoW) is a representation of text that describes **the occurrence of words** within a document. It involves two things:
  - A vocabulary of known words.
  - A measure of the presence of known words.

- It is called a "**bag**" of words, because any information about the **order or structure of words in the document is discarded.** The model is only concerned with whether known words occur in the document.

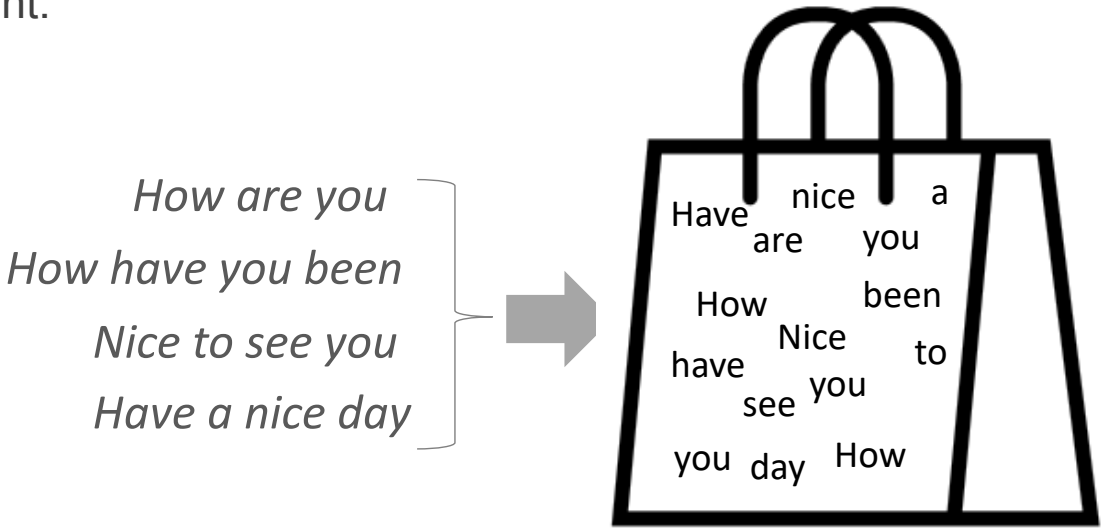


COUNT based WORD REPRESENTATION

Bag of Words (BOW)

- A bag-of-words model (BoW) is a representation of text that describes **the occurrence of words** within a document. It involves two things:
  - A vocabulary of known words.
  - A measure of the presence of known words.

- It is called a "**bag**" of words, because any information about the **order or structure of words in the document is discarded.** The model is only concerned with whether known words occur in the document.



4 COUNT based WORD REPRESENTATION

THE UNIVERSITY OF SYDNEY

Bag of Words (BOW)

A vocabulary of known words

a	are	been	day	have	how	nice	see	to	you

* **WO** = occurrence of words

[WO*a*, WO*are*, WO*been*, WO*day*, WO*have*, WO*how*, WO*nice*, WO*see*, WO*to*, WO*you*]

How are you = [0, 1, 0, 0, 0, 1, 0, 0, 0, 1]

How have you been = [0, 0, 1, 0, 1, 1, 0, 0, 0, 1]

Nice to see you = [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]

Have a nice day = [1, 0, 0, 1, 1, 0, 1, 0, 0, 0]

***Remember, I use complete vectors here just to explain the concept. In practise, a form of sparse vector is used.**

Total Frequency = [1, 1, 1, 1, 2, 2, 2, 1, 1, 3]

a	are	been	day	have	how	nice	see	to	you
1	1	1	1	2	2	2	1	1	3

4 COUNT based WORD REPRESENTATION

THE UNIVERSITY OF SYDNEY

Why use BoW?

- The intuition is that documents are similar if they have similar content. Further, that from the content alone we can learn something about the meaning of the document.

Problem with BoW

- Discarding word order ignores the context, which can dramatically change the meaning.

