

COMP 8420 Advanced NLP

WEEK 1: NLP FUNDAMENTALS

Date: 28th Feb 2025



NLP Fundamentals

REVIEW OF NATURAL LANGUAGE PROCESSING

Date: 28th Feb 2025



NLP Fundamentals

YOU SHOULD ***HAVE KNOWN
IT***, OR YOU CAN LEARN IT IN
A COUPLE OF WEEKS.

Start from string

- What are texts represented in programs?

“I love NLP!”

[‘I’, ‘ ’, ‘l’, ‘o’, ‘v’, ‘e’, ‘ ’, ‘N’, ‘L’, ‘P’, ‘!’]

- ASCII and UTF-8

“I love NLP!”

- Chunking by white spaces and symbols:
[‘I’, ‘love’, ‘NLP’, ‘!’]
- Sometimes you have tokens like:
[‘I’, ‘lo#’, ‘#ve’, ‘NLP’, ‘!’]

Tokenisation

Example:

```
>>> sentence = """At eight o'clock on Thursday morning  
... Arthur didn't feel very good."""  
>>> tokens = nltk.word_tokenize(sentence)  
>>> tokens  
['At', 'eight', 'o'clock', 'on', 'Thursday', 'morning',  
'Arthur', 'did', 'n't', 'feel', 'very', 'good', '.']
```

Part-of-Speech Tagging

- Now we have: “I love NLP!”
-> [‘I’, ‘love’, ‘NLP’, ‘!']
- What are the grammatical roles:
[‘I’, ‘love’, ‘NLP’, ‘!']
[‘Pronoun’, ‘Verb’, ‘Noun’, ‘Symbol’]

Demo: <https://parts-of-speech.info/>

Part-of-Speech Tagging

Example:

```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
 'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'),
 ('Thursday', 'NNP'), ('morning', 'NN')]
```


- Bag of words:
 - ['I', 'love', 'NLP', '!']
 - {'I', 'love', 'NLP', '!'}
- N-gram:
 - uni-gram: {'I', 'love', 'NLP', '!'}
 - bi-gram: {'I_love', 'love_NLP', 'NLP_!'}
- Other combinations:
 - uni-gram+POS: {'I_Pronoun', 'love_verb', 'NLP_noun'}

- Pros of high-order features:
 - More features (e.g., for classification)
 - Disambiguation (e.g. on semantic)
- Cons of high-order features:
 - Overfit training set
 - Time complexity
 - Storage space

Common NLP Libraries

- NLTK: www.nltk.org
 - Pip/conda install nltk
 - nltk.download()
 - punkt, wordnet, gutenbergl (corpus), etc.
- spaCy: <https://spacy.io>
 - POS
 - Parsing
 - Word vectors

Two applications

INFORMATION RETRIEVAL

TEXT CLASSIFICATION

Information Retrieval

- Google
- Bing
- Baidu
- ...

- Text (Document / Query) -> Vector
 - One-hot
 - TF-IDF
 - Word vector
- Similarity: cosine similarity and others
- Evaluation

Document Similarity

Doc 1:

Macquarie University, located in Sydney, Australia, is renowned for its innovative research, vibrant campus culture, and interdisciplinary academic programs. Emphasizing practical learning, the institution fosters collaboration between students and industry leaders. Its commitment to sustainability, community engagement, and global partnerships distinguishes Macquarie as a leading modern university with academic excellence.

Doc 2:

Macquarie University excels in blending tradition and innovation on its picturesque campus near Sydney's suburbs. It offers a wide range of courses, world-class research opportunities, and state-of-the-art facilities. Students enjoy diverse extracurricular activities, multicultural experiences, and career development programs that prepare them for global challenges, nurturing future innovative leaders worldwide.

Document Similarity

Doc 1:

Macquarie University, located in Sydney, Australia, is renowned for its innovative research, vibrant campus culture, and interdisciplinary academic programs. Emphasizing practical learning, the institution fosters collaboration between students and industry leaders. Its commitment to sustainability, community engagement, and global partnerships distinguishes Macquarie as a leading modern university with academic excellence.

Doc 2:

Harvard University is a prestigious Ivy League institution known worldwide for academic excellence, innovative research, and influential alumni. Founded in 1636, Harvard offers diverse programs, world-class faculty, and a vibrant campus community. Its commitment to scholarship, leadership, and public service shapes future generations and drives global progress with enduring impact.

