COMP4650/6490 Document Analysis

Course Review & Final Exam

ANU School of Computing

Administrative matters

- SELT evaluations
- Quiz 3
  - Closes: 5pm on Thursday 26 October
- Practice exercise solution
  - Will be available on Tuesday
- Assignment 3
  - Results will be released later this week
- Final exam coversheet
  - Will be released on Wattle later this week
- Drop-in sessions
  - 1pm - 3pm Thursday 26 October
  - 1pm - 2pm Friday 27 October
  - Location: Room 3.41, Level 3, Hanna Neumann Building
- NLP-related guest lecture - Dr. Zheng Yuan (King’s College London)
  - 1pm - 2pm Monday 23 October, Manning Clark Hall, Kambri

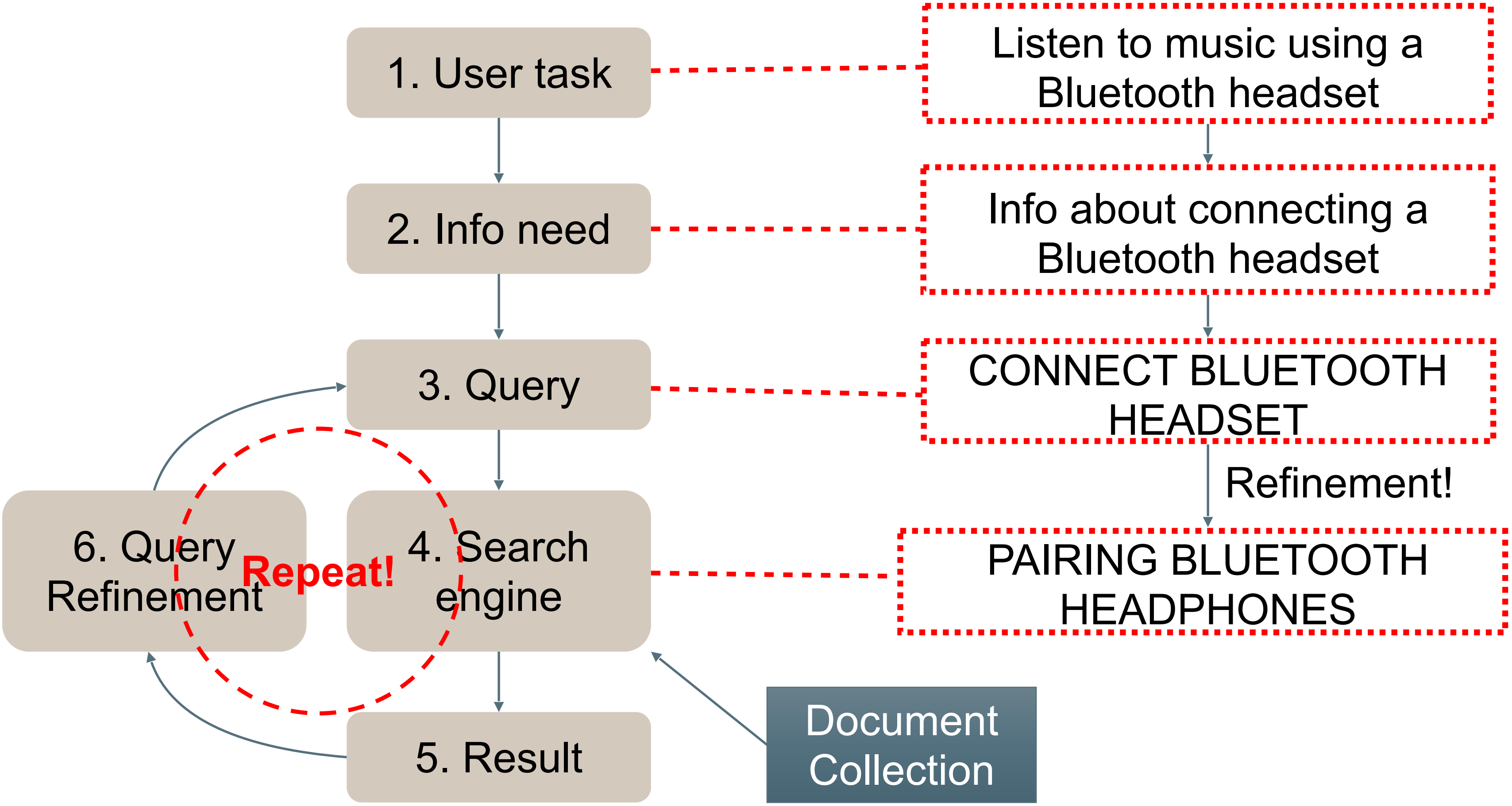
Course Summary

- Information retrieval (IR)
  - How can computers identify relevant information?
- Machine learning (ML) for NLP
  - How can computers learn from data?
- Natural language processing (NLP)
  - How can computers understand human language?

Course Summary

IR: Introduction

Classic search model


1. User task
2. Info need
3. Query
4. Search engine
5. Result
6. Query Refinement
Repeat!
Document Collection
Listen to music using a Bluetooth headset
Info about connecting a Bluetooth headset
CONNECT BLUETOOTH HEADSET
Refinement!
PAIRING BLUETOOTH HEADPHONES

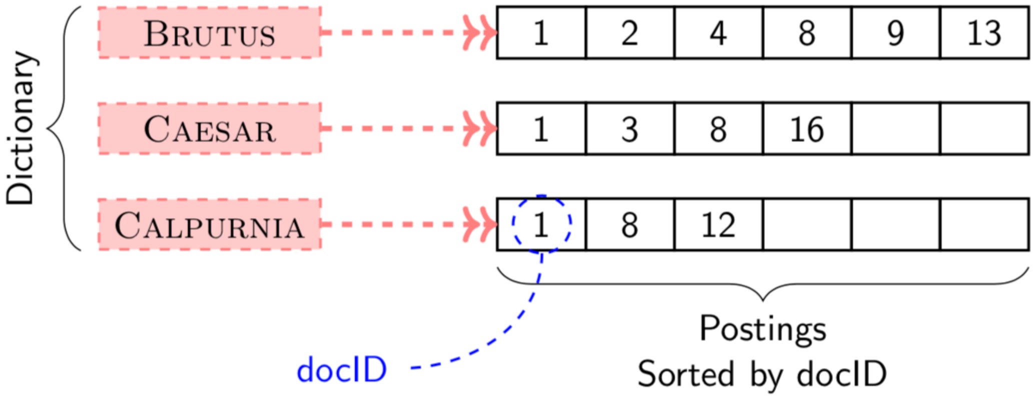
