COMP5046
Natural Language Processing

Lecture 1: Introduction

Semester 1, 2020
School of Computer Science
The University of Sydney, Australia

THE UNIVERSITY OF SYDNEY

Dr Caren Han
Caren.Han@sydney.edu.au

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

INTRODUCTION

Dr Caren Han

- **Education**

  - B CompScience (1st Class Honours)

  - PhD Computer Science (Artificial Intelligence)

- **Australian Young Achiever Teaching Excellence Award**

  - Artificial Intelligence, Natural Language Processing, etc.

- **NLP Research Experience (7 years experience)**

  - **U.S. Air Force**: Robotics Natural Language Processing Interface

  - **U.S. Navy**: Deep Learning with NLP

  - **Hyundai Co.**: Car Navigation Dialog System

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, text categorisation and syntactic parsing.

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?

CUSP - COMP5046
https://cusp.sydney.edu.au/students/view-unit-page/alpha/COMP5046

Canvas – COMP5046
https://canvas.sydney.edu.au/courses/15253

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 visual question answering (QA) and machine translation.

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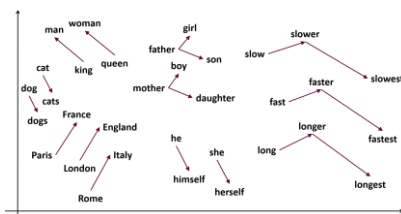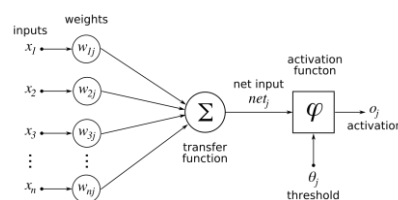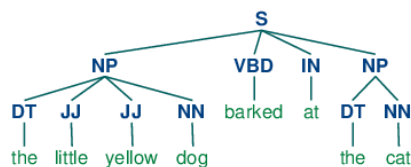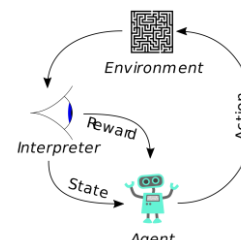

word2vec
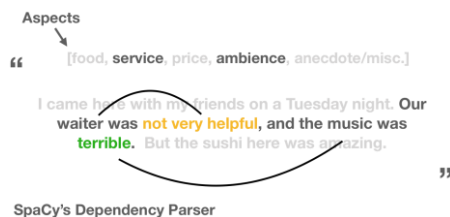
NN

RL

Dependency Parsing

Part-of-Speech Tagging
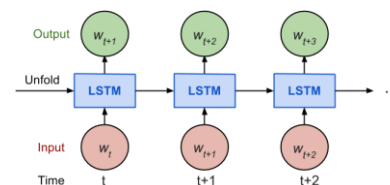
Language Modelling

INTRODUCTION

What will you learn in this course?

Week 1: Introduction to Natural Language Processing (NLP)

Week 2: Word Embeddings (Word Vector for Meaning)

Week 3: Word Classification with Machine Learning/Deep Learning I

Week 4: Word Classification with Machine Learning/Deep Learning II

NLP and Machine Learning

Week 5: Language Fundamental

Week 6: Part of Speech Tagging

Week 7: Dependency Parsing

Week 8: Language Model

NLP Techniques

Week 9: Information Extraction I: Named Entity Recognition

Week 10: Information Extraction II: Relation Extraction

Week 11: Application I: Question and Answering

Week 12: Application II: Machine Translation

Advanced Topic

Week 13: Future of NLP and Exam Review

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

- 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 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
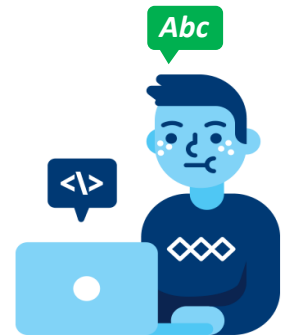
EXPECTATIONS

