Chapter 3: Visualization of Unstructured Data

In the midst of chaos, there is also opportunity–
Sun-Tzu

Learning Objective

- Learn the importance of text data visualization.
- Learn the challenges of handling text data and various techniques of text analysis.
- Explore various types of text data and understand the text pre-processing pipeline.
- Understand how we use visualizations to comprehend text data using an example dataset.
- Learn about exploratory analysis of text data using various tools, python libraries and
- techniques of text data analysis.
- Build and understand model interpretation on the example dataset.

Introduction

- Data produces information that leads to value creation in the form of data driven decisions.
- It can be generally divided into two categories: structured data and unstructured data.
- Structured data, also known as quantitative data, is a format that fits well into a relational database.
- Unstructured data, also called qualitative data, cannot be displayed in a readable format by computers or relational databases.
- For example, webpages, text data, emails, customer review data, social media data, chatbot conversations, and so on.

Importance of Text Data Visualization

- International Data Corporation (IDC) estimates that 80% of all the data will be unstructured by 2025 (Anon, 2017).
- This trend is primarily driven by increase in digitalization. Social media platforms like Facebook, Twitter and Instagram, and increased interest in online shopping continue to the rapid growth of unstructured data.
- Between 2017 and 2025, the global **text analytics** market is projected to grow from USD 3.23 Billion in 2017 to USD 18.28 Billion by 2025 at a CAGR of 24.2%.[1]

[1] Source: https://www.theinsightpartners.com/reports/text-analytics-market

Importance of Text Data Visualization

- Quantzig's social media analytics solutions helped a leading entertainment and sports company achieve a 65% reduction in survey expenditures and improved customer satisfaction by three times (Anon, 2020).
- Many organizations extract hidden patterns in their customer feedback data.
- Insurance companies are using text data to identify fraud.
- The healthcare industry is using a huge amount of published literature in clinical practices.

Importance of Text Data Visualization

- Due to its readability for both humans and machines, text data visualization is an important tool in Data Science.
- Interesting enough, the human brains are not good at processing such a large amount of information.
- It can process a visual much faster than text.
- Text data visualization helps in summarizing and providing the gist of a document by showing the most relevant keywords.
- It can reveal hidden patterns and trends across documents over time using line charts.

Harry Potter's Spells

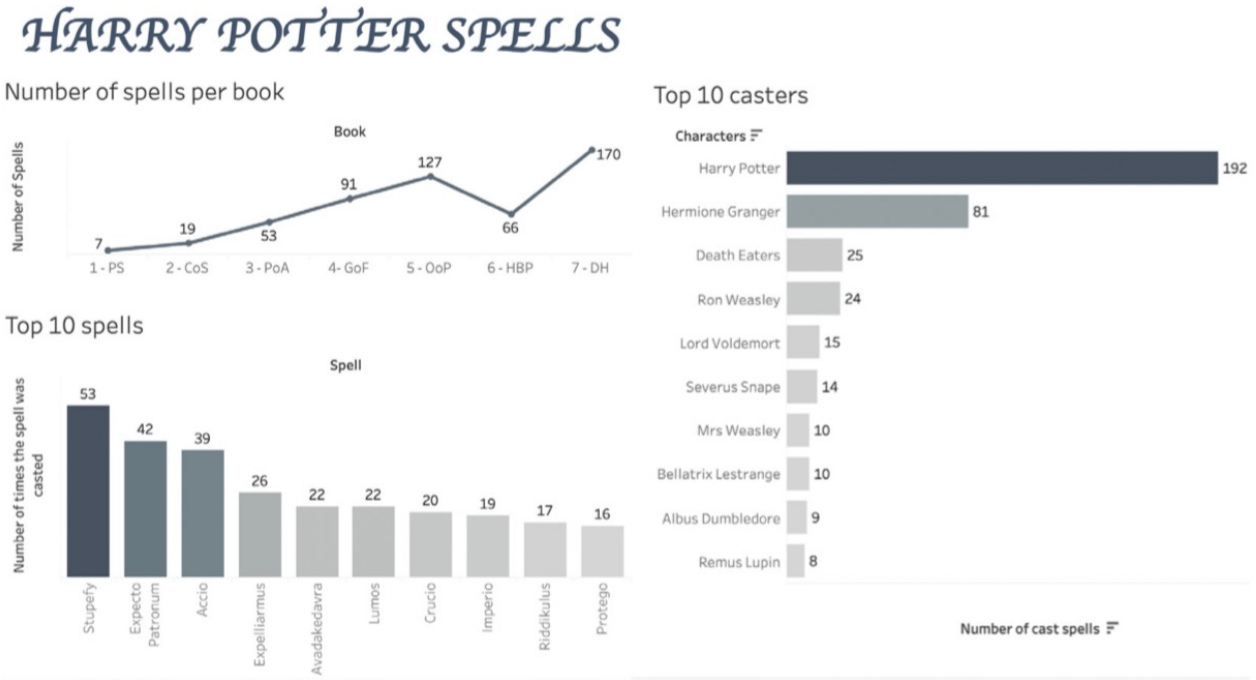


Figure 4.2 Harry Potter spells.
Source: https://public.tableau.com/app/profile/chiara.pedrazzoli/viz/HarryPotterSpells_15986333032660/HPSpellsOverview

- The visualization summarizes analysis of various spells in the Harry Potter book series [Pedrazzoli, 2020].
    1. We can check the number of unique spells throughout the series,
    2. the Top 10 spells that were used,
    3. the Top 10 casters who used these spells, and
    4. the number of spells per book.

Harry Potter's Spells

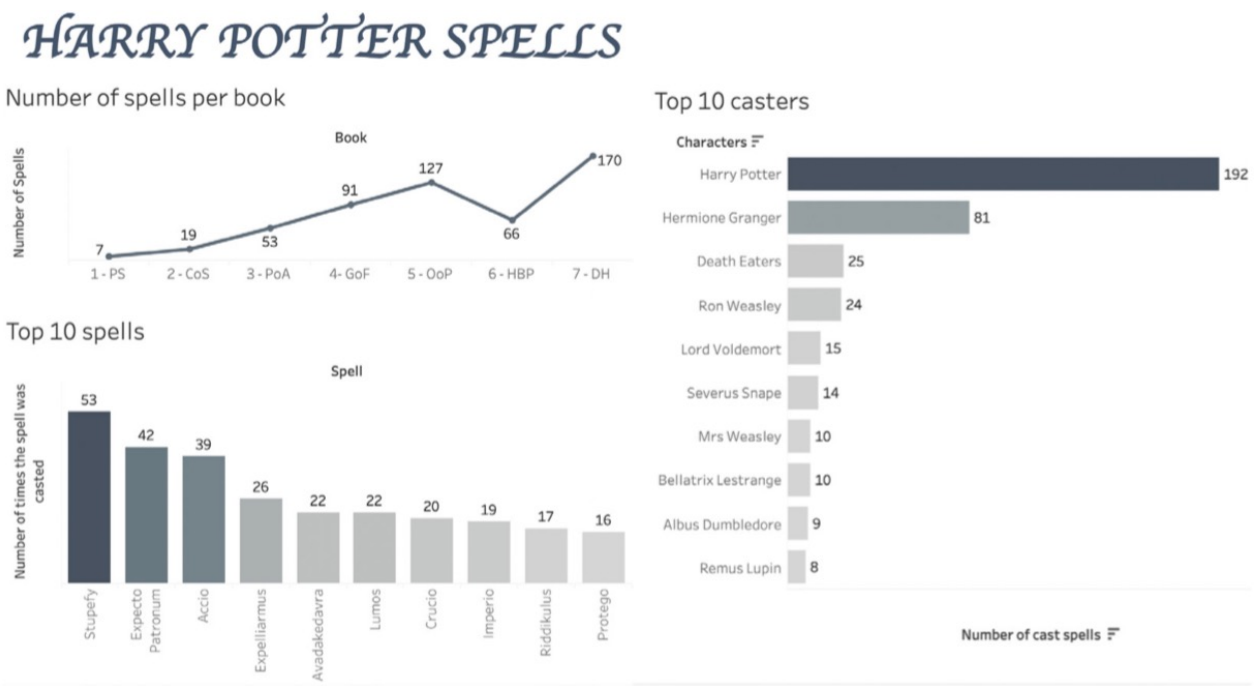


Figure 4.2 Harry Potter spells.
Source: https://public.tableau.com/app/profile/chiara.pedrazzoli/viz/HarryPotterSpells_15986333032660/HPSpellsOverview

- This graph helps us compare the use of spells throughout the book series.
- Interestingly, "Harry Potter and The Deathly Hallows" has the highest number of spells throughout the Harry Potter series, and
- Harry Potter is the top caster among all.

Mr. Bill Clinton's vs Mr. Obama's Speech

Figure 4.3 World cloud of Mr. Bill Clinton's speech on healthcare (1993).

Figure 4.4 World cloud of Mr. Barack Obama's speech on healthcare (2009).

- Figures show the word clouds of President Bill Clinton's speech in 1993, and
- President Barack Obama's speech to the Congress in 2009 regarding the need for healthcare reform in the United States [Rampell, 2009] .
- Word clouds make it easier for us to get the gist of the data by high-lighting the most important keywords in their speech data.

Mr. Bill Clinton's vs Mr. Obama's Speech

Figure 4.3 World cloud of Mr. Bill Clinton's speech on healthcare (1993).

Figure 4.4 World cloud of Mr. Barack Obama's speech on healthcare (2009).

- We can infer, the focus in Clinton's speech (Fig. 4.3) is towards "healthcare", "people", "doctor" and "American", compared to Obama's speech (Fig. 4.4) which focuses on "insurance", "job", "coverage", "pay" along with "healthcare" and "Nation".
- So, while Obama's speech focused more on the Insurance coverage, Clinton's views on insurance are slightly blurred.

Challenges of Text Data Visualization

- One of the most challenging problems with text data is that it is huge.
- Harry Potter is a fantasy novel consists of seven books, total about 3500 pages.
- Mr. Bill Clinton's speech and Mr. Barack Obama's speech on healthcare has about 7000 and 600 words respectively.
- There are more than 155.3 million items in the Library of Congress[2], which is almost impossible to read for a human in a lifetime.
- Global healthcare data generated in 2020 was estimated to be 2,314 exabytes [Zwolenski, 2014].

[2] Source: Discover 10 treasures from Library of Congress - https://constitutioncenter.org/

Challenges of Text Data Visualization

- The human brain is wired to understand, speak, and write a language, but a computer's native language is different.
- Computers communicate not with words but with millions of combinations of ones and zeroes resulting in logical actions.
- Machines can understand numbers, but not natural language.
- When humans read text, they see words and understand their meaning, while a computer sees a word as a vector of dummy variables or a sequence of tokens.
- This makes visualizing text data extremely challenging.

Various Forms of Text Data

A few examples of text data are mentioned below:

1. **Documents:** An organization's annual report, legal documents, presentations, books, literature, web pages, blogs, news articles, research publications, etc. Investors may like to extract important information from such documents to make investment decisions.
2. **Emails:** Business communications via emails can be translated into relevant statistics, insights, and entities. Especially since the Covid 19 crisis, which led to the work from home phenomenon and social distancing norms, digital channels have become an essential for communication.
3. **Social media data:** The increased digitalization, in today's world has led people to share their views and feedback on several social media sites such as WhatsApp, Instagram, LinkedIn, Facebook, etc.

Text Data Pre-processing Pipeline

- The process of converting text data into numeric form which machines can understand is called pre-processing.
- It is a process of breaking the text into numerical features that machines can understand.
- The text pre-processing pipeline includes:

1. Converting all letters to one single case, either lowercase or uppercase.
2. Removing numbers or converting them to words.
3. Removing punctuation and special characters such as !/@#$%\^&<>?*”

Text Data Pre-processing Pipeline

4. **Tokenization:** It is a process of breaking text into smaller units called tokens.
   - It can be at word or character level.
   - A text data can be converted into a Bag-of-Words (BoW).
   - Frequently BoW is represented as a set listing the words often along with its frequency of occurrence in a document.
   - For example:

   a. Input: Text data visualization can be challenging.
   b. BoW Output: [‘text’, ‘data’, ‘visualization’, ‘can’, ‘be’, ‘challenging’]

Text Data Pre-processing Pipeline

5. Removal of non-value adding information such as:

a. **Stop words**- These are commonly occurring words in a language that does not add any valuable information to the text.

- For example, ‘the’, ‘a’, ‘and’, etc. Removing stop words helps us focus on more important words in the text that can give us useful insights.

Text Data Pre-processing Pipeline

b. **Stemming**- Many words may convey the same meaning.

- Enjoyed → Enjoy
- Enjoyable → Enjoy

- In stemming, we retain the common portion, that is, ‘enjoy’ is retained in both cases.
- However, this may not work always.
- For example, consider the words ‘Happy’ and ‘Happiness’.
- The following root words do not have any meaning and thus are not useful.

i. Happy → Happ
ii. Happiness → Happ

Text Data Pre-processing Pipeline

c. **Lemmatization:** It is a process of converting a word into its root form using morphological analysis, thus an improvement over stemming.

For example:

i. Increase → increase

ii. Increased → increase

iii. Increasing → increase

Visualizing Text Data

Table 4.1 Parts of Panchatantra

Book Subtitle	Translation
Mitra-bheda	The Separation of Friends
Mitra-labha	The Gaining of Friends
Kakolukiyam	War and Peace (Crows and Owls)
Labdhapranasam	Loss of Gains
Apariksitakarakam	Ill-Considered Actions

- We will use an ancient Indian collection of inter-related animal fables arranged within a story frame, known as the Panchatantra.[3]
- In each part, there is a main story, called frame story, and within that main story, several stories are embedded.
- This literature consists of five parts as listed in Table 4.1.

