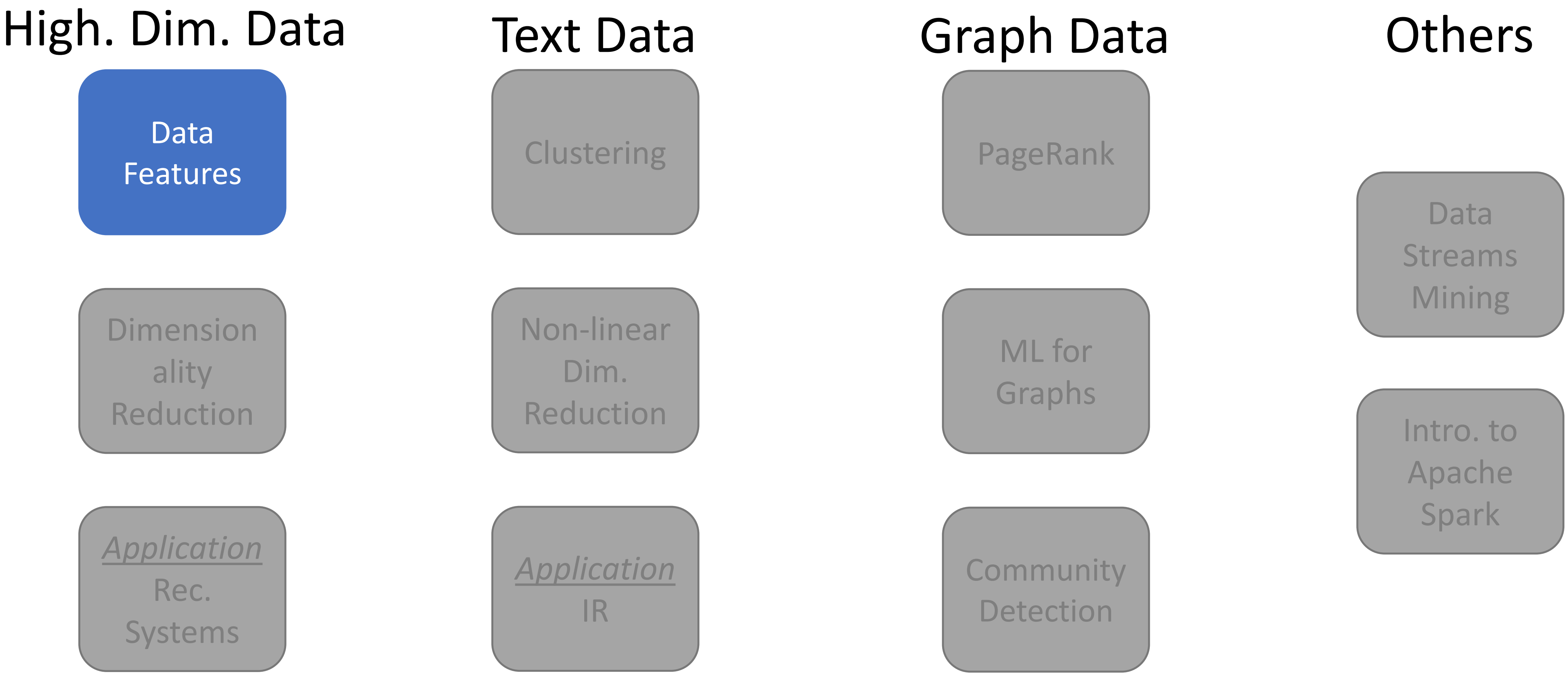
CS 5683: Big Data Analytics

Data Featurization

Arunkumar Bagavathi
Department of Computer Science
Oklahoma State University

Topics Overview



High. Dim. Data

- Data Features
- Dimensionality Reduction
- *Application* Rec. Systems

Text Data

- Clustering
- Non-linear Dim. Reduction
- *Application* IR

Graph Data

- PageRank
- ML for Graphs
- Community Detection

Others

- Data Streams Mining
- Intro. to Apache Spark

Data of (ML/AI) Models

- All data-driven models rely on the given dataset **X** to perform optimization for the given underlying task (e.g. classification task **t: X → y**), where X is a set of numerical values

- Ideally, the given raw can be ***structured*** or ***unstructured***

- Examples of unstructured data
  - Text
  - Images & Videos
  - Graphs
  - Timeseries
  - Audio
  - Spatial (lat. and long.)

