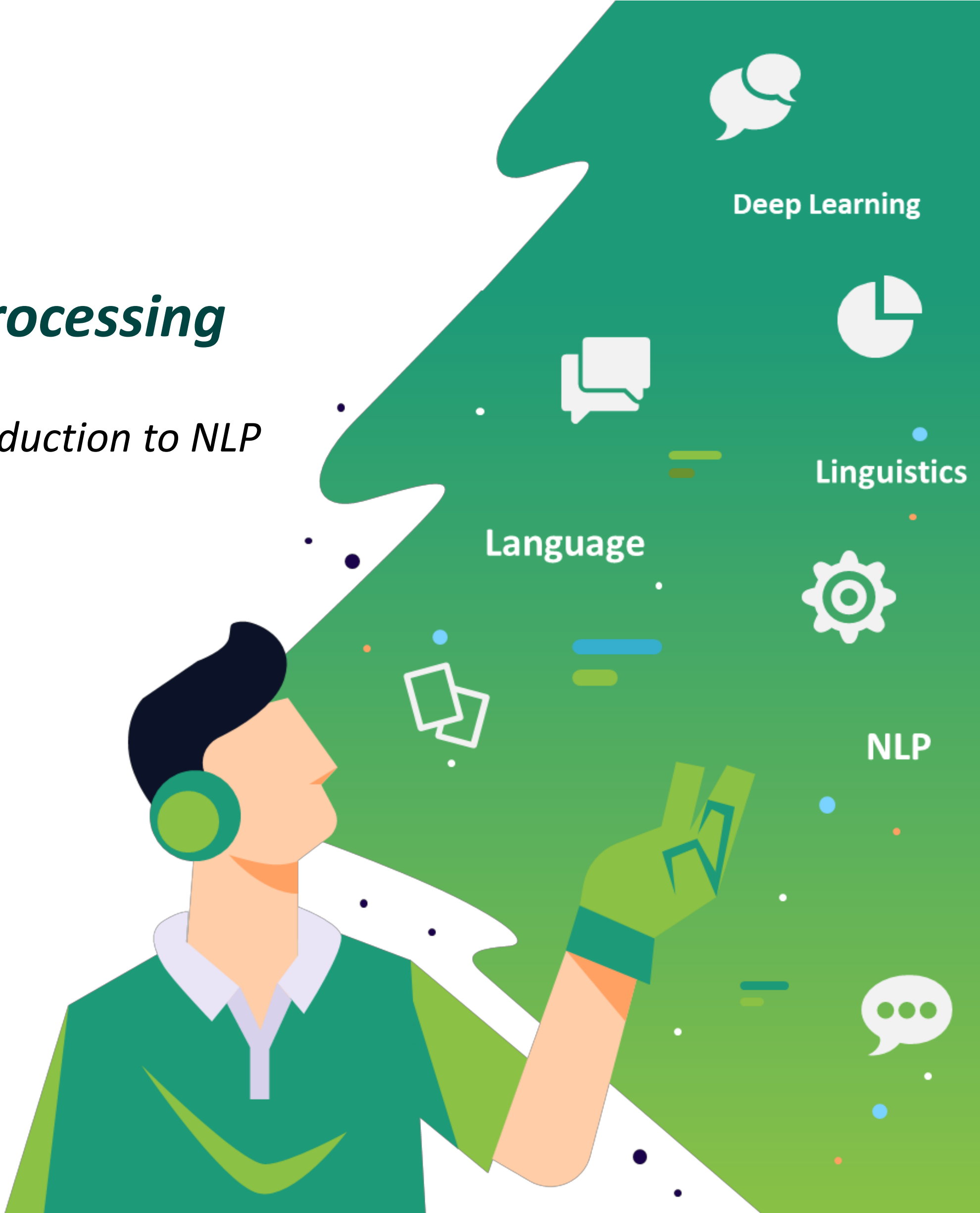
COMP5046
Natural Language Processing

Lecture 1: Course Info & Introduction to NLP



Dr. Caren Han

Semester 1, 2022
School of Computer Science,
University of Sydney

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*

1 INTRODUCTION

Dr Caren Han

Education

- B Computer Science (1st Class Honours, Medal)
- PhD Computer Science (Artificial Intelligence)

Teaching

- Received **Australian Young Achiever Teaching Excellence Award 2018**
- Received **Teacher of the Year 2020 Award**
- Received **Dean's Outstanding Teaching Award 2021**
- Teaching Natural Language Processing, Machine Learning, and Introduction to Programming

Research& Programming Experience

- Published 46 papers (conference and journal articles)
- Received **NLP Top-tier Conferences Best Paper Award/Best Area Paper Award**
- Worked in Industry Projects**: Google, Australia Defence Department, Thales, U.S. Air Force, U.S. Navy, NASA, Samsung, Hyundai etc.**

1 INTRODUCTION

COMP5046 Natural Language Processing

This unit introduces computational linguistics and the statistical techniques and algorithms used to automatically process natural languages. It will review the core statistics and information theory, and the basic linguistics, required to understand natural language processing (NLP).

NLP is used in a wide range of applications, including information retrieval and extraction; question answering; machine translation; and classifying and clustering of documents. This unit will explore the key challenges of natural language to computational modelling, and the state-of-the-art approaches to the key NLP sub-tasks, including tokenisation, morphological analysis, word sense representation, part-of-speech tagging, named entity recognition and other information extraction.

Students will implement many of these sub-tasks in labs and assignments, that can be used in the real-world cases. The unit will also investigate the annotation process that is central to creating training data for interesting application. With this unit, students can develop the innovative application that can be used in the real world.

Where to find the course information?

Unit Outline - COMP5046
https://www.sydney.edu.au/units/COMP5046

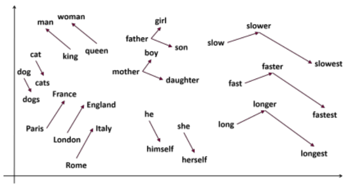
Canvas – COMP5046
https://canvas.sydney.edu.au/courses/39694

1 INTRODUCTION

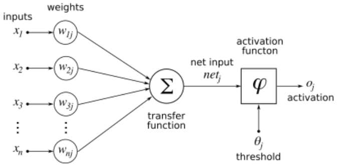
What will you learn in this course?

The focus of this course is on the review and comparison of models and methods that have achieved state-of-the-art results on various NLP tasks such as question answering (QA) and machine translation.

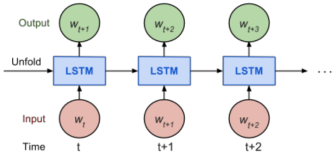
In this comprehensive review, students will get a detailed understanding of the past, present, and future of NLP. In addition, students will learn some of the current best practices for applying deep learning in NLP



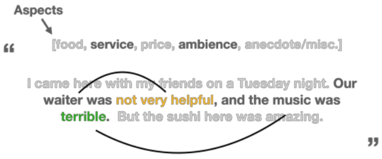
word2vec



NN

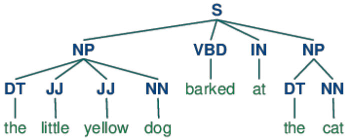


Language Modelling

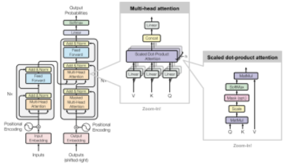

Aspects
[food, service, price, ambience, anecdote/misc.]
" I came here with my friends on a Tuesday night. Our waiter was not very helpful, and the music was terrible. But the sushi here was amazing. "

SpaCy's Dependency Parser

Dependency Parsing



Part-of-Speech Tagging



Transformer

What will you learn in this course?