[3] Source:Wikipedia - https://en.wikipedia.org/wiki/Panchatantra

Data Description - Panchatantra

```
text[0:1000]
```

```
TextBlob("The Jackal and the Drum. A hungry jackal ended up at an abandoned battlefield while in search of food. Here
, he encountered loud and strange sounds and he thought "I must run away from this place before the person who is mak
ing such sounds gets me". But soon after, he realized, "It will not be proper for me to run away without knowing the
cause of the sounds. Whether it is fear, or happiness, one must know the reason behind it, otherwise there will be no
regret. Let me look for the source of these scary noises". Warily, the jackal marched in the direction of the sounds
and found a drum there. It was this drum, which was sending the sounds whenever the branches of the tree above brushe
d against it. Relieved, the jackal began playing the drum and thought that there could be food inside it. The jackal
entered the drum by piercing its side. He was disappointed to find no food in it. Yet he consoled himself saying that
he rid himself of the fear of sound. "Therefore", Damanaka told king Pin")
```

Figure 4.5 Snippet from Panchatantra.txt.

- The story data was scraped from the web, Tales of Panchatantra, and stored in two formats.

1. “.txt format”: .txt format consists of plain text.
a. Panchatantra.txt contains all the five parts of the literature in one file.

Data Description- Panchatantra

Table 4.2 Panchatantra.xlsx dataset

Variable	Variable Type	Description
Title	Categorical	Story title
Chapter	Numerical	Chapter number of story for each part
Story	Categorical	Full text of the story
Strategy	Categorical	Name of different parts of the literature, grouped under following categories, 1. Mitra-bheda (The Separation of Friends) 2. Mitra-labha (The Gaining of Friends) 3. Kakolukiyam (War and Peace) 4. Labdhapranasam (Loss of Gains) 5. Apariksitakarakam (Ill-considered Actions)
Strategy Number	Numerical	Strategy number

b. Excel format:

- Panchatantra_Data.xlsx format consists of data in MS-Excel format.
- The data consists of five columns containing both categorical and numerical variables.
- Data description is provided below (Table 4.2).

Word Cloud

- The simplest and most common form of text visualization is word cloud.
- It is a visual representation of words of text data arranged in various sizes, colours, or categories.
- Size encoding is generally used to represent the frequency of words in the data.
- The larger the size of a word in a word cloud, the larger is its frequency within a given text.
- Whether we are analysing customer review data or social media posts, word cloud gives us a sense of what is there in the text data.

Query: What could be the major characters or themes which would feature in Panchatantra stories?

Figure 4.6 Word cloud- visualize the most prominent words in Panchatantra.txt.

- We can build this chart using Tableau.wordClouds.com, WordArt, TagCrowd and Tagxedo are a few other word cloud generating tools.
- We have used Python library word cloud to plot world cloud in Fig. 4.6.

Query: What could be the major characters or themes which would feature in Panchatantra stories?

Figure 4.6 Word cloud- visualize the most prominent words in Panchatantra.txt.

- **Inference:** As we can see, “king’, “friend”, “lion”, “jackal”, “monkey”, and “crow” are a few words highlighted the most.
- From this, we can infer that the story text consists of different tales about friends, animals, and king.

Query: what are the prominent words in each Panchatantra strategy book.

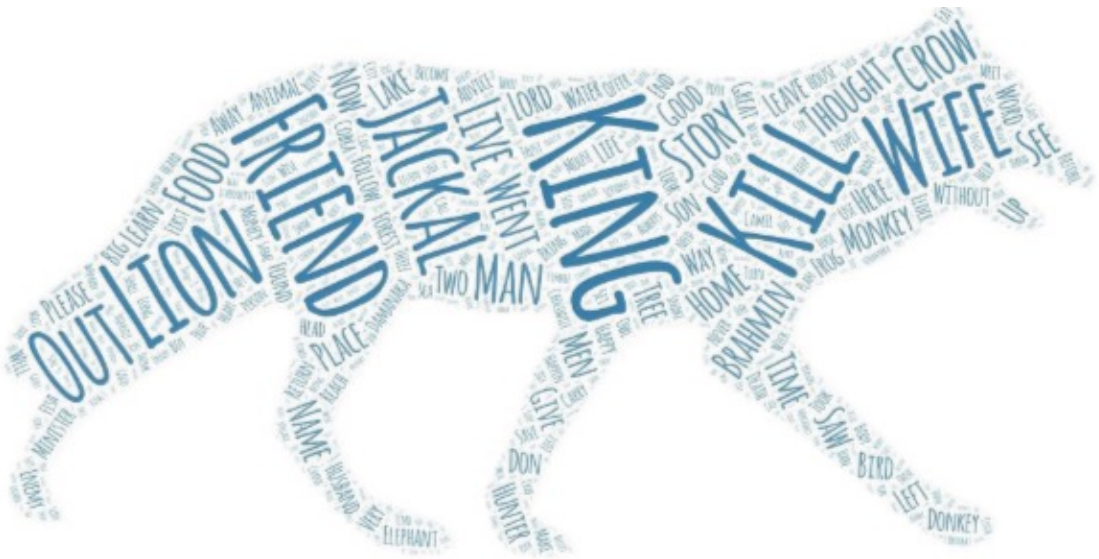


Figure 4.7 Mitra-bheda– the separation of friends.

- **Inference:** Figure 4.7 depicts Strategy 1 **Mitra-bheda– The separation of friends.**
- It consists of words like “King”, “jackal”, “friend”, “kill”.
- In this, the stories are woven with Lion king, Jackal and breaking up friendships as a result of a prominent theme of killing.

Query: what are the prominent words in each Panchatantra strategy book.

Figure 4.8 Mitra-labha– the gaining of friends.

- **Inference:** Figure 4.8 depicts Strategy 2 **Mitra-labha– the gaining of friends.**
- The graph highlights terms such as “Crow”, “Mouse”, “Deer”, “Friend”.
- Hence, the stories for this strategy of friendship are woven with characters like Crow, mouse, and deer.

Query: what are the prominent words in each Panchatantra strategy book.

Figure 4.9 Kakolukiyam– war and peace.

- **Inference:** Figure 4.9 depicts Strategy 3 **Kakolukiyum– War and peace**.
- It highlights terms such as “Crow”, “Owl”, “Bird”, “Minister”.
- Choosing animals/ birds to represent good and evil is a metaphor for war between good and evil.

Query: what are the prominent words in each Panchatantra strategy book.

Figure 4.10 Labdhapranasam– loss of gains.

- **Inference:** Figure 4.10 depicts Strategy 4 **Labdhapranasam– Loss of gains**.
- Here, we see a story about the loss of what is attained told through characters such as “Jackal”, “Crocodile”, and “Donkey”.

Query: what are the prominent words in each Panchatantra strategy book.

Figure 4.11 Apariksitakarakam– Ill-Considered actions.

- **Inference:** Figure 4.11 depicts Strategy 5 **Apariksitakarakam– Ill-considered actions**.
- The stories for this strategy are told using characters such as “Brahmin”, “Wife”, “King”, “Barber” and “Monkey”.
- Interestingly, we can also notice that unlike the previous four parts, this part has lesser number of animal characters involved.

Bar Chart

- Once we have pre-processed text data, we can also use various other charts like bar chart, line chart etc. for exploratory analysis.
- Whenever we work with text data to build any **natural language processing** use cases, the parts of speech of words are identified and labelled accordingly.
- This process of tagging words would further help us in understanding the context of a sentence.
- We have tagged the words of Panchatantra stories data into the following word types.

Bar Chart

- Once we have pre-processed text data, we can also use various other charts like bar chart, line chart etc. for exploratory analysis.
- Whenever we work with text data to build any **natural language processing** use cases, the parts of speech of words are identified and labelled accordingly.
- This process of tagging words would further help us in understanding the context of a sentence.

Query: What are the top 20-word types like Noun, Verb etc., which are used in the Panchatantra stories?

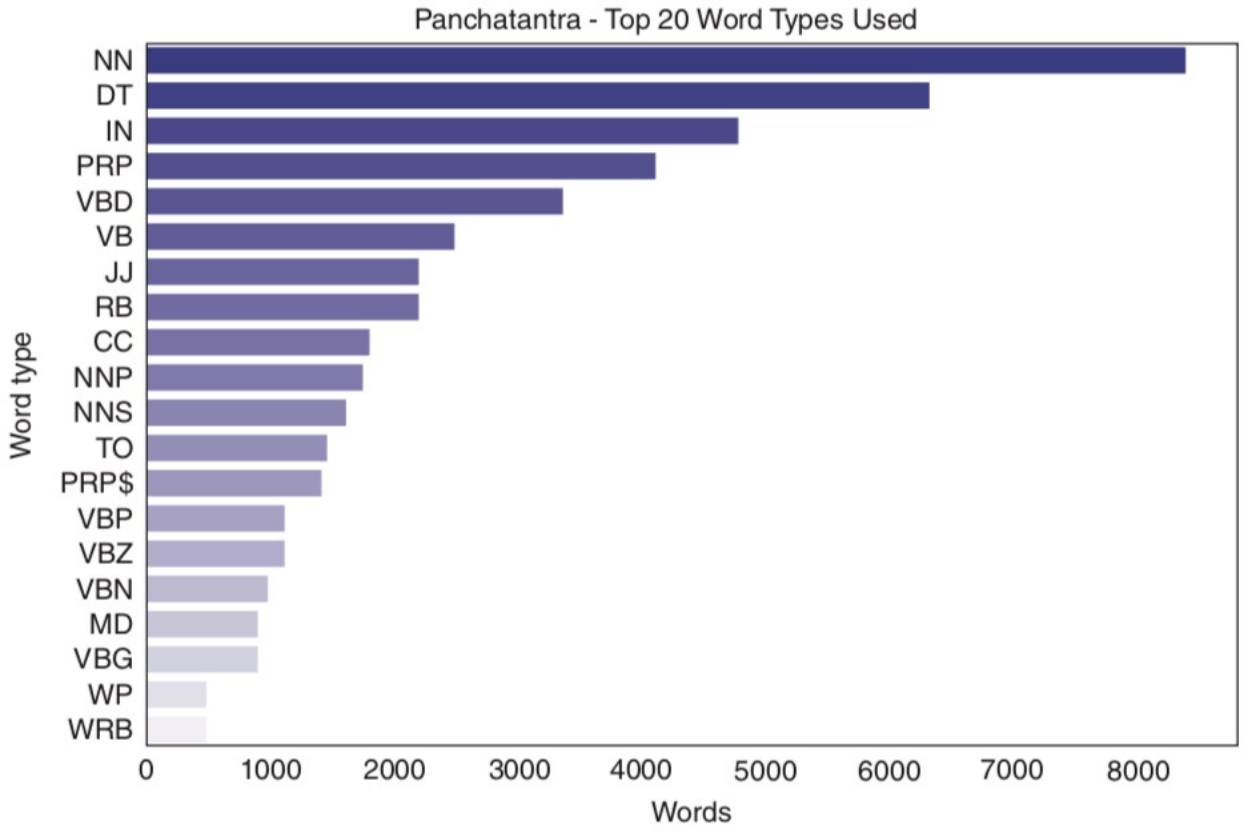


Figure 4.12 Horizontal bar chart- top 20-word types used in Panchatantra text.

- **Visualize:** Since we would like to compare and visualize categorical variables, let us create a bar chart. Refer to Fig. 4.12.
- **Inference:** Singular Nouns is the topmost word type in the Panchatantra stories data, followed by Determiner, which modifies noun or noun phrase.
- We have the least number of adverbs in the text.

Query: What are the most frequently used nouns in Panchatantra stories?

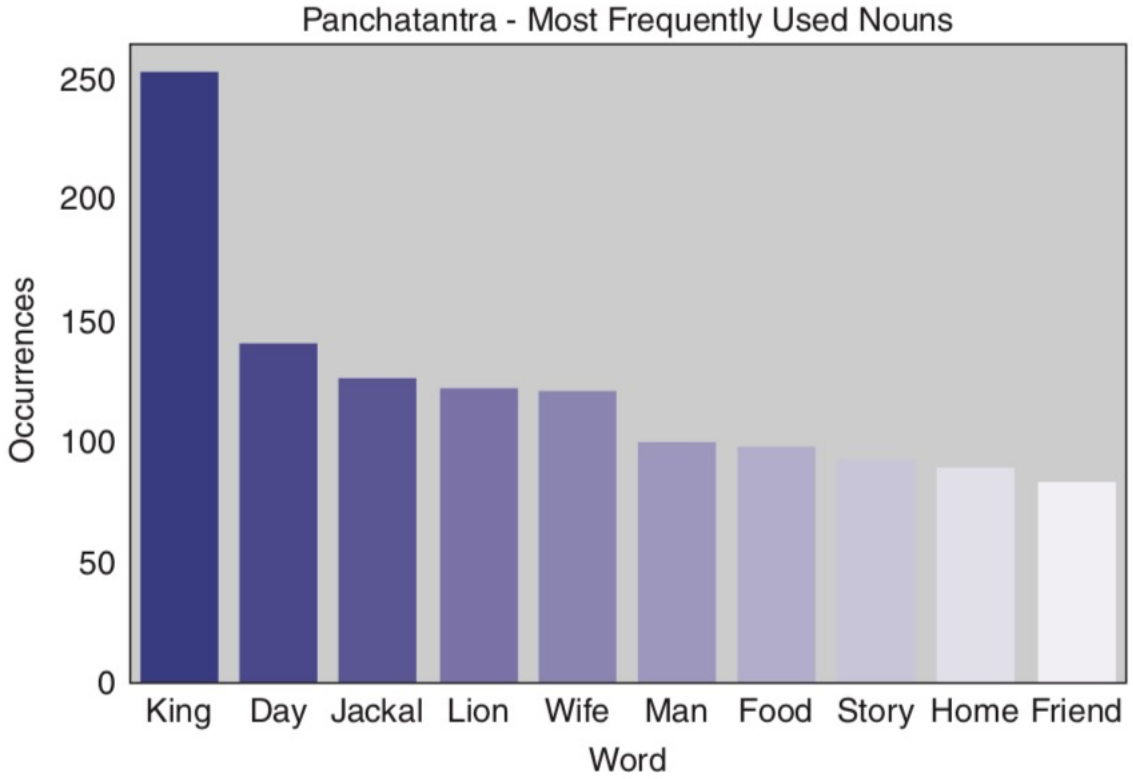


Figure 4.13 Vertical bar chart- most frequently used nouns in Panchatantra text.

