

**LECTURER: TAI LE QUY**

**MACHINE LEARNING**

**UNSUPERVISED LEARNING  
AND FEATURE ENGINEERING**

INTRODUCTION TO UNSUPERVISED MACHINE LEARNING AND FEATURE ENGINEERING	1
CLUSTERING	2
DIMENSIONALITY REDUCTION	3
FEATURE ENGINEERING	4
FEATURE SELECTION	5
AUTOMATED FEATURE GENERATION	6

**UNIT 4**

# **FEATURE ENGINEERING**



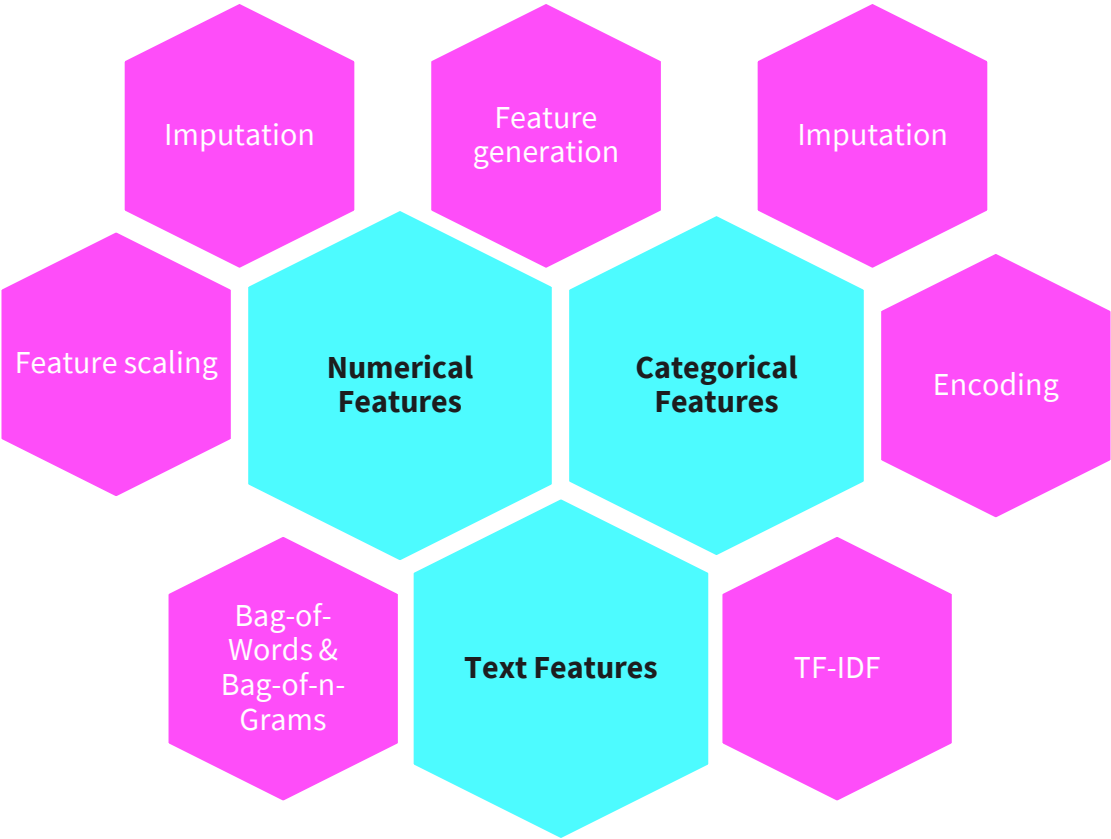
- Explain the difference between **numerical**, **categorical**, and **text** features.
- **Clean, scale, encode, or transform** these features.
- Generate **new features** by **transforming, splitting,** or **grouping** existing features as interaction features.



1. Name and explain **three different methods** to deal with **missing values** in a dataset.
2. Explain why, during data preprocessing, **features** are often **scaled**.
3. Name **three methods** that can be applied to **count words** in text data.