NLP and Machine Learning

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

NLP Techniques

- Week 5: Language Fundamental
- Week 6: Part of Speech Tagging
- Week 7: Dependency Parsing
- Week 8: Language Model and Natural Language Generation

Advanced Topic

- Week 9: Information Extraction: Named Entity Recognition
- Week 10: Advanced NLP: Attention and Reading Comprehension
- Week 11: Advanced NLP: Transformer and Machine Translation
- Week 12: Advanced NLP: Pretrained Model in NLP

- 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 (esp.) statistics
  - linguistics and intuition about language
  - analytical thinking

1 EXPECTATIONS

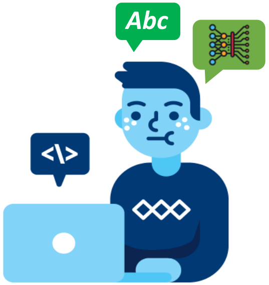
I DO NOT assume you are a linguist

- 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 computational models capture **aspects of language**

1 EXPECTATIONS

I DO NOT assume you are a deep learning researcher

- But you do need to know (really) **roughly how machine learning works**.
- We will think critically **how to use text data and embeddings**
- and about how deep learning models capture **aspects of language (context)**



1 The NLP Big Picture

The purpose of Natural Language Processing: Overview

Application

Understanding

- ***Sentiment Analysis***
- ***Topic Classification***
- ***Topic Modelling***
- ***....***

- ***Generation***
- ***Translation***
- ***Dialog***

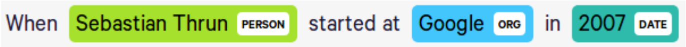
Searching

- ***Search***
- ***....***

NLP Stack

Entity Extraction

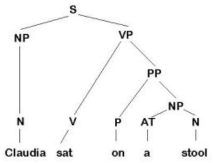
When Sebastian Thrun …



When Sebastian Thrun PERSON started at Google ORG in 2007 DATE

Parsing

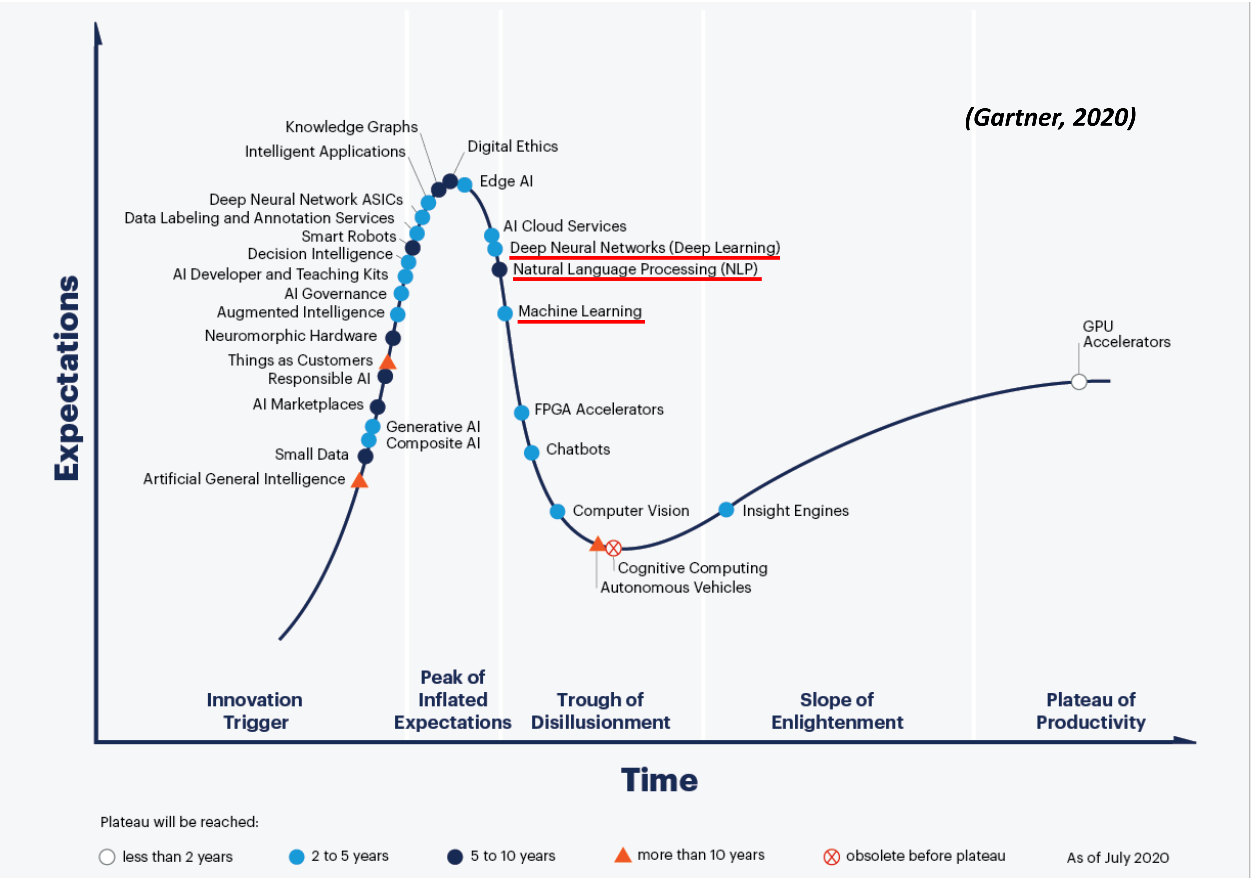
Claudia sat on a stool



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

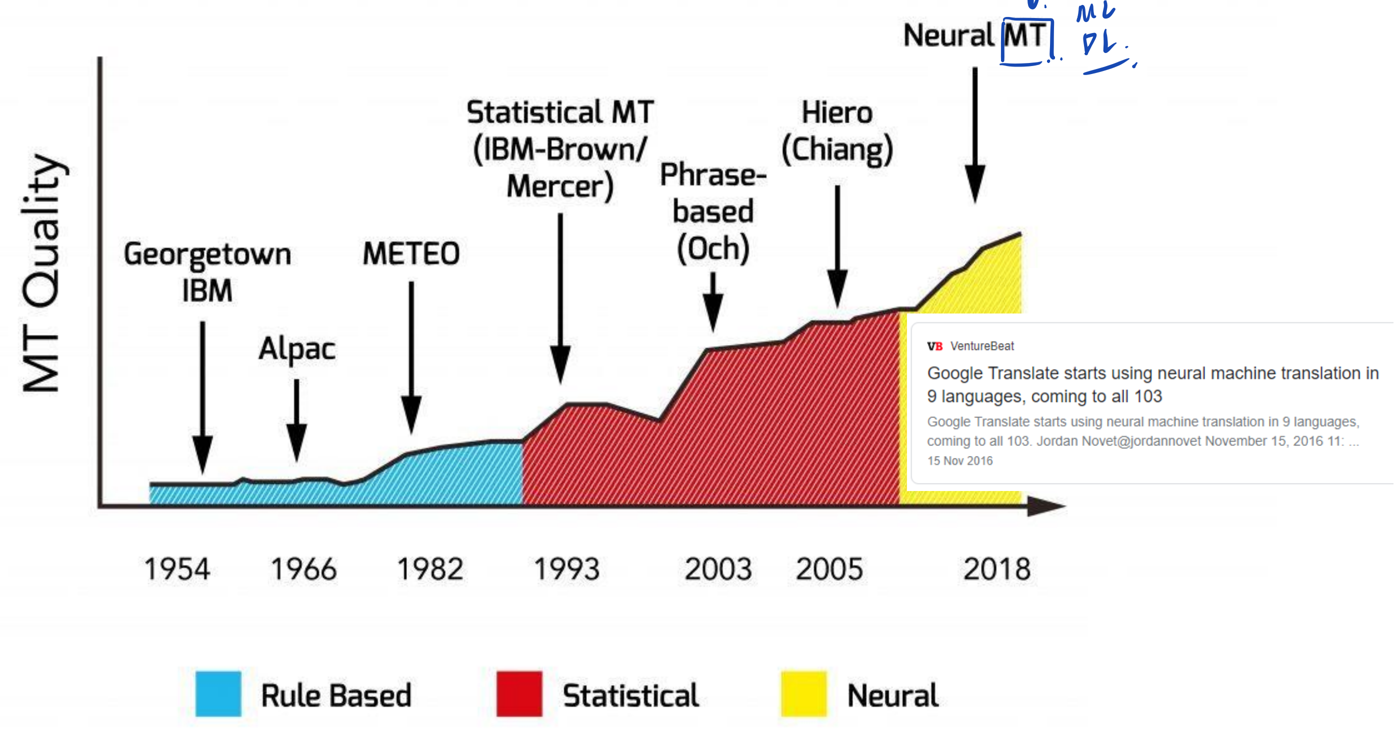
Hype Cycle for Artificial Intelligence



(Gartner, 2020)