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)
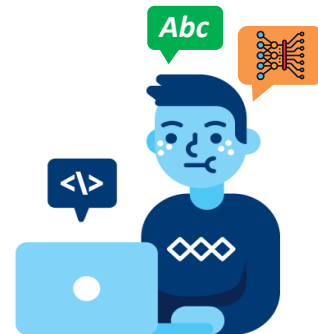
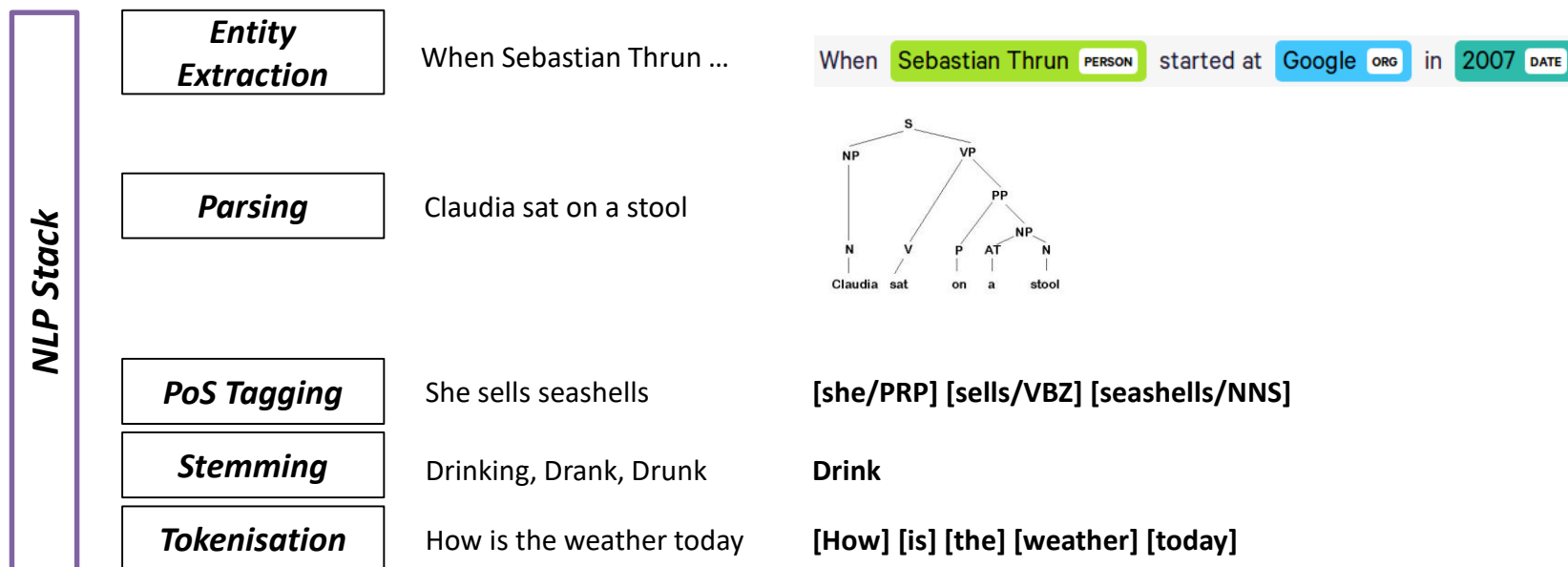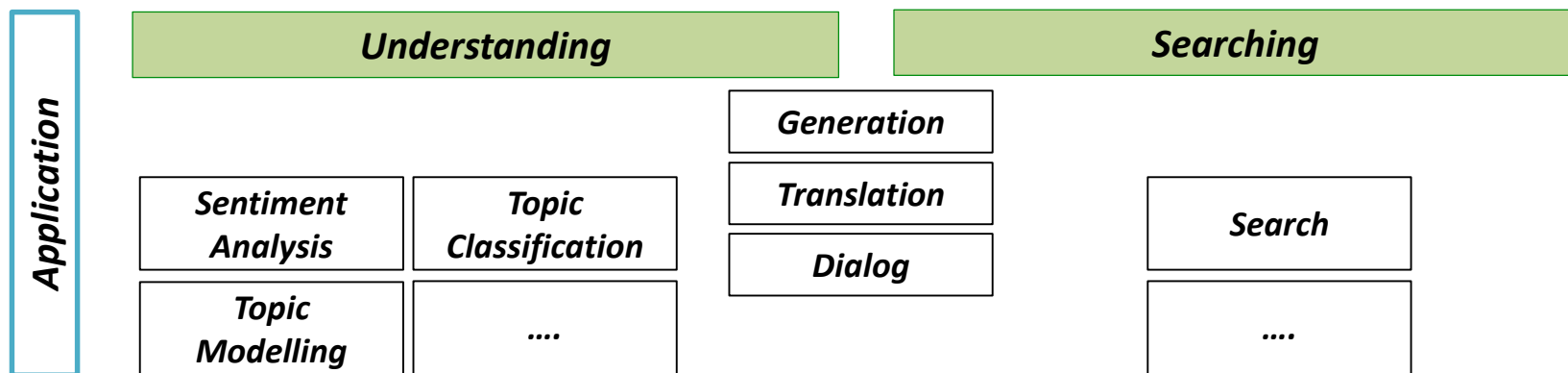# The NLP Big Picture

The purpose of Natural Language Processing: Overview

Application

Understanding		Searching

Generation

Sentiment Analysis	Topic Classification	Translation	Search

Dialog

Topic Modelling	….		….