*S1= I **love** you but you **hate** me*

*S2= I **hate** you but you **love** me*

COUNT based WORD REPRESENTATION

Term Frequency-Inverse Document Frequency

- Term Frequency-Inverse Document Frequency (TF-IDF) is a way of representing ***how important a word*** *is to a document in a collection or corpus*.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$w_{i,j}$ = *weight of term i in document j*

$tf_{i,j}$ = *number of occurrences of term i in document j*

N = *total number of documents*

df_i = *number of documents containing term i*

- The ***Term Frequency*** is a count of how many times a word occurs in a given document
- The ***Document Frequency*** is the number of times a word occurs in a corpus of documents

4 COUNT based WORD REPRESENTATION

Term Frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = *number of occurrences of term i in document j*

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

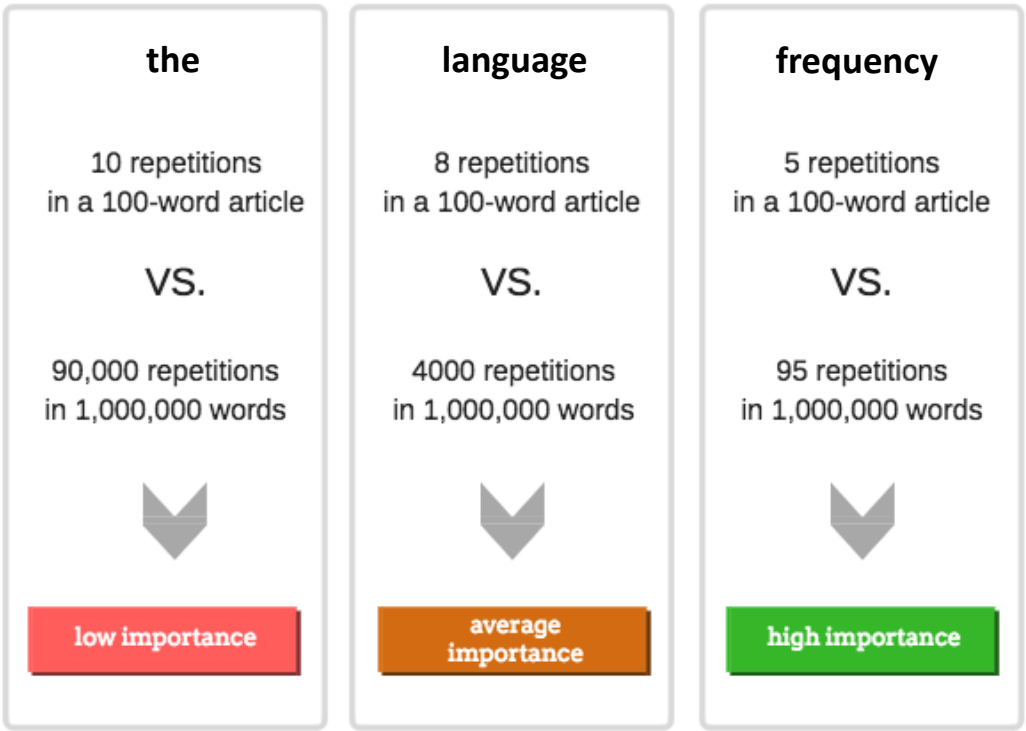
	and	apple	banana	fruit	I	like	sweet	yellow
D#1	0	1	0	0	1	1	0	0
D#2	0	0	1	0	1	1	0	0
D#3	1	0	2	0	0	0	1	1
D#4	0	0	0	1	0	0	1	0

4 COUNT based WORD REPRESENTATION

THE UNIVERSITY OF SYDNEY

What if we just use Term Frequency Only?