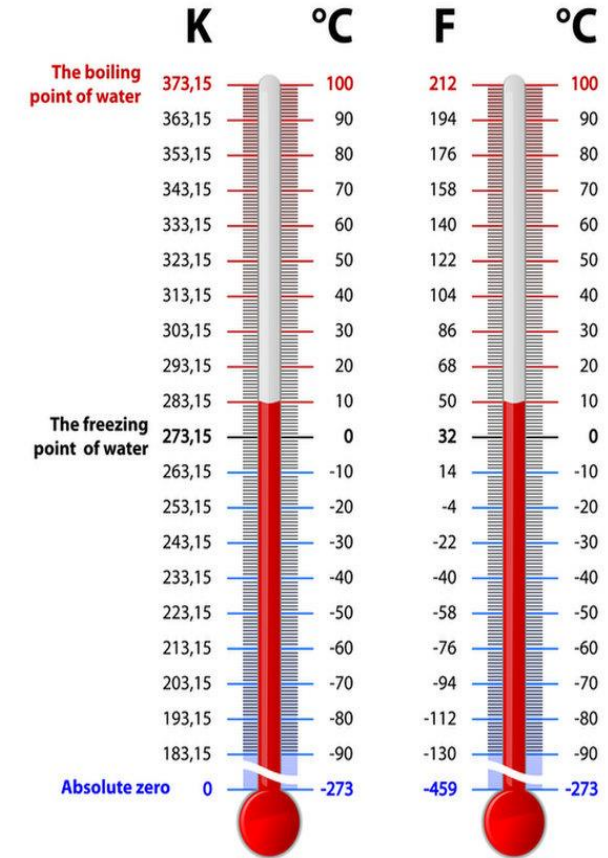
# UNIT CONTENT

Image 1: Unit content - Feature engineering






## NUMERICAL FEATURES

- Interval scales
  - Differences between values are meaningful
    - The difference between  $90^{\circ}$  and  $100^{\circ}$  temperature is the same as the difference between  $40^{\circ}$  and  $50^{\circ}$  temperature.
    - Examples: Calendar dates , Temperature in Fahrenheit or Celsius
    - Ratios till has no meaning
      - A temperature of  $2^{\circ}$  Celsius is not much different than a temperature of  $1^{\circ}$  Celsius
      - The issue is that the  $0^{\circ}$  point of the Celsius scale is in a physical sense arbitrary and therefore the ratio of two Celsius temperatures is not physically meaningful.
- Ratio scales
  - Both differences and ratios have a meaning
  - E.g., age, weight, length, number of sales
    - A 100 kgs person is twice heavy as a 50 kgs person
    - Temperature in Kelvin: When measured on the Kelvin scale, a temperature of  $2^{\circ}$  is, in a physical meaningful way, twice that of a  $1^{\circ}$ . The zero value is absolute 0, represents the complete absence of molecular motion



## Types of missing values

- Missing **Completely At Random** (MCAR) 
- Missing **At Random** (MAR) 
- **Not Missing At Random** (NMAR) 



# IMPUTATION

## Types of imputation

- List-wise imputation / complete list
- Mean /median /default substitution
- k-Nearest Neighbors
- Regression
- Multiple imputation
  - Multiple Imputation by Chained Equations (MICE)

Image 2: Filling the gap



## FEATURE SCALING

### Min-max scaling /normalization

- Resulting range [0, 1]
- $$Z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$
- Susceptible to outliers

Image 3: Scaling



## FEATURE SCALING

### Variance scaling / standardization

- $\text{Mean}(z) = 0, \text{Std}(z) = 1$
- $$Z_i = \frac{x_i - \text{mean}(x)}{\text{std}(x)}$$
- Susceptible to outliers

Image 3: Scaling





## FEATURE SCALING

### Robust scaling

- Similar to standarization

- $$Z_i = \frac{x_i - \text{median}(x)}{IQR(x)}$$

- Not susceptible to outliers

Image 3: Scaling



COMPARISON BETWEEN FEATURE SCALING TECHNIQUES IN THE CASE OF SPARSE DATASET

- Robust scaling has the advantage of being more adapted to scale-sparse data points than min-max scaling and standardization (use of the median and IQR of data points)
- Min-max scaling and standardization compress most of the data points to a narrow range, while robust scaling does much better at keeping the spread of the data points.

