MDS651
Unit 5 – Text and Document Visualization

Dipesh Koirala

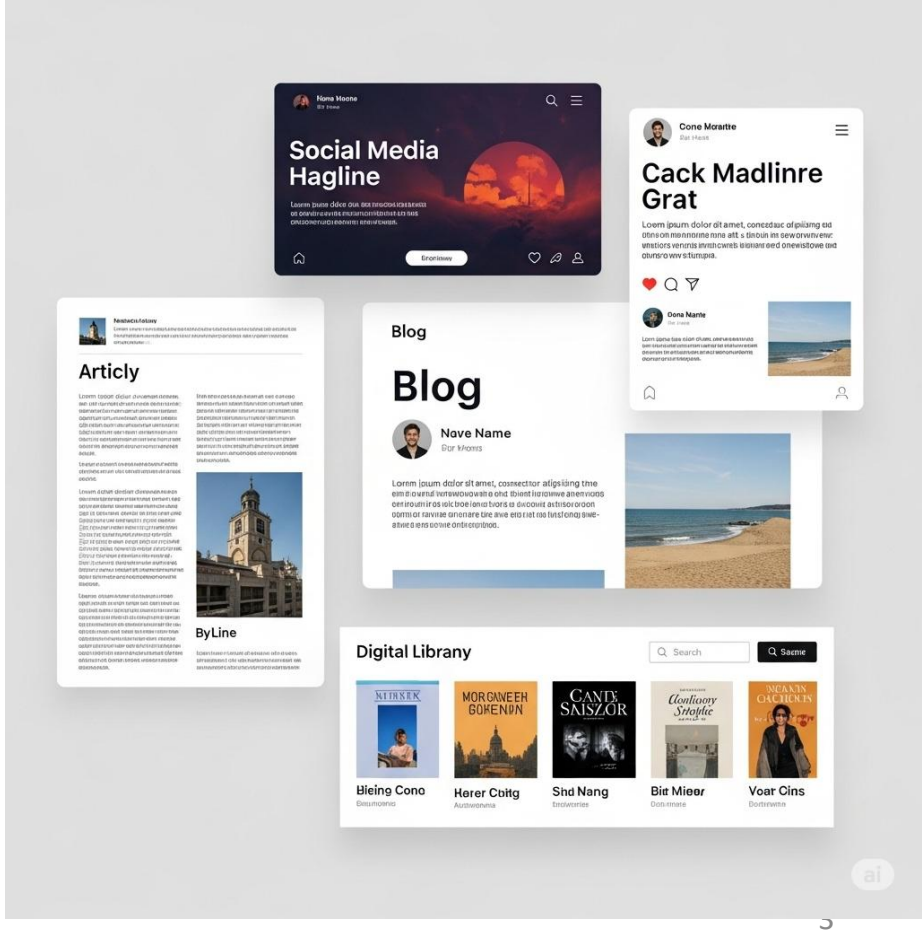
Outline

- Text and Document Data
- Levels of Text Representation
- The Vector Space Model
- Visualizations of a Single Text Document
- Word Cloud
- Word Tree
- Text Arc
- Themes and Self Organizing Maps

Text and Document Data

- Text and document data are a form of **unstructured or minimally structured data** that consist of language.