1 The NLP ERA

NLP Techniques – with the Trend of Machine Translation



VB VentureBeat

Google Translate starts using neural machine translation in 9 languages, coming to all 103

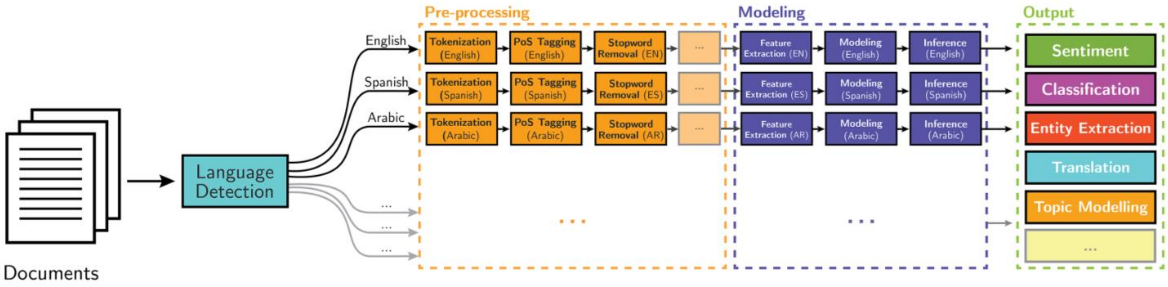
Google Translate starts using neural machine translation in 9 languages, coming to all 103. Jordan Novet@jordannovet November 15, 2016 11: ...

15 Nov 2016

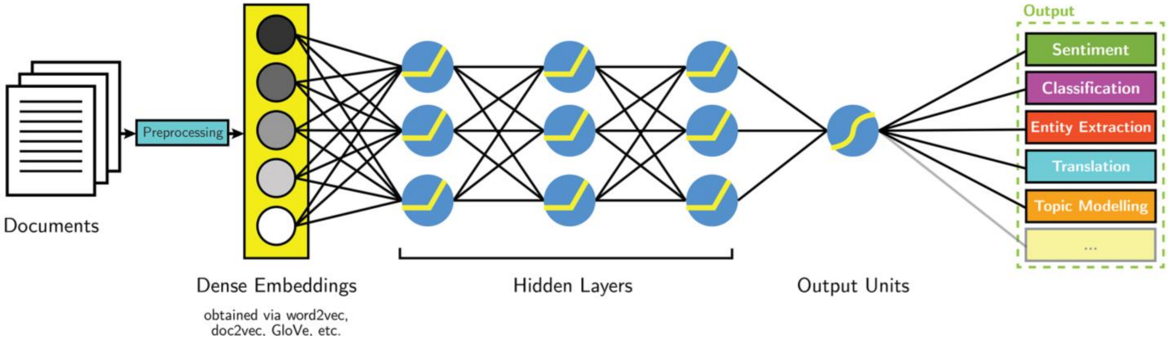
1 The NLP ERA

Language Modelling using Deep Learning-based NLP Techniques

Classical NLP



Deep Learning-based NLP



1 ASSESSMENT

Assessment Overview

Assessment	Weight	Due
Lab Exercise	10%	Multiple Weeks
Assignment 1	20%	Week 8 (Friday 11.59pm – AU time)
Assignment 2	20%	Week 14 (Friday 11.59pm – AU time)
Final Exam	50%	Exam Period

Lab Exercises

- Programming tasks done in fortnightly computer labs

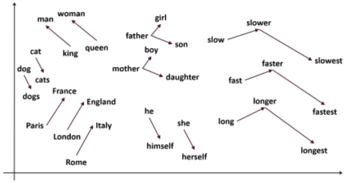
Assignments

- Take place through the teaching period
- Implementation and Documentation

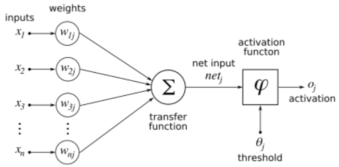
1 ASSESSMENT

Lab Exercise – 10%

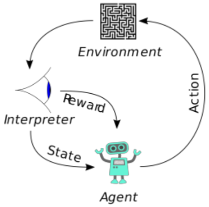
- In the Lab, students need to do the small tasks (2-3% for each week).
- 2-3 tasks are given based on what you learned in the previous lectures.
- You must have been assessed as having completed 5 out of 6 in order to get the 10% for lab exercise.



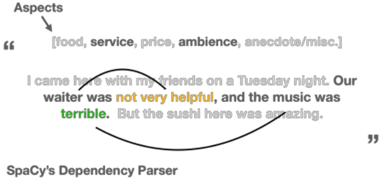
word2vec



NN

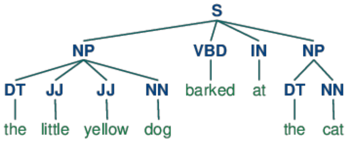


RL

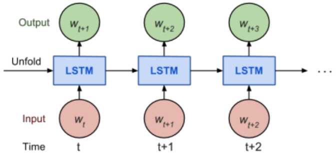


Aspects
[food, service, price, ambience, anecdote/misc.]
“
I came here with my friends on a Tuesday night. Our waiter was not very helpful, and the music was terrible. But the sushi here was amazing.
”
SpaCy's Dependency Parser