Give examples of ML tasks for each unstructured data

Types of variables in X: https://statsandr.com/blog/variable-types-and-examples/

Data Featurization

- *Engineering* or *Learning* features from the raw data $\tilde{\mathbf{X}}$

- **Feature Engineering:**
  - Manually collecting the features required for the tasks
  - Mostly independent features
  - ***Example:*** animal features for animal classification problem

1.0 2.0 1.0 0.0 1.0…….

- **Feature (Representation) Learning:**
  - Mapping the feature extraction as a learning task $t': \tilde{X} \rightarrow X$
  - Extracts both dependent and independent features
  - Captures syntactic and contextual relationships
  - ***Example:*** learning animal features directly from images

0.382 -1.234 1.456 0.004…….

This Lecture

What features to initialize for different data modalities for any given learning task?

Note-1: *These methods can be assumed as engineering*

Note-2: *Assume the data is cleaned & pre-processed*

Generic & Simple Approach

- Initialize with random vectors where an entry in the vector is [0.0,1.0] or [-1.0,1.0]

- This applies to all data modalities and all learning tasks

- Let the representation learning model optimize the initialized random vectors for the given task

- **Potential risks:** Poor/slow model convergence, skewed towards the randomness, and loss of generalization

(i) Text Data - 1

- Task: Text Classification/Anomaly detection/Document similarity

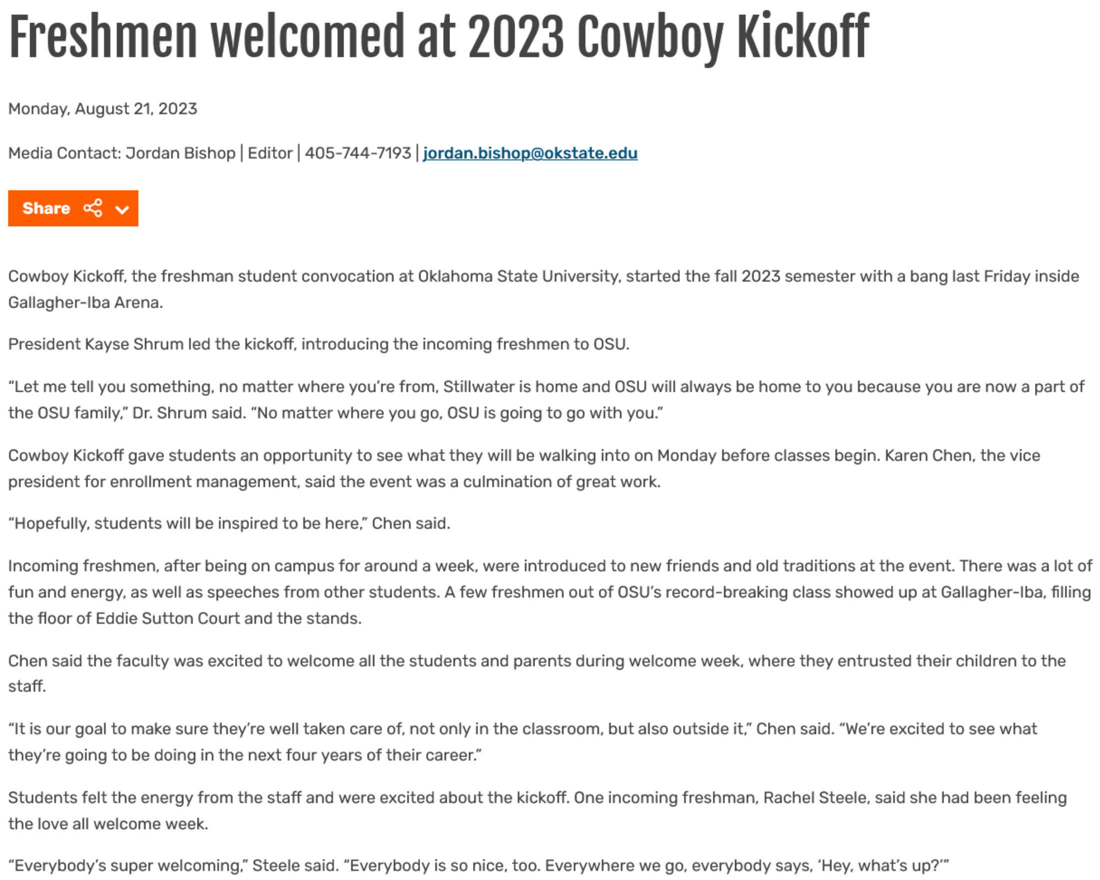
- Given: Cleaned text data (**‘d’** sentences/documents)