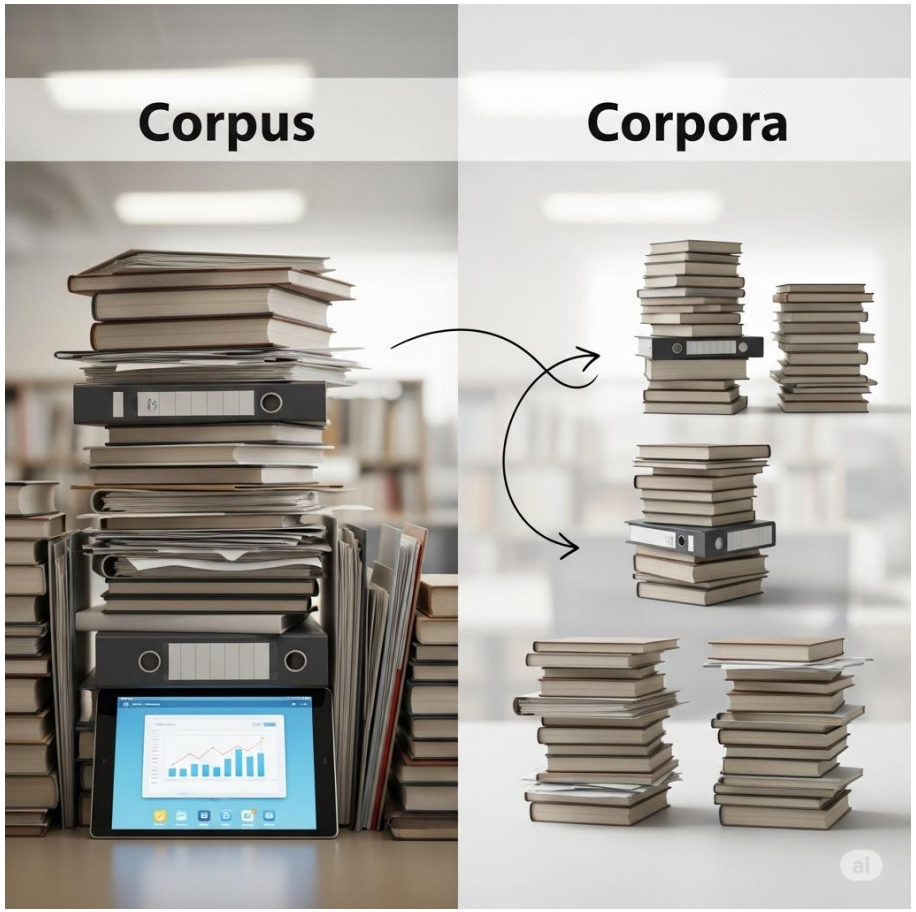
- It includes all forms of written communication, such as emails, social media posts, articles, books, and customer reviews, a blog, a wiki, a twitter feed, billions of words, a collection of papers or a digital library.



Text and Document Data

- The collection of documents is defined as a corpus (plural corpora). Document visualization deals with objects within corpora.

- **These objects** can be words, sentences, paragraphs, documents, or even collections of documents.



Text and Document Data

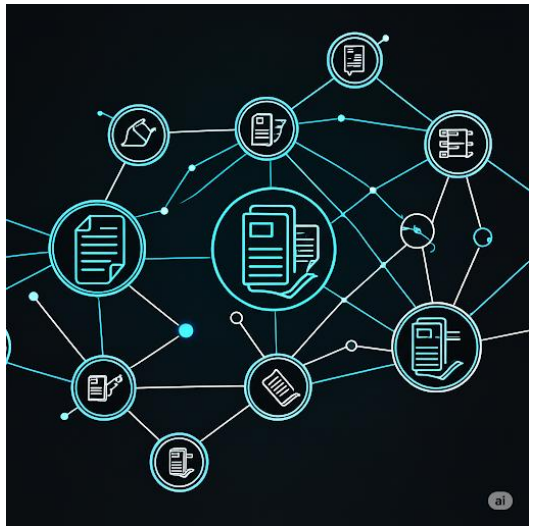
- Visualizing text and document data is the process of using visual representations to make sense of large volumes of text.
- This helps in identifying patterns, topics, and relationships that would be difficult to discover by simply reading the text.

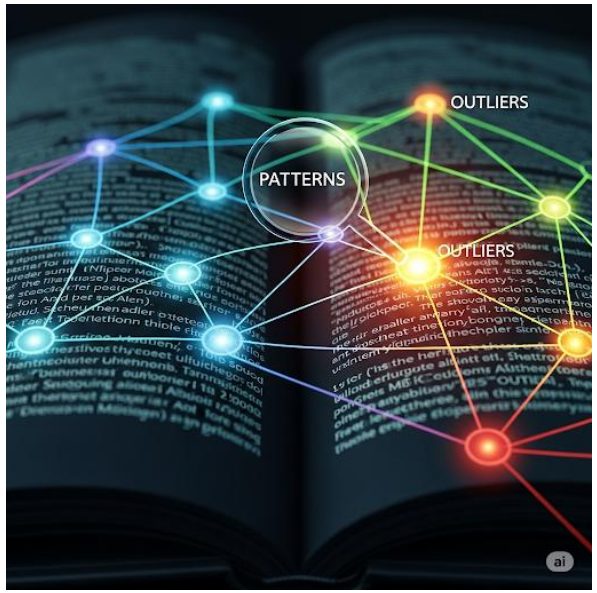
- For text and documents, the most obvious tasks are searching for a word, phrase, or topic.

