text analytics:

* activities:
  + information retrieval
  + text mining
  + web mining
  + natural language process
  + document classification
  + document clustering
  + sentiment analysis
  + topic analysis
* aim to extract useful knowledge from text data

terminology:

* document: a piece of text
* term/token: usually word
* corpus: collection of all documents to be analysed
* dictionary (feature set): all features in the corpus

vector space model:

* document text is converted to a bag of words (tokens) which simplified to a term-document matrix
* word count in each document is then treated as orthogonal vectors in n-dimensional space
* the angle between documents indicates their degree of similarity
* ‘bag of words’ approach:
  + each document assumed to be just a collection of words
  + make implicit assumptions that the order of the words in a document does not matter
  + syntactically similar documents are semantically similar

term-document matrix:

A picture containing text, shoji, crossword puzzle, image

Description automatically generated

Refine ‘bag of words’:

* upper and lower case words usually have the same meaning
* some frequently occurring words are not useful to discriminate between documents
* punctuation not useful
* tense may make similar words appear differently
* groups of words may be important for meaning

**extract structure from text:**

1. tokenise: break up text into tokens

* split into a stream of words
* remove all punctuation marks
* replace tabs and other non-text characters by single white spaces
* merge all remaining words from all documents: form dictionary of corpus

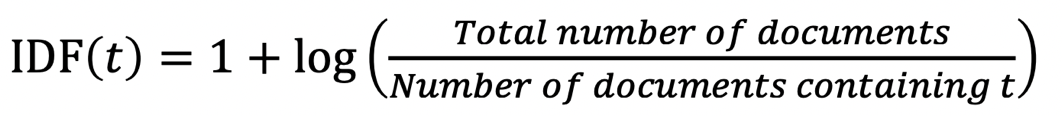
1. convert case
2. remove stop words:

the, of, and, to, a, in, for, is, on, that…

1. stem
2. lemmatise
3. create n-grams

analyse text & documents:

* term importance:
  + term document matrix (TDM): (frequency of a word for a specific document)
    - usually very sparse, most entries=0
    - not too common
    - no too infrequent
  + inverse document frequency (an extension of term document frequency considers the relative number of documents in which a word occurs)
    - assume a word appearing in fewer documents is more likely to be important when it does occur



t: term

* document similarity
  + cosine distance

**TFIDF:**

TFIDF(t,d)=TF(t,d) \* IDF(t)

TF(t,d): term frequency specific to one document (the number of times term t appears in document d)

Cosine distance similarity:

Text, whiteboard

Description automatically generated