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]**

The NLP ERA

NLP Techniques – with the Trend of Machine Translation
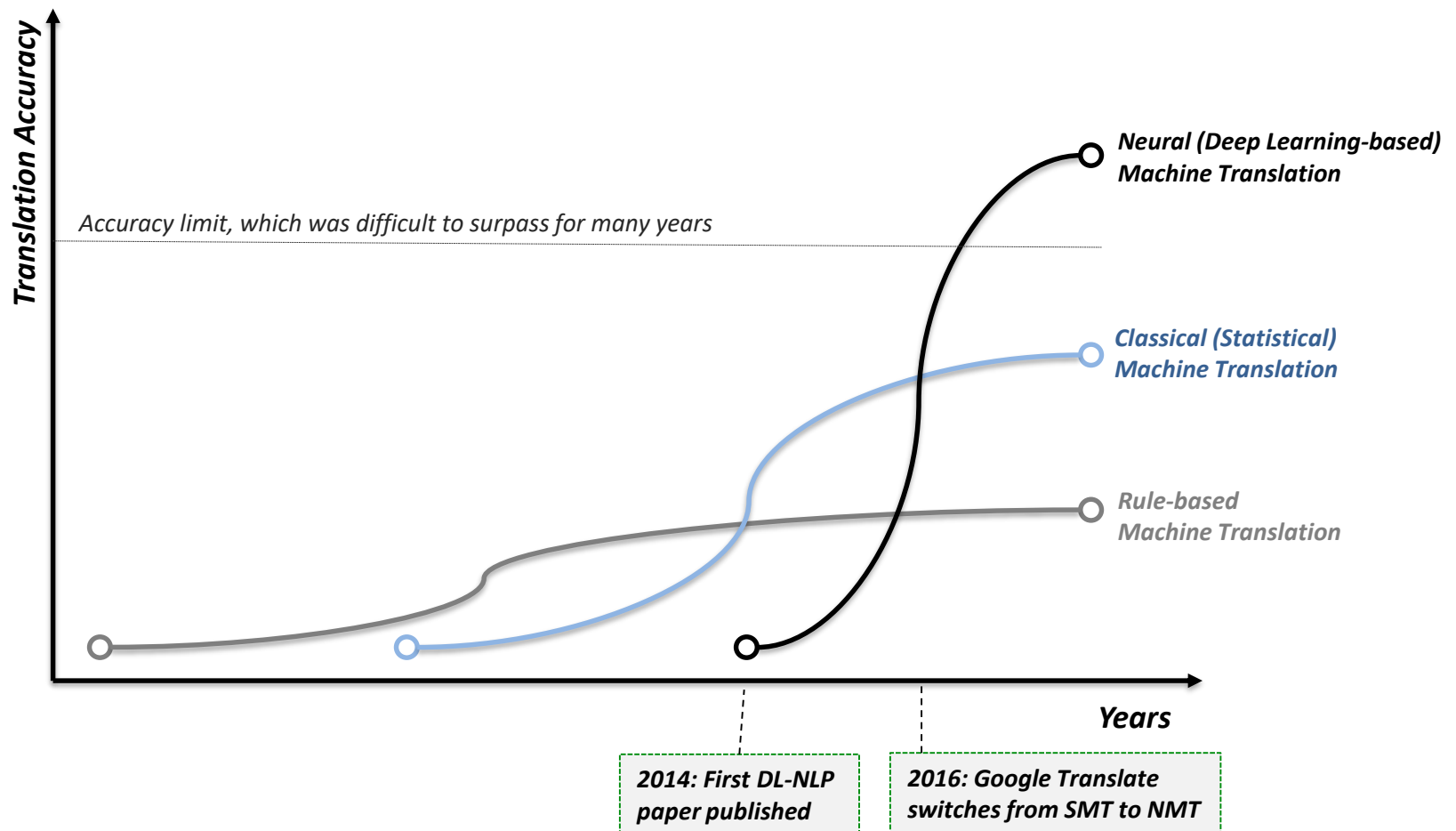
ASSESSMENT