C-ID	Consumption	Min-max scaling of consumption	Standardization of consumption	Robust scaling of consumption
1	70	0.004	−0.50	0
2	140	0.015	−0.47	0.93
3	65	0.003	−0.50	−0.06
4	40	0.000	−0.51	−0.4
5	6500	1	2	85.73

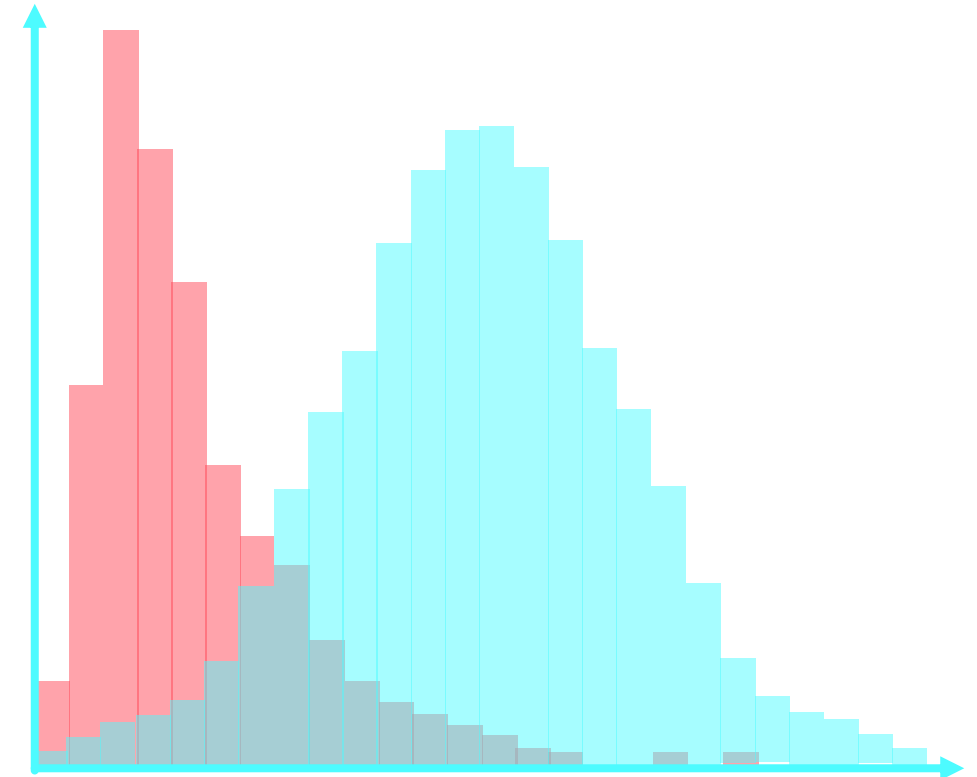
## FEATURE GENERATION

New features can be generated as a transformation or combination of existing features