(i) Count vectors or Bag-of-Words (BoW)

(ii) Term Frequency Inverse Document Frequency (TF-IDF)

(iii) n-gram features

(iv) Syntactic features



Freshmen welcomed at 2023 Cowboy Kickoff

Monday, August 21, 2023

Media Contact: Jordan Bishop | Editor | 405-744-7193 | jordan.bishop@okstate.edu

Cowboy Kickoff, the freshman student convocation at Oklahoma State University, started the fall 2023 semester with a bang last Friday inside Gallagher-Iba Arena.

President Kayse Shrum led the kickoff, introducing the incoming freshmen to OSU.

“Let me tell you something, no matter where you’re from, Stillwater is home and OSU will always be home to you because you are now a part of the OSU family,” Dr. Shrum said. “No matter where you go, OSU is going to go with you.”

Cowboy Kickoff gave students an opportunity to see what they will be walking into on Monday before classes begin. Karen Chen, the vice president for enrollment management, said the event was a culmination of great work.

“Hopefully, students will be inspired to be here,” Chen said.

Incoming freshmen, after being on campus for around a week, were introduced to new friends and old traditions at the event. There was a lot of fun and energy, as well as speeches from other students. A few freshmen out of OSU’s record-breaking class showed up at Gallagher-Iba, filling the floor of Eddie Sutton Court and the stands.

Chen said the faculty was excited to welcome all the students and parents during welcome week, where they entrusted their children to the staff.

“It is our goal to make sure they’re well taken care of, not only in the classroom, but also outside it,” Chen said. “We’re excited to see what they’re going to be doing in the next four years of their career.”

Students felt the energy from the staff and were excited about the kickoff. One incoming freshman, Rachel Steele, said she had been feeling the love all welcome week.

“Everybody’s super welcoming,” Steele said. “Everybody is so nice, too. Everywhere we go, everybody says, ‘Hey, what’s up?’”

https://news.okstate.edu/articles/communications/2023/freshmen_welcomed_at_2023_cowboy_kickoff.html

(i) Text Data - 2

Doc	*word1*	*word2*	*word3*	*word4*	*word5*	*word6*	.....
1	*5.0*	*0.0*	*1.0*	*0.0*	*1.0*	*0.0*	.....
2	*0.0*	*1.0*	*0.0*	*0.0*	*0.0*	*0.0*	.....
3	*2.0*	*0.0*	*6.0*	*9.0*	*0.0*	*1.0*	.....
…							
…							