- **Visualize:** Refer to Fig. 4.13.
- **Inference:** It is not surprising to see King, Jackal and Lion as the most used nouns in our dataset.
- It is said that Panchatantra stories were written to educate the sons of royalty and interestingly, we see from Fig. 4.13 that King is used almost twice as much as any other word.

Query: Compare the number of words in each Panchatantra strategy.

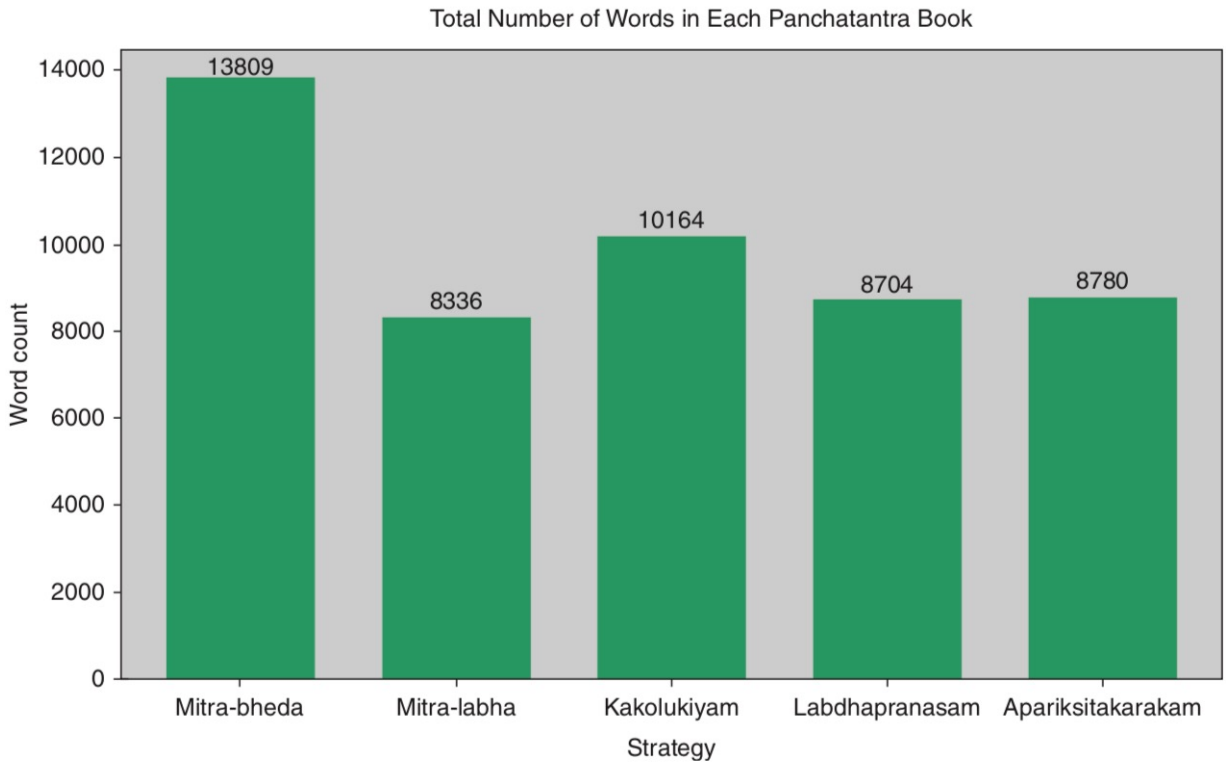


Figure 4.14 Vertical bar chart- total number of words in each Panchatantra book.

- **Visualize:** Refer to Fig. 4.14.
- **Inference:** Story collection on Mitra-bheda (The Separation of Friends), has the highest number of words, followed by Kakolukiyam (War and peace).
- Mitra-labha (The Gaining of Friends), is the part with the lowest number of words.

Word Tree

- Word tree is a text data visualization of a set of words which depicts multiple parallel sequences of words.
- Word tree helps visualize how words are used in different sentences in the text data.
- In addition, it shows the words that follow or precede a target word or phrase.
- In this visualization, as the usage frequency of a word increases, the size of the word increases as well.

Query: Can we visualize and gather information on how the word “lion king” is used in Panchatantra?

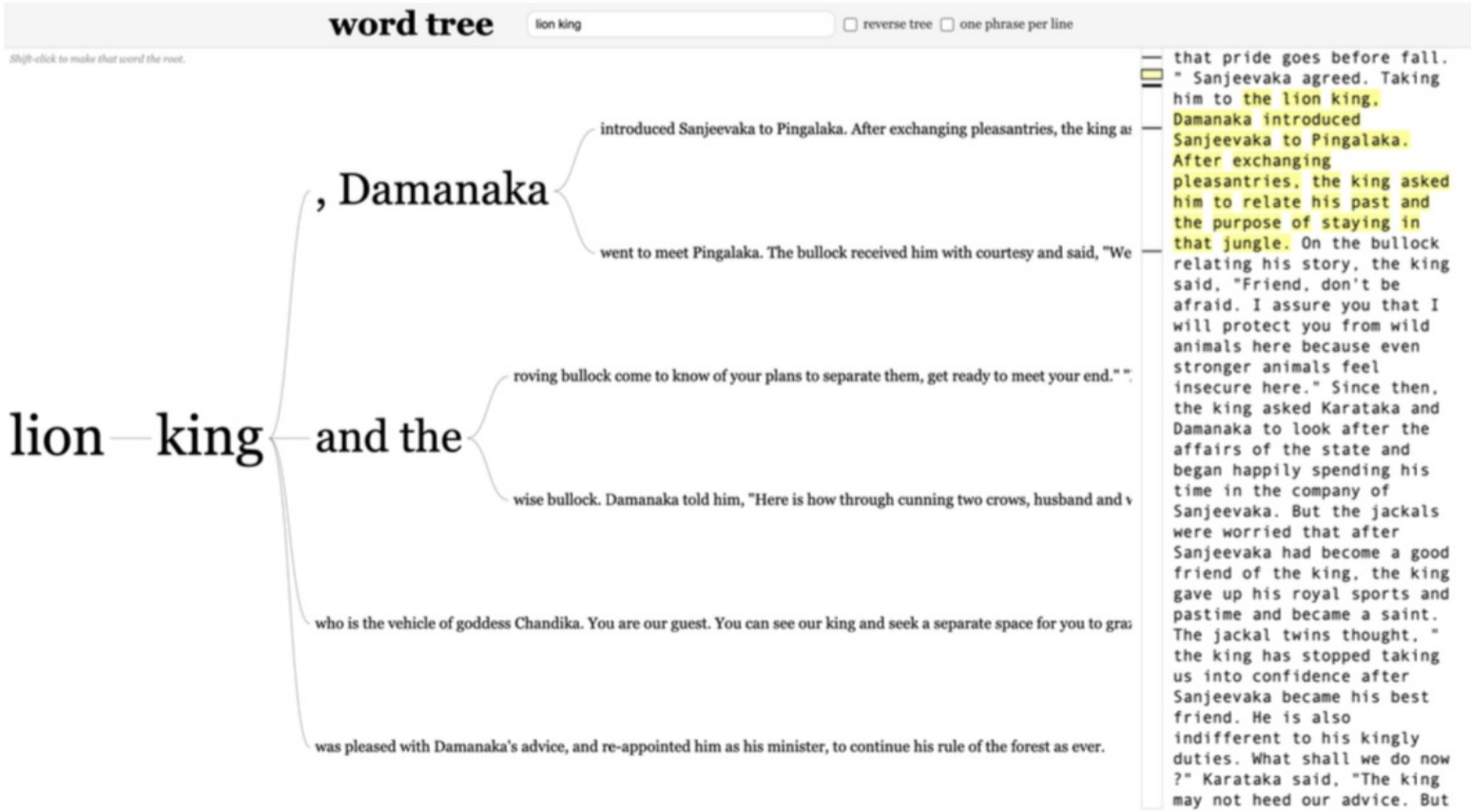


Figure 4.15 Word tree.

Query: Can we visualize and gather information on how the word “lion king” is used in Panchatantra?

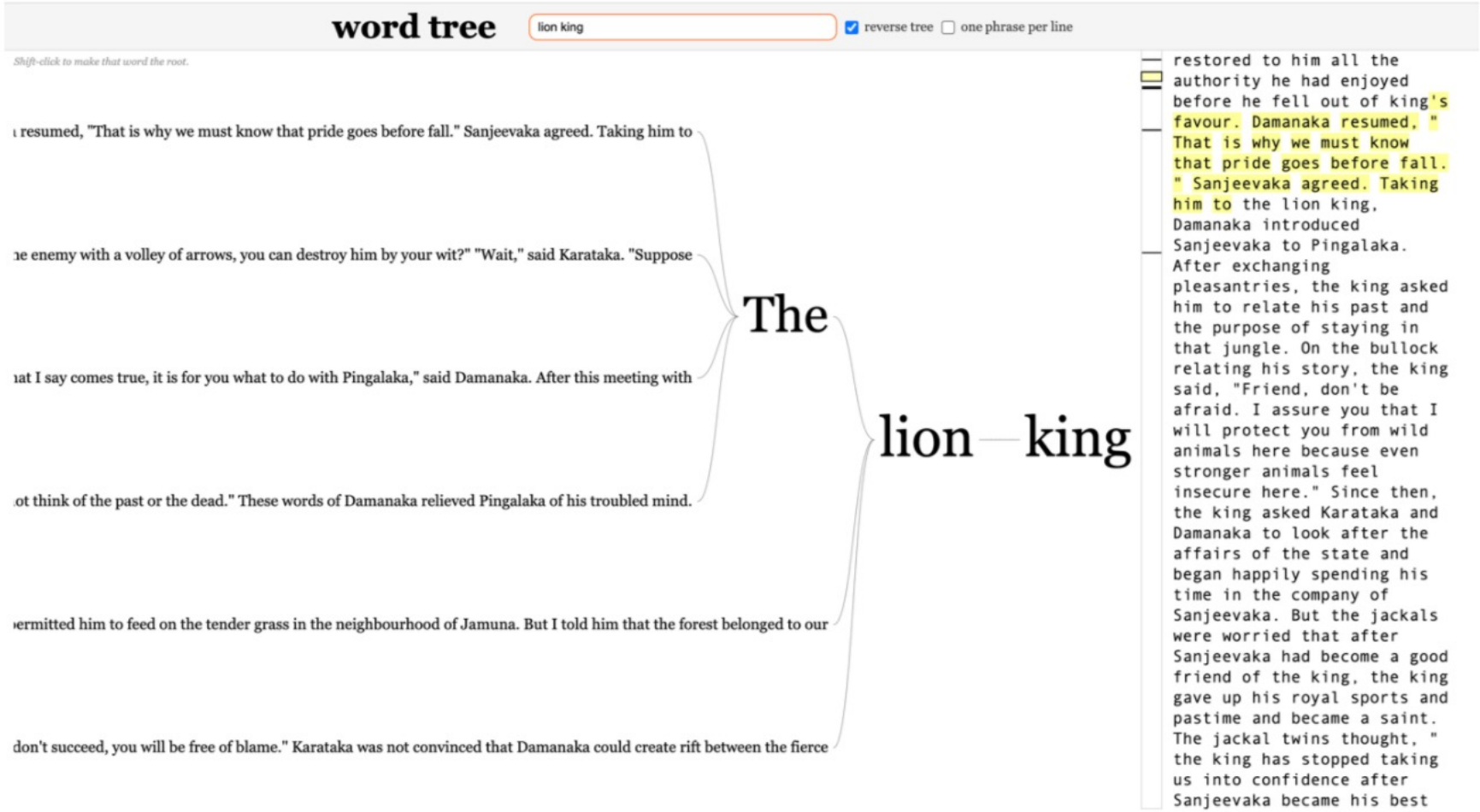


Figure 4.16 Reverse word tree.

Query: Can we visualize and gather information on how the word “lion king” is used in Panchatantra?

- **Visualize:** We can create word trees using a tool for word tree created by Jason Davies.[4] Figure 4.15 shows words that follow “lion king”.
- Figs. 4.15 and 4.16. Figure 4.15 shows words that follow “lion king” whereas Fig. 4.16 shows words that precede it.
- **Inference:** We can infer from word tree that the word “lion king” is used at the beginning as well as the end of a sentence.
- The identical phrases containing the term “lion king” are grouped into nodes.
- Hovering over a node will show all the related sentences in the data.

[4]Source: https://www.jasondavies.com/wordtree/

Line Chart

- When we work with text data, sentiment analysis is one of the important and common use cases.
- It helps determine the sentiment in text dataset as positive, negative, or neutral.
- Applications for sentiment analysis include:
    1. Analysing text data from reviews or feedbacks to determine a user's sentiment. It helps to understand why customers are not interested in a certain product or brand.
    2. Governments use sentiment analysis to analyse public sentiment during election campaigns.
    3. Organizations use sentiment analysis to make better policies.