Text and Document Data

- For partially structured data, the task may be search for relationships between words, phrases, topics, or documents.

- For structured text or document collections, the key task is most often searching for patterns and outliers within the text or documents.





Levels of Text Representation

- Three levels of text representation are defined, each requires us to convert the unstructured text to some form of structured data:
- Lexical,
- Syntactic and
- Semantic

Lexical Level

- The lexical level is concerned with transforming a string of characters into a sequence of atomic entities, called tokens.

Levels of Text Representation

Lexical Level

- The lexical level is concerned with transforming a string of characters into a sequence of atomic entities, called tokens.
- Lexical analyzers process the sequence of characters with a given set of rules into a new sequence of tokens that can be used for further analysis.

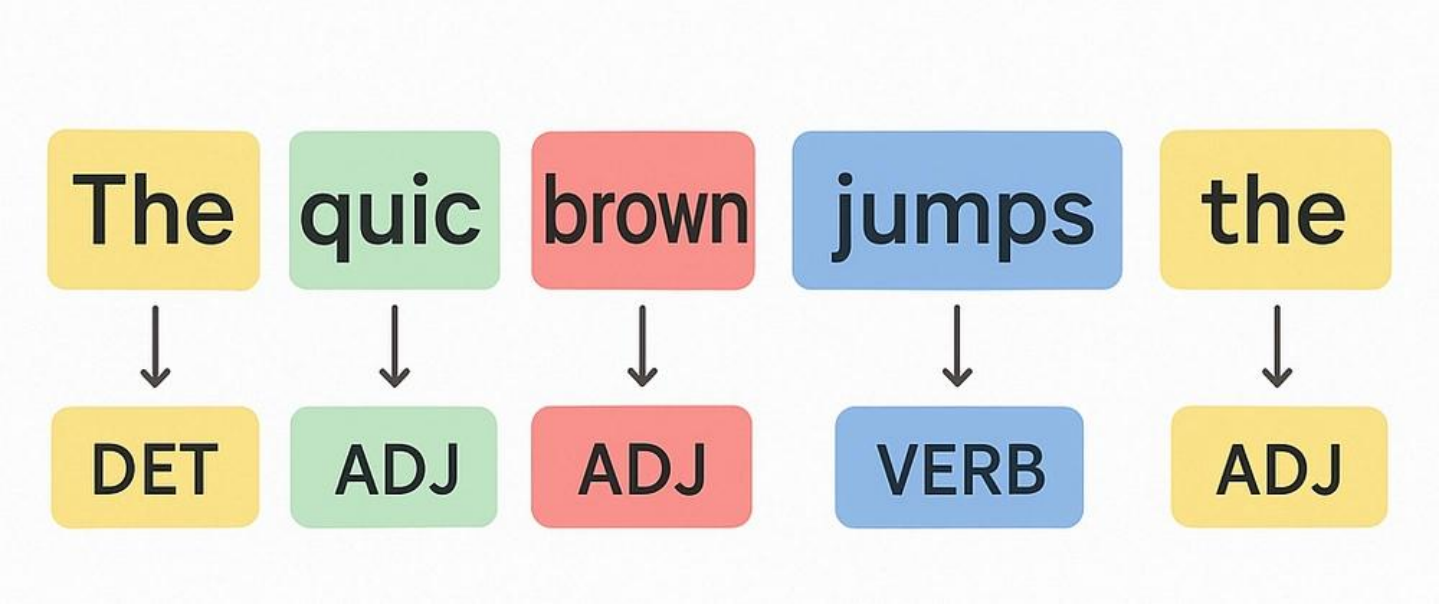
- Tokens can include characters, character n-grams, words, word stems, lexemes, phrases, or word n-grams, all with associated attributes.

["The", "quick", "brown", "fox", "jumped", "over", "the", "lazy", "dog", "."]

Levels of Text Representation

Syntactic Level

- The syntactical level deals with identifying and tagging (an notating) each token's function.
- Various tags are assigned, such as sentence position or whether a word is a noun, expletive, adjective, dangling modifier, or conjunction.
- Tokens can also have attributes such as whether they are singular or plural, or their proximity to other token



Levels of Text Representation

Semantic Level

- The semantic level encompasses the extraction of meaning and relationships between pieces of knowledge derived from the structures identified in the syntactical level.
- The goal of this level is to define an analytic interpretation of the full text within a specific context, or even independent of context.
- Includes meaning extraction, relationship analysis , conceptual relationships.