Course Summary

IR: Boolean Retrieval

- Term-Document Incidence Matrix

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

- Inverted Index
- Boolean retrieval



IR: Ranked Retrieval

Definition (TF-IDF)

The tf-idf weight of term t in document d is as follows:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

Definition (Inverse Document Frequency (IDF))

Let df_t be the number of documents in the collection that contain a term t. The inverse document frequency (IDF) can be defined as follows:

$$\text{idf}_t = \log \frac{N}{\text{df}_t}$$

where N is the total number of documents.

Vector Space Model

- Representing both documents and queries as vectors (e.g. TF-IDF vectors)
- Ranking documents using the similarities (e.g. cosine similarity) between the query vector and the vectors of documents

Course Summary

IR: Evaluation

- Evaluation of unranked retrieval results (e.g. Boolean retrieval)
  - Precision, Recall and F-measure
  - Accuracy is not appropriate for IR
- Evaluation of ranked retrieval results
  - Precision-Recall Curve (Precision and Recall of the top-k retrieved documents)
  - Interpolated Precision

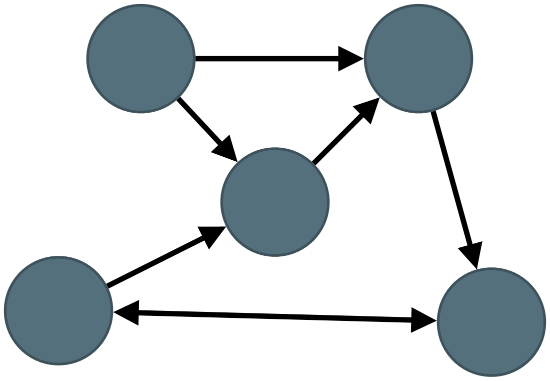
    $$p_{\text{interp}}(r) = \max_{r' \geq r} p(r')$$

  - Average Precision, MAP, Mean Reciprocal Rank, etc.