Dependency Parsing



Part-of-Speech Tagging

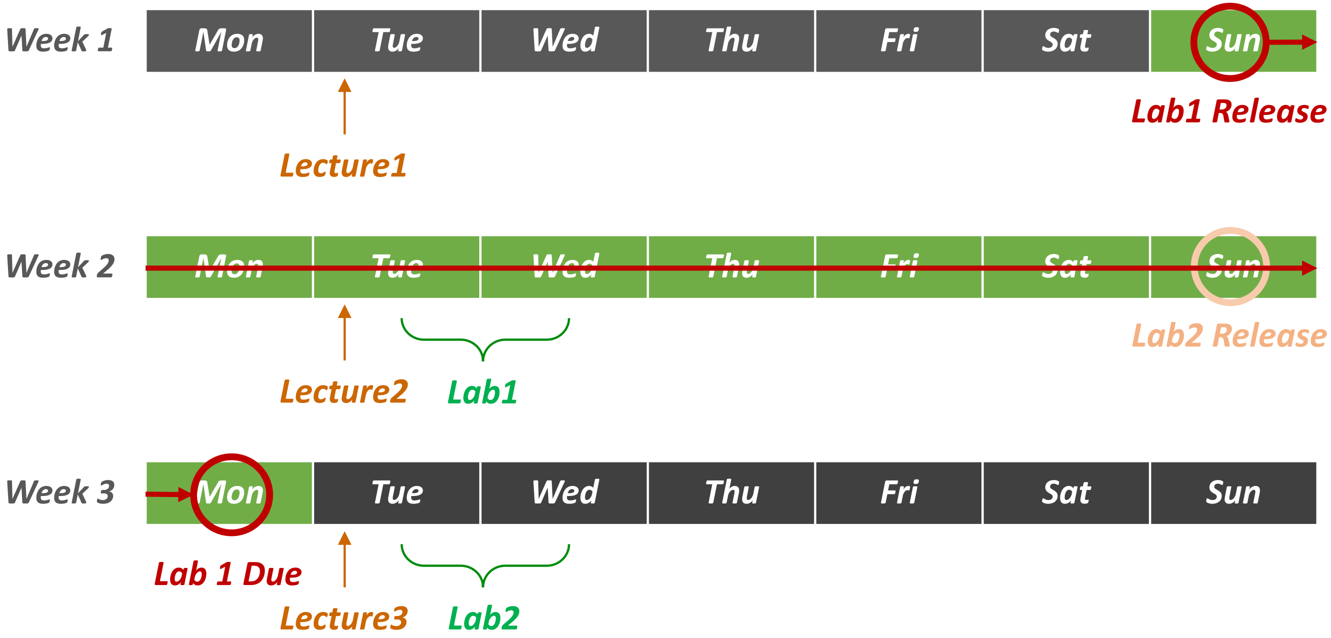


Language Generation

1 ASSESSMENT

Lab Exercise – 10%

- When to submit the Fortnightly Lab Exercise (e.g. Lab1 Release and Submission) Please check the course contents page in the canvas.

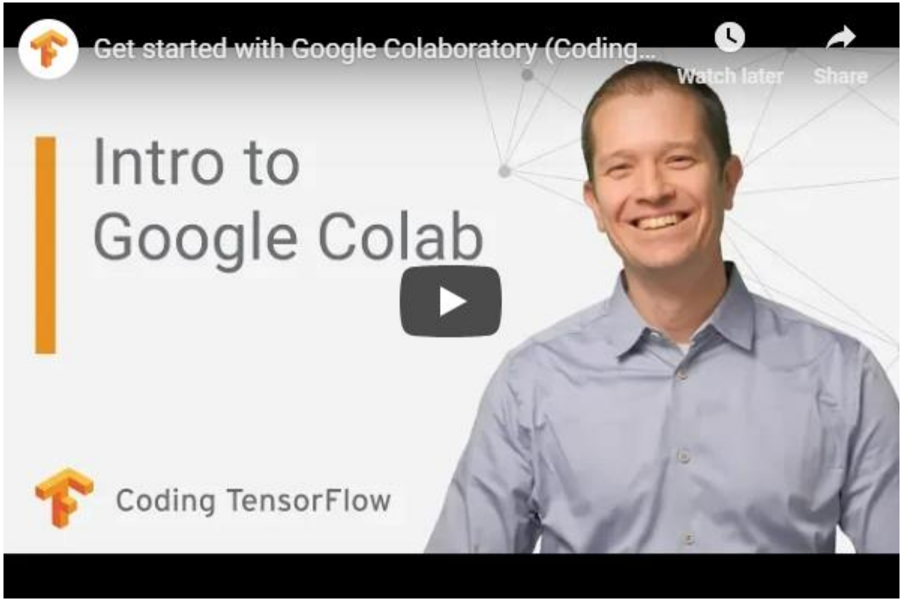


Week 1	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week 2	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week 3	Mon	Tue	Wed	Thu	Fri	Sat	Sun

What do we do during Labs?

In Labs, Students will use Google CoLab

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.



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

1 ASSESSMENT

Assignment 1 (20%)

The focus of this assignment 1 is implementing the text/word embeddings/language models, which ***1)understands the language*** and ***2)produces the detection/prediction/generation*** decision.

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

Assignment 2 (20%)

The focus of this assignment 2 is proposing a natural language model to produce high performance (such as accuracy, consistency, plausibility, validity, and distribution) in different NLP tasks. The different NLP tasks can be Question Answering, Named Entity Recognition, Text Classification, or etc.

NOTE: *Assignment 2 would be a group assignment (2 people in group).*

However, you can do individually only if you want.

Final Exam (50%)

No ProtoU

The final exam will be a short take-home exam hosted on Canvas (3 hours duration).
You will be asked to answer variety of theoretical questions.
The sample exam questions will be shared in the week 13 lecture.

1 ASSESSMENT

ASSESSMENT Due

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

- Week 5: Language Fundamental
- Week 6: Part of Speech Tagging
- Week 7: Dependency Parsing
- Week 8: Language Model and Natural Language Generation — ***Assignment 1 Due***

- Week 9: Information Extraction: Named Entity Recognition
- Week 10: Advanced NLP: Attention and Reading Comprehension
- Week 11: Advanced NLP: Transformer and Machine Translation
- Week 12: Advanced NLP: Pretrained Model in NLP

- Week 13: Future of NLP and Exam Review

Assignment 2 Due (Week 14)

1 ASSESSMENT

Start assignments early!

- All assignments involve coding and report writing
- Reports are the primary deliverable
- Though we will check implementations for correctness
- Assignments will Be very different from last year’s
- Reports will be submitted to Turnitin through Canvas
- Code is also submitted (for assignments 1 and 2) and retained
- We will use code plagiarism detection tools
- Clearly reference any copied/adapted code portions and cite their origins

1 ASSESSMENT

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

1 TIMETABLE

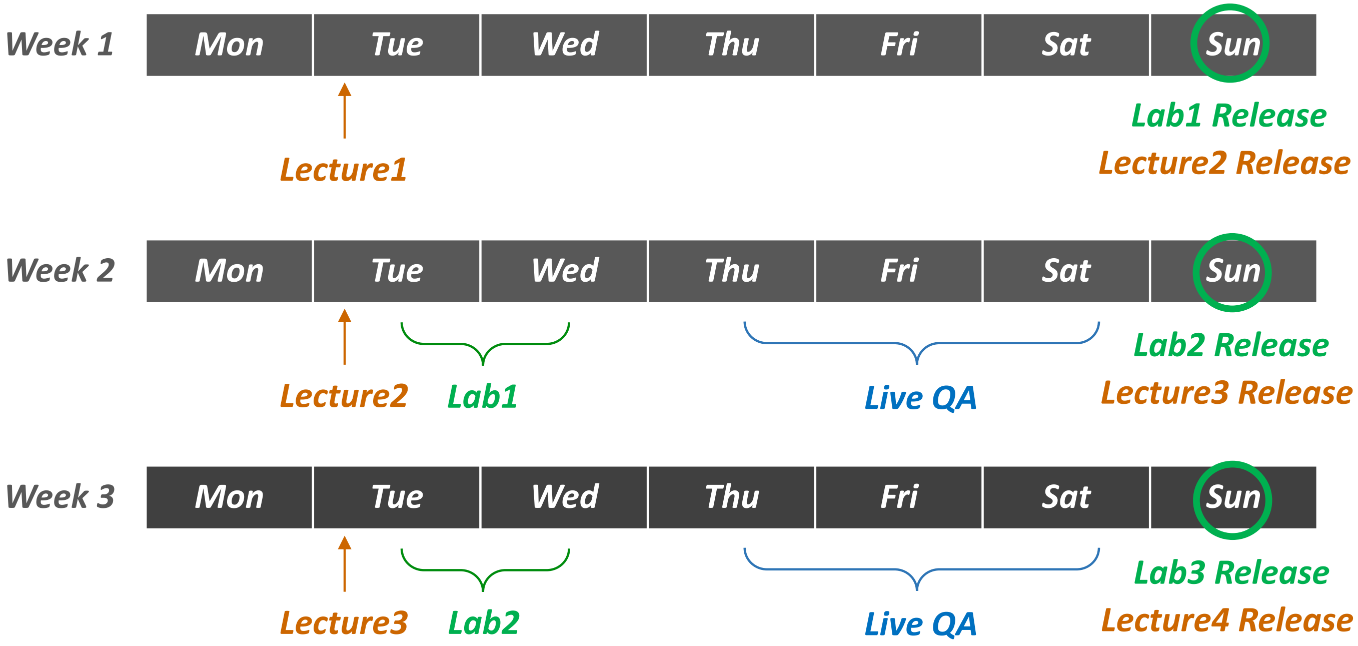
Working Hours

- Work 12 hours per week for this course (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/laboratory time
- Participate respectfully in discussions in lectures and labs;
- Complete all assessment tasks on time.