The Vector Space Model

- The vector space model is a way to represent text data in a mathematical form.

- In the vector space model , **a term vector** for an object of interest (paragraph, document, or document collection) is a vector in which each dimension represents the weight of a given word in that document.

- Typically, to clean up noise, stop words (such as “the” or “a”) are removed (filtering), and words that share a word stem are aggregated together (stemming).

The Vector Space Model

- A vector space is a collection of vectors, characterized by their dimension.

Term-Document Matrix

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- Column-wise: Document as vector
- Row-wise: Word as vector

The Vector Space Model

- A term vector is a representation of a document (or any text object like a paragraph) as a vector. Here's how it works:
  - **Dimensions:** Each dimension of the vector corresponds to a unique word from the entire collection of documents (corpus).
  - **Weights:** The value in each dimension represents the weight of the corresponding word in the document. This weight indicates the importance or relevance of the word in that document.

The Vector Space Model

Tf-Idf

- Provides a weighting scheme for assigning weights to terms in a document.
  - Tf (w) = term frequency or number of times that word w occurred in the document,
  - Df(w) = document frequency (number of documents that contain the word).
  - N = number of documents.

- Tf-Idf(w) is defined as:

$$TfIdf(w) = Tf(w) * \log\left(\frac{N}{Df(w)}\right).$$

$$tf(w,d) = \frac{occurence\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

$$idf(w,D) = ln(\frac{total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w})$$

The Vector Space Model

Tf-Idf

- **Text A**: Jupiter is the largest planet
- **Text B**: Mars is the fourth planet from the sun

Words	TF (A)	TF (B)	IDF	TFIDF (A)	TFIDF (B)
jupiter	1/5	0	ln (2/1)=0.69	0.138	0
is	1/5	1/8	ln (2/2)=0	0	0
the	1/5	2/8	ln (2/2)=0	0	0
largest	1/5	0	ln (2/1)=0.69	0.138	0
planet	1/5	1/8	ln (2/2)=0	0.138	0
mars	0	1/8	ln (2/1)=0.69	0	0.086
fourth	0	1/8	ln (2/1)=0.69	0	0.086
from	0	1/8	ln (2/1)=0.69	0	0.086
sun	0	1/8	ln (2/1)=0.69	0	0.086

The Vector Space Model

Tf-Idf

- Provides a relative importance of the word in the document.
- A word is more important if it appears several times in a single target document (larger Tf).
- As well as, the fewer documents it appears in (lower Df).

- Said another way, we are more interested in words that appear often in a document, but not often in the collection.

- Such words are intuitively more important, as they are differentiating, separating or classifying words.

The Vector Space Model

Tf-Idf

id	men	entered	bank	charlotte	missiles	masks	aryan	guns	witnesses	reported	silver	suv	august
seg1.txt	0.239441	0	0.153457	0.195243	0	0.237029	0	0.195243	0.237029	0.140004	0.195243	0.237029	0
seg13.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg14.txt	0	0.192197	0	0	0	0	0	0	0	0	0	0	0.172681
seg15.txt	0	0	0	0	0	0	0	0	0	0	0	0	0.149652
seg16.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg17.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg18.txt	0	0.158432	0	0	0	0	0	0	0	0	0	0	0
seg19.txt	0	0	0	0.197255	0	0	0	0	0	0.141447	0	0	0.155038
seg2.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg20.txt	0	0.234323	0	0	0	0	0	0	0	0	0	0	0
seg21.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg22.txt	0	0	0	0	0.139629	0	0.127389	0	0	0	0	0	0
seg23.txt	0	0	0	0	0	0	0	0	0	0.180656	0	0	0
seg24.txt	0	0	0	0	0	0	0.117966	0	0	0.117966	0	0	0
seg25.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg26.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg27.txt	0	0	0.235418	0	0	0	0.214781	0	0	0	0	0	0
seg28.txt	0	0	0	0	0.151753	0	0	0	0	0	0	0	0
seg29.txt	0	0	0	0	0	0	0.129852	0	0	0	0	0	0.142329
seg3.txt	0	0	0	0	0.18432	0	0	0	0	0	0	0	0
seg30.txt	0.078262	0	0	0	0	0	0	0	0	0	0	0	0
seg31.txt	0	0	0.213409	0	0	0	0.194701	0	0	0	0	0	0
seg32.txt	0	0	0	0	0	0	0	0	0	0	0	0	0

Visualizations of a Single Text Document

Word Clouds

- also known as a tag cloud or text cloud, is a visual representation of text data where the size of each word indicates its frequency or importance within the text.
- Larger, bolder words appear more often, while smaller words appear less frequently.
- **wordle** are examples of visualizations that use only term frequency vectors and some layout algorithm to create the visualization.