Course Summary

IR: Web search

- Documents on the web are linked by hyperlinks
- Authorities and Hubs
- Hyperlink-Induced Topic Search (HITS) algorithm
- PageRank



Course Summary

ML: Basics

- Linear models
  - Linear Regression, (Multinomial) Logistic Regression
  - MSE and cross entropy loss
  - Gradient descent
- Practical considerations in ML
  - Train, validation, test setup
  - Feature standardisation
  - Bias-Variance trade-off and Regularisation
  - Hyper-parameter tuning

Course Summary

ML: Representation

- Simple document representation (BoW model)
  - Binary occurrence, word count, TF-IDF vectors
- Word representation
  - One-hot, word co-occurrence (weighted by Positive Pointwise Mutual Information, PPMI)
  - Word2Vec

Course Summary

ML: Clustering

- Unsupervised learning
- Flat vs. Hierarchical clustering
- Hard vs. Soft clustering
- K-Means algorithm
- Evaluating clustering
  - Internal criteria: RSS in K-Means
  - External criteria: Evaluate w.r.t. human-defined classification, e.g. Purity

Course Summary

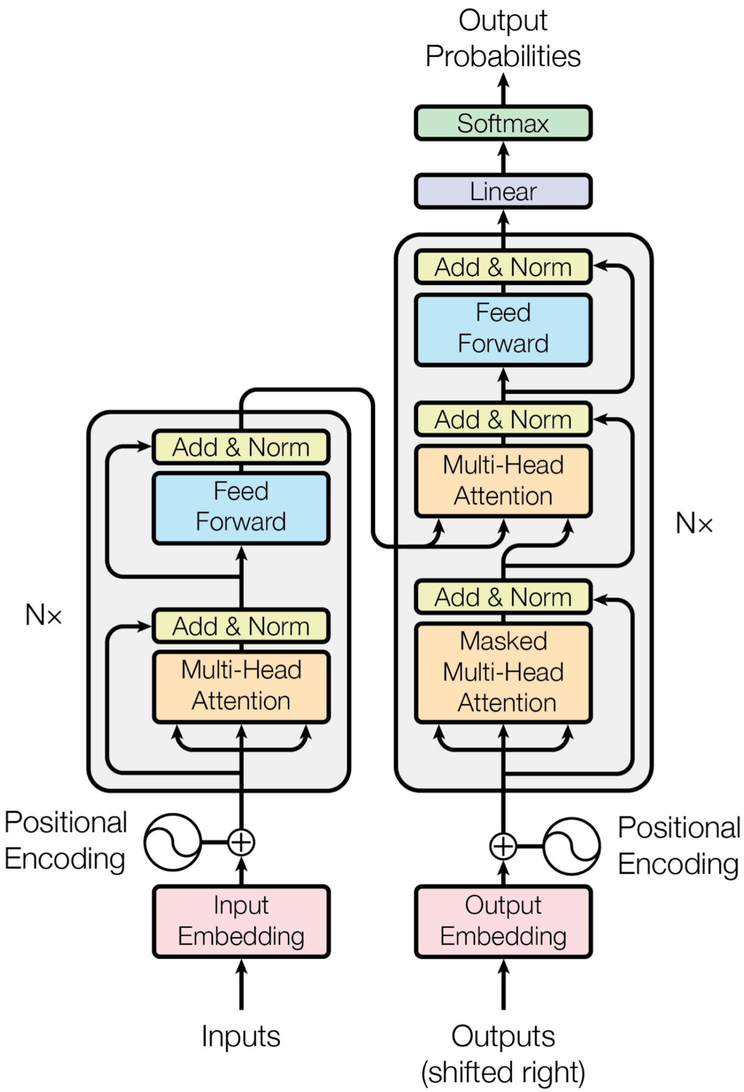
ML: Deep Neural Networks

- Feedforward neural network
  - From logistic regression to Feedforward NN
  - Non-linear activation functions
- DNN training
  - Computation graph, back-propagation, SGD
- Recurrent Neural Network
  - Simple RNN for sequences
  - Back-propagation through time
  - Vanishing/exploding gradients
  - Better architectures (e.g. GRU, LSTM) allow learning of longer temporal dependencies