Line Chart

TABLE 4.3 Panchatantra dataset

Variable	Variable Type	Description
Chapter	Numerical	Chapter number of story for each strategy
Strategy Number	Numerical	Strategy number
Sentence	Categorical	Story divided in each sentence
CompoundScore	Numerical	Compound score – Normalized score of the sum of valence computed based on some heuristics and a sentiment lexicon.
PositiveFlag	Boolean	1 if overall sentiment is Positive, else 0
NegativeFlag	Boolean	1 if overall sentiment is Negative, else 0
NeutralFlag	Boolean	1 if overall sentiment is Neutral, else 0
StrategyTitle	Categorical	Name of different parts of the literature, grouped under the following categories, 1. Mitra-bheda (The Separation of Friends) 2. Mitra-labha (The Gaining of Friends) 3. Kakolukiyam (War and Peace) 4. Labdhapranasam (Loss of Gains) 5. Apariksitakarakam (Ill-Considered Actions)

- Once we do sentiment analysis on the Panchatantra stories data, we get output data with each book tagged with determined sentiment details.
- Data description is provided in Table 4.3.

Line Chart

- We can categorize the sentiments in three categories – positive, negative, and neutral based on their sentiment scores.
- Sentiment scores will have a range between −1 to +1.
- An extreme positive sentiment will have sentiment score close to +1.0 and extreme negative sentiment will have sentiment score close to −1.0.
- Sentiment scores for data which we are analysing ranges between −0.2 to +0.2.
- This indicates that the sentiment throughout the Panchatantra book stays either slightly positive or slightly negative.
- The stories do not seem to have extreme sentiments.

Query: What are the sentiments throughout the book and for each strategy?

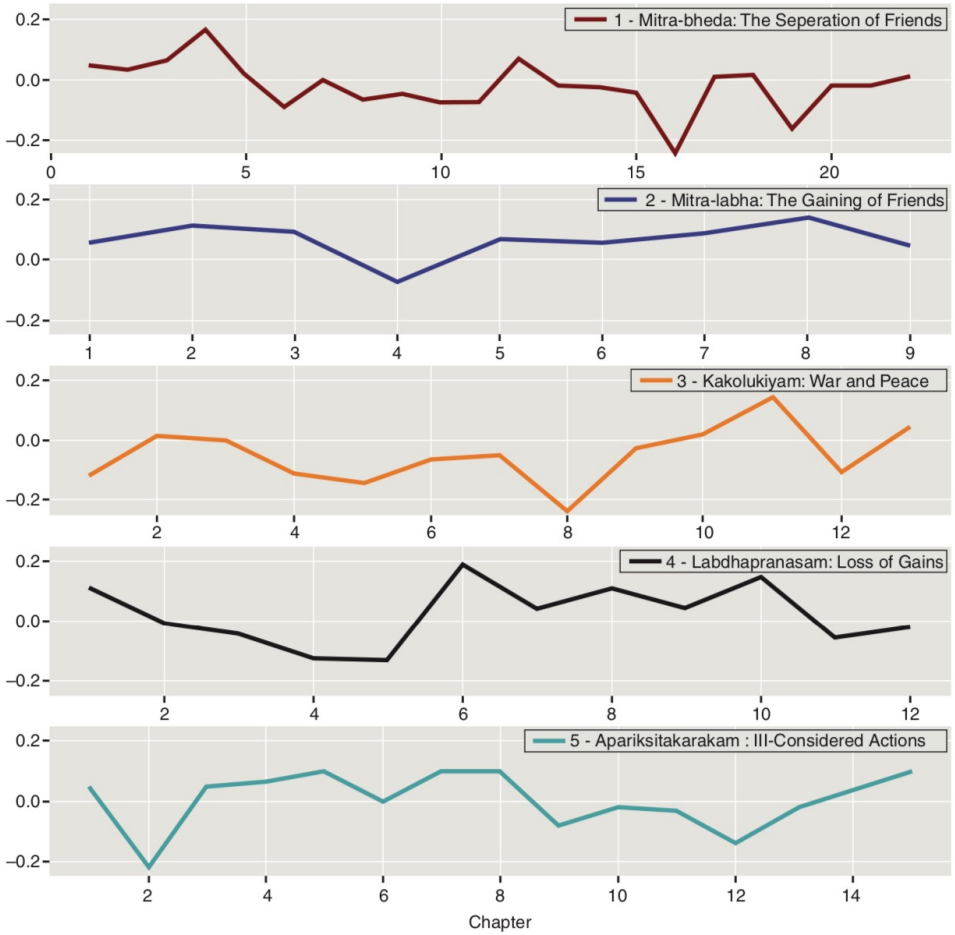


Figure 4.17 Line chart- compound sentiment score in each Panchatantra book.

- **Visualize:** Refer to Fig. 4.17. The sentiment scores are shown in the vertical axis and the chapter numbers are in the horizontal axis.
- **Inference:** Strategy 1 shows a greater number of negative sentiments than positive sentiments.
- As we can see in the first line chart, from chapter 5 to chapter 22 the sentiment score is having either 0 or negative value (except chapter 12).

Query: What are the sentiments throughout the book and for each strategy?

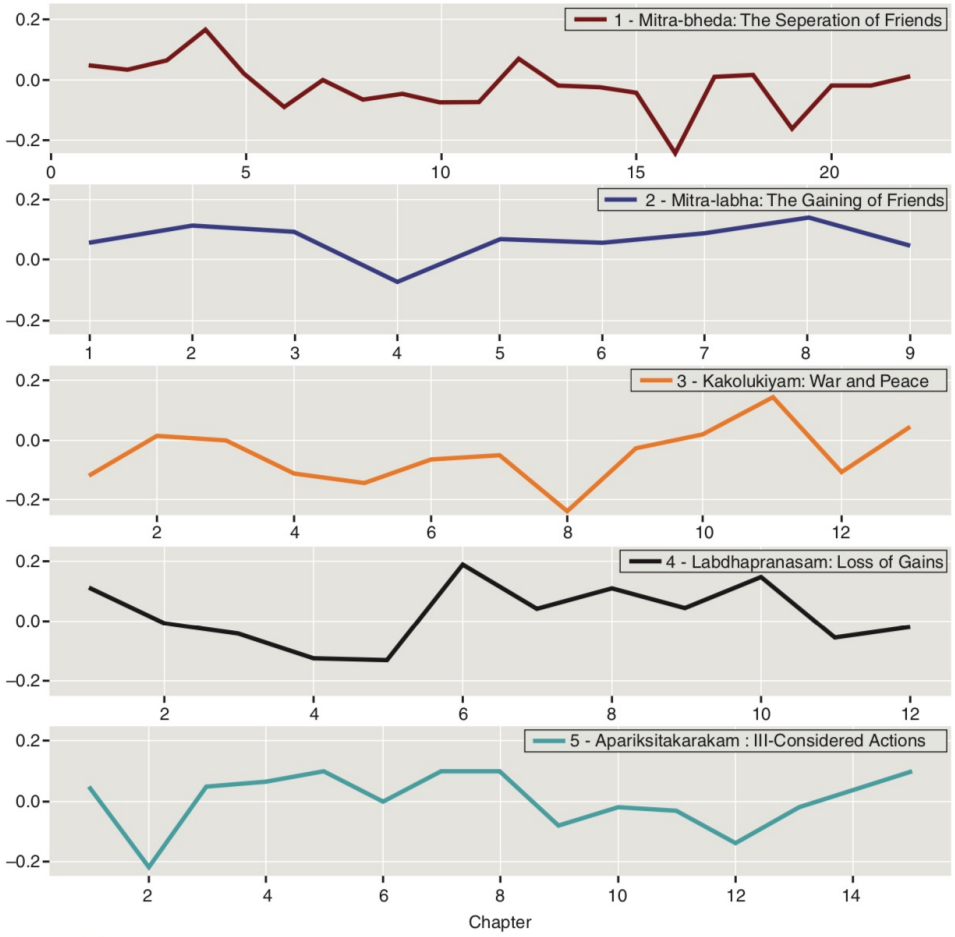


Figure 4.17 Line chart- compound sentiment score in each Panchatantra book.

- Similarly, in the line chart of Strategy 3, nine out of thirteen chapters have negative sentiment score.
- The chapters with positive sentiment score are chapter 2, 10, 11, and 13.
- It is obvious as these strategies are about ‘The Separation of Friends’, and ‘War and Peace’, hence, they must be having a greater number of negative sentiments than positive sentiments.

Query: What are the sentiments throughout the book and for each strategy?

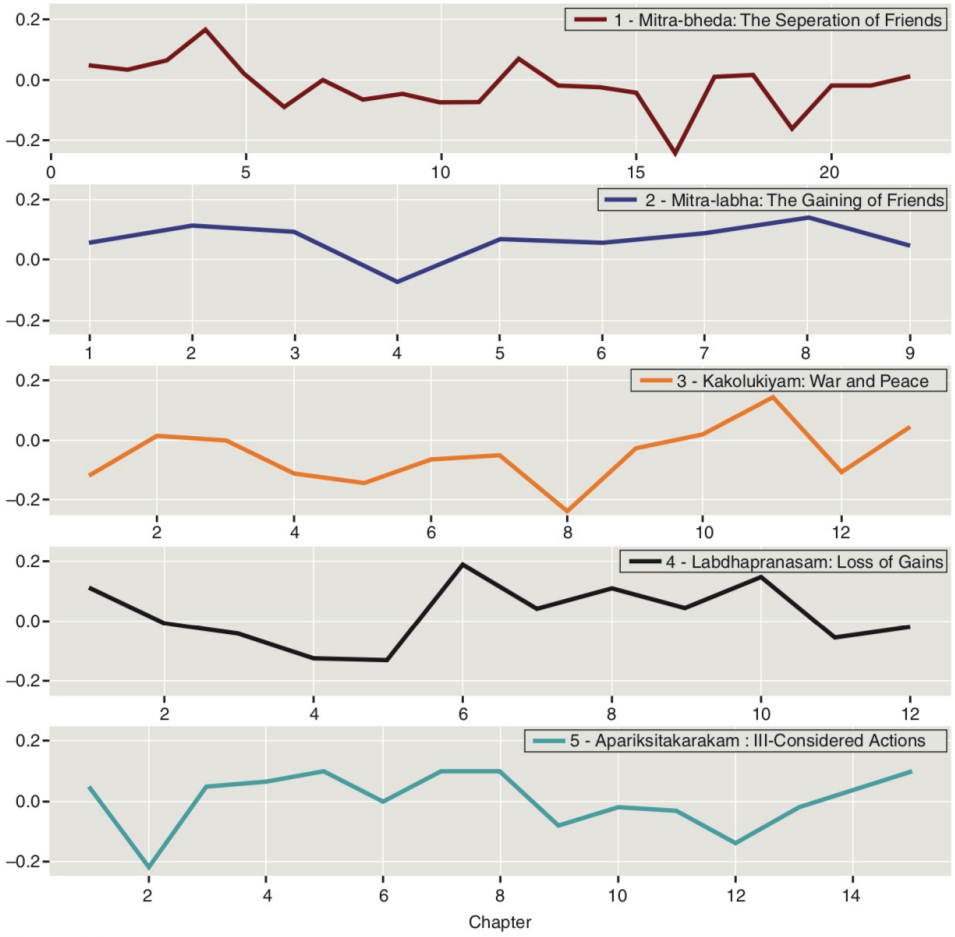


Figure 4.17 Line chart- compound sentiment score in each Panchatantra book.

- On the other hand, in Strategy 2, The Gaining of Friends, except chapter 4, all the chapters have positive sentiment scores which depicts a positive sentiment pattern.
- Strategy 4, Loss of Gains, and Strategy 5, Ill-Considered Actions, shows very slightly positive sentiment pattern across their stories.

Query: What are the sentiments throughout the book and for each strategy?

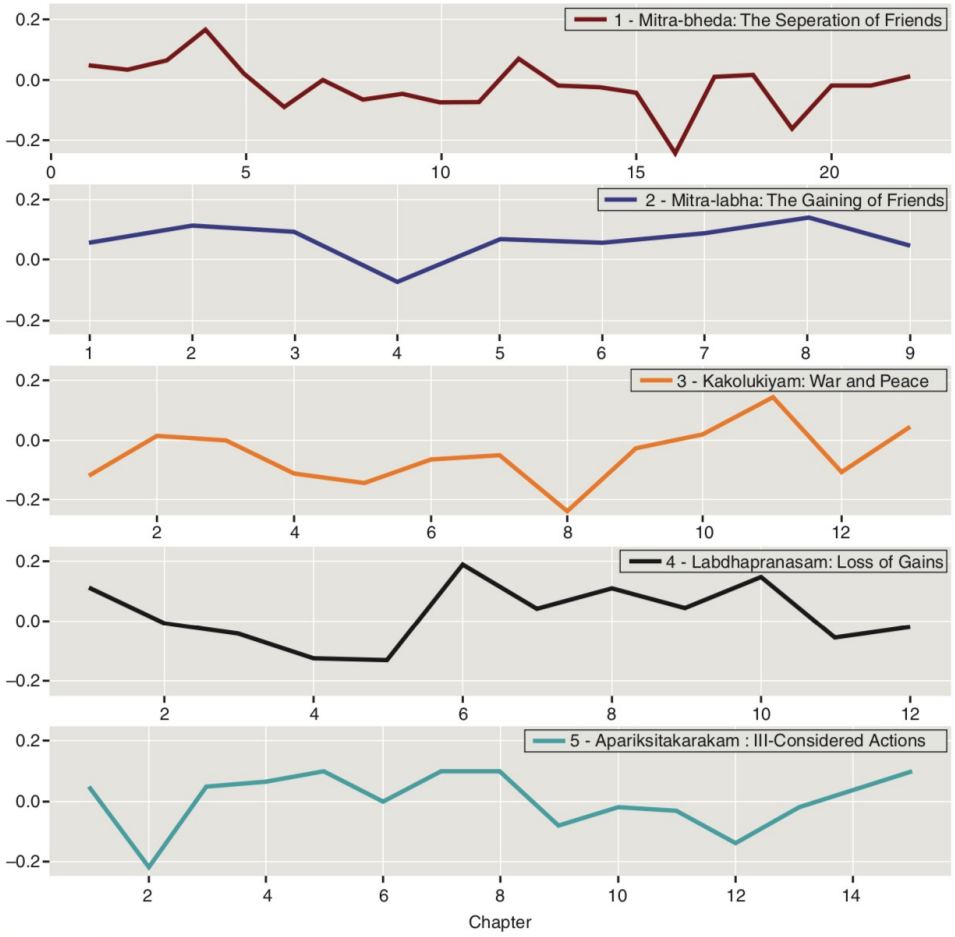


Figure 4.17 Line chart- compound sentiment score in each Panchatantra book.

