Master's in Data Sciences (MDS)

# DATA VISUALIZATION

## Chapter(5)

### Text and Document Visualization

Lecturer: Er. Saroj Ghimire
Qualification: Msc.CSIT,
BE(COMPUTER)

# OVERVIEW

- Text and document data
- Levels of text representation,
- The vector space model,
- Visualizations of a single text document
- Word cloud, Word tree, Text arc
- Them escapes and self organizing maps

# TEXT AND DOCUMENT VISUALIZATION

✓Text and document visualization is a process of transforming textual information and documents into visual representations that aid in understanding, analysis, and communication.

1. Visualizations reveal patterns and relationships within the text that may not be apparent in raw form.
2. Visual representations provide an intuitive way to comprehend textual content, presenting it in a structured and organized manner.
3. Interactive exploration of visualizations allows users to delve into specific details and gain deeper insights.
4. Visualizations enhance communication by presenting textual information in an engaging and memorable manner.

# EXAMPLES OF TEXT AND DOCUMENT VISUALIZATION TECHNIQUES

1. Word Cloud:

   Data: A collection of customer reviews for a product.

   Visualization: A word cloud representing the most frequently mentioned words in the reviews, with the size of each word indicating its frequency.

2. Bar Chart:

   Data: Sales figures for different product categories in a retail store.

   Visualization: A bar chart showing the sales amounts for each category, with the length of each bar representing the sales volume.

3. Tree Map:

   Data: File sizes in a computer directory.

   Visualization: A tree map representing the directory structure, with each file represented as a rectangle, and the size of the rectangle indicating the file's size

# TEXT

- Typography:
  - typefaces (serif, sans-serif, **bold**, *italic*)
  - point size (10pt, 12pt, 24pt, 36pt.. ) - nowadays: 1/72 inch
  - line length (alignment: left, right, justified)
  - vertical: line spacing (leading)
  - horizontal: spaces between groups of letters (tracking)
  - space between pairs of letters (kerning)[For example, letters like "AV" or "To" may appear visually unbalanced if the default spacing is applied.]
  - combining letters to a glyph ligatures {glyph ligatures is a typographic technique where two or more letters are visually merged into a single interconnected character} Ligatures can be created when specific letter combinations, such as "fi," "fl," or "ff," occur in a word

# TEXT/DOCUMENT AS VIS

Typography:

- typefaces (serif, sans-serif, **bold**, *italic*)
- point size (10pt, 12pt, 24pt, 36pt.. ) - nowadays: 1/72 inch
- line length (alignment: left, right, justified)
- vertical: line spacing (leading)
- horizontal: spaces between groups of letters (tracking)
- space between pairs of letters (kerning)
    - combining letters to a glyph ligatures


ß

fi → fi

fl → fl

# LEVELS OF TEXT REPRESENTATION

- Levels of text to be represented:
  - Lexical level -- Simple grouping of characters into "tokens" which are typically words, but word stems, phrases, word n-grams and character n-grams may be beneficial
  - Syntactic level --Parsing purpose of token, grammatical category, tense, plurality, in the context of the phrase, sentence and paragraph
  - Semantic level -- Extract meaning of the syntactic structure with the tokens using fuller analysis of the context

# VECTOR SPACE MODEL

- Analysis of the words in a document and determine their value in contribution and significance to the document.
- Removal of noise words ("a", "an", "the", "that") and punctuation, and stemming (collecting roots of words) are typical of preprocessing.
- Simple frequency counts of significant words ordered by decreasing frequency is a simple vector.
- Here is a web site to generate this vector:
  http://www.wordcounter.com/ It has options to remove noise and do stemming.
- Plugging the above Introduction text into this web site, we get the following vector:

# TEXT VISUALIZATION

- The text visualization chart is the graphical representation of qualitative data frequency, such as keywords or customer feedback.
- The graph gives greater prominence to words that appear more frequently in a source text. The larger the word, the higher its frequency.
- You can use the chart to perform exploratory textual analysis by identifying words that frequently appear in a set of interviews, documents, or other text.
- Also, you can use it to communicate the most salient points or themes in the reporting stage

# TEXT VISUALIZATION

The uses of text visualization charts below:

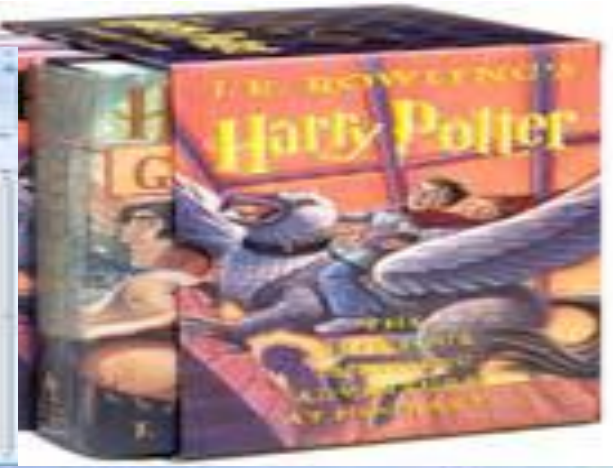
1. Summarize Large Amounts of Text
   - Automatically highlight key terms in a series of texts, and categorize text by topic, sentiment, and more, saving hours of reading time.
   - With a text visualization or data visualization dashboard, you can understand text data at a glance.
2. Make Text Data Easy to Understand
   - Our brains process visual data 60,000 times faster than texts and numbers. Text visualization examples effectively simplify complex data and communicate ideas and concepts to team managers.
3. Find Insights in Qualitative Data
   - Customer feedback holds a trove of insights. Through text visualization examples, you can get an overview of the features, products, and topics that are most important to your customers.
4. Discover Hidden Trends and Patterns
   - You can easily analyze and visualize insights over time to detect fluctuations, and quickly find the root cause.
   - Extracting reliable insights from qualitative data sets, such as keywords, should never be an Achilles Heel for you. Keep reading because we'll address the following question: why do we need text visualization?

# WHY VISUALIZE TEXT?

- **Understanding** – get the "gist" of a document
- **Grouping** – cluster for overview or classification
- **Compare** – compare document collections, or
- inspect evolution of collection over time
- **Correlate** – compare patterns in text to those in
- other data, e.g., correlate with social network

# WHAT IS TEXT DATA?

- Documents
  - Articles, books and novels
  - E-mails, web pages, blogs
  - Tags, comments
  - Computer programs, logs
- Collection of documents
  - Messages (e-mail, blogs, tags, comments)
  - Social networks (personal profiles)
  - Academic collaborations (publications)




# WHY DO WE NEED TEXT VISUALIZATION?

1. Text Visualization can help reveal your audience's thoughts

   You can use the chart to understand your audience's feelings about a topic/situation. Besides, you can leverage the chart to summarize data-driven views. The chart can help you summarize the market feedback using first-hand data.

2. Quick and informative:

   You can easily get live feedback from your audience in real-time

3. Exciting and emotional:

   The chart can help audiences feel part of your data story.

4. Engaging

   The Word cloud is incredibly engaging and visually appealing to many audiences. The chart can be an icebreaker or an entry point for a topic of discussion. Our brains process visual content 60,000 times faster than texts and numbers. This provides a logical rationale for using the Word Cloud generator to analyze your textual data for actionable insights.

# TEXT DATA VISUALIZATION EXAMPLES

- Word Clouds are charts that display insights into qualitative data frequency.
- The visualization design gives greater prominence to words that appear more frequently in a source text. The larger the word, the higher its frequency.
- You can use the chart (one of the text visualization examples) to perform exploratory textual analysis by identifying words that frequently appear in a set of interviews, documents, or other text.
- Also, you can use it to communicate the most salient points or themes in your data stories

# TEXT DATA VISUALIZATION EXAMPLES

- **Tag clouds or text clouds** are ideal if your goal is to pull out the most pertinent parts of textual data, from blog posts to databases.
- You can use the tag cloud as a text visualization tool to compare and contrast two different pieces of text for similarities and differences.