One-hot

Does the word appear in the document? [1: yes and 0: no]

Vectorise the representation.

Macquarie University, located in Sydney, Australia, is renowned for its innovative research, vibrant campus culture, and interdisciplinary academic programs. Emphasizing practical learning, the institution fosters collaboration between students and industry leaders. Its commitment to sustainability, community engagement, and global partnerships distinguishes Macquarie as a leading modern university with academic excellence.

‘Macquarie’: ?,
‘university’: ?,
‘academic’: ?
‘car’: ?

‘Macquarie’: 1,
‘university’: 1,
‘academic’: 1
‘car’: 0

Term Frequency

Doc 1: ‘Macquarie’: ?,
 ‘university’: ?,
 ‘academic’: ?

Macquarie University, located in Sydney, Australia, is renowned for its innovative research, vibrant campus culture, and interdisciplinary academic programs. Emphasizing practical learning, the institution fosters collaboration between students and industry leaders. Its commitment to sustainability, community engagement, and global partnerships distinguishes Macquarie as a leading modern university with academic excellence.

Doc 2: ‘Macquarie’: ?,
 ‘university’: ?,
 ‘academic’: ?

Macquarie University excels in blending tradition and innovation on its picturesque campus near Sydney’s suburbs. It offers a wide range of courses, world-class research opportunities, and state-of-the-art facilities. Students enjoy diverse extracurricular activities, multicultural experiences, and career development programs that prepare them for global challenges, nurturing future innovative leaders worldwide.

Term Frequency

Doc 1:

‘Macquarie’: ?,
‘university’: ?,
‘academic’: ?

Macquarie University, located in Sydney, Australia, is renowned for its innovative research, vibrant campus culture, and interdisciplinary academic programs. Emphasizing practical learning, the institution fosters collaboration between students and industry leaders. Its commitment to sustainability, community engagement, and global partnerships distinguishes Macquarie as a leading modern university with academic excellence.

Doc 2:

‘Macquarie’: ?,
‘university’: ?,
‘academic’: ?

Harvard University is a prestigious Ivy League institution known worldwide for academic excellence, innovative research, and influential alumni. Founded in 1636, Harvard offers diverse programs, world-class faculty, and a vibrant campus community. Its commitment to scholarship, leadership, and public service shapes future generations and drives global progress with enduring impact.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

where

TF (term frequency) is the relative frequency of term t within document d

$$\text{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

IDF (inverse document frequency) is a measure of how much information the word provides

$$\text{idf}(t, D) = \log \frac{N}{|\{d : d \in D \text{ and } t \in d\}|}$$

Cosine Similarity

$$\cos(v, u) = \frac{\langle v, u \rangle}{\|v\| \cdot \|u\|}$$

where

$$\langle v, u \rangle = \sum_i v_i u_i$$
$$\|v\| = \sqrt{\langle v, v \rangle}$$

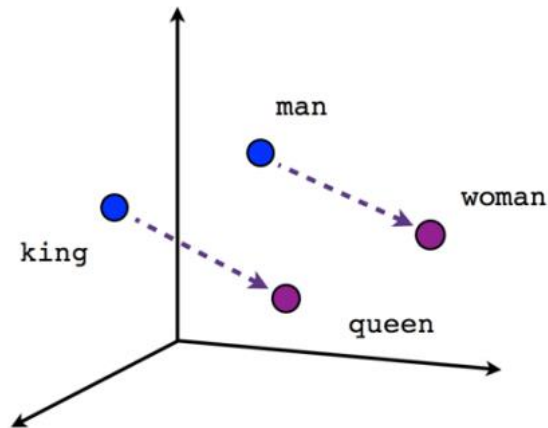
Text Classification

- Text (Document) -> Vector
 - One-hot
 - TF-IDF
 - Word vector
- Machine Learning (ML) models:
 - Linear regression
 - Logistic regression
- Evaluation