- Strategy 4 has six chapters with positive sentiment scores (1, 6–10), five chapters with negative sentiment score (2–5, 11, 12) and one chapter with neutral sentiment score.
- Similarly, Strategy 5 has eight chapters with positive sentiment score (1, 2–8, 14, 15) and six chapters with negative sentiment score (2, 9–13) and chapter 6 having neutral sentiment score.
- This implies that across all the chapters the strategies are having extremely slightly positive sentiment.

Query: What are the sentiments throughout the book and for each strategy?

- Table 4.4 shows the calculated sentiment scores for each strategy.

Table 4.4 Sentiment scores in each strategy

Strategy	Sentiment Score
Mitra-bheda (The Separation of Friends)	–0.018707
Mitra-labha (The Gaining of Friends)	0.069727
Kakolukiyam (War and Peace)	–0.040799
Labdhapranasam (Loss of Gains)	0.013179
Apariksitakarakam (Ill-Considered Actions)	0.015234

Joint Plot

- TextBlob library is used to perform sentiment analysis.
- In sentiment analysis, a sentiment is defined by its semantic orientation and intensity of each word in the sentence.
- The output of a sentence in TextBlob generates two scores, polarity, and subjectivity.
- Polarity score can take any value ranges from −1.0 to +1.0, where +1 indicates a very positive sentiment, 0 indicates neutral sentiment, and −1.0 indicates a highly negative sentiment.
- Subjectivity score ranges from 0.0 to 1.0 where 0.0 is considered as highly objective meaning the sentences are factual, while 1.0 indicates highly subjective meaning the sentences express various feelings like opinions, allegations, suspicion, beliefs, etc.

Joint Plot

	Polarity	Subjectivity
count	3493.000000	3493.000000
mean	0.050813	0.260003
std	0.262651	0.328517
min	−1.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.053247	0.500000
max	1.000000	1.000000

Figure 4.18 Statistic of the Panchatantra text.

- The dataset contains 3,493 sentences in total.
- Figure 4.18 provides the statistics of the dataset.
- A Polarity score mean of 0.0508 is considered a slight positive sentiment.
- Subjectivity scores mean 0.260 indicates that the data is slightly subjective.
- Let us look at how we can draw more insights from this data.

Query: How does sentiment vary throughout the Panchatantra book?

- **Visualize:** Here, we would like to find if there is any relationship between polarity and index (sequential number of all the sentences).
- We can create a joint plot to uncover the pattern (Refer to Fig. 4.19).
- A joint plot comprises of three plots:
    1. The first plot is a bivariate graph which shows the relationship between two variables.
    2. The second plot shows the distribution of the variable defined on the horizontal axis placed horizontally at the top of the bivariate graph.
    3. On the right margin of the bivariate graph, the third plot shows the distributions of variables along the vertical axis (y-axis) with the orientation set to vertical.

Query: How does sentiment vary throughout the Panchatantra book?

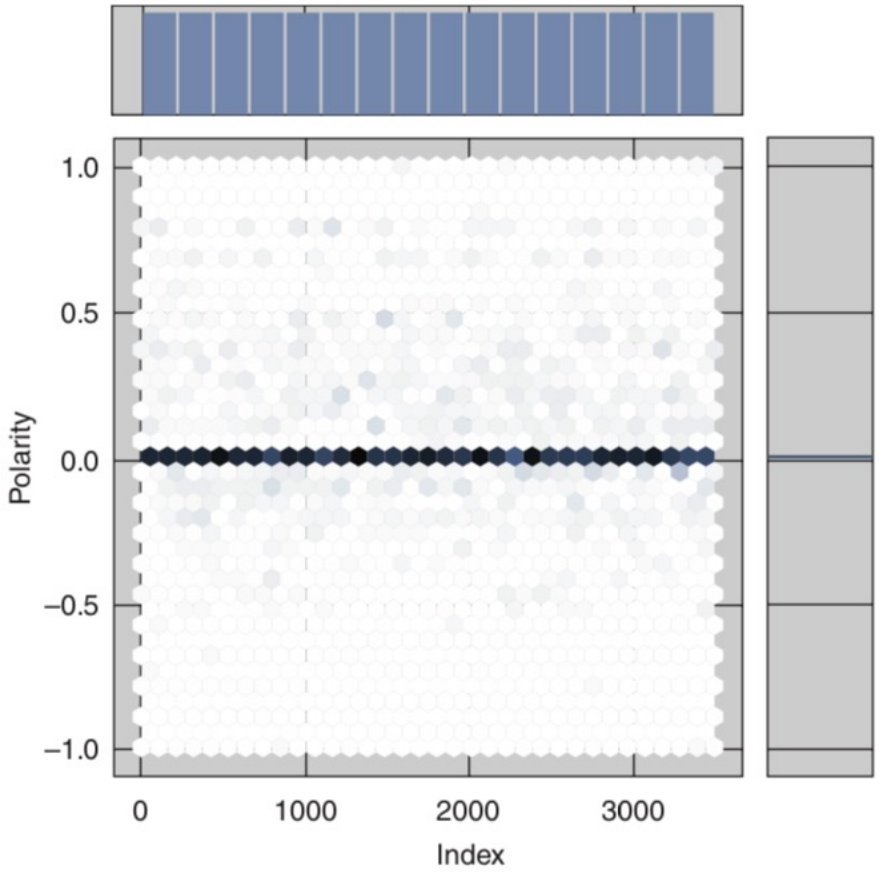


Figure 4.19 Joint plot- Index vs Polarity.

- **Inference:** Figure 4.19 shows that most of the sentences have been classified as neutral.
- This is evident from the joint plot.

Query: How does the sentiment vary once we filter out the sentences having non-zero polarity?

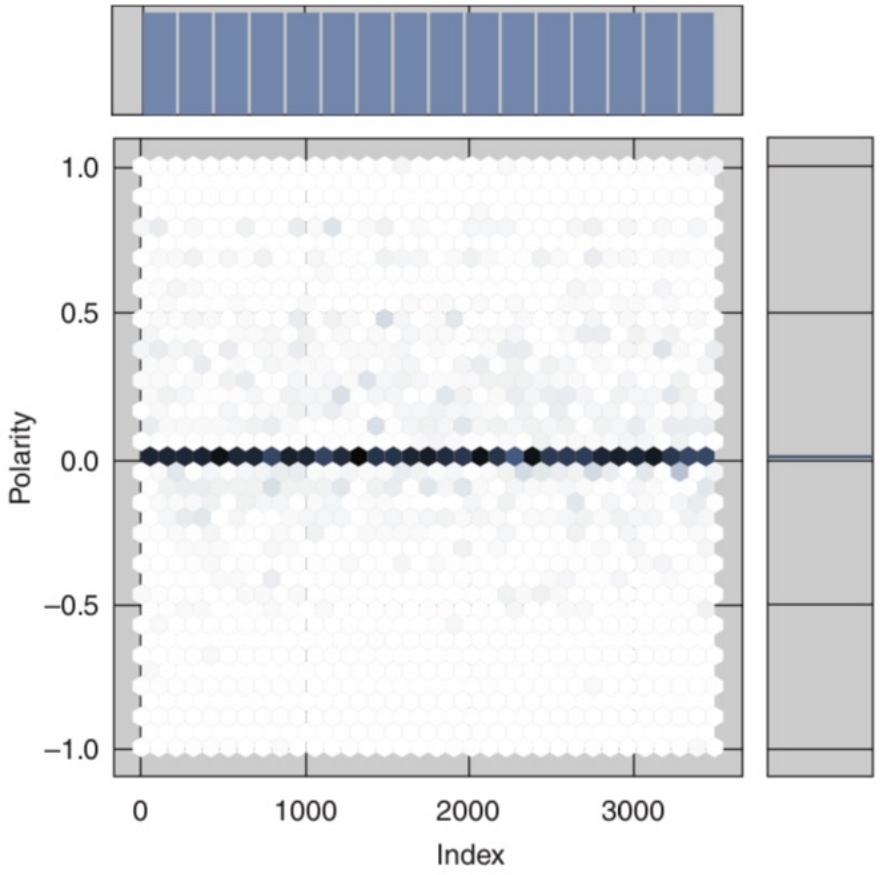


Figure 4.19 Joint plot- Index vs Polarity.

- **Visualize:** Refer to Fig. 4.20.
- **Inference:** At vertical axis (*y*-axis) histogram, we can see a bimodal distribution of polarity.
- In the bivariate graph we can see sentences numbered between 0 to 1000 having slight negative polarity (hexagon having darker shade of blue colour).
- While the sentences numbered greater than 1000 are having majorly positive polarity.

Query: How does the sentiment vary once we filter out the sentences having non-zero polarity?

Print 900th sentence

```
text.sentences[900]
```

```
Sentence("Karataka then told Damanaka, "You have done this foul deed because you were jealous of the king's friendshi
p with Sanjeevaka.")
```

Figure 4.21 900th sentence of the Panchatantra text.

- If we check the data and see sentences where these spikes occurred (Refer to Fig. 4.21).
- We can see in the 900th sentence (Fig. 4.21), strong negative emotions such as “foul deed” and “jealous” have been mentioned, which explains a negative sentiment spike.

Query: How is the relationship between subjectivity and index?

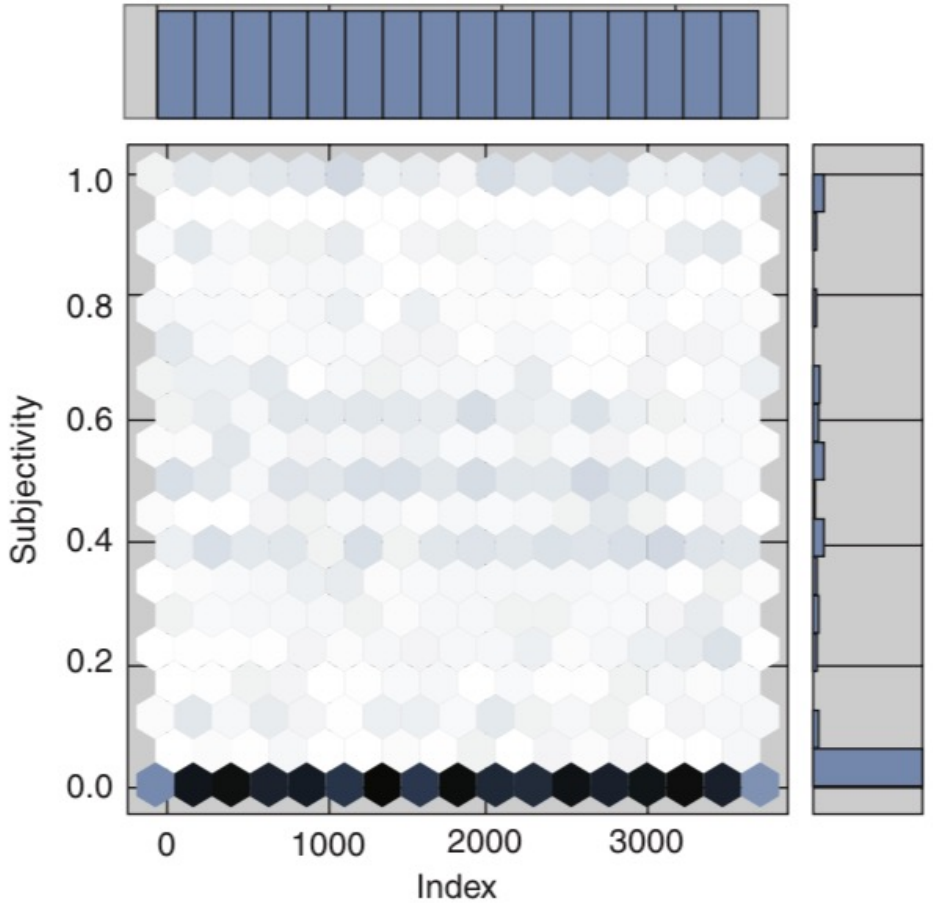


Figure 4.22 Joint plot- Index vs. Subjectivity.

- **Visualize:** Refer to Fig. 4.22.
- **Inference:** The plot reveals that the subjectivity score of 0.0 is dominant.

Query: Let us check for sentences having non-zero subjectivity score.