Visualizations of a Single Text Document

Tag Cloud

- A tag cloud is a visual representation of text data which is often used to depict keyword metadata on websites, or to visualize free form text.
- Tags are usually single words, and the importance of each tag is shown with font size or color.
- Tags are sorted alphabetically or according to their relevance, frequency or similarity.

Visualizations of a Single Text Document

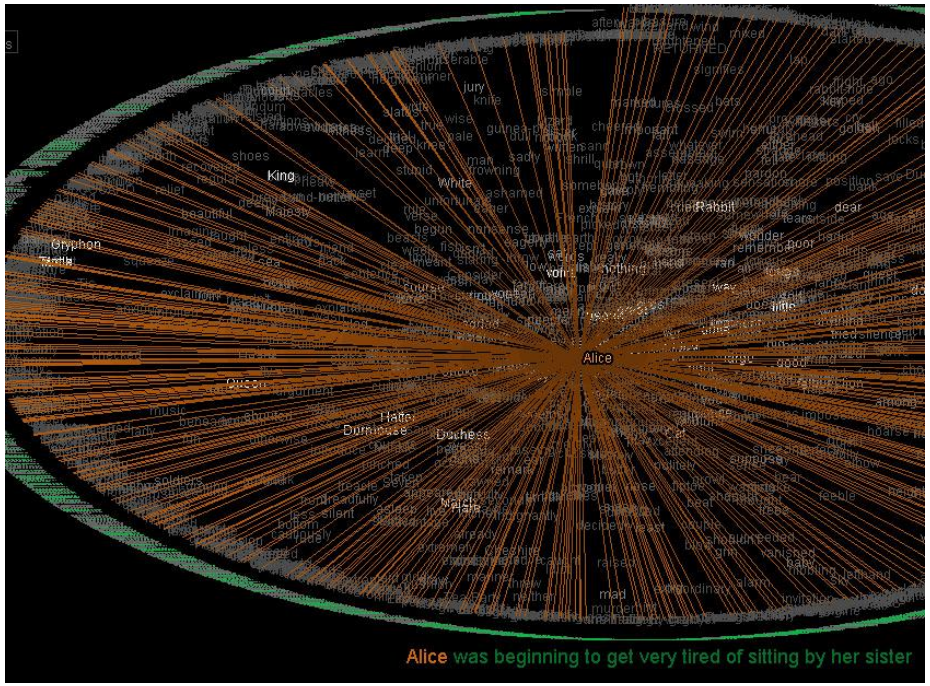
Word Tree

- The Word Tree visualization is a visual representation of both term frequencies, as well as their context.
- Size is used to represent the term or phrase frequency.
- The root of the tree is a user-specified word or phrase of interest, and the branches represent the various contexts in which the word or phrase is used in the document.

Visualizations of a Single Text Document

TextArc

- TextArc is a visual representation of how terms relate to the lines of text in which they appear
- The entire text is mapped onto a large **ellipse** or arc.
- Each individual word is represented as a small line or tick mark, placed sequentially around the curve.
- Within the ellipse, the most frequent or significant words are pulled toward the center.
- These words are connected by **arcs** or straight lines to their corresponding tick marks on the outer ellipse.



Visualizations of a Single Text Document

TextArc

- The power of TextArc lies in its ability to show **connectivity** and **context.**
- The size and brightness of a word reflect its frequency, similar to a word cloud.