1 TIMETABLE

Classes and Release Date

- When the Class and Release Dates are



Week 1

Mon Tue Wed Thu Fri Sat Sun

Lecture1

Lab1 Release

Lecture2 Release

Week 2

Mon Tue Wed Thu Fri Sat Sun

Lecture2 Lab1 Live QA

Lab2 Release

Lecture3 Release

Week 3

Mon Tue Wed Thu Fri Sat Sun

Lecture3 Lab2 Live QA

Lab3 Release

Lecture4 Release

1

TIMETABLE

Full Course Timetable at course website:

- **Lecture:** Tue 5-7pm
- **Tutorial:** Tue/Wed (depends on your selection)
- **LiveQA**: Will vote this in the Week 1 Lecture together!

1 STAFF

Who are we?

Unit Coordinator & Lecturer

- Caren.Han@sydney.edu.au
- No fixed consultation hour; please arrange a time to see me.

Teaching Assistants & Tutors

- They will introduce themselves really soon!

For Qs related to course content, please post in Ed.

For Qs related to admin, please contact to the unit coordinator

- Please put [COMP5046] in the title of the email

1 READINGS (OPTIONAL)

No Textbook Recommended, but if you really want to read some

Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.

Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. " O'Reilly Media, Inc.".

Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

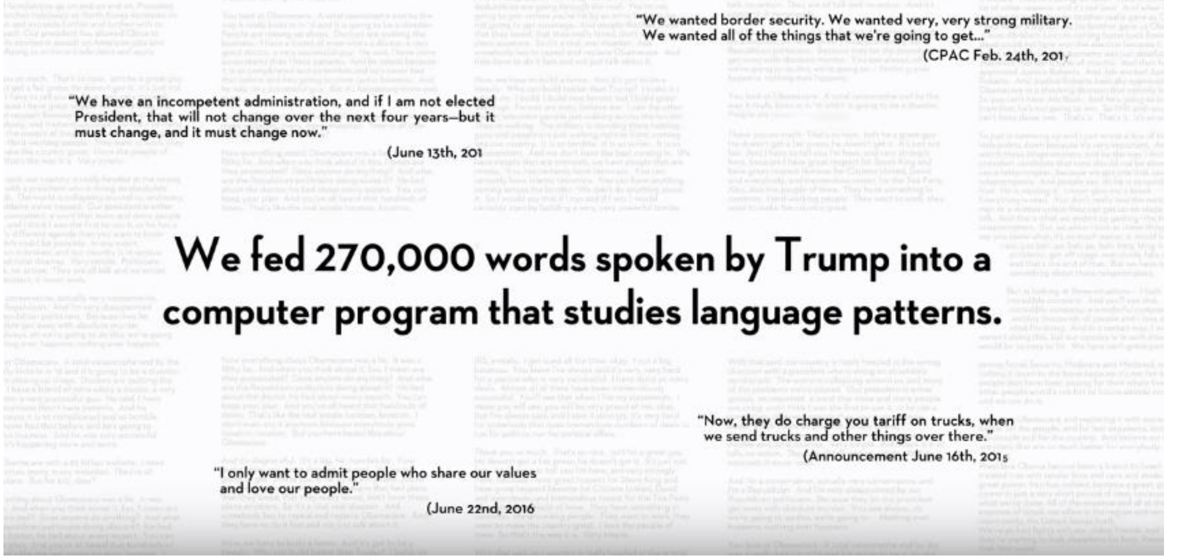
1 SELF TEST

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 do you do if you get sick during semester?
- What is Turnitin?
- What programming language do you need to know?

1 What do we learn? (NLP and Deep Learning)

https://www.youtube.com/watch?v=u5T5Q9yzXnQ



"We wanted border security. We wanted very, very strong military. We wanted all of the things that we're going to get..."
(CPAC Feb. 24th, 201

"We have an incompetent administration, and if I am not elected President, that will not change over the next four years—but it must change, and it must change now."
(June 13th, 201

We fed 270,000 words spoken by Trump into a computer program that studies language patterns.

"Now, they do charge you tariff on trucks, when we send trucks and other things over there."
(Announcement June 16th, 2015

"I only want to admit people who share our values and love our people."
(June 22nd, 2016

0 LECTURE PLAN

wed 9-10 pm AU time → Beijing time Wed. 6-7 pm

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

Why Process Language?

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

2 LANGUAGE

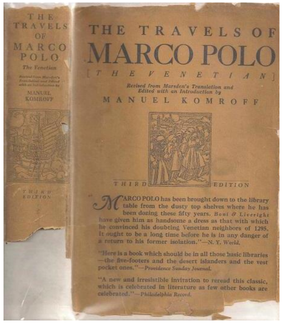
Why Process Language?

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

2 LANGUAGE

Why Process Language?

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



THE TRAVELS OF MARCO POLO

The Venetian

Revised from Marsden's Translation and Edited with an Introduction by

MANUEL KOMROFF

THIRD EDITION

THE TRAVELS OF MARCO POLO

[THE VENETIAN]

Revised from Marsden's Translation and Edited with an Introduction by

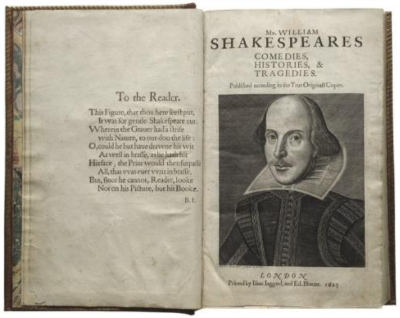
MANUEL KOMROFF

THIRD EDITION

"MARCO POLO has been brought down to the library table from the dusty top shelves where he has been dozing these fifty years. *Boni & Liveright* have given him as handsome a dress as that with which he convinced his doubting Venetian neighbors of 1295. It ought to be a long time before he is in any danger of a return to his former isolation."—N. Y. *World.*

"Here is a book which should be in all those basic libraries—the five-footers and the desert islanders and the vest pocket ones."—*Providence Sunday Journal.*

"A new and irresistible invitation to reread this classic, which is celebrated in literature as few other books are celebrated."—*Philadelphia Record.*



To the Reader.

This Figure, that thou here ſeeſt put,
It was for gentle Shakeſpeare cut;
Wherein the Grauer had a ſtrife
with Nature, to out-doo the life:
O, could he but haue drawne his wit
As well in braſſe, as he hath hit
His face; the Print would then ſurpaſſe
All, that vvas euer vvrit in braſſe.
But, ſince he cannot, Reader, looke
Not on his Picture, but his Booke.

B. I.

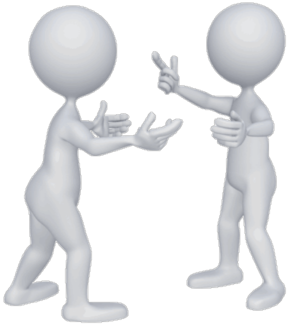
Mr. WILLIAM
SHAKESPEARES
COMEDIES,
HISTORIES, &
TRAGEDIES.
Publiſhed according to the True Originall Copies.

LONDON
Printed by Iſaac Iaggard, and Ed. Blount. 1623.

2 LANGUAGE

Why Process Language?

- language stores knowledge
- language communicates new knowledge
- language is a key to 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 developing systems that allow computers to communicate with people using everyday language

Computational Linguistics

It concerns how computational methods can aid the understanding of language

Communication

The goal in the production / comprehension of language is communication.