Language Modelling using Deep Learning-based NLP Techniques
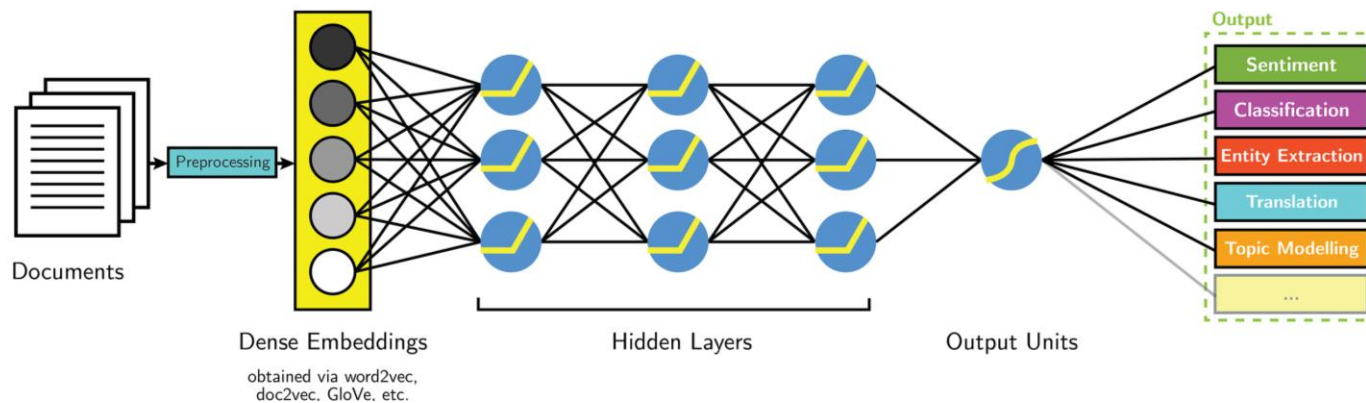
ASSESSMENT

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

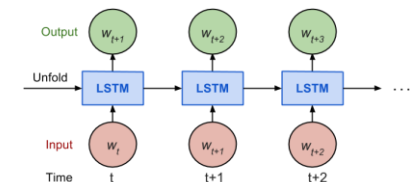- Programming tasks done in weekly computer labs

Assignments

- Take place through the teaching period

- Implementation and Documentation

Penalties for lateness

- 10% of the available marks per day late; maximum 7 days late

ASSESSMENT

Lab Exercise – 10%

- In the Lab, students need to do the small tasks (1% for each week).

- 2-3 tasks are given based on what you learned in the previous lecture.

- You must have been assessed as having completed 10 out of 11 in order to get the 10% for weekly lab exercise.

word2vec

NN

RL

Dependency Parsing

Part-of-Speech Tagging

Language Generation

1 ASSESSMENT
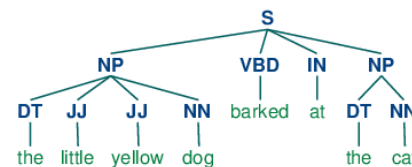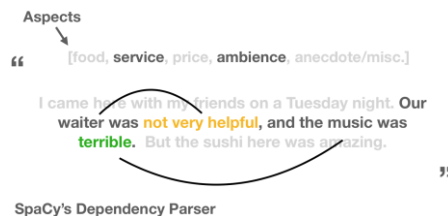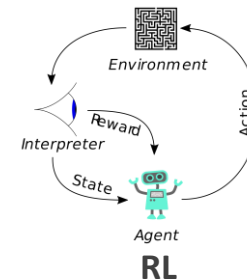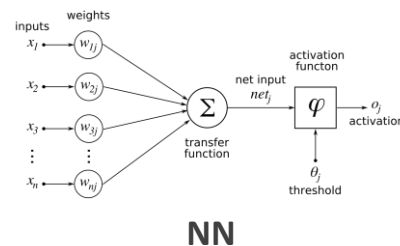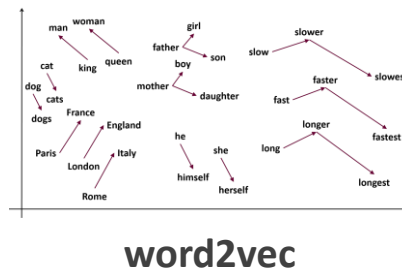
What do we do during Labs?