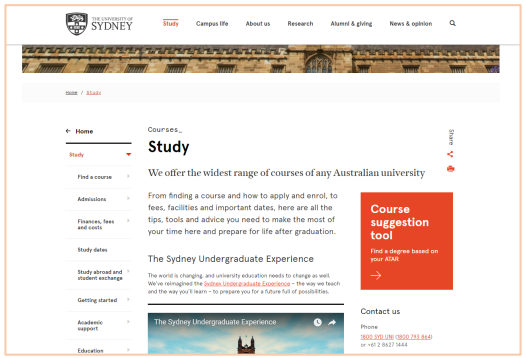
- It is known that certain terms, such as "is", "of", and "that", may appear many times but not be informative. TF/IDF is a principled way to reweigh words based on their use throughout a document collection



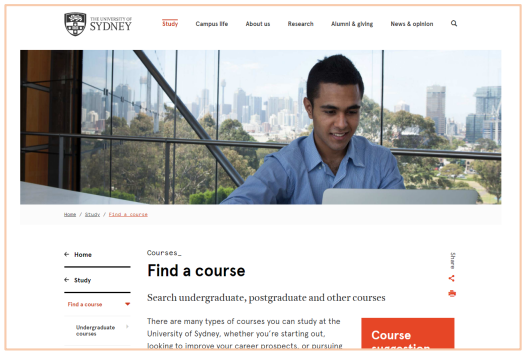
4 COUNT based WORD REPRESENTATION

What if we just use Term Frequency Only?

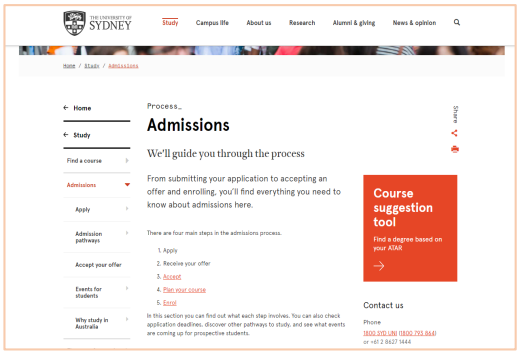
University of Sydney Website



Webpage#1



Webpage#2



Webpage#3

4 COUNT based WORD REPRESENTATION

Inverse Document Frequency

Why do we need log?

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

N = total number of documents

df_i = number of documents containing term i

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

N = 4

	and	apple	banana	fruit	I	like	sweet	yellow
df	1	1	2	1	2	2	2	1

THE UNIVERSITY OF SYDNEY

Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

N = total number of documents

df_i = number of documents containing term i

With log

$$idf(d,t) = log(n/df(t))$$

$$n = 1,000,000$$

	$df(t)$	$idf(d,t)$
word1	1	6
word2	100	4
word3	1,000	3
word4	10,000	2
word5	100,000	1
word6	1,000,000	0

Without log

$$idf(d,t) = n/df(t)$$

	$df(t)$	$idf(d,t)$
word1	1	1,000,000
word2	100	10,000
word3	1,000	1,000
word4	10,000	100
word5	100,000	10
word6	1,000,000	1

4 COUNT based WORD REPRESENTATION

THE UNIVERSITY OF SYDNEY

Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$1+df_i$ *sometimes, why?*

N = total number of documents

df_i = number of documents containing term i

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

N = 4

	and	apple	banana	fruit	I	like	sweet	yellow
df	1	1	2	1	2	2	2	1

COUNT based WORD REPRESENTATION

Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$1+df_i$

N = total number of documents

df_i = number of documents containing term i

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

N = 4

	and	apple	banana	fruit	I	like	sweet	yellow
df	1	1	2	1	2	2	2	1
idf (with 1+*dfi*)	Inv(4/(1+1)) =0.693147	Inv(4/(1+1)) =0.693147	Inv(4/(2+1)) =0.287682	Inv(4/(1+1)) =0.693147	Inv(4/(2+1)) =0.287682	Inv(4/(2+1)) =0.287682	Inv(4/(2+1)) =0.287682	Inv(4/(1+1)) =0.693147