2 Natural Language Processing (NLP)

Communication for the speaker:

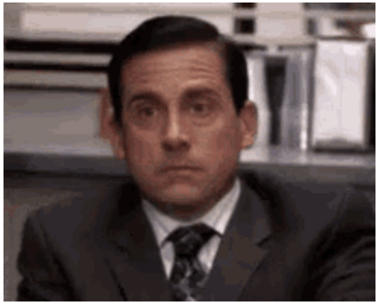
- Intention: **Decide when and what information** should be transmitted (a.k.a. strategic generation). May require planning and reasoning about agents' goals and beliefs.

- Generation: **Translate the information to be communicated** (in internal logical representation or "language of thought") into string of words in desired natural language (a.k.a. tactical generation).

- Synthesis: **Output the string** in 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 (parsing): Find the correct parse tree showing the phrase structure of the string.
  - Semantic Interpretation: Extract the (literal) meaning of the string .
  - Pragmatic Interpretation: Consider 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 add it to the Knowledge Base.

2 SPECIAL ABOUT NLP

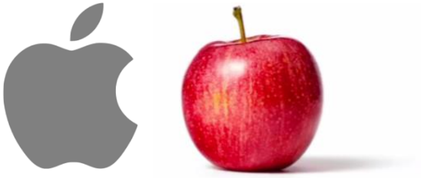
What is special about NLP?

- 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 just symbols for an extra-linguistic entity : the word is a signifier that maps to a signified (idea or thing).

“Computer”

“Apple”



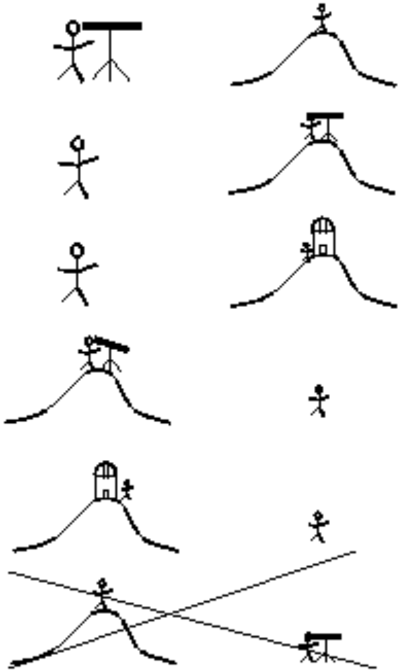
“Whaaaaaa”

???????

2 SPECIAL ABOUT NLP

Ambiguity

I saw the man on the hill with a telescope.



2 SPECIAL ABOUT NLP

Ambiguity is Explosive

Ambiguities compound to generate enormous numbers of possible interpretations.

In English, a sentence ending in n prepositional phrases has over 2n interpretations.

"I saw the man with the telescope": 2 parses

{ I, with the telescope
the man, with the telescope.

"I saw the man on the hill with the telescope.": 5 parses

"I saw the man on the hill in Texas with the telescope": 14 parses

"I saw the man on the hill in Texas with the telescope at noon.": 42 parses

"I saw the man on the hill in Texas with the telescope at noon on Monday" 132 parses

2 SPECIAL ABOUT NLP

Ambiguity is Ubiquitous

Speech Recognition

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

Syntactic Analysis

- “I ate spaghetti with chopsticks” vs. “I ate spaghetti with meatballs.”

tool

food.

Semantic Analysis

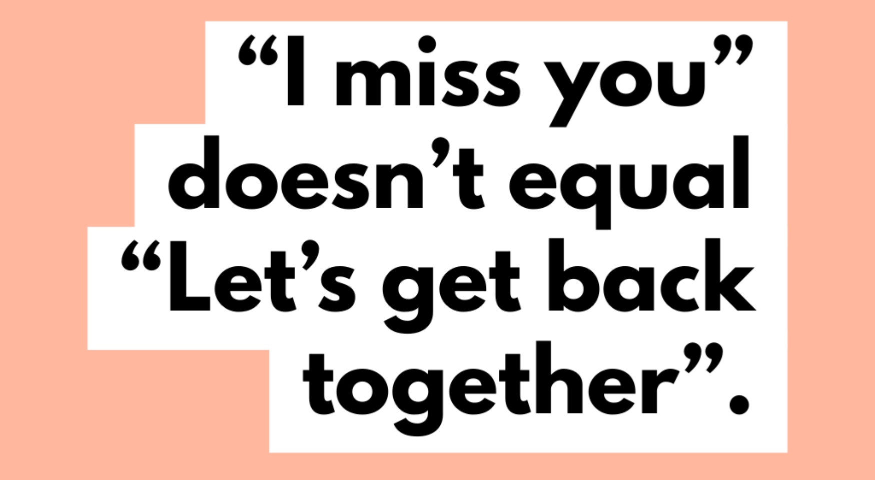
- “The dog is in the pen.” vs. “The ink is in the pen.”
- “I put the plant in the window” vs. “Ford put the plant in Mexico”

↓

factory

2 SPECIAL ABOUT NLP

Even human struggle with understanding



"I miss you" doesn't equal "Let's get back together".

??????

2 SPECIAL ABOUT NLP

The difficulty level in various NLP tasks

Easy

- Spell Checking
- Keyword Search
- Finding Synonyms

Medium

- Extracting Information from documents (including websites)

Difficult

- Semantic Analysis (What is the meaning of query statement?)
- Machine Translation
- Coreference Resolution
- Question Answering

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

3 WORD REPRESENTATION

How to represent the meaning of the 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.

signifier (symbol) ⟺ signified (idea or thing)

“Apple”

3 WORD REPRESENTATION

How do we have usable meaning in a computer?

- Common solution: Use e.g. WordNet, a thesaurus containing lists of synonym sets and hypernyms ("is a" relationships).
- http://wordnetweb.princeton.edu/perl/webwn