In Labs, Students will use Google Co Lab

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

ASSESSMENT
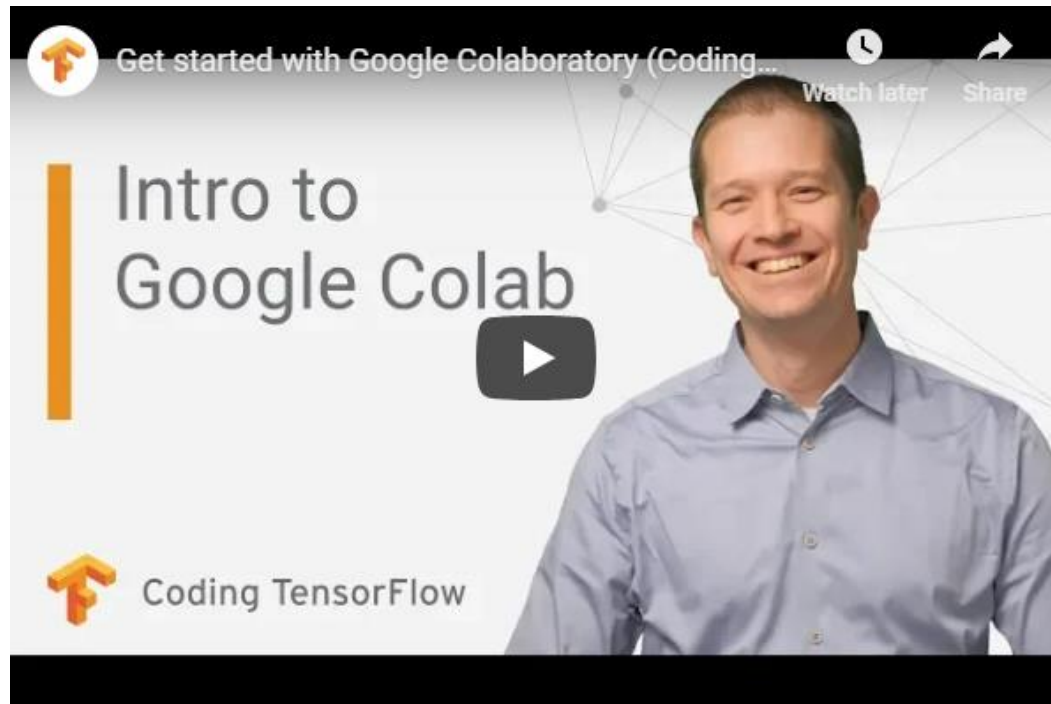
Assignment 1 (20%)

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

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.

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/Deep Learning I

Week 4: Word Classification with Machine Learning/Deep Learning II

Week 5: Language Fundamental

Week 6: Part of Speech Tagging

Week 7: Dependency Parsing

Week 8: Language Model

Assignment 1 Due

Week 9: Information Extraction I: Named Entity Recognition

Week 10: Information Extraction II: Relation Extraction

Week 11: Application I: Question and Answering

Week 12: Application II: Machine Translation

Week 13: Future of NLP and Exam Review

Assignment 2 Due (Week 14)

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

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

TIMETABLE

Working Hours

- Work 12 hours per week for this course (including 3 contact hours);

- Attend 2 hours of lectures per week:

  - Tuesday 2-4pm, Chemistry Lecture Theatre 3

  - 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.

TIMETABLE

Full Course Timetable at course website:

	Tuesday			Wednesday			
14	Lecture (Chemistry Lecture Theatre 3)						
15							
16	Lab T16A SIT Lab 114	Lab T16B SIT Lab 130B	Lab T16C SIT Lab 118	Lab W16A CS Lab 457			
17	Lab T17A SIT Lab 114	Lab T17B SIT Lab 130B	Lab T17C SIT Lab 130A	Lab W17A SIT Lab 118	Lab W17B CS Lab 457	Lab W17C SIT Lab 116	Lab W17D SIT Lab 130A
18	Lab T18A SIT Lab 117	Lab T18B SIT Lab 118					

STAFF

Who are we?

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

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!

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.

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?

What do we learn? (NLP and Deep Learning)

We're going to make America **BIGGER** again.

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, 201_

"I only want to admit people who share our values and love our people."

(June 22nd, 2016

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

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

LANGUAGE

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

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

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

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

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

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 human language

Communication

The goal in the production and comprehension of natural language is communication.

Natural Language Processing (NLP)

Communication for the speaker:

Listen to me!

- 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.

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.

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"**   **"Whaaaaaaa"**

???????