### Transformation

- Better interpretability
- Meet statistical assumptions

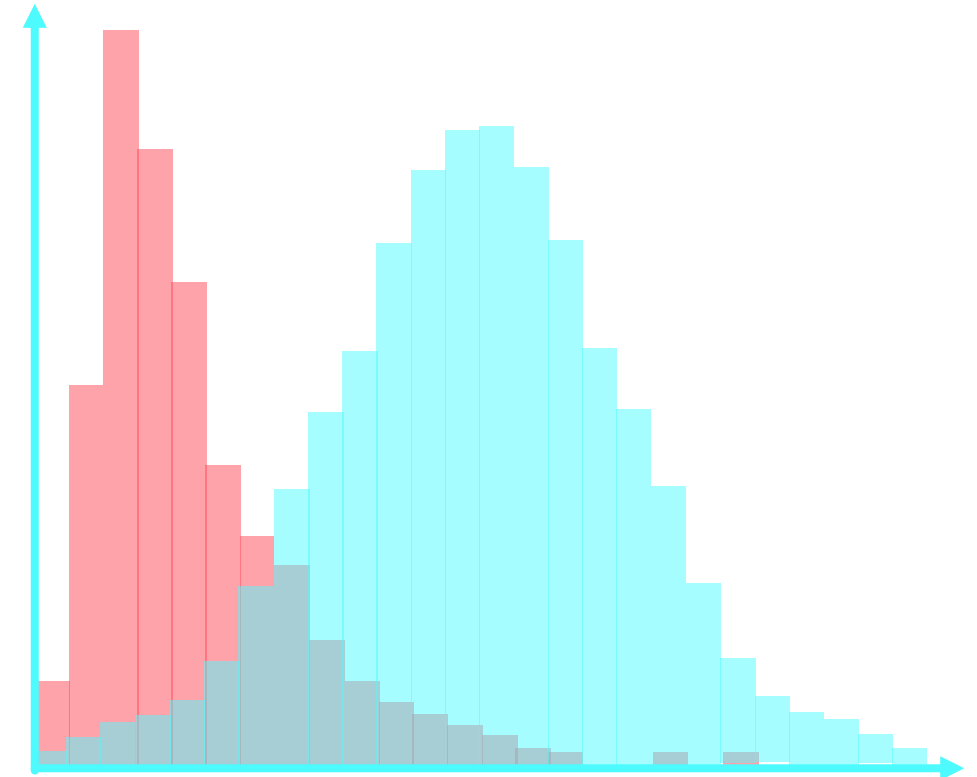
Image 4: Feature transformation



# Transformation types

- Square transformation
- Exponential transformation
- Reciprocal transformation  
Eg. population density → area per person
- Log transformation  
Make a nonlinear relationship more linear  
e.g.,  $\text{Log}(x^2) = 2\text{Log}(x)$   
Transform skewed data points to approximately conform to normal distribution
- Others: min, max, variance, mean, median, sum, count

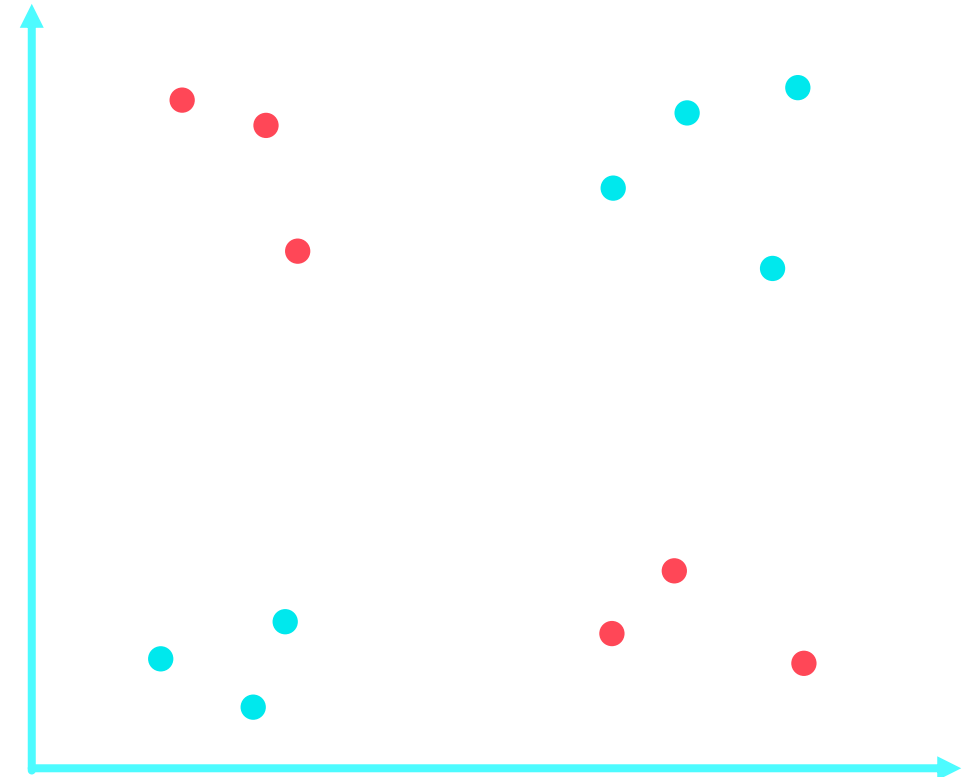
Image 4: Feature transformation



### Interaction features

- Combining information from two or more features
- Polynomial combinations
- Boolean operators