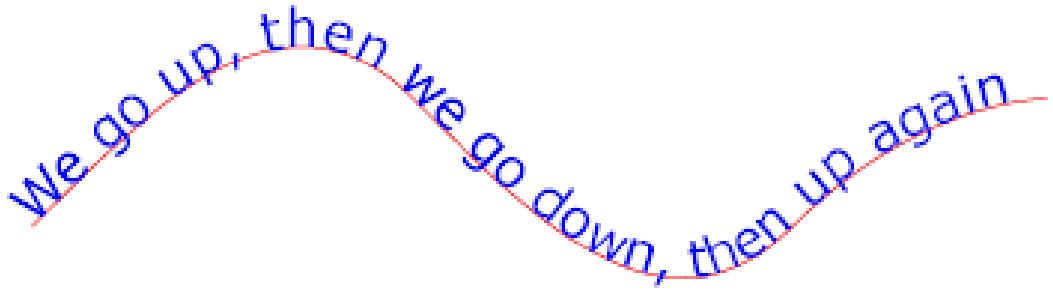
- By looking at a word inside the ellipse, it can be instantly seen all the places it appears in the text, revealing its patterns and relationships to surrounding words.
- It provides a visual answer to questions like "Where does this character appear most often?“

- The original **TextArc** project (by W. Bradford Paley) takes an entire book, arranges all its words in a spiral or oval arc, and draws lines between repeated words.
- For example: The entire text of *Alice's Adventures in Wonderland* is placed in an arc.

Visualizations of a Single Text Document

TextArc

- Represents textual information along a curved or circular arc.
- Text arc technique involves placing individual characters or words along the arc, following the curvature of the path.
- Text arcs are often employed in various types of visualizations, such as pie charts, radial charts, or circular timelines.



Document collection Visualizations

- In most cases of document collection visualizations, the goal is to place similar documents close to each other and dissimilar ones far apart.
- The similarity between all pairs of documents are computed and determine a layout.
- Documents with similar content are placed close to each other, forming distinct clusters.

Document collection Visualizations

Themes

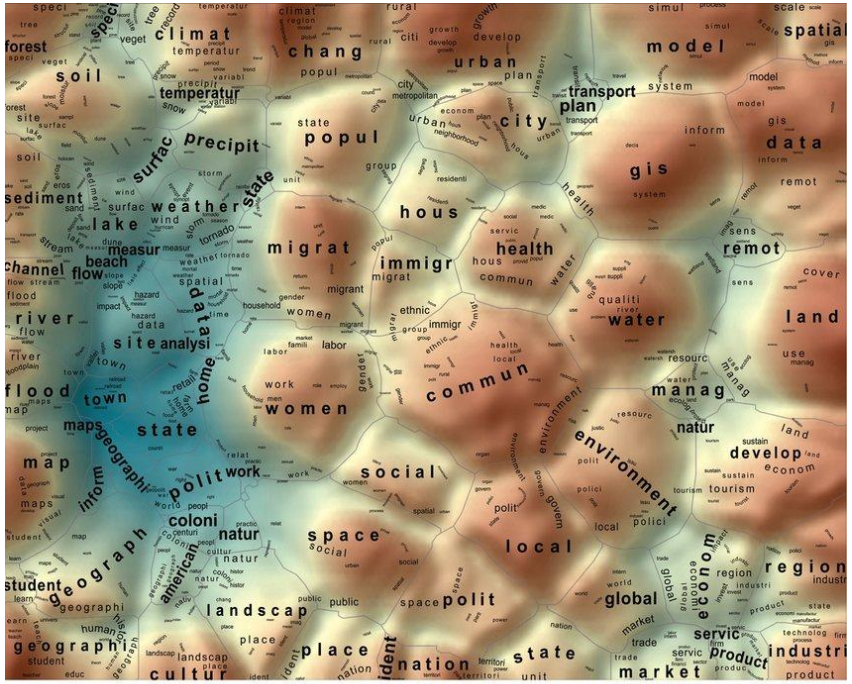
- **themes** refer to recurring topics, concepts, or subject matter that appear throughout a document. They are broader ideas that are typically represented by a group of related words.

- **For e.g.,** in a novel, "love" or "betrayal" could be a theme.
- In a scientific paper, "quantum mechanics" or “AI" could be a theme

Document collection Visualizations

Themescapes

- are a method of visualizing a collection of documents as a 3D landscape.
- are summaries of corpora using abstract 3D landscapes in which height and color are used to represent density of similar documents.
- The taller mountains represent frequent themes in the document corpus (height is proportional to number of documents relating to the theme).
- Convey relevant information about the topic or themes.



Börner, Katy. *Atlas of Science: Visualizing What We Know*. (2010). The MIT Press. Pg 38.