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')]
```

3 WORD REPRESENTATION

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.
  - e.g. “glad” can be synonym for “fXXX off”???
- Missing new meanings of words
  - e.g., wicked, badass, nifty, wizard, genius, ninja, bombast
  - Impossible to keep up-to-date! always new words. eg. google Microsoft.
- Subjective
- Requires human labor to create and adapt
- Can’t compute accurate word similarity know they are similar but how much do they similar.

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

4 COUNT based WORD REPRESENTATION

One-hot Encoding

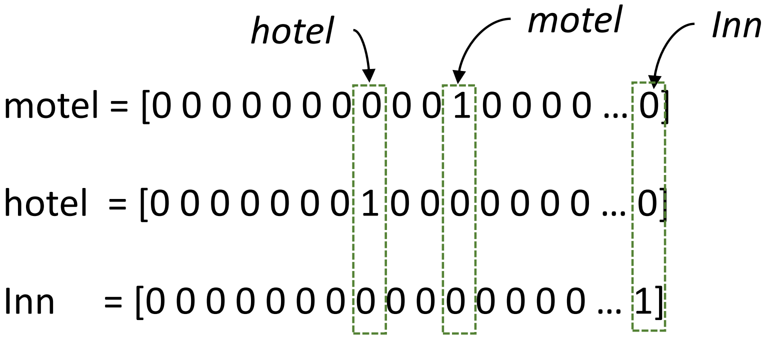
- In traditional NLP, we regard words as discrete symbols.

Hot (True) ***Cold (False)***

Means one 1, the rest 0s

Words can be represented by **one-hot vectors**:

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.



motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 … 0]

hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 … 0]

Inn = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 1]

Vector dimension = number of words in vocabulary

4 COUNT based WORD REPRESENTATION

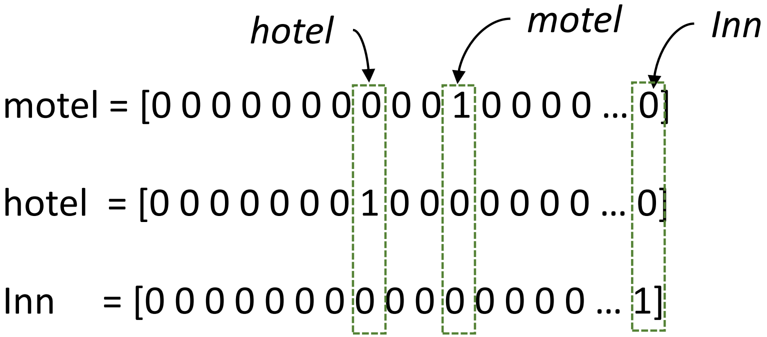
Baseline.

Problem with one-hot vectors

Problem #1. No word similarity representation

No similarity between 2 words.

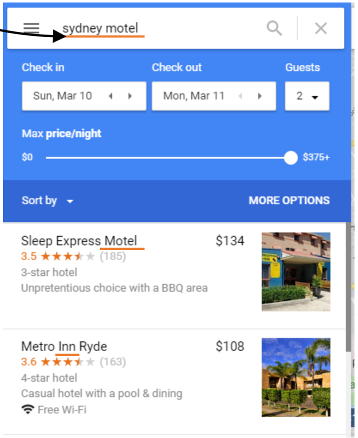
Example: in web search, if user searches for "Sydney motel", we would like to match documents containing "Sydney Inn"



motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 … 0]

hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 … 0]

Inn = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 1]



sydney motel

Check in: Sun, Mar 10
Check out: Mon, Mar 11
Guests: 2

Max price/night: $0 – $375+

Sort by | MORE OPTIONS

Sleep Express Motel — $134
3.5 ★★★★★ (185)
3-star hotel
Unpretentious choice with a BBQ area

Metro Inn Ryde — $108
3.6 ★★★★★ (163)
4-star hotel
Casual hotel with a pool & dining
Free Wi-Fi

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

Problem #2. Inefficiency

Vector dimension = number of words in vocabulary

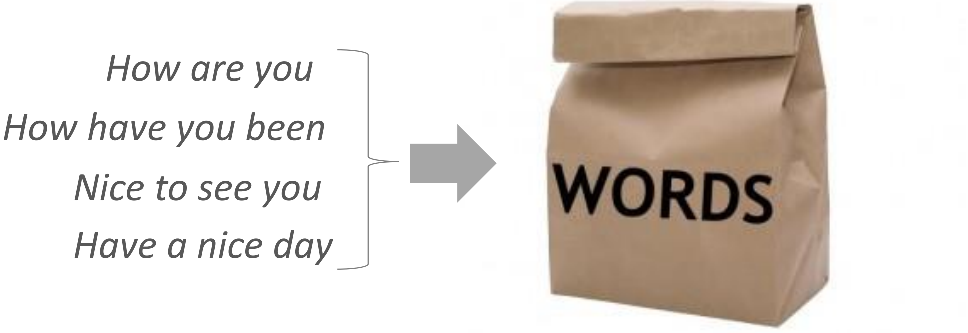
Each representation has only a single '1' with all remaining 0s.

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, not where in the document.



How are you

How have you been

Nice to see you

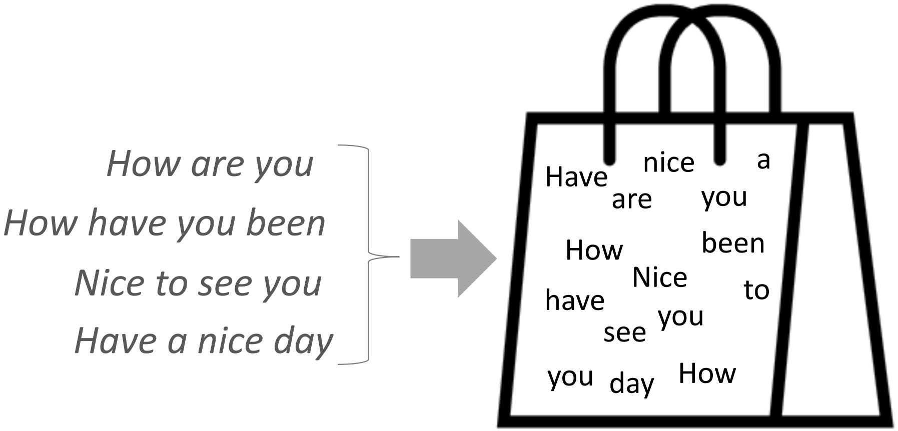
Have a nice day

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**



How are you

How have you been

Nice to see you

Have a nice day

4 COUNT based WORD REPRESENTATION

Bag of Words (BOW)

similar ⇒ word occurrence is similar.

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]

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

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, and in turn meaning of words in the document (semantics). Context and meaning can offer a lot to the model, that if modeled could tell the difference between the same words differently arranged (“this is interesting” vs “is this interesting”).

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

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

4 COUNT based WORD REPRESENTATION

THE UNIVERSITY OF SYDNEY

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 (synonymous with bag of words)
- The ***Document Frequency*** is the number of times a word occurs in a corpus of documents

4 COUNT based WORD REPRESENTATION

Term Frequency

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

Like BoW

$tf_{i,j}$ = 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

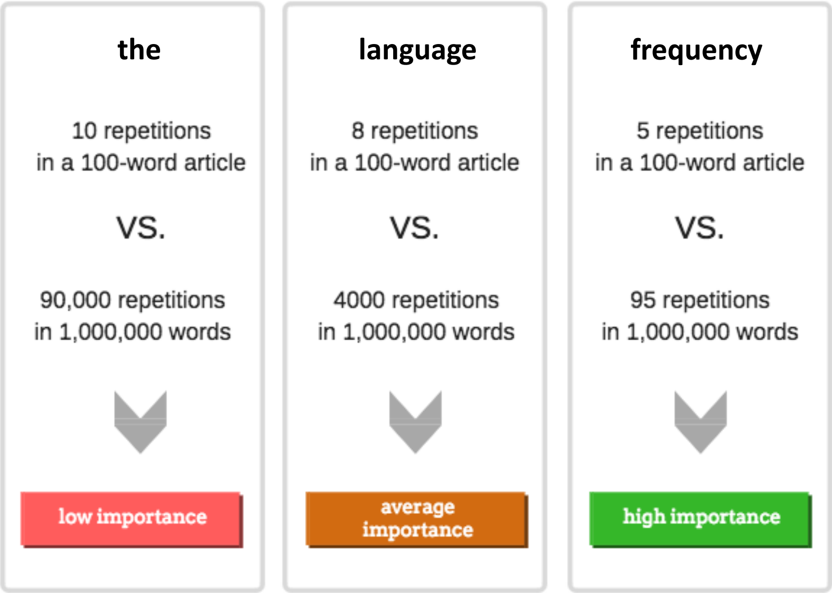
	and	apple	banana	fruit	I	like	sweet	yellow
D#1	0	1	0	0	1	1	0	0
D#2	0	0	1	0	1	1	0	0
D#3	1	0	2	0	0	0	1	1
D#4	0	0	0	1	0	0	1	0

4

COUNT based WORD REPRESENTATION

What if we just use Term Frequency Only?

- It is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms while scale up the rare ones.



the

10 repetitions in a 100-word article

VS.

90,000 repetitions in 1,000,000 words

low importance

language

8 repetitions in a 100-word article

VS.

4000 repetitions in 1,000,000 words

average importance

frequency

5 repetitions in a 100-word article

VS.

95 repetitions in 1,000,000 words

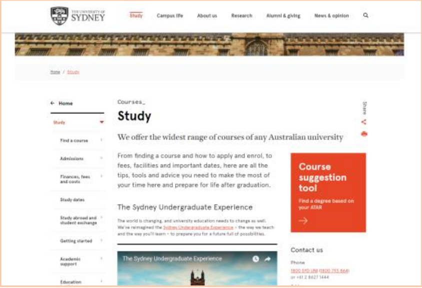
high importance

4 COUNT based WORD REPRESENTATION

Can we use Term Frequency Only?

then Sydney university will appear frequently.
(rather than stopwords)

University of Sydney Website



Courses_

Study

We offer the widest range of courses of any Australian university