4 COUNT based WORD REPRESENTATION

Term Frequency Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$ ← $1+df_i$

$w_{i,j}$ = *weight of term i in document j*

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

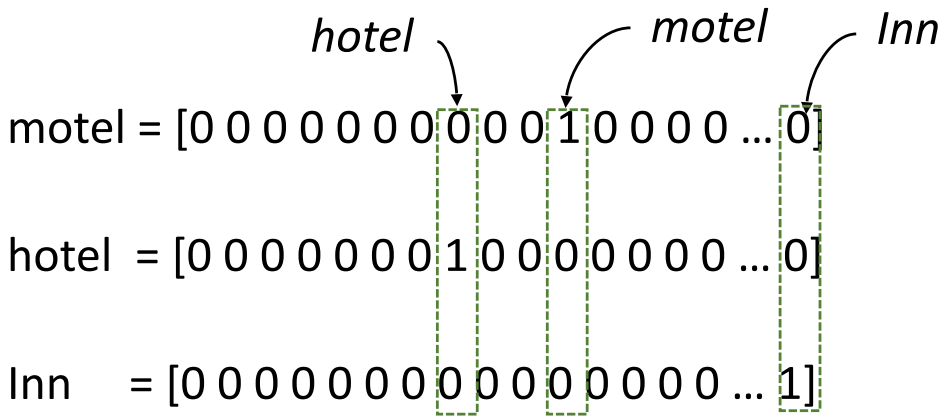
Document #4: Sweet fruit

	and	apple	banana	fruit	I	like	sweet	yellow
D#1	0	0.693147	0	0	0.287682	0.287682	0	0
D#2	0	0	0.287682	0	0.287682	0.287682	0	0
D#3	0.693147	0	0.575364	0	0	0	0.287682	0.693147
D#4	0	0	0	0.693147	0	0	0.287682	0

COUNT based WORD REPRESENTATION

Sparse Representation

The representations we have discussed are considered ‘sparse’ because the vector is mainly 0s:



In contrast, ‘dense’ representations have fewer dimensions (typically) but each word has a representation with various values in every element of the vector.

0 LECTURE PLAN

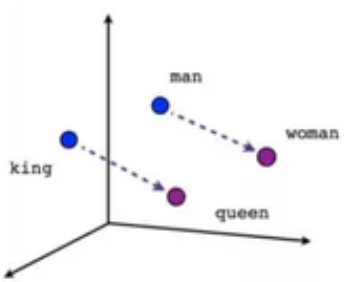
THE UNIVERSITY OF SYDNEY

Lecture 1: Introduction to Natural Language Processing

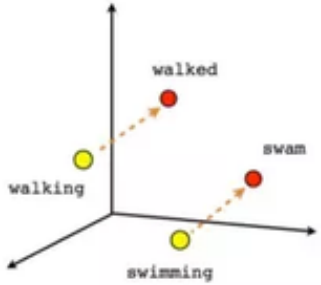
1. Course Introduction
2. Overview of Natural Language Processing (NLP)
3. Word Meaning and Representation
4. **Count-based Word Representation**
   - One-hot Encoding
   - Bag of Words
   - Term Frequency-Inverse Document Frequency
5. Next Week Preview

5 NEXT WEEK PREVIEW…

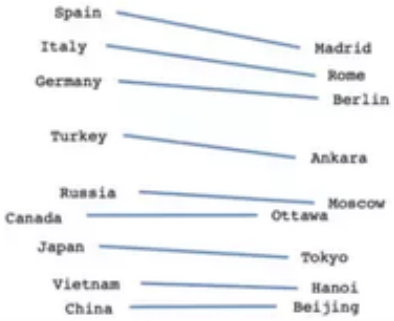
Dense vectors and how they capture similarity!



Male-Female



Verb tense



Country-Capital

Word Algebra

Enter all three words, the first two, or the last two and see the words that result.

shanghai + (australia - sydney) = Get result

china 0.7477672216910414

Reference: http://turbomaze.github.io/word2vecjson/

5 CS All Student & Staff 2023 - Meet & Greet

Tuesday 28th February
3:30 - 5:30 PM
J12 - Wintergarden

Join and meet your School of Computer Science lecturers, staff, and peers

Everyone is welcome!