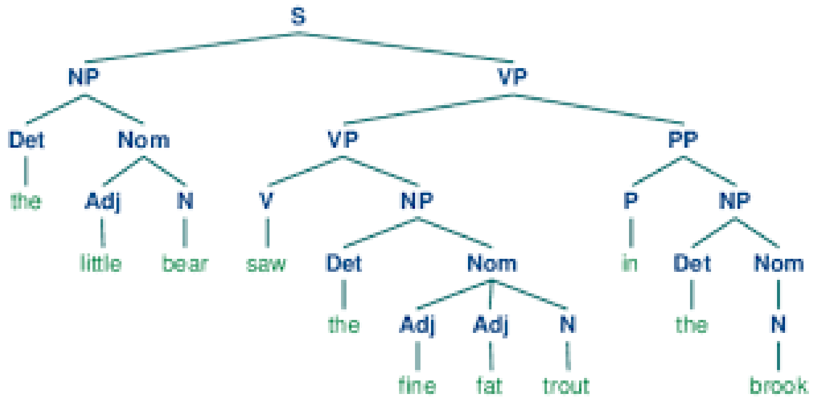
- Both BoW and TF-IDF generate features of size $\boldsymbol{d * n}$, where ‘n’ is the global number of words in our repository
- **BoW**: Count of words present in the document
- **TF-IDF**: TF * IDF, where
  - $TF_w = \frac{Count(w)}{No.\,of\ words\ in\ doc.}$
  - $IDF_w = log \frac{N}{No.of\ docs\ with\ w}$ captures the relevance/importance of words

- What happens when there is a new word in testing phase?
- What will be the size of ‘n’ in Wikipedia document classification?

(i) Text Data - 3

- **n-gram features**: Similar to count features, but count phrases containing ‘n’ words rather than independent words

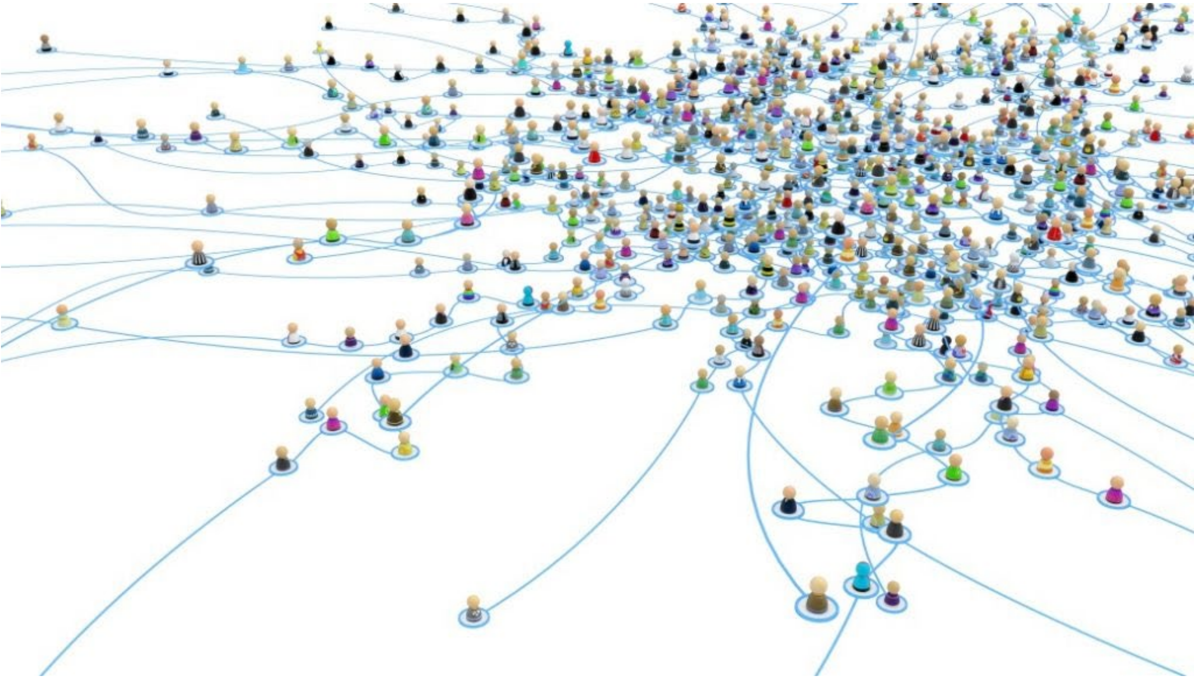
Doc	*w1+w2+w3*	*w2+w3+w4*	*w3+w4+w5*	*w4+w5+w6*	*w5+w6+w7*	*.....*
1	*5.0*	*0.0*	*1.0*	*0.0*	*1.0*	*.....*
2	*0.0*	*1.0*	*0.0*	*0.0*	*0.0*	*.....*
3	*2.0*	*0.0*	*6.0*	*9.0*	*0.0*	*.....*
…						
…						