Image 5: Interaction features

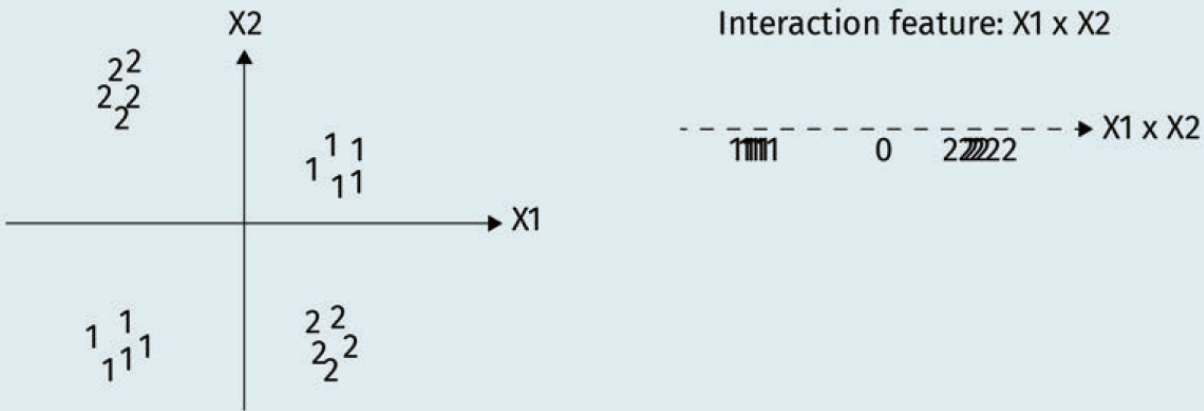




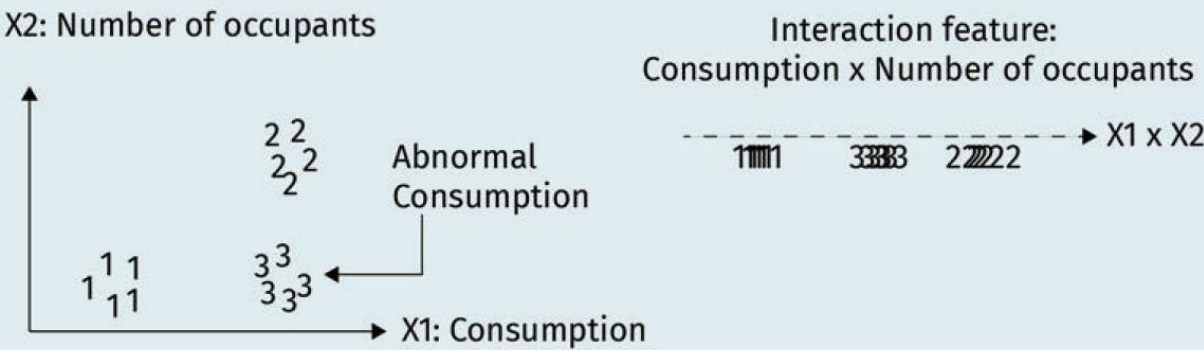
Interaction features

Motivation to Use Cross Product Features for Energy Consumption Example		
2	70	140
4	140	560
2	65	130
1	40	40
2	65	130

Use of an Interaction Feature to Discriminate Two Different Clusters



Interest of the Use of Interaction Features to Discriminate The Abnormal Consumption Cluster



CATEGORICAL FEATURES

Feature Cleaning (Imputing Missing Values)