Course Summary

ML: Attention and Transformers

- Attention mechanism
  - A neural network layer that learns to select out relevant parts of the input
  - Attention in Seq2Seq (encoder-decoder) models
  - Query-Key-Value attention
- Transformers
  - Self-attention
  - (Masked) Multi-head attention
  - Positional encodings
  - Residual connection
  - Layer normalisation



ML: Pre-training & Transfer Learning

- Neural Language Models
  - Generating text (auto-regressively)
  - Computing the likelihood of a text sequence
- Self-supervised learning
  - Use naturally existed supervision signals for training
- Pre-trained language models
  - Context specific word representations
  - ELMo, BERT, GPT models
- Transfer learning through fine-tuning
  - Task-specific layers on top of a pre-trained model
  - Fine-tuning using a small amount of (task-specific) labelled data
  - Freeze or make minimal adjustments to parameters of pre-trained models (e.g. using a small learning rate)

Course Summary

NLP: Language Modelling & Smoothing

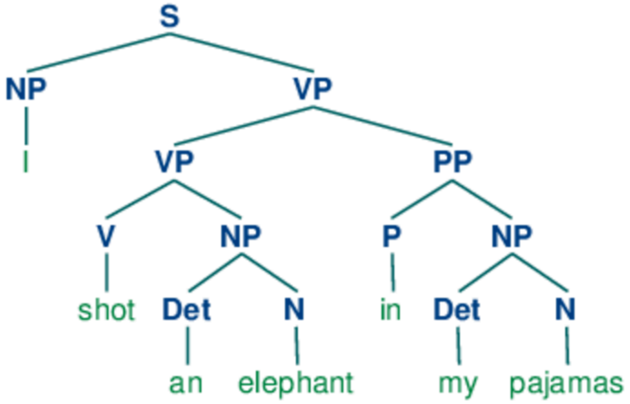
- N-gram language models
  - Markov assumption and N-gram LMs
  - Maximum likelihood estimation of N-gram probabilities
- Smoothing
  - Deal with overfitting (zero probability)
  - Adjusting low probabilities upwards and high probabilities downwards
  - Interpolation, Absolute Discounting, Kneser-Ney Smoothing, Stupid Backoff
- Evaluation of language models
  - Extrinsic evaluation: Put model in a task
  - Intrinsic evaluation:

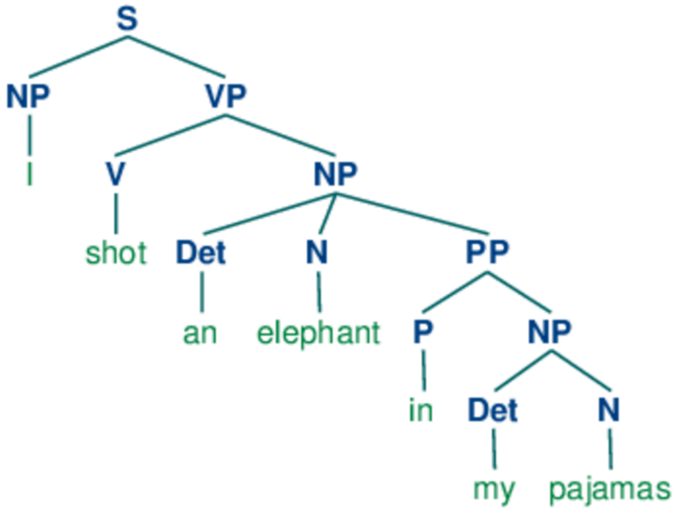
    e.g. Perplexity $PP(x_{1:L}) = P(x_{1:L})^{-\frac{1}{L}} = \sqrt[L]{\frac{1}{P(x_{1:L})}}$

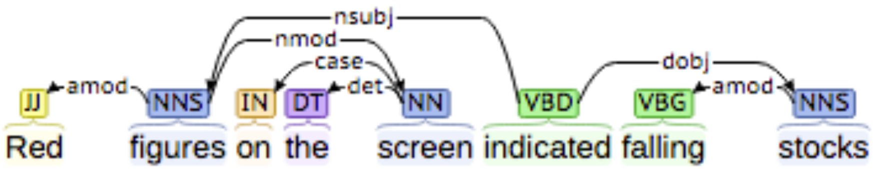
Course Summary

NLP: Syntactic Parsing

- Constituency parsing
  - (Probabilistic) Context Free Grammar
  - Nodes represents phrases in a phrase structure tree
  - Structural ambiguity
- Dependency parsing
  - Dependencies between words
  - Nodes represent words
  - Edges represent dependencies