# VISUALIZATIONS OF A SINGLE TEXT DOCUMENT

- Visualizations of a single text document refer to the use of graphical representations or interactive displays to visually convey information, patterns, or insights contained within the document

# WORD CLOUD

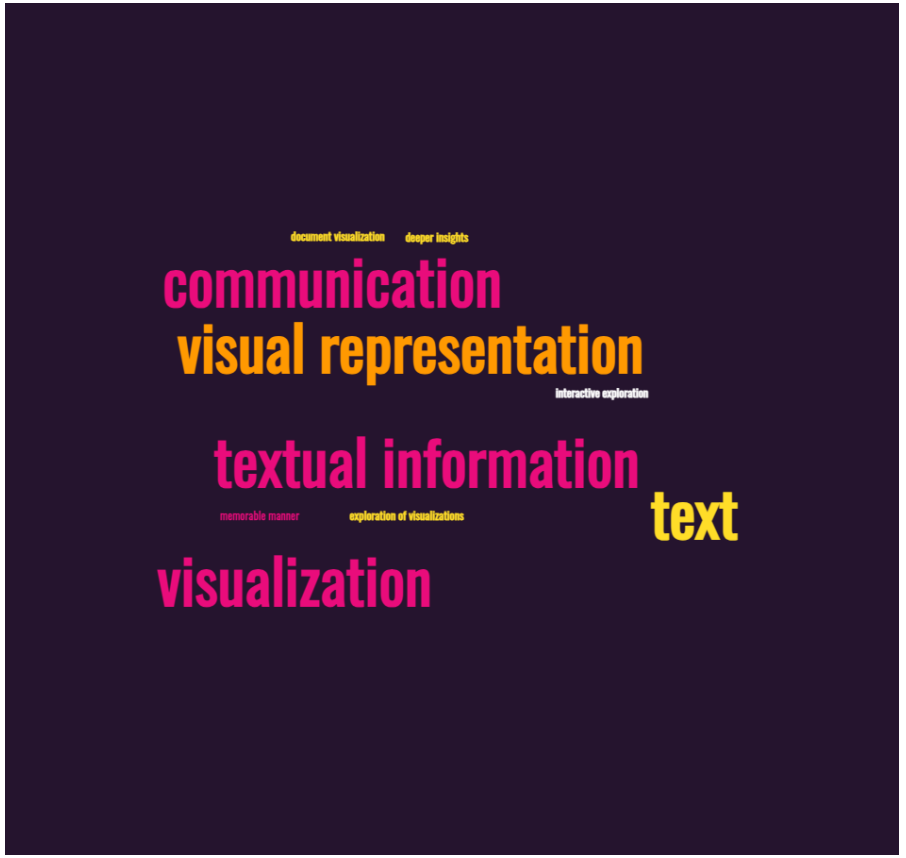
- A word cloud (also known as a tag cloud) is a visual representation of words.
- Cloud creators are used to highlight popular words and phrases based on frequency and relevance.
- They provide you with quick and simple visual insights that can lead to more in-depth analyses.

https://monkeylearn.com/

Word Clouds have become a popular way of displaying the most common words and phrases occurring in large volumes of text.

# WORD CLOUD: EXAMPLE

Text and document visualization is a process of transforming textual information and documents into visual representations that aid in understanding, analysis, and communication. Visualizations reveal patterns and relationships within the text that may not be apparent in raw form. Visual representations provide an intuitive way to comprehend textual content, presenting it in a structured and organized manner. Interactive exploration of visualizations allows users to delve into specific details and gain deeper insights. Visualizations enhance communication by presenting textual information in an engaging and memorable manner.