SPECIAL ABOUT NLP
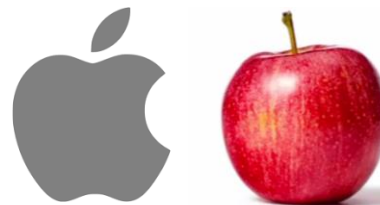
Ambiguity

I saw the man on the hill with a telescope.

SPECIAL ABOUT NLP
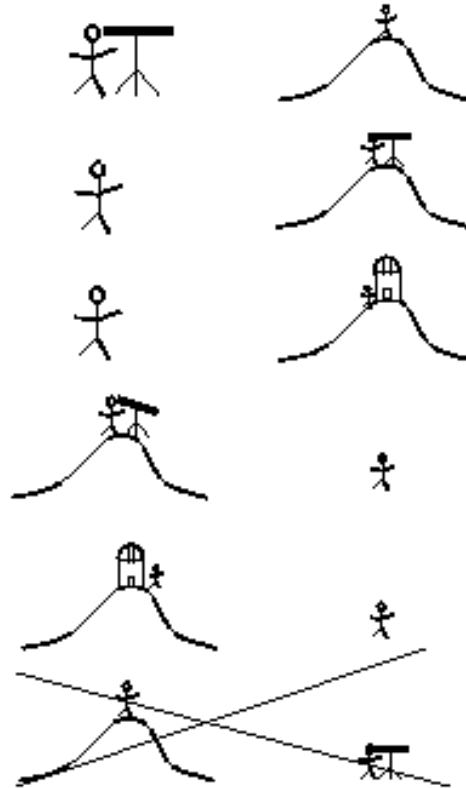
Ambiguity is Explosive

Ambiguities compound to generate enormous numbers of possible interpretations.

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

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

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."

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"

SPECIAL ABOUT NLP

Even human struggle with understanding

??????

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
- Question Answering

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

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.
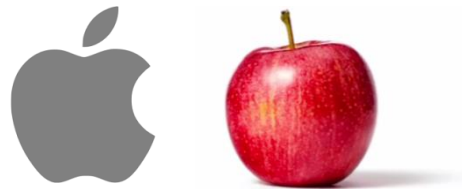- the thoughts or ideas that are intended to be expressed by it.

Commonest linguistic way of thinking of meaning:

signifier (symbol) ⟺ signified (idea or thing)

"Computer"   **"Apple"**

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')]

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!

- Subjective

- Requires human labor to create and adapt

- Can't compute accurate word similarity

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

COUNT based WORD REPRESENTATION
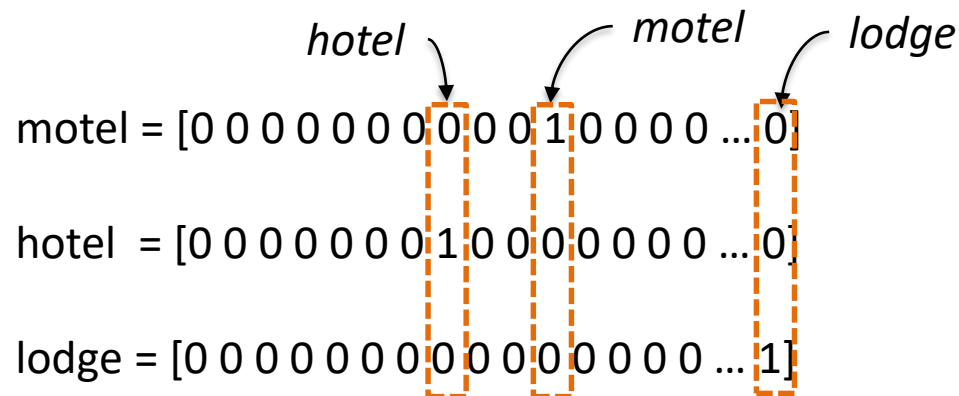
One-hot Encoding

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

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.

hotel   *motel*   *lodge*

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]

lodge = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 1]