Course Summary

NLP: Semantics

- Meaning representation
  - Unambiguous, Linking to external knowledge,
    Supporting computational inference, Sufficiently expressive
- Logical semantics
  - Using λ-calculus:
    e.g. Alex likes Sam → $(\lambda x\,.\,\text{LIKES}(x, \text{SAM}))@\text{ALEX}$ → LIKES(ALEX, SAM)
- Predicate-argument semantics
  - A light semantic representation
    e.g. (arg1: someone) read (arg2: something)
- Lexical semantics
  - What is the meaning of words
  - How are the meanings of different words related
- Coreference Resolution

Course Summary

NLP: Additional Topics

- Evaluation in NLP
  - Classification metrics (e.g. precision, recall, F-measure, and their macro-/micro-averaged counter-parts)
  - Threshold free metrics (e.g. ROC-AUC)
  - Term overlap metrics (e.g. BLEU and ROUGE)
  - What to compare: algorithms, features sets, baselines; ablation studies; different datasets.
- Multi-lingual NLP
- Low resource NLP

Final Exam

- Monday 6 November, 2:00pm - 4:15pm AEDT
- Duration:
  - Writing time: 120 minutes (more for students with EAPs + SEAs)
  - Reading time: 15 minutes
- Venue:
  - CSIT N111, N112, N113, N114, N115-N116
  - Hanna Neumann Lab 1.23 & 1.24
  - Room and Seat allocation communicated through the ANU Examinations Office
- Lab exam
  - Please log into the Student Registration and Marks System (https://cs.anu.edu.au/streams/index.php) at least once before the exam
  - This registers your UID so that you will be able log into a lab computer
  - Necessary if you have never logged into a lab computer

Final Exam

- Closed-book lab exam
  - Access to course notes (including lecture and lab notes, quizzes and assignments) on Wattle is permitted
  - Can post questions privately to all instructors/staff on the course forum
  - Access to other materials or websites (except the course Wattle site and the course forum) is NOT permitted
- Exam will be centrally invigilated by the ANU with support from the course team
- Make sure to carefully read the coversheet on Wattle (To be released)
- Do NOT discuss exam questions nor answers during nor after the exam (some students will likely have special exam times)
- Any such behaviour is academic misconduct and treated as a very serious matter

Final Exam

- The exam is set as an online quiz on Wattle, with a mix of multiple choice, text, and numerical questions
- Main focus is on understanding concepts and techniques
- No programming is required as answers
- Some (simple) calculations will be required (using a non-programmable calculator is permitted)

Final Exam

- Questions will cover:
  - All lecture material
  - All tutorial / lab material
  - All online quizzes
  - Material from assignments
- Weighting of questions roughly corresponds to the coverage of a topic in the course

Final Exam

- Keep textual answers short and to the point
- Read questions carefully – do not answer X if the question is Y
- Write in clear English – if we cannot understand your writing you might lose marks
- Answers must be written in your own words – if you directly copy sentences from lecture slides or other sources you might not receive marks (and we can detect this)

Final Exam

- Exam is worth 60% of final mark
- To pass the course, you need a total mark of at least 50 out of 100
- You will also need to obtain at least 50% in the final exam (hurdle assessment)
- Your final course mark consists of:
  - Three Assignment marks, each is worth 10%
  - Quiz marks: Q1 (3%), Q2 (4%), Q3 (3%)
  - Exam mark (60%)
- Final mark is (**possibly, we aim not to**) subject to some scaling as a result of school or college academic review

Final Exam

- Supplementary examination will be offered to any student who has
  - passed the hurdle assessment AND has achieved a final overall mark of at least 45% and less than 50%; OR
  - achieved between 45% and 49% for the hurdle assessment, and if that assessment item were passed, would otherwise pass the course

Thank you

Best of luck with the final exam

We hope you enjoyed the course and learned something new