# WORD TREES

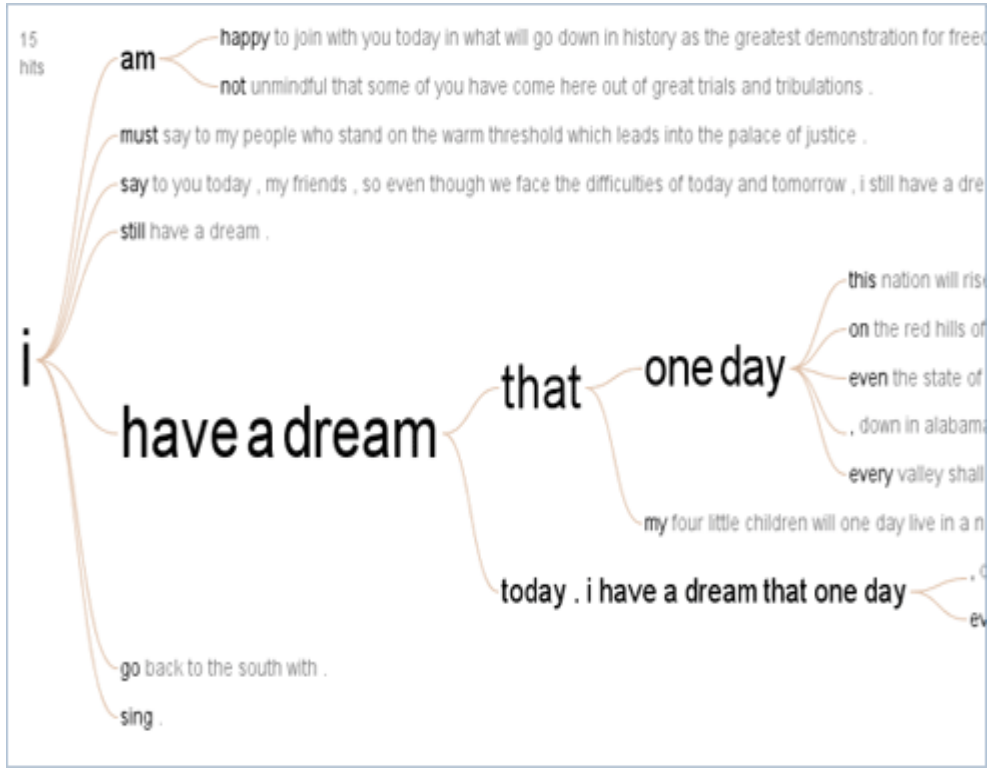
- Word trees use a visual branching structure to show how a pre-selected word(s) is connected to other words
- **Word Tree** is a visualization of a set of words, i.e. text data, in a hierarchical way.
- Invented by the duo Martin Wattenberg and Fernanda Viégas in 2007, the **Word Tree** chart type can be helpful in displaying which words most often follow or precede a target word or phrase (e.g. "AnyChart is...") and in showing a hierarchy of terms.
- Words in a **Word Tree** chart are shown as branches going from the root word. The font size of each word represents its weight determined by the frequency of occurrence / number of children
- https://www.jasondavies.com/wordtree/

# WORD TREES

- Here is a step-by-step explanation of how a Word Tree is constructed:

1. Starting point: You choose a root word or phrase that serves as the starting point of the Word Tree. This can be any word or phrase of interest in your dataset.
2. Branches: From the root word, branches emerge representing different paths or sequences of words that occur in the dataset. Each branch represents a specific word or phrase that follows the preceding words in the path.
3. Hierarchy: The Word Tree builds a hierarchical structure by displaying the relationships between words. The root word is at the top, and subsequent words or phrases are displayed as branches descending from the root. These branches can have additional branches branching out from them, representing further sequences or paths of words.
4. Font size: The font size of each word in the Word Tree is not typically used to indicate its weight or frequency of occurrence. Instead, it may be used to emphasize certain words or to convey additional information. For example, you could use different font sizes to represent the importance or significance of specific terms within the hierarchy.
5. Visualization: The Word Tree is usually displayed as a visual diagram, where the branches and words are arranged in a way that allows for easy exploration and interpretation. The layout can vary, but commonly used formats include horizontally oriented trees, radial trees, or stacked trees.

# EXAMPLE




- A word tree is a visual search tool for unstructured text, such as a book, article, speech or poem.
- It lets you pick a word or phrase and shows you all the different contexts in which it appears. The contexts are arranged in a tree-like branching structure to reveal recurrent themes and phrases.
- The image above is a word tree made from Martin Luther King's famous "I have a dream" speech, using the search term "I." Font sizes show frequency of use, so you can see that among King's many uses of "I," the most frequent context is the phrase "I have a dream."