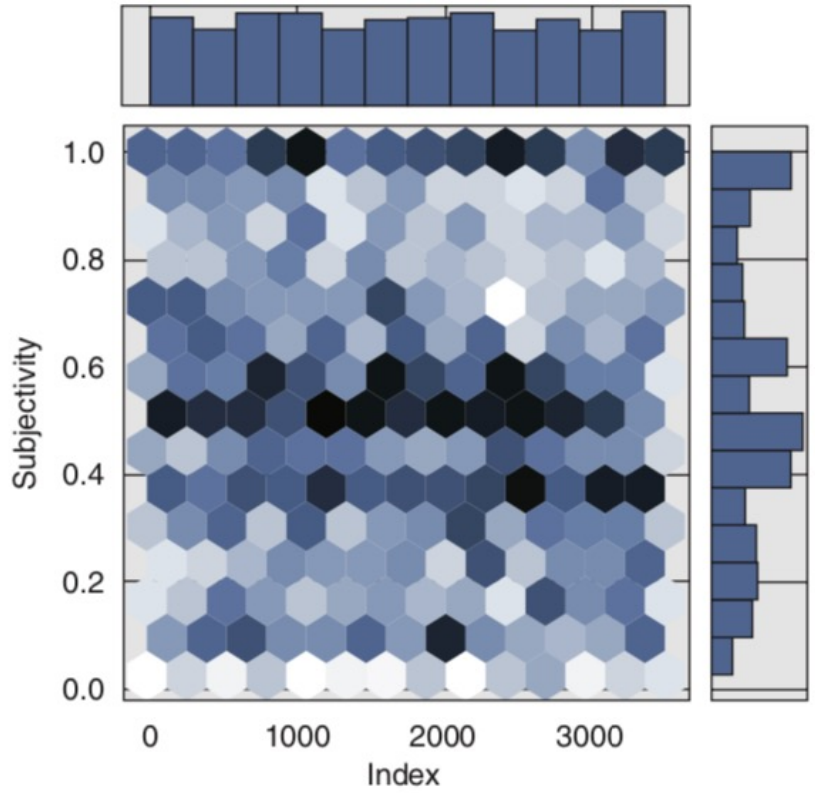


Figure 4.24 Joint plot- Subjectivity vs Index (Subjectivity not equal to 0).

- **Visualize:** Refer to Fig. 4.24.
- **Inference:** The subjectivity score 0.5 is maintained throughout the book, most prominently for the sentences numbered between 1000 to 3000, whereas the subjectivity score around 0.4 and 1.0 can be seen in the later part of the book (for the sentences numbered greater than 2500).
- Sentences numbered around 1000 and 2500 are shown to have high subjectivity i.e., subjectivity score 1.0.

Query: What can be inferred about subjectivity and polarity distribution?

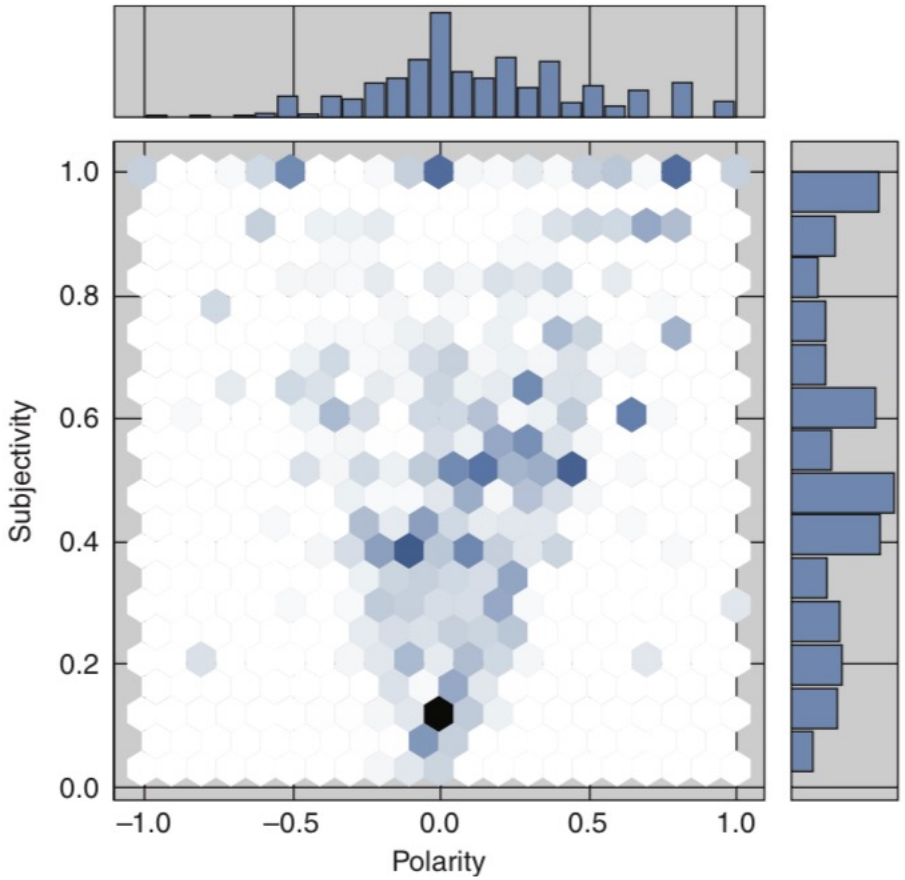


Figure 4.25 Joint plot- Subjectivity vs. Polarity.

- **Visualize:** Create a joint plot between subjectivity and polarity (Refer to Fig. 4.25).
- **Inference:** From the plot, we can see a funnel like shape.
- We can infer that as the subjectivity of the sentence increases, the more polarized it is.

Histogram Plot:

Query: How is the subjectivity score distributed throughout the Panchatantra text?

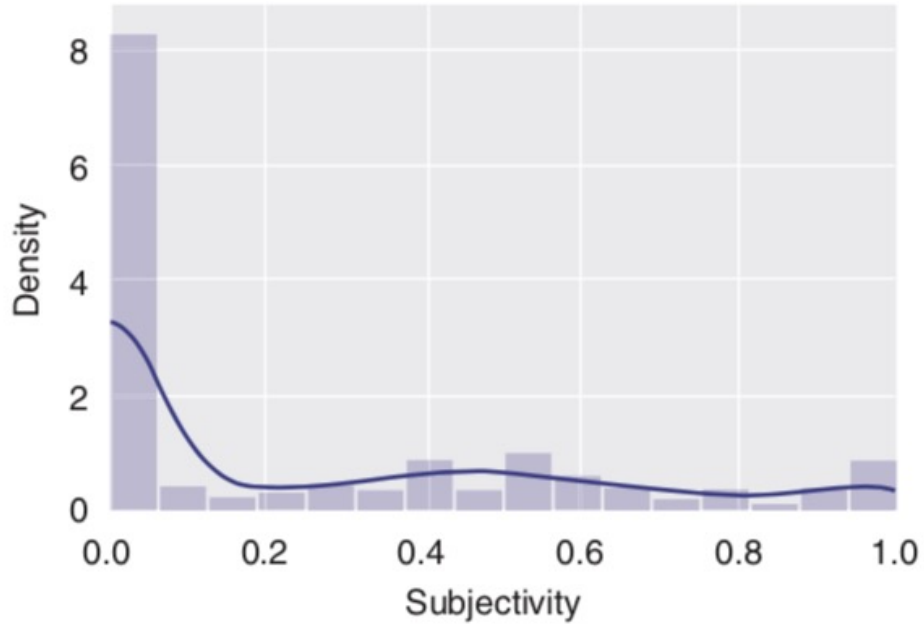


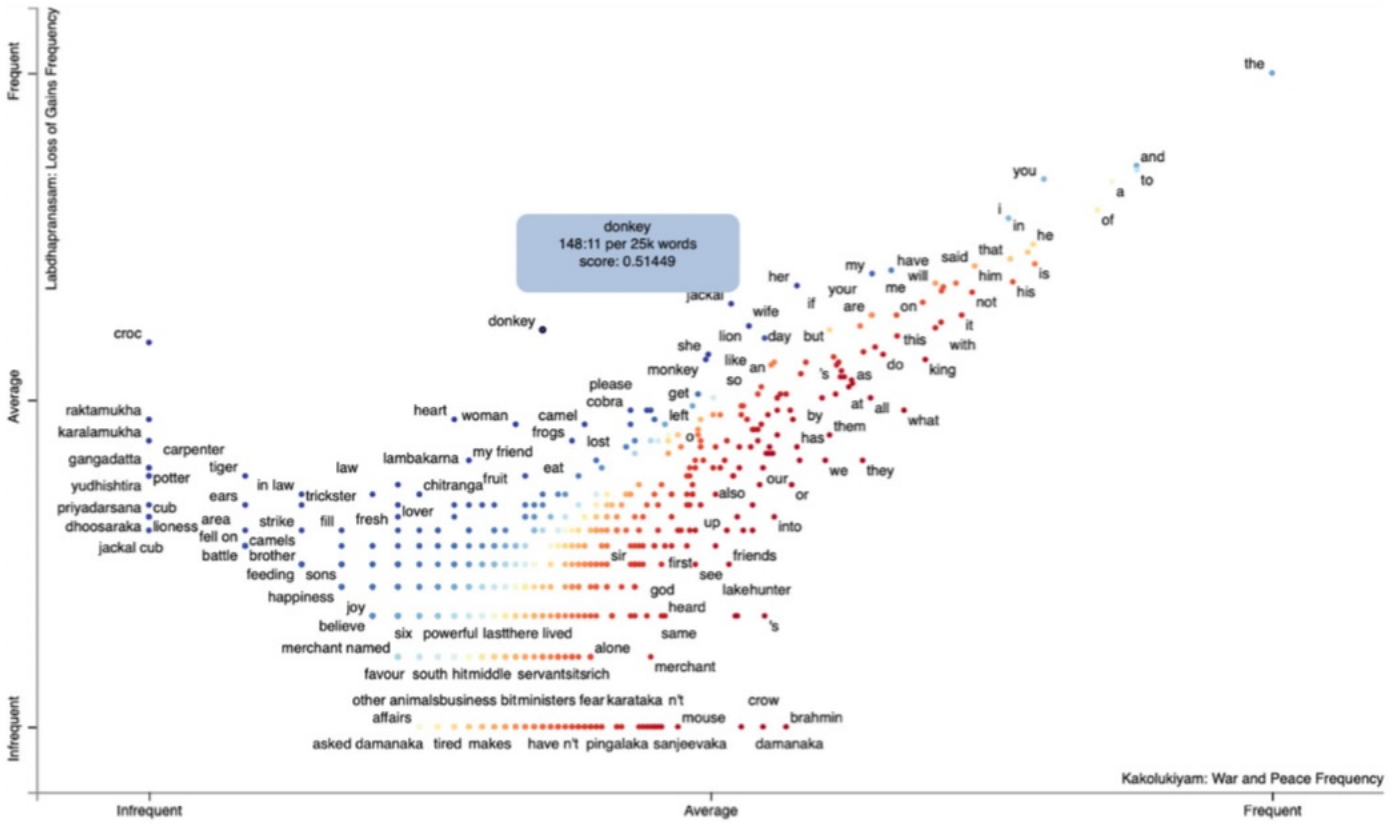
Figure 4.23 Compound sentiment score in each Panchatantra text.

- **Visualize:** We can create a histogram to check the distribution of subjectivity scores. (Refer to Fig. 4.23)
- **Inference:** The subjectivity score of 0.0 is clearly dominating.

Scattertext Visualization

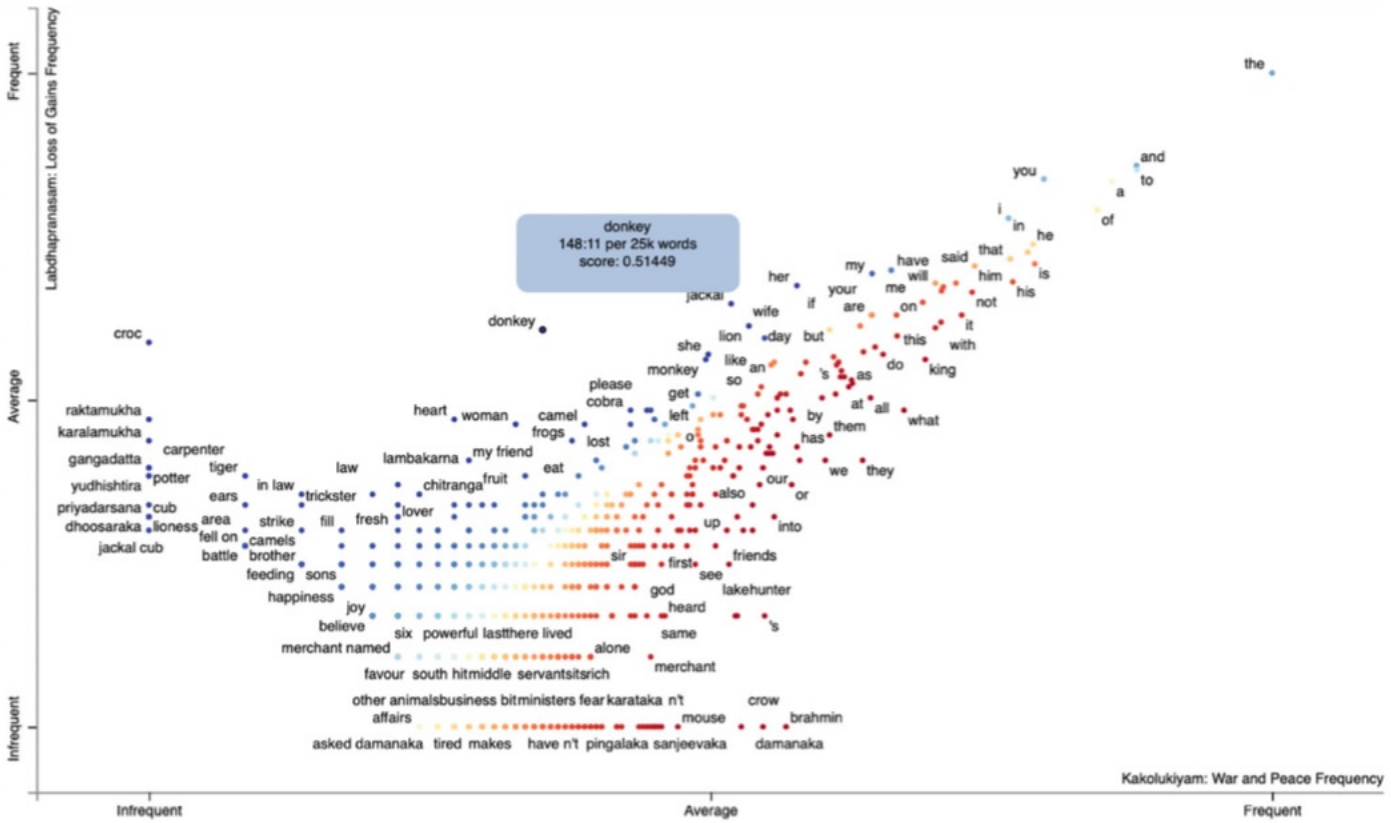
- Scattertext is a python package that shows us how two documents or two categories in a text are similar or different.
- Scattertext allows us to visualize which words or phrases are more characteristic of a particular category or document.
- It focuses on visualizing how words are used across documents/categories.
- Scattertext visualization is interactive.
- We can click on a dot to get statistics about a word’s relative use, as well as excerpts from the strategy where that word appears.