Document collection Visualizations

Themescapes

- Using a vector space model, each document is converted into a vector. Then the similarity (or dissimilarity) between every pair of documents is calculated. This is often done using a measure like cosine similarity.

- The similarity data is used to position the documents in a 2D or 3D space, with similar documents placed closer together.

- The density of documents in a given area determines the elevation of the landscape at that point, creating the peaks and valleys.

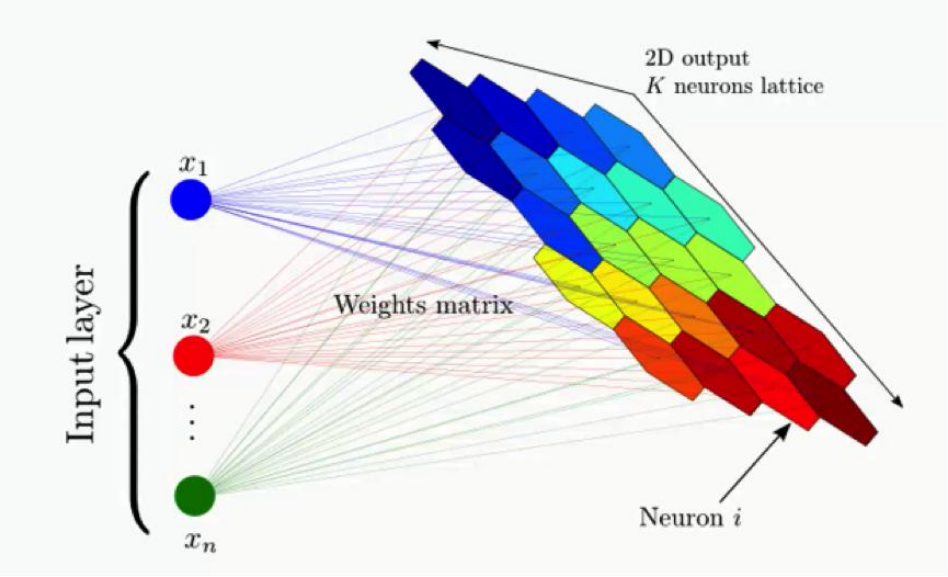
- The peaks are automatically labeled with key terms or phrases that are most representative of the documents within that cluster, providing a summary of the theme.

Document collection Visualizations

Self-Organizing Maps

- A self-organizing map (SOM) is an unsupervised learning algorithm using a collection of typically 2D nodes, where documents will be located.

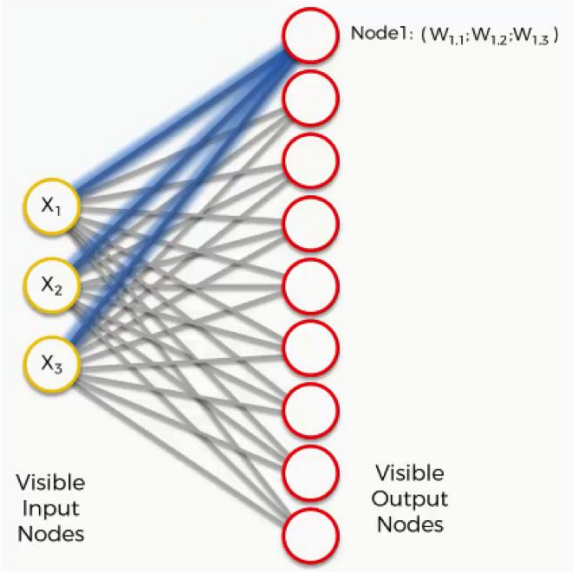
- It is particularly useful for visualizing high-dimensional data by projecting it onto a lower dimensional grid, typically a 2D map.
- This process reveals clusters and patterns within the data that would be difficult to discern otherwise.



