# IN4080 Exam prep: Vector semantics

## Lexical semantics

Which lexical relationships exist between the words in the following pairs:

- car-vehicle
- car-automobile
- car-engine
- tall-high
- tall-gall
- tall-short

## Vectors, TF-IDF and Cosine

Consider the following documents, each consisting of two sentences:

- Document 1:
  - "john had a ball at the party"
  - "he said it was a great party"
- Document 2:
  - " peter hit the ball with the bat"
  - "he said it was a great hit"
- Document 3:
  - "paul dressed up as a bat at the party"
  - "he said the party was a hit"

a) Construct a term document matrix for the three documents with raw counts.
Write it up as a table with the terms as rows and documents as columns.

b) Construct an IDF vector for the terms, using the following formula for IDF: $\log \frac{N}{n_t}$,
where $N$ is the size of the corpus, and $n_t$ is the number of documents that contain .
Write it up as a table with the terms as rows and the IDF values as a single column.

c) Construct a TFIDF weighted term document matrix for the three documents. For the TF part of TF-IDF, use raw counts, i.e. no normalization for the term frequencies.

d) Using the TFIDF weighted term document matrix and the formula for cosine similarity, which of the documents are most similar.

e) Some of the terms do not affect the cosine similarities between the documents. List these terms.

# Word2Vec

Negative sampling is an important part of the Word2Vec algorithm. Negative sampling is done by sampling from the probabilities of the words in the vocabulary, given by this formula:

$$P_\alpha(w) = \frac{count(w)^\alpha}{\sum_w count(w)^\alpha}$$

a) What happens when you change $\alpha$? Explain in your own words
b) What is the motivation of changing the original word probabilities?