Query: How can we compare the word frequency in any two strategies of Panchatantra text.



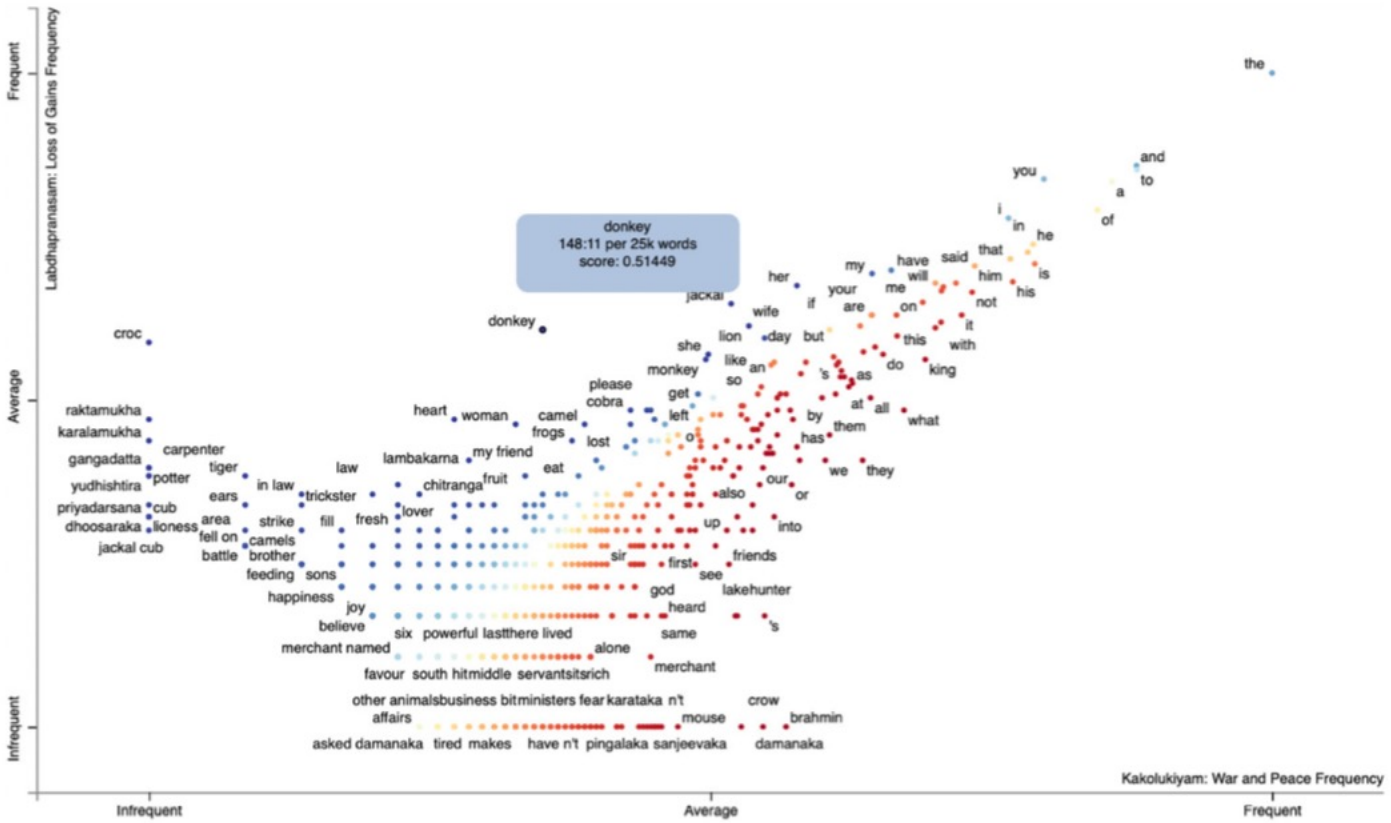
- **Visualize:** We use Scattertext to compare two parts (Kakolukiyam- War and Peace, and Labdhapranasam- Loss of Gains) of Panchatantra strategies.
- **a.** The horizontal axis (*x*-axis) indicates word frequency in strategy “War and Peace” and vertical axis (*y*-axis) indicates word frequency in strategy “Loss of Gains”.

Query: How can we compare the word frequency in any two strategies of Panchatantra text.



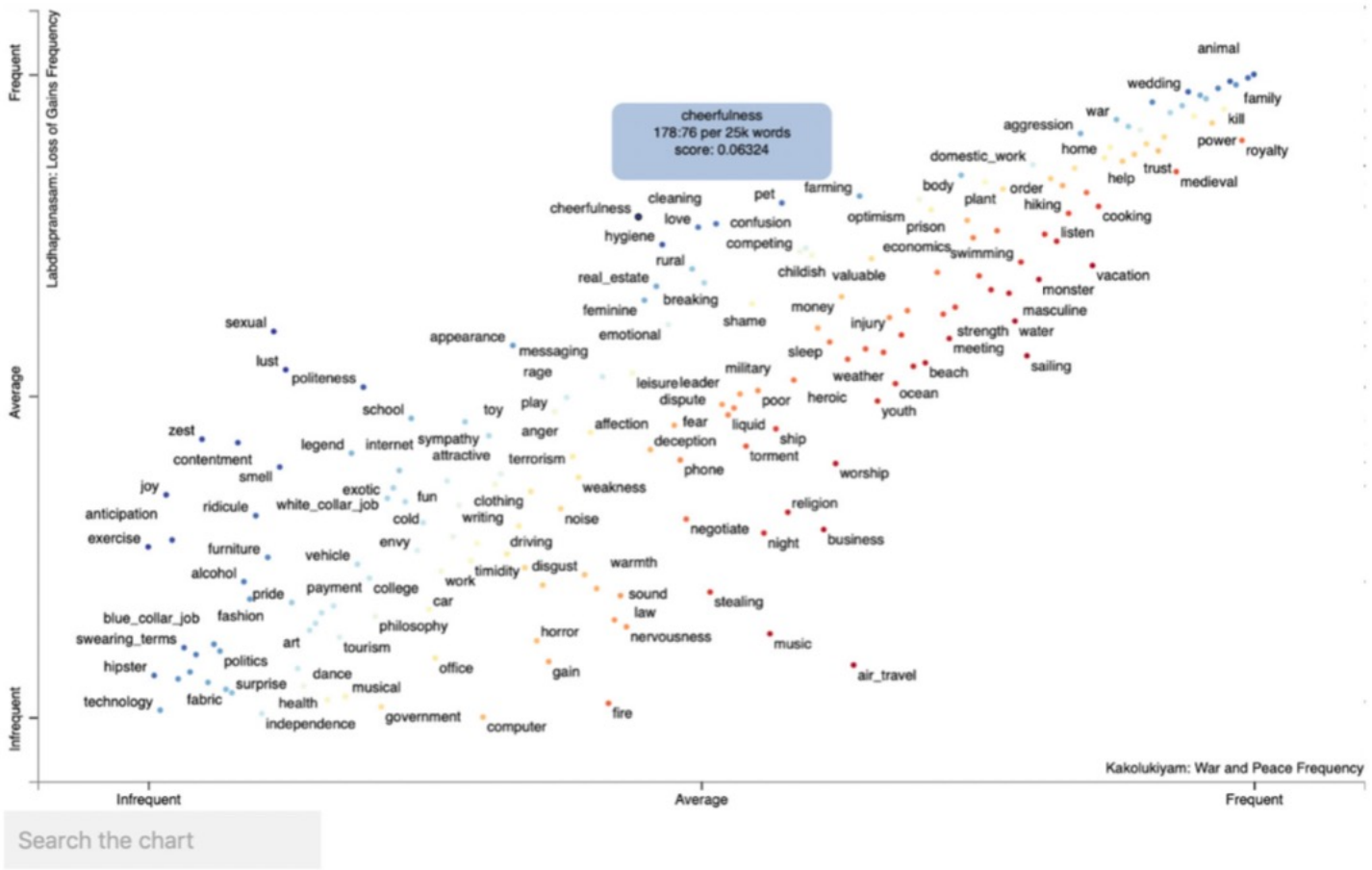
- If a term is on the top of the vertical axis, the greater is the number of times it is used in strategy “Loss of Gains”.
- If a term is further right on the horizontal axis, the more it is used in strategy “War and Peace”.
- The red coloured dot shows a stronger association with the horizontal axis while blue-coloured dots show stronger association with the vertical axis.

Query: How can we compare the word frequency in any two strategies of Panchatantra text.



- **Inference:**
- In strategy Kakolukiyam: War and Peace, words like “crow”, “mouse”, “brahmin”, “damanaka” are used, while strategy Labdhapranasam: Loss of Gains uses words such as “croc”, “donkey”, “raktamukha”, “jackal”, etc.
- It’s interesting to note that there are not many similar words between both strategies as the top right corner of the visualization shows only most commonly used words (i.e., stop words) like “the”, “and”, “to” and so on.

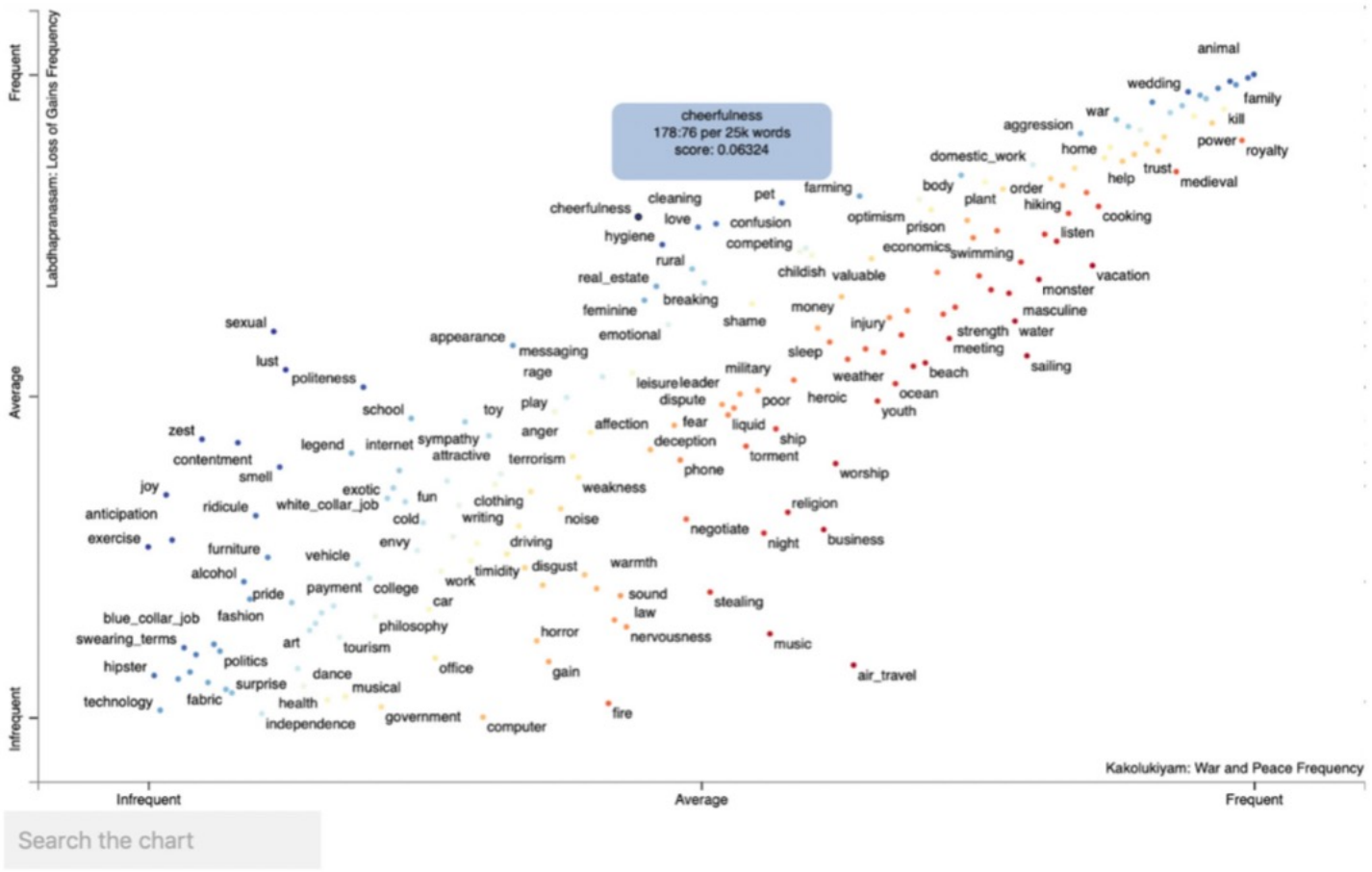
Query: How do we compare the sentiments in two strategies of Panchatantra text?