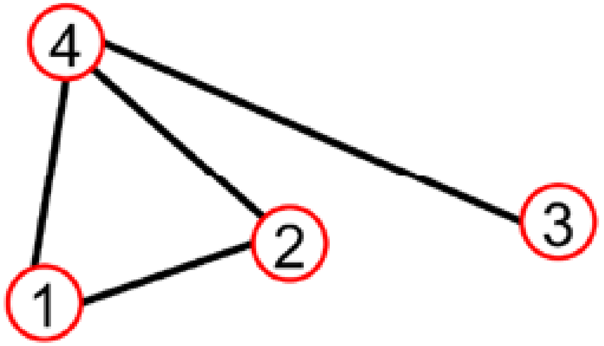
- **Syntactic features**: Giving importance to syntactic or grammatic details of the language of the text
  - Nouns, verbs, adjectives, pronouns, adverbs, prepositions, etc.
  - Usually done with Parts-of-Speech (PoS) tagging tools like StanfordNLP

(ii) Graph Data - 1

- Graph data is denoted as $G = (V, E)$, where V is a set of ‘k’ nodes, E is a set of ‘m’ edges, and $\mathrm{m} \gg \mathrm{k}$. In general, $e_i = (v_a, v_b)$, where $\mathrm{e_i} \in \mathrm{E}$ is a node pair representing a link between nodes $\mathrm{v_a}$ and $\mathrm{v_b}$
- *Task:* Node classification/Edge prediction
  - *Applications: bot detection and friend recommendations*



(ii) Graph Data - 2



Adjacency matrix (.txt)

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

- Methods to represent a graph in program readable files
  - Adjacency matrix
  - Edge list
  - Adjacency list

Adjacency list (.txt or .csv)

1,2,4
2,1,4
3,4
4,1,2,3

Edge list (.txt or .csv)

1,2
1,4
2,4
3,4

- Adjacency matrix can be considered as initial node features. This captures structural connectivity. Node features are usually massive of size $k * k$

- Graphs may also have node and edge properties
  - Features extracted for graph nodes must incorporate these properties
  - *Easy option:* simply append this with initial node features