Document collection Visualizations

Self-Organizing Maps

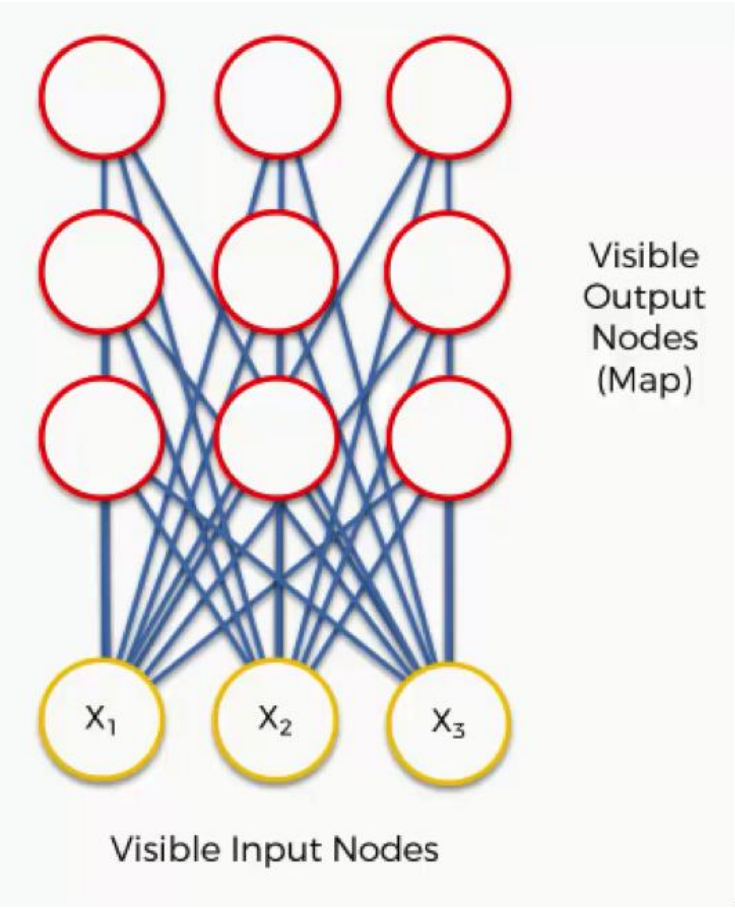
- A grid of neurons is created, each with a randomly initialized weight vector. The number of dimensions in each weight vector is the same as the number of features in the input data.
- A data point (input vector) is selected from the dataset.
- The algorithm then calculates the distance between this input vector and the weight vector of every neuron on the grid.



Document collection Visualizations

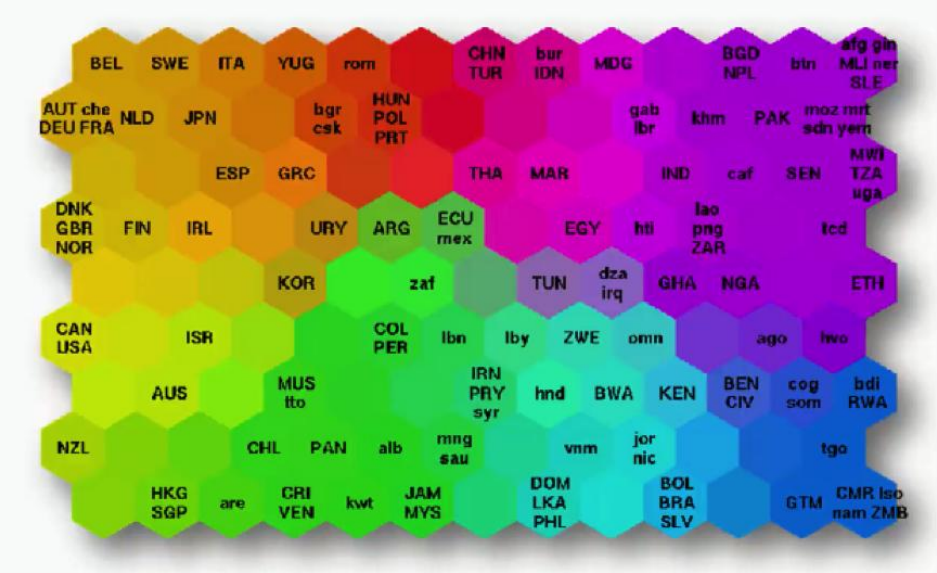
Self-Organizing Maps

- The neuron with the weight vector closest to the input vector is declared the Best Matching Unit (BMU).
- Once the BMU is found, its weights and the weights of its neighboring neurons are updated.
- The weights are adjusted of the closest nodes (within a particular radius), making each closer to the input vector, with the higher weights corresponding to the closest selected node.



Document collection Visualizations

Self-Organizing Maps



- Self Organizing Maps (SOM's) - How do Self-Organizing Maps Learn? (Part 1) - Blogs - SuperDataScience | Machine Learning | AI | Data Science Career | Analytics | Success

End of Unit 5

Thank you