# EXAMPLE


Shift-click to make that word the root.
cried
the
Mouse,
in a shrill, passionate voice. `Would YOU like cats if you were me?' `Well, perhaps not,' said Alice in a soothing tone: `don't be angr
who was trembling down to the end of his tail. `As if I would talk on such a subject!
sharply and very angrily. `A knot!' said Alice, always ready to make herself useful, and looking anxiously about her.
Gryphon
, and
the Mock Turtle had just begun to repeat it, when a cry of `The trial's beginning!
, taking Alice by the hand, it hurried off, without waiting for the end of the song.
. `Of course,' the Mock Turtle said: `advance twice, set to partners--' `--change lobsters, and retire in same order,' continued the
Mock Turtle
. `Seals, turtles, salmon, and so on; then, when you've cleared all the jelly-fish out of the way--' `THAT generally takes so
, capering wildly about. `Change lobsters again!' yelled the Gryphon at the top of its voice.
Alice
,
with a sudden burst of tears, `I do wish they WOULD put their heads down! I am so VERY tired of being all alone here!
jumping up and down in an agony of terror. `Oh, there goes his PRECIOUS nose'; as an unusually large saucepan flew close by it, and very near
quite forgetting in the flurry of the moment how large she had grown in the last few minutes, and she jumped up in such a hurry that she tipped
(she was so much surprised, that for the moment she quite forgot how to speak good English); `now I'm opening out like the largest telescope that ever w
hastily, afraid that she had hurt the poor animal's feelings. `I quite forgot you didn't like cats.
again, for this time the Mouse was bristling all over, and she felt certain it must be really offended.
in a sorrowful tone, `I'm afraid I've offended it again!' For the Mouse was swimming away from her as hard as it could go, and making quite a commotion
.
`Come, there's no use in crying like that!' said Alice to herself, rather sharply; `I advise you to leave off this minute!
`Wake up, Dormouse!' And they pinched it on both sides at once. The Dormouse slowly opened his eyes.
out
when they saw Alice coming. `There's PLENTY of room!' said Alice indignantly, and she sat down in a large arm-chair at one end of the table.
, `Silence in the court!' and the King put on his spectacles and looked anxiously round, to make out who was talking.
so much!' said Alice, as she swam about, trying to find her way out. `I shall be punished for it now, I suppose, by being drowned in my own tears!
`We indeed!' cried the Mouse , who was trembling down to the end of his tail. `As if I would talk on such a subject! Our family always HATED cats: nasty, low, vulgar things! Don't let me hear the name again!'
`I won't indeed!' said Alice , in a great hurry to change the subject of conversation. `Are you--are you fond--of--of dogs?' The Mouse did not answer, so Alice went on eagerly: `There is such a nice little dog near our house I should like to show you! A little bright-eyed terrier, you know, with oh, such long curly brown hair! And it'll fetch things when you throw them, and it'll sit up and beg for its dinner, and all sorts of things--I can't remember half of them--and it belongs to a farmer, you know, and he says it's so useful, it's worth a hundred pounds! He says it kills all the rats and--oh dear!' cried Alice in a sorrowful tone, `I'm afraid I've offended it again!' For the Mouse was swimming away from her as hard as it could go, and making quite a commotion in the pool as it went.
So she called softly after it, `Mouse dear! Do come back again, and we won't talk about cats or dogs either, if you don't like them!' When the Mouse heard this, it turned round and swam slowly back to her: its face was quite pale (with passion, Alice thought), and it said in a low trembling voice, `Let us get to the

# TEXT ARCS

- Text arcs are a type of data visualization technique that represents textual information along a curved or circular arc. This technique is commonly used to display and analyze sequential or time-based data, such as narratives, timelines, or event sequences.
- The text arc technique involves placing individual characters or words along the arc, following the curvature of the path.
- The curvature can be a complete circle or a partial arc, depending on the specific visualization design and the purpose of the text placement.
- Text arcs are often employed in various types of visualizations, such as pie charts, radial charts, or circular timelines.
- They can provide additional context or explanation for the data points or segments being represented. For example, in a pie chart, the labels for each slice can be placed along the arc corresponding to its position

# TEXT ARCS

- Here's an explanation of how text arcs work in data visualization:

✓Arc representation:

In a text arc visualization, a curved or circular arc serves as the visual container for displaying the textual information.

The arc can be a complete circle or a portion of it, depending on the specific requirements of the visualization.

✓Text placement:

The text is positioned along the arc in a way that represents the sequential or temporal order of the data.

Each text element, such as a word, phrase, or sentence, is placed along the arc, following the direction of the sequence.

# TEXT ARCS

- ✓Arc length and positioning:

  The length of the arc segment assigned to each text element can be proportional to its duration or significance.

  For example, longer arcs can represent longer time periods or more important events. The positioning of the text elements along the arc can be evenly spaced or adjusted based on specific criteria or data attributes.

- ✓Formatting and visualization enhancements:

  Various formatting techniques can be applied to the text to enhance the visualization and convey additional information.

  For instance, different font sizes, colors, or styles can be used to represent different categories, emphasize important elements, or indicate specific attributes associated with the text data

# TEXT ARCS