

School of Computing and Information Systems  
The University of Melbourne  
COMP90042 NATURAL LANGUAGE PROCESSING (Semester 1, 2021)

Workshop exercises: Week 12

### Discussion

1. What is **Summarisation**?
  - (a) What is **log likelihood ratio**, and how is it useful for summarisation?
  - (b) What is the log likelihood ratio of word  $w$  in document  $d$  if:  $F_d(w) = 1$ ,  $F_b(w) = 20$ ,  $N_d = 30$ , and  $N_b = 4000$ , where  $b$  denotes the background corpus,  $F$  the frequency and  $N$  the total number of word tokens.
2. You have an idea of building a **comment generation system for news articles**. Training data can be created by mining news articles and their comments on the web. What are the ethical implications of such application? Discuss.

### Programming

1. In the iPython notebook `13-embedding-bias`, we build a sentiment classifier using pre-trained word embeddings, and explore the hidden biases contained in the embeddings.
  - Modify the code to remove stopwords and only consider non-stopwords when computing the mean sentiment of a sentence. Does that change the results?
  - Test Word2Vec pre-trained embeddings and see if they contain similar biases. Pre-trained Word2Vec embeddings can be downloaded here (note: it's 1.5GB in size).
  - Revisit the notebook `10-bert`, where we build a sentiment classifier using pre-trained BERT. Test whether this BERT-based sentiment classifier contains similar biases and explain your observations.