



Extracting Features from Texts and Images

Saptarshi Pyne, Assistant Professor
Department of Computational and Data Sciences (CDS)
Indian Institute of Science Education and Research (IISER) Kolkata
Mohanpur, West Bengal, India 741246

CS5103 Applied Machine Learning Lectures 15-17
September 23, 24, 26; 2025

Feature Extraction from Text Documents

A few techniques for extracting features from text documents:

Sections 7.2.1, 7.2.2, 7.2.3,

https://scikit-learn.org/stable/modules/feature_extraction.html

An application: Categorizing text documents by topics.

https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html#classification-of-text-documents-using-sparse-features

Intuition Behind the Tf-Idf of a Term

Suppose, 't' is a term.

$df(t)$ = What is the probability of the term appearing in a document.

$idf(t)$

= Inversely proportional to $df(t)$

= How rare it is for the term to appear in a document.

$tf(t, d)$ = How many times the term appears in document 'd'.

Intuition Behind the Tf-Idf of a Term (contd.)

$\text{tf-idf}(t, d)$

$= \text{idf}(t) * \text{tf}(t, d)$

= How rare it is for the term 't' to appear in a document * How many times does the term 't' appear in document 'd'

= How specific is term 't' for the document 'd'.

Intuition Behind the Tf-Idf of a Term (contd.)

Suppose, the term 't' appears in a very small number of documents.

$df(t)$ = very small.

$idf(t)$ = very large = Term 't' is found very rarely in a document.

Although, the term 't' is found rarely in a document, let us assume that it is found multiple times in document 'd'.

Therefore, the term 't' is extremely specific to document 'd'. We should assign the same class to document 'd' as the class where the training documents having multiple occurrences of term 't' belong to.

Feature Extraction from Image Files

A few techniques for extracting features from image files:

Section 7.2.4,

https://scikit-learn.org/stable/modules/feature_extraction.html

An application: Given an image with multiple coins, detect the coins.

https://scikit-learn.org/stable/auto_examples/cluster/plot_coin_ward_segmentation.html#a-demo-of-structured-ward-hierarchical-clustering-on-an-image-of-coins

Keep Calm
and
Learn Machine Learning

Thank You