From finding a course and how to apply and enrol, to fees, facilities and important dates, here are all the tips, tools and advice you need to make the most of your time here and prepare for life after graduation.

The Sydney Undergraduate Experience

The world is changing, and university education needs to change as well. We've reimagined the Sydney Undergraduate Experience – the way we teach and the way you'll learn – to prepare you for a future full of possibilities.

The Sydney Undergraduate Experience

Course suggestion tool

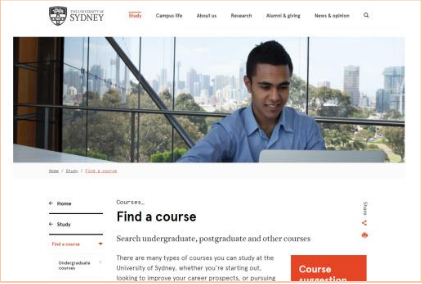
Find a degree based on your ATAR

Contact us

Phone

1800 SYD UNI (1800 793 864)
or +61 2 8627 1444

Webpage#1



Courses_

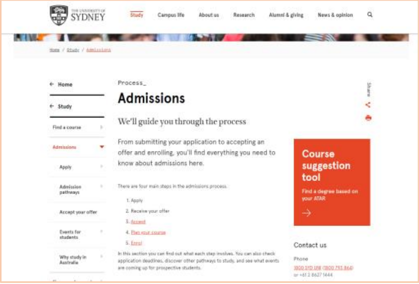
Find a course

Search undergraduate, postgraduate and other courses

There are many types of courses you can study at the University of Sydney, whether you're starting out, looking to improve your career prospects, or pursuing

Course suggestion

Webpage#2



Process_

Admissions

We'll guide you through the process

From submitting your application to accepting an offer and enrolling, you'll find everything you need to know about admissions here.

There are four main steps in the admissions process.

1. Apply
2. Receive your offer
3. Accept
4. Plan your course
5. Enrol

In this section you can find out what each step involves. You can also check application deadlines, discover other pathways to study, and see what events are coming up for prospective students.

Course suggestion tool

Find a degree based on your ATAR

Contact us

Phone

1800 SYD UNI (1800 793 864)
or +61 2 8627 1444

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

4 COUNT based WORD REPRESENTATION

Inverse Document Frequency

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

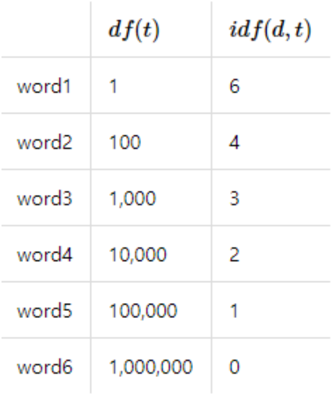
N = total number of documents

df_i = number of documents containing term i

$$n = 1,000,000$$

With log

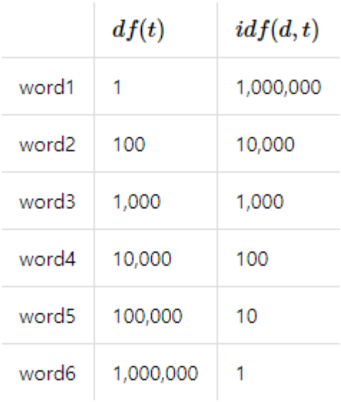
$$idf(d,t) = log(n/df(t))$$



	$df(t)$	$idf(d,t)$
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

Inverse Document Frequency

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

$1+df_i$ sometimes, why?

may come out with 0.

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

4 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+*df*i)	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

We use a natural logarithm to the base of the mathematical constant e.
where e is an irrational and transcendental number approximately equal to 2.718281828459

4 COUNT based WORD REPRESENTATION

Term Frequency Inverse Document Frequency

more documents

→ will perform well

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

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

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

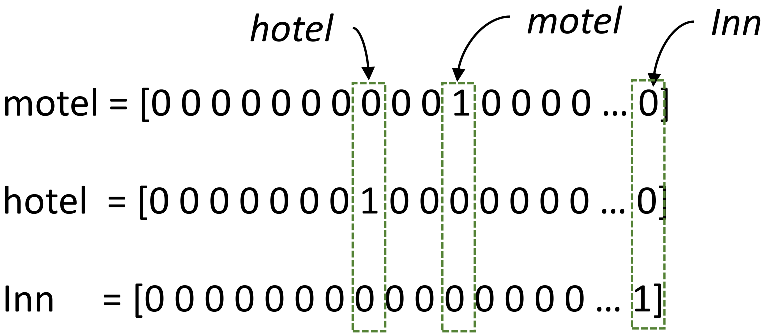
Document #4: Sweet fruit

	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

4 COUNT based WORD REPRESENTATION

Sparse Representation

With COUNT based word representation (especially, one-hot vector), linguistic information was represented with sparse representations (high-dimensional features)



hotel motel Inn

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 … 0]

hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 … 0]

Inn = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 1]

However, with the recent popularity and success of **word embeddings (low dimensional, distributed representations)**, **neural-based models have achieved superior results on various language-related tasks** as compared to traditional machine learning models with high-dimensional features.

0 LECTURE PLAN

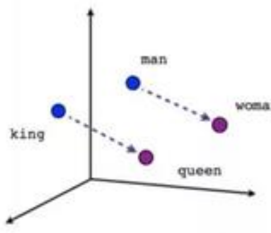
Lecture 1: Introduction to Natural Language Processing

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

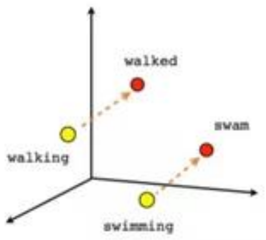
5 NEXT WEEK PREVIEW…

How to Represent the Word Similarity!

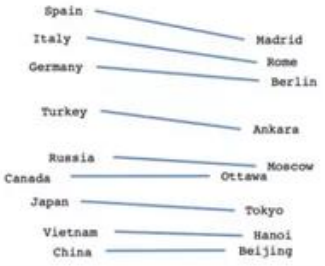
- How to represent the word similarity with dense vector



Male-Female

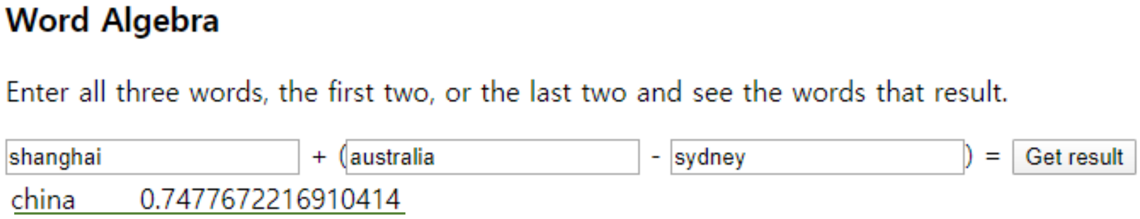


Verb tense



Country-Capital

- Try this with word2vec



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/

/

Reference

Reference for this lecture

- Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.
- Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. " O'Reilly Media, Inc.".
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- Manning, C 2017, Introduction and Word Vectors, Natural Language Processing with Deep Learning, lecture notes, Stanford University
- Mooney, RJ 2000, Natural Language Processing Introduction, CS388, lecture notes, University of Texas at Austin