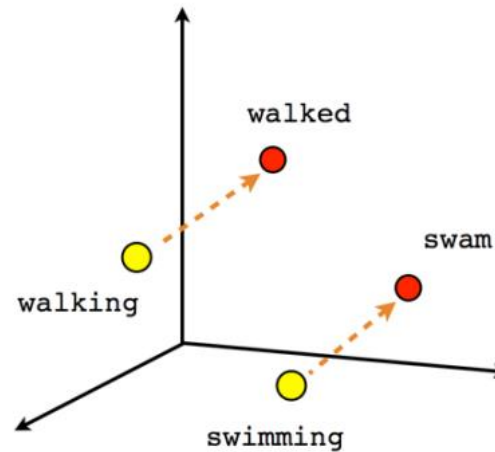
Word Vectors

		Dimensions					
Word vectors	dog	-0.4	0.37	0.02	-0.34	animal	
	cat	-0.15	-0.02	-0.23	-0.23	domesticated	
	lion	0.19	-0.4	0.35	-0.48	pet	
	tiger	-0.08	0.31	0.56	0.07	fluffy	
	elephant	-0.04	-0.09	0.11	-0.06		
	cheetah	0.27	-0.28	-0.2	-0.43		
	monkey	-0.02	-0.67	-0.21	-0.48		
	rabbit	-0.04	-0.3	-0.18	-0.47		
	mouse	0.09	-0.46	-0.35	-0.24		
	rat	0.21	-0.48	-0.56	-0.37		