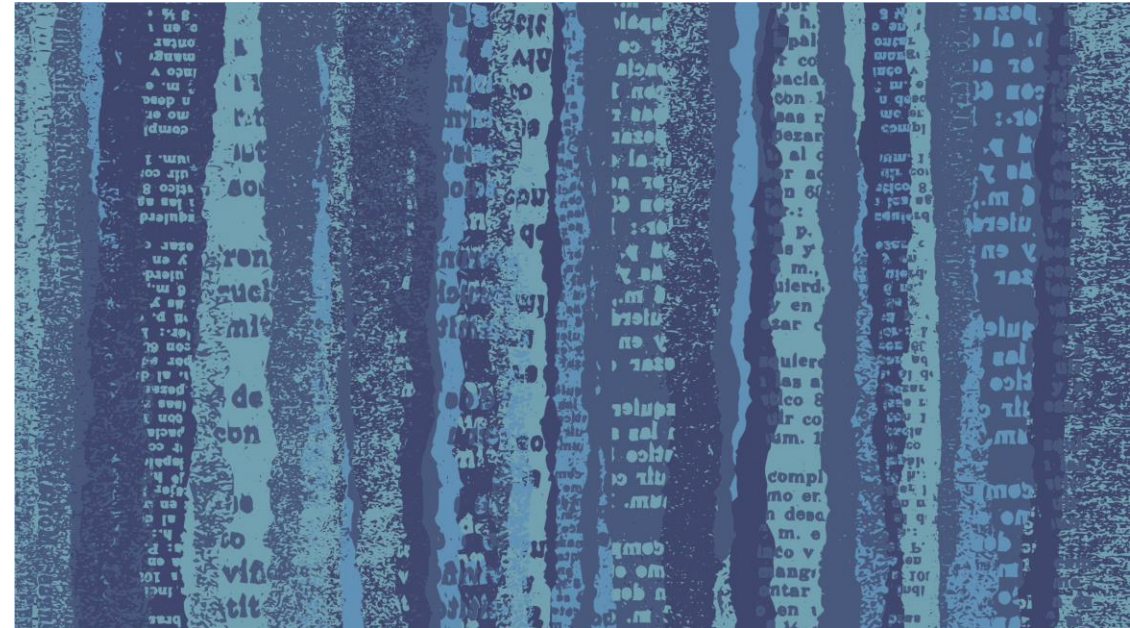
- Missing values can be replaced with the maximum occurred value in the corresponding column (feature)

Imputing a Missing Value in the Gender Feature of the Energy Consumption Example		
C-ID	Gender before imputing	Gender after imputing by replacing by the maximum occurred value
1	M	M
2	NA	M
3	M	M
4	F	F
5	F	F

## FEATURE GENERATION

- **Splitting / spreading**
- **Grouping**
- **Date extraction**
  - Year, month, quarter, day of the week
  - Elapsed time
  - Weekdays, holidays

Image 6: Information in columns



## IMPUTATION FOR CATEGORICAL DATA

### Types of imputation

- List-wise imputation / complete list
- Mode /default substitution
- k-Nearest Neighbors
- Classification
- Multiple imputation
  - Multiple Imputation by Chained Equations (MICE)

Image 2: Filling the gap





FEATURE ENCODING

- **Translate** each category to a numeric value
- **One-hot encoding**
  - One column for each unique category
  - 1 for applicable
  - 0 for non-applicable

Image 7: Zeros and ones



Source of the image: Microsoft Archive.

One-Hot Encoding for Gender and Work-types Features of a Table				
Gender-M	Gender-F	Work-type-1	Work-type-2	Work-type-3
1	0	1	0	0
0	1	0	1	0

FEATURE GENERATION

- **Cross products features (interaction features)**
- Certain features grouped together can generate useful information

Example Showing Whether the Gender Affects the Status of Receiving a Grant	
Gender	Scientific-Grant
M	Y
F	N
M	Y
F	N
F	Y
M	N
F	N
M	N

Source of the image: Müller-Kett, 2023.

## BAG-OF-WORDS (BOW)

- **Tokenizer / naïve scoring**
- **Word counts**
  - Each unique word in a text or document represents one dimension or feature
- **Stop words** must be removed
  - “a,” “an,” “the,” “they,” “where,” “etc.,” punctuation, white spaces, and so on
- Does not preserve **semantics** or **categorical hierarchies**
  - “dog toy” and “toy dog” have different semantics while sharing the same BoW

Source of the image: Microsoft Archive.