Vector dimension = number of words in vocabulary

COUNT based WORD REPRESENTATION

Problem with one-hot vectors

Problem #1. No word similarity representation

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

$$\text{motel} = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ \dots\ 0]$$

$$\text{hotel}\ = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ \dots\ 0]$$

$$\text{Inn}\ \ \ = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ \dots\ 1]$$
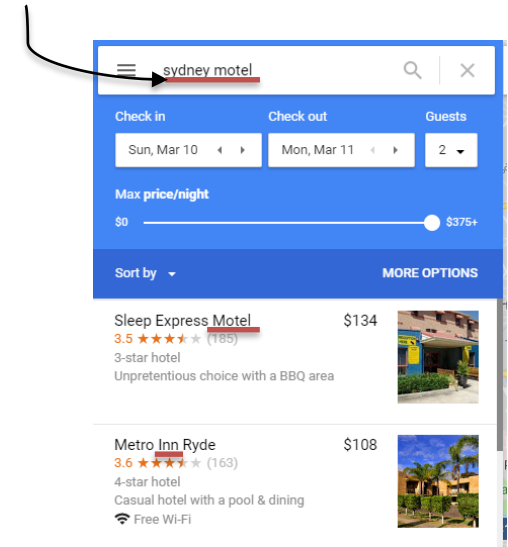
There is no natural notion of similarity for one-hot vectors!

Problem #2. Inefficiency

Vector dimension = number of words in vocabulary

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

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

WORDS

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

COUNT based WORD REPRESENTATION
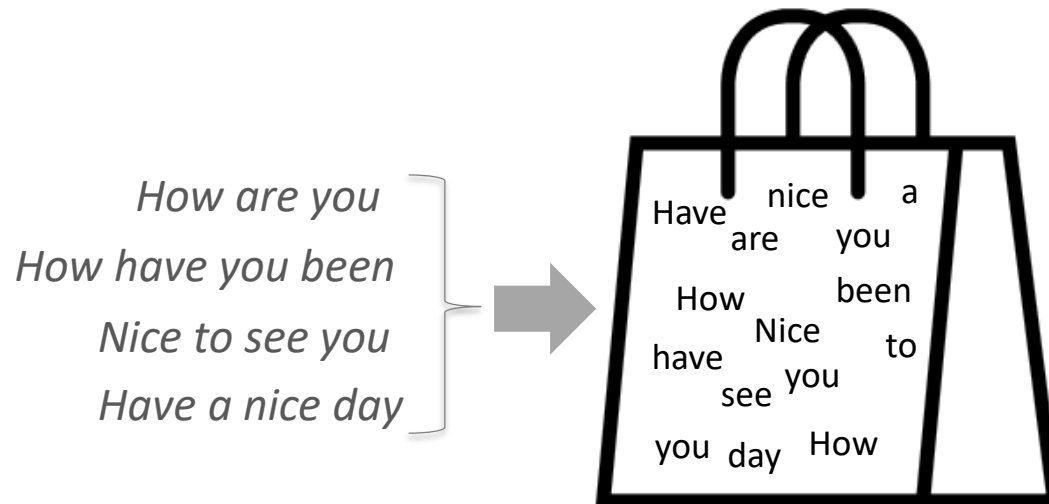
Bag of Words (BOW)

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]

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

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*

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 occurances 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 **Inverse Document Frequency** is the number of times a word occurs in a corpus of documents

COUNT based WORD REPRESENTATION

Term Frequency

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

Like BoW

tf_{i,j} = number of occurances 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

COUNT based WORD REPRESENTATION

Can we 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.
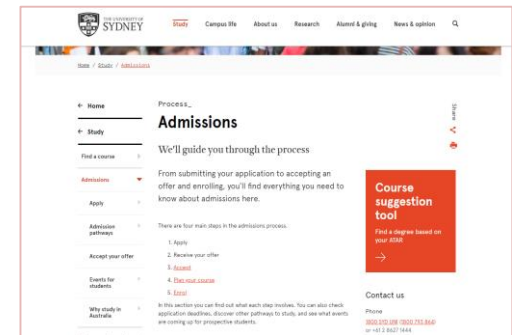

