

# COMP348

## Summary: First Half of Unit

---

## Unit Overview: Concepts

---

- range from preprocessing (low-level) to applications (high-level)
- topics:
  - overview of LT
  - tokenisation (P)
  - sentence segmentation (P)
  - morphological analysis (P)
  - statistics and evaluation
  - text classification (A)
  - machine translation (A)

# Unit Overview: Technical Detail

---

- topics:
  - code
  - regular expressions
  - Porter Stemmer
  - evaluation via precision, recall, accuracy
  - frequency counts (n-grams)
  - conditional probability
  - classification via Naïve Bayes
  - classification via kNN
  - use of Information Gain / Mutual Information in feature selection
  - BLEU and machine translation evaluation
  - constructing grammars and parses
  - constructing rules for transfer-based machine translation
  - statistical machine translation