Image 8: Text data



“Martin is not bad person. Kevin, Martin’s brother, is bad person.”

The text feature is [`bad`, `brother`, `is`, `Kevin`, `Martin`, `not`, `person`]

BoW =  $\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}$ .

## BAG-OF-N-GRAMS

- Counts **n successive words**
  - one word is 1-gram or unigram, two successive words is 2-grams or bigram
- **Stop words** must be removed
- **Punctuation** is regarded words
- Aims to preserve **semantics** to some extent
  - The higher the number of successive words to consider n, the better the conserved information or semantic
  - A feature text of p individual words, there are  $p^2$  bigrams

Source of the image: Microsoft Archive.

Image 8: Text data



"Martin is not bad person. Kevin, Martin's brother, is bad person."

Bag-of-2-Grams feature text ['Kevin Martin', 'Martin brother', 'Martin is', 'bad person', 'brother is', 'is bad', 'is not', 'not bad']

BoW vector [[0 0 1 1 0 0 1 1] [1 1 0 1 1 1 0 0]].



## TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

- Counts the words **per document**
  - weights the words by their occurrence frequency in the text
- **Term Frequency (TF)**
  - Number of word occurrences in a document
- **Inverse Document Frequency (IDF)**
  - Number of documents and number of overall word occurrences
  - Often logarithmized
- **Often normalized** by all TF-IDF values

Image 8: Text data



## TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

$$TF - IDF = BoW(w, d) \cdot \frac{N}{n_w} = TF \cdot IDF$$

- $BoW(w, d)$  be the number of times that the word  $w$  appeared in a document  $d$  (the frequency term (TF) of word  $w$  in document  $d$ ).
- Inverse Document Frequency (IDF) is the number of documents  $N$  in the dataset divided by the number  $n_w$  that the word  $w$  occurred in the documents contained in the dataset

### Example:

1. The first document  $d_1$  is "Martin is good person"
2. The second document  $d_2$  is "Kevin is bad person"

Feature vector: ['Kevin', 'Martin', 'bad', 'good', 'is', 'person'].

TF (Kevin,  $d_1$ ) = 0, TF (Martin,  $d_1$ ) = 1, TF (bad,  $d_1$ ) = 0, TF (good,  $d_1$ ) = 1, TF (person,  $d_1$ ) = 1  
TF (Kevin,  $d_2$ ) = 1, TF (Martin,  $d_2$ ) = 0, TF (bad,  $d_2$ ) = 1, TF (good,  $d_2$ ) = 0, TF (person,  $d_2$ ) = 1

IDF(Kevin,  $d_1$ ) =  $(N=2/n_w=1)=2$ , IDF(Martin,  $d_1$ ) =  $(2/1)=2$ , IDF(bad,  $d_1$ ) =  $(2/1)=2$ , IDF(good,  $d_1$ ) =  $(2/1)=2$ , IDF(person,  $d_1$ ) =  $(2/2)=1$ , IDF(Kevin,  $d_2$ ) =  $(2/1)=2$ , IDF(Martin,  $d_2$ ) =  $(2/1)=2$ ,  
IDF(bad,  $d_2$ ) =  $(2/1)=2$ , IDF(good,  $d_2$ ) =  $(2/1)=2$ , IDF(person,  $d_2$ ) =  $(2/2)=1$

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

Often, we take the log of IDF instead of the raw IDF

$$TF-IDF = TF \cdot IDF = BoW(w, d) \cdot \log_2\left(\frac{N + 1}{n_{w,d} + 1} + 1\right)$$

TF(Kevin,d<sub>1</sub>) = 0, IDF(Kevin,d<sub>1</sub>) = log<sub>2</sub>((2+1)/(2+1))+1 = 1, therefore, TF-IDF = 1