the	language	frequency
10 repetitions in a 100-word article	8 repetitions in a 100-word article	5 repetitions in a 100-word article
VS.	VS.	VS.
90,000 repetitions in 1,000,000 words	4000 repetitions in 1,000,000 words	95 repetitions in 1,000,000 words
low importance	average importance	high importance
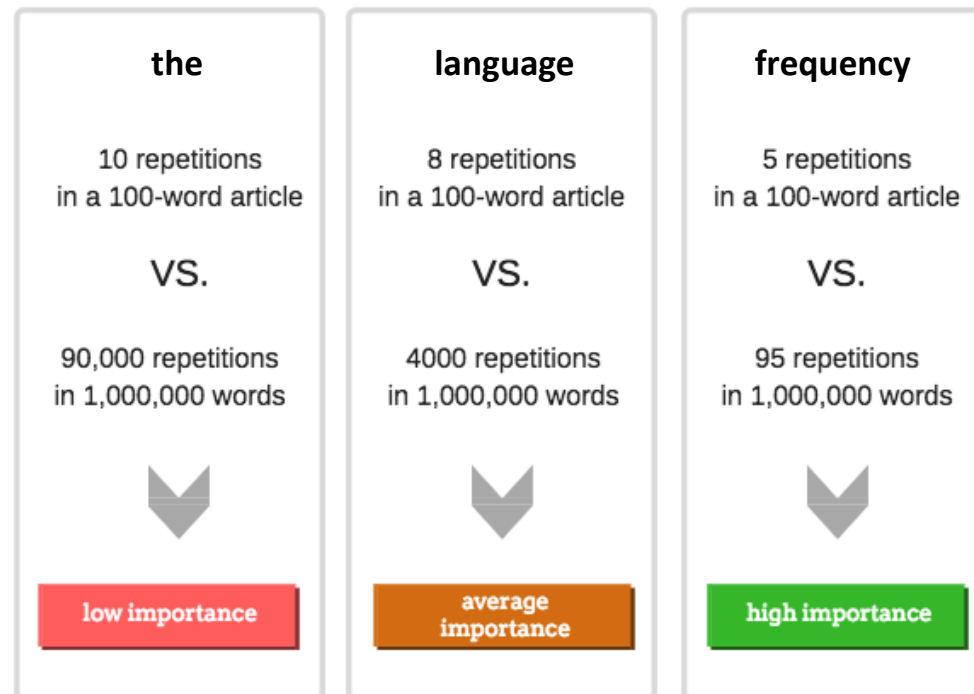
Can we use Term Frequency Only?

University of Sydney Website

Webpage#1

Webpage#2

Webpage#3

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?

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+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

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

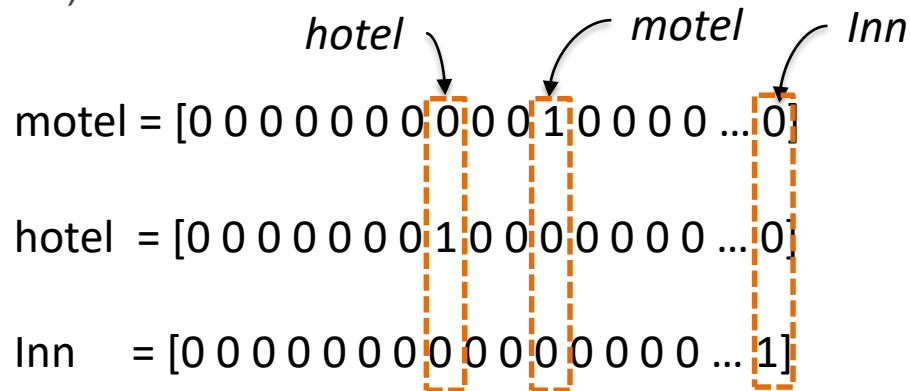*Document #3: Sweet and yellow banana banana*

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

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.

LECTURE PLAN

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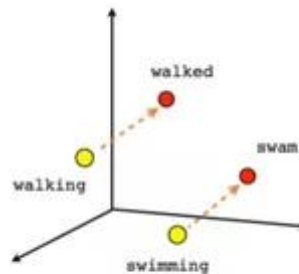
1. Course Introduction
2. Overview of Natural Language Processing (NLP)
3. Word Meaning and Representation
4. **Count-based Word Representation**
   - One-hot Encoding
   - Bag of Words
   - Term Frequency-Inverse Document Frequency
5. Next Week Preview
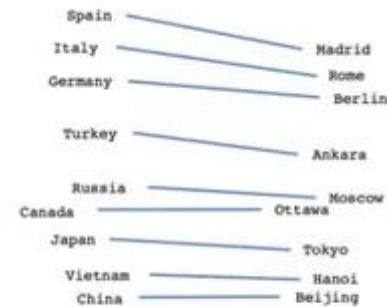
How to Represent the Word Similarity!

- How to represent the word similarity with dense vector
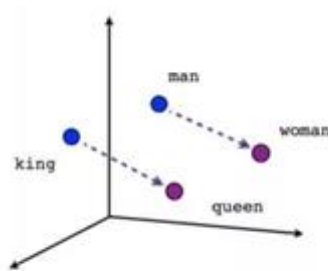
Male-Female   Verb tense   Country-Capital

- 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