(ii) Graph Data - 3

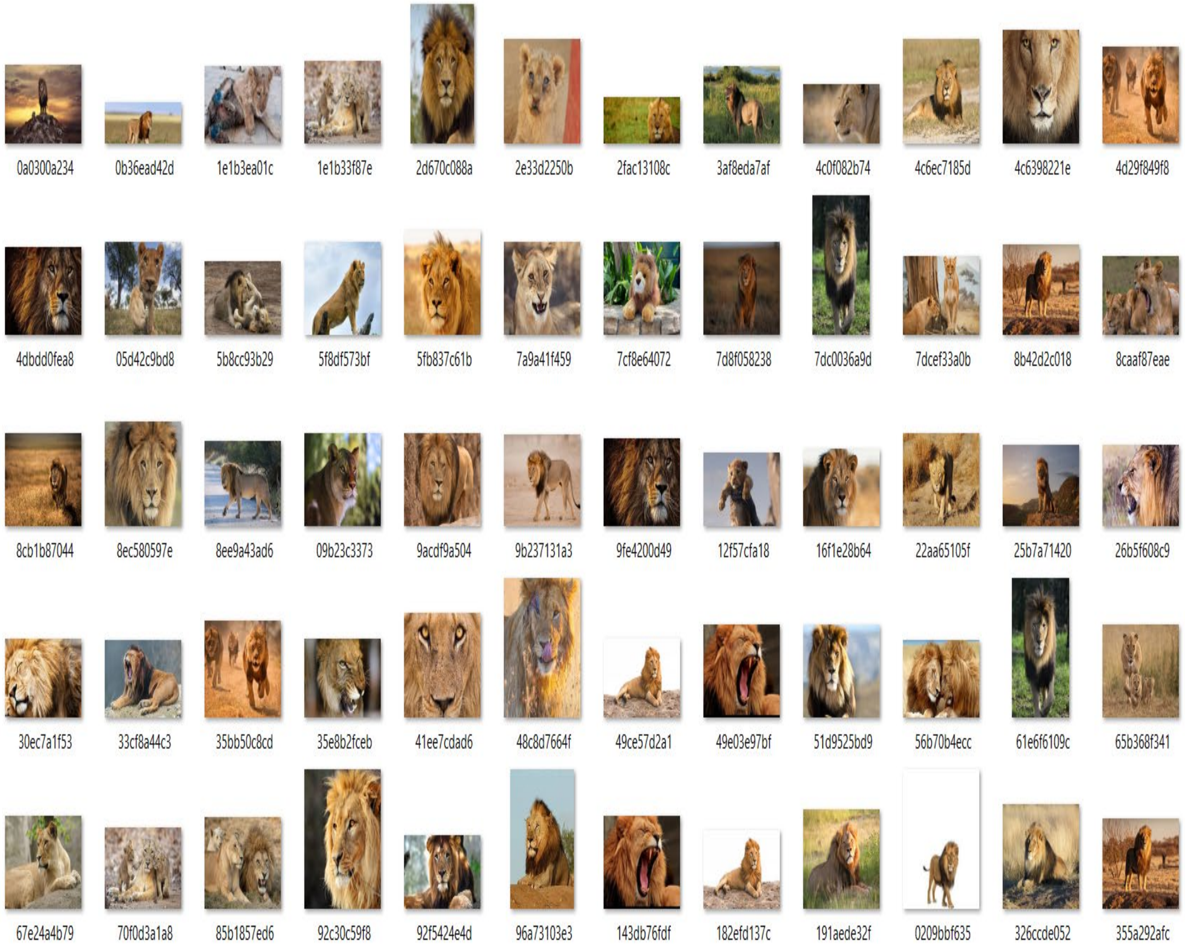
- Other measures to add on the node features in Adjacency matrix:
  - Structural features to measure node importance – degree, centrality, and PageRank
  - Neighborhood information with nth order neighbors and random walks
  - Spectral methods like Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and non-Negative Matrix Factorization (NNMF)

(iii) Images Data - 1

- Task: Image classification/Object detection/image captioning
- Given: raw images (**‘d’** images)

(i) Flattening

(iii) Data augmentation



https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals

(iii) Images Data - 2

- Pixel value [0,255]
- Pre-processing before extracting features
  - Resize images to same size
  - Normalize pixel values
- **Flattening:** Transform 2D pixel values to 1D format
- **Data augmentation:** Augment the input data by applying transformations like rotating, flipping, cropping, and scaling

Methods to Learn Features

- *Linear:* Using linear combinations of addition, multiplication, and division
  - **Example:** Principal Component Analysis (PCA) and TF-IDF

- *Non-linear:* Using non-linear functions like sigmoid $f(x) = \frac{1}{1+e^{-x}}$, logarithmic $f(x) = a.log_b(x)$, rectified linear unit (RELU) $f(x) = \max(0, x)$, or other non-linear functions
  - **Example:** Neural networks

What's next?

How does ML/AI tasks handle these high-dimensional datasets?

Dimensionality reduction while capturing semantic relations

Questions???