TF(good,d<sub>1</sub>) = 1, IDF(good,d<sub>1</sub>) = log<sub>2</sub>((2+1)/(1+1))+1 = 1.4, TF-IDF = 1.4

TF-IDF for a Simple Dataset of Two Simple Documents					
Word	TF		IDF	TF-IDF	
	d <sub>1</sub> (0)	d <sub>2</sub> (1)		d <sub>1</sub> (0)	d <sub>2</sub> (1)
Kevin (0)	0	1	Log <sub>2</sub> (3/2)+1	0	1.4
Martin (1)	1	0	Log <sub>2</sub> (3/2)+1	1.4	0
bad (2)	0	1	Log <sub>2</sub> (3/2)+1	0	1.4
good (3)	1	0	Log <sub>2</sub> (3/2)+1	1.4	0
is (4)	1	1	Log <sub>2</sub> (3/3)+1	1	1
person (5)	1	1	Log <sub>2</sub> (3/3)+1	1	1

d<sub>1</sub>: “Martin is good person”  
d<sub>2</sub>: “Kevin is bad person”

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

- The computed TF-IDF is not scaled or normalized
- It is very useful to scale the TF-IDF to obtain unbiased TF-IDF with respect to the document length
- There are several norms used to normalize TF-IDF. One of the most used norms is l2 norm.

$$(TF-IDF(w_i, d))_{l_2} = TF-IDF(w_i, d) / \sqrt{\sum_{w_k \in d} (TF-IDF(w_k, d))^2}$$

Normalized TF-IDF According to l2 Norm for a Simple Dataset of Two Simple Documents						
	Kevin (0)	Martin (1)	bad (2)	good (3)	is (4)	person (5)
d1 (0)	0	0.57	0	0.57	0.40	0.40
d2 (1)	0.57	0	0.57	0	0.40	0.40



- Explain the difference between **numerical**, **categorical**, and **text** features.
- **Clean, scale, encode, or transform** these features.
- Generate **new features** by **transforming, splitting,** or **grouping** existing features as interaction features.

SESSION 4

# TRANSFER TASK

## TRANSFER TASKS

A start-up that sells **sustainable products in smaller stores worldwide** has been very successful in recent years. As a Data Scientist, you and your team are **tasked with tidying up the data** stored in the company's databases. During a quick scan, you noticed the following issues:

- There are **gaps** in the data. Consulting the ones who collected the data, you learn that some gaps are the result of a longer **power shortage** and others were not filled out, as this information is **evident from other columns**
- The columns contain values with highly **different ranges**, some with values between 0.1 and 1 and others between 100 and 1000
- There seem to be some **extreme values** that are **not outliers** to be removed but correct values
- Some columns are **hard to interpret**. For example, there is a column containing the persons per squared sales for some reason
- The **stores are encoded as integers** (1 for the first store, 2 for the next store, ...)
- The **product names** and **IDs** are stored in **one column**
- You are also assigned the task of **sorting the products into categories**. You noticed a **textual description** for each product that might be for this.

**Describe how you would approach these tasks.**

TRANSFER TASK  
PRESENTATION OF THE RESULTS

Please present your  
results.

The results will be  
discussed in plenary.







1. Which of the following scale levels is a person's age?
  - a) an interval scale
  - b) a ratio scale
  - c) a nominal scale
  - d) an ordinal scale



2. Which of the following scale levels is customer satisfaction (low, medium, high)?
- a) an ordinal scale
  - b) a ratio scale
  - c) an interval scale
  - d) a ratio scale



3. Which of the following scale levels are car brands?
- a) an ordinal scale
  - b) a ratio scale
  - c) a nominal scale
  - d) an interval scale

## LIST OF SOURCES

### Text

Badr, W. (2019). 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples). Popular strategies to statistically impute missing values in a dataset, Towards Data Science, <https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>

### Images

Müller-Kett, 2021.

Müller-Kett, 2023.

Microsoft Archive.

© 2022 IU Internationale Hochschule GmbH

This content is protected by copyright. All rights reserved.

This content may not be reproduced and/or electronically edited, duplicated, or distributed in any kind of form without written permission by the IU Internationale Hochschule GmbH.