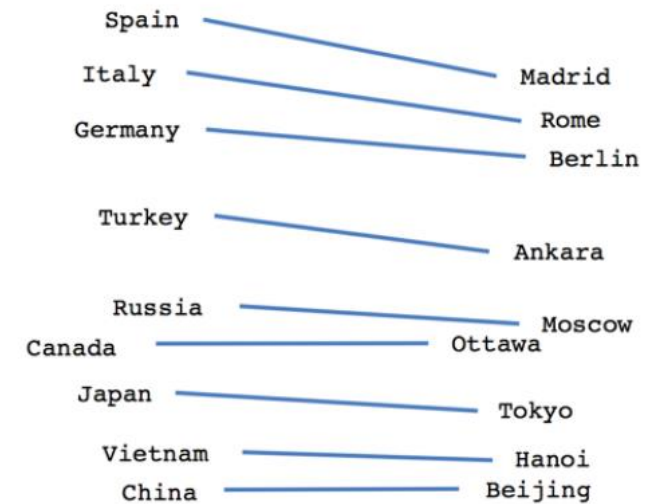
Word Vectors



Male-Female

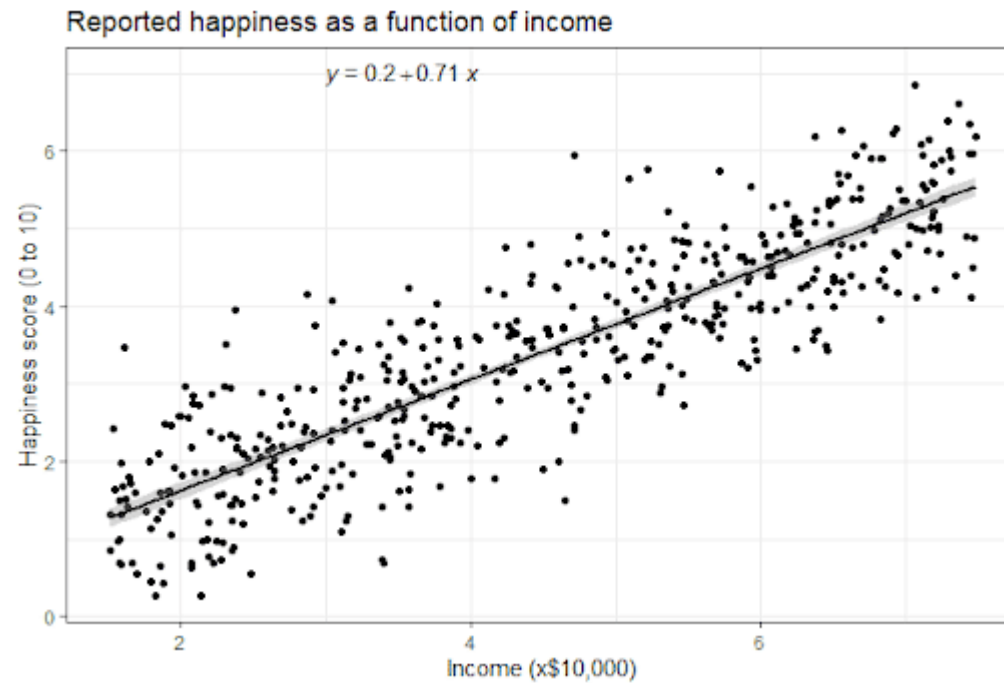


Verb tense



Country-Capital

Linear Regression



Value for prediction

$$Y = \mathbf{W} \mathbf{X} + b$$

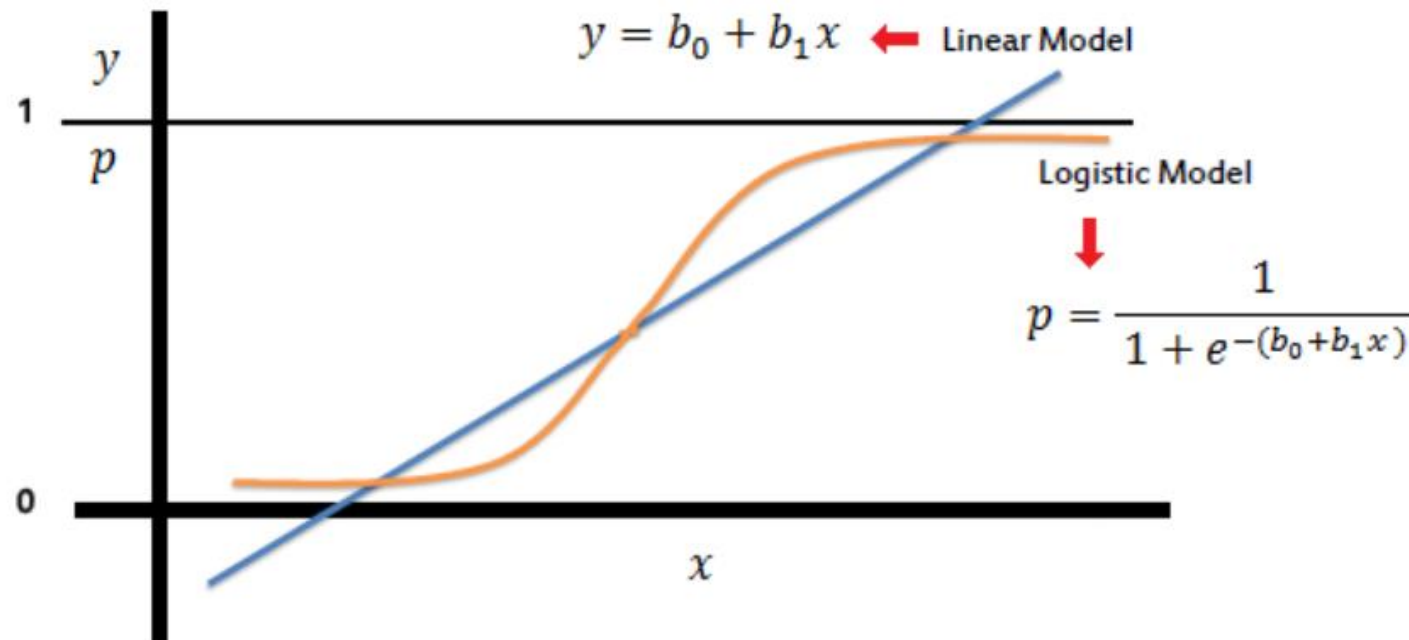
↑

→ Bias

↓ Input values

↓ Weights

Logistic Regression



This material is provided to you as a Macquarie University student for your individual research and study purposes only. **You cannot share this material publicly online without permission.** Macquarie University is the copyright owner of (or has licence to use) the intellectual property in this material. Legal and/or disciplinary actions may be taken if this material is shared without the University's written permission.

Common ML Libraries

- Scikit-learn: https://scikit-learn.org/stable/getting_started.html
 - Regression models
 - Pipeline for processing datasets
 - Analysis
- Gensim: <https://pypi.org/project/gensim/>
 - Word vectors
 - Topic model and latent semantic models

Evaluation – Accuracy & F1

$$\text{Acc} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1} = 2 \text{ P} * \text{R} / (\text{P} + \text{R})$$

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN


Connection to Assignment 1

MOST OF THE TECHNIQUES WILL BE USED
IN SOLVING QUESTIONS IN ASSIGNMENT 1.

YOU SHOULD HAVE BEEN FAMILIAR WITH
THESE CONCEPTS AND RELATED TOOLS.

Connection to Major Project

THEY ARE BASIC TOOLS FOR NLP, WHICH
YOU CAN USE TO SOLVE YOUR BUSINESS
CHALLENGES IN YOUR MAJOR PROJECT.



Workshops:
15:00-17:00 06EaR 118
17:00-19:00 06EaR 206

Questions & Answers