- **Visualize:** Empath[5] is a tool which analyses the text across topics.
- In addition, it can create new categories based on text. For example, “bleed” or “punch” words give category as “violence” [Chi, 2016].
- A comparison of Empath topics that are drawn from Strategy Kakolukiyam - War and Peace, and Labdhapranasam - Loss of Gains are shown in the Figure.
- The horizontal and vertical coordinates of a dot represent how many times that sentiment appears in the strategy.

[5]Source: http://empath.stanford.edu/

Query: How do we compare the sentiments in two strategies of Panchatantra text?



- **Inference:** Some insights we gather are:
- Top sentiments in Labdhapranasam: Loss of Gains strategy are “lust”, “cheerfulness”, “zest”, “anticipation” and “contentment”, which are more of positive sentiments.
- A few top sentiments in Kakolukiyam: War and Peace strategy are “trust”, “aggression”, “family” and “kill”, which indicates negative sentiment.

Visualizing Conversations

- We understand that analysing text data can be challenging but analysing conversations can be even more difficult.
- In conversations, people tend to use their own vocabulary or different ways of communicating their feelings.
- Hence, analysing these patterns sometimes needs another level of expertise and techniques beyond traditional ones.
- Another reason is that the whole conversation cannot be viewed at once; in different timeframes and with different people, emotions, and the tone change.
- Hence, it becomes easy to analyse if this information can be visualized.
- Visualizing a conversation can tell us how one person's emotions vary with time and the people with whom they are communicating.

Visualizing Conversations

- In this section, we will learn to visualize conversations based on the following dimensions:

1. Visualizing Timeline (When)
2. Visualizing People (Who)
3. Visualizing Information Flow (Who * When)
4. Visualizing Network (Who * Who)
5. Visualizing Content (What)

Enron Email Dataset

- We will try to build some machine learning models and understand the model diagnostics through various types of interactive visualizations.
- We took the **Enron Email dataset**.[6]
- It is one of the largest public datasets available, which consists of 150 users and about 0.5 million messages, mostly from the senior management of Enron.
- Enron Corporation was an American energy, commodities and service company based in Houston, Texas.

[6]Source: https://www.cs.cmu.edu/~enron/

Enron Email Dataset

- Fortune magazine named Enron as “America’s most innovative company” for six consecutive years, from 1996 to 2001.
- Enron Corp., which used to be referred to as “Wall Street Darling” filed for bankruptcy in December 2001.
- Its share prices dropped from US$ 90.56 to less than 30 cents in a few months.
- It is till date the largest bankruptcy in the history of the United States.

Enron Email Dataset- Description

	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...
2	allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...
4	allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...

Figure 4.31 Enron email data before cleaning.

- Initially, the dataset consists of two columns – file and message.
- Figures 4.31 show the top five rows of dataset before pre-processing the raw data.
- Pre-processing of the data includes separating text from email, splitting email addresses, extracting information such as from whom the emails have been sent, who received it, username, date, message-ID and so on (emails.csv).

Enron Email Dataset- Description

```
data_poi.head()
```

Message-ID	Date	From	To	Subject	X-From	X-To	X-cc	X-bcc	X-Folder	X-Origin	X-FileName	content	user
<33030902.1075854478703.JavaMail.evans@thyme>	2000-12...	(david...	(brian...	Re: ME...	David ...	Janet ...			\David...	Delain...	ddelai...	Guys, ...	delain...
<12969872.1075854479049.JavaMail.evans@thyme>	2000-12...	(david...	(greg....		David ...	Greg W...	Mark E...		\David...	Delain...	ddelai...	guys, ...	delain...
<30501751.1075854481441.JavaMail.evans@thyme>	2000-11...	(david...	(mike....	Wind Dash	David ...	Mike C...			\David...	Delain...	ddelai...	Mike, ...	delain...
<31880823.1075854481463.JavaMail.evans@thyme>	2000-11...	(kay.c...	(dlhar...	Re: Co...	Kay Ch...	dlhart...			\David...	Delain...	ddelai...	David ...	delain...
<23822759.1075854481485.JavaMail.evans@thyme>	2000-11...	(david...	(paul....	Insura...	David ...	Paul C...			\David...	Delain...	ddelai...	Paul, ...	delain...

Figure 4.32 Enron email data after cleaning.

- After pre-processing, the final dataset consists of 14 columns, each depicting different information from the original message body of the email such as Date, From, to, subjects, user, content and so on.

Enron Email Dataset- Description

```
print(data['message'][450000])
```

```
Message-ID: <21899142.1075862394894.JavaMail.evans@thyme>
Date: Mon, 19 Nov 2001 13:34:59 -0800 (PST)
From: no.address@enron.com
Subject: Holiday Party - Canceled
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Ken Lay- Chairman of the Board & CEO@ENRON
X-To: All Enron Houston@ENRON <??SAll Enron Houston@ENRON>
X-cc:
X-bcc:
X-Folder: \TSTAAB (Non-Privileged)\Staab, Theresa\Inbox
X-Origin: Staab-T
X-FileName: TSTAAB (Non-Privileged).pst


I know that this is a difficult time for all of us.  With everything going on inside the company as well as in the wo
rld around us, we have been carefully considering whether a holiday celebration is appropriate this year.  To be hone
st, employee feedback has been mixed.  Many viewed the holiday party as a unique opportunity for us to come together
as Enron employees to share the spirit of the season.  Others felt a holiday party would be improper given the compan
y's current circumstances.

After weighing these points of view, we have ultimately decided to cancel the all-Enron holiday party that was schedu
led for December 8.  Given what has transpired over the past month, it could be considered imprudent for Enron to inc
ur the expense of such an event.  I regret that this action is necessary because I recognize that your hard work thro
ughout the year merits a holiday celebration and so much more.  We will attempt to find other, more appropriate ways
to recognize your outstanding contributions as we move into the holiday season.

Ken Lay
```

Figure 4.33 A raw email from Enron email corpus.

- Figure 4.33 provides a snapshot of one email in raw data.
- This email was sent by former Enron corporation founder and CEO Kenneth Lay on 19th November 2001, a month before the company declared bankruptcy.
- The company was in a crisis, and everyone was aware of it by then.

Enron Email Dataset- Description

- The complete Enron email dataset is huge, consisting of about 500,000 messages.
- Using the full dataset would be time consuming.
- Hence, we will take a smaller subset of this dataset for further analysis.
- In order to make the Enron email dataset more interesting, instead of choosing a random sample, we will consider five Persons of Interest and analyse their emails for the reminder of the analysis.
- “Persons of Interest” (POI) are all the people who were either charged with a crime or people who settled without admitting guilt in the Enron scandal.

Enron Email Dataset- Description

- These persons of interest are chosen by reading various news articles, blogs, and case studies about the Enron scandal.
- The introduction of these POI's is described below:

1. Kenneth Lay – He was the founder and CEO of Enron Corp. He was convicted on six counts of fraud and conspiracy and four counts of bank fraud[7]

Username – "lay-k"

2. Jeffery Skilling – Former Enron CEO (hired by Kenneth Lay in 2000) and Chief Operating Officer. He was convicted of 19 counts against him, including fraud, conspiracy, and insider trading.

Username – "skilling-j"

[7]Source: https://www.justice.gov/archive/opa/pr/2006/May/06_crm_328. html

Enron Email Dataset- Description

3. David Delainey – Former CEO, Enron energy services. He manipulated earnings by participating in fraudulent schemes to please Wall Street. He was sentenced to two and a half years in prison for insider trading, which led to the Enron's financial collapse.

Username – "delainey-d"

4. John M. Forney – Former Enron Corporation trader. He was found guilty on one count of wiring fraud and one count of conspiracy. The FBI arrested him saying that in 2000 and 2001, when California's energy crisis happened, he was the key architect of the electricity trading schemes.

Username – "forney-j"

Enron Email Dataset- Description

5. James Derrick – Enron's General Counsel.

- Though the examiner did not find any grounds to attribute any intentional wrongdoings to Derrick, a few things about him make him an interesting person who needs to be analysed.
- He oversaw informing other executives what was and was not legal.[8] But the question went unanswered as to how much he knew about Enron's problems. Username – "derrick-j"

[8]Source: The case of Enron's Top Lawyer - https://www.bloomberg.com/ news/articles/2002-12-18/the-case-of-enrons-top-lawyer

Visualizing Timeline

Query: When did the maximum number of communications occur?

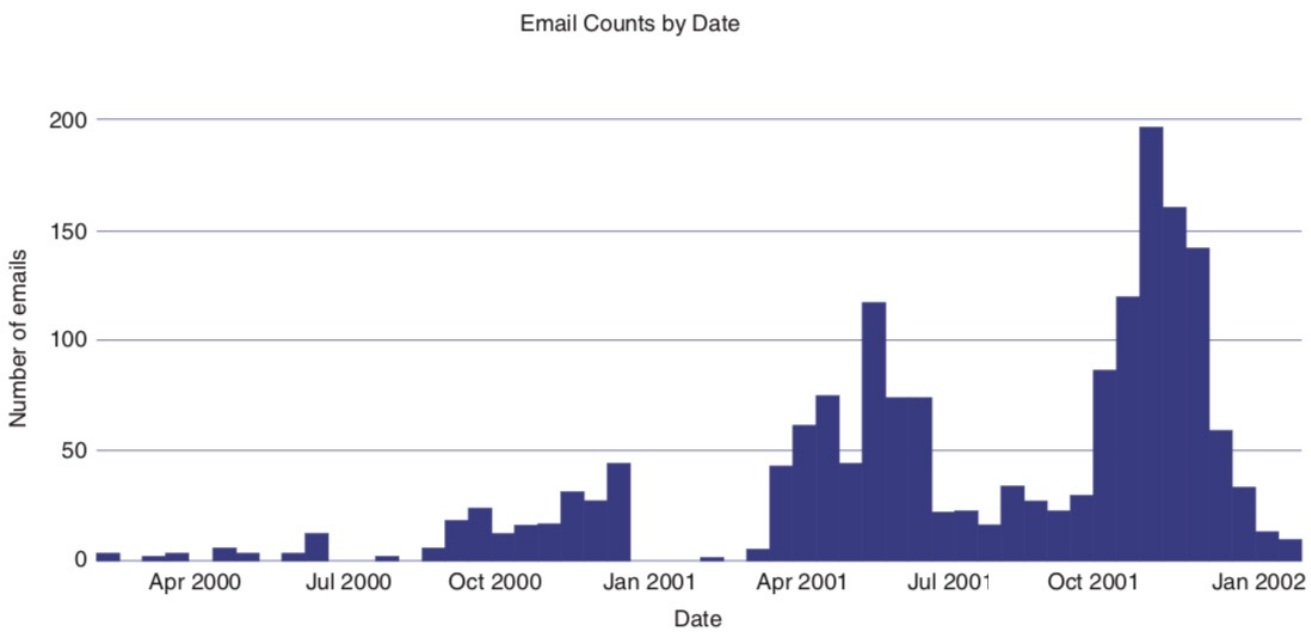


Figure 4.34 Email counts by date.

- **Visualize:** Refer to Fig. 4.34, histogram with time on horizontal axis.
- **Inference:** The plot shows how emails are spread across the timeline from 2000 to 2002.
- We can see that the maximum number of conversations happened during November–December 2001, i.e., the toughest year for Enron when it declared bankruptcy.

Query: What is the pattern of email counts of Enron employees over a month or week?

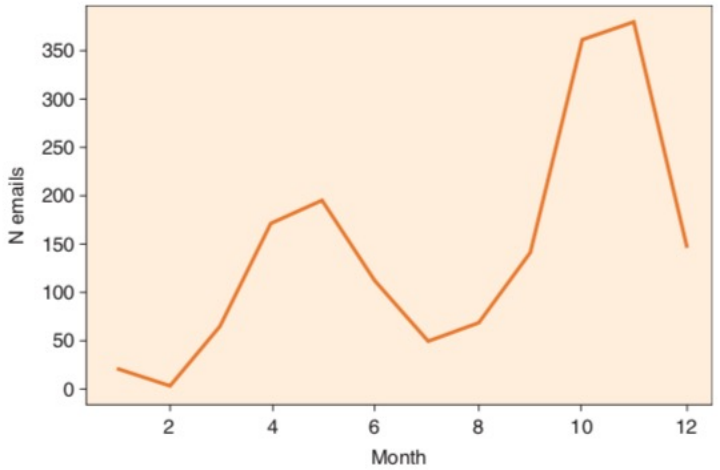


Figure 4.35 Email counts by month.

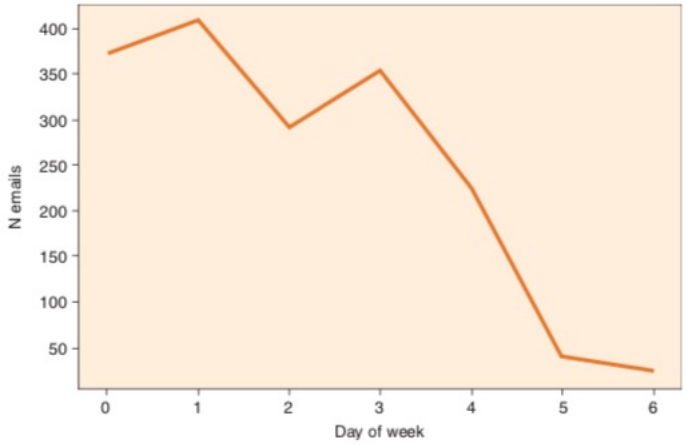


Figure 4.36 Email counts by day of week.

- **Visualize:** Refer to Figs. 4.35 and 4.36 (Line charts).
- **Inference:** We can infer that end of the year usually was the busiest time when most of the communications happened.
- If you look at the weekly pattern, the beginning of the week is the busiest period.

Visualizing People
Query: What is the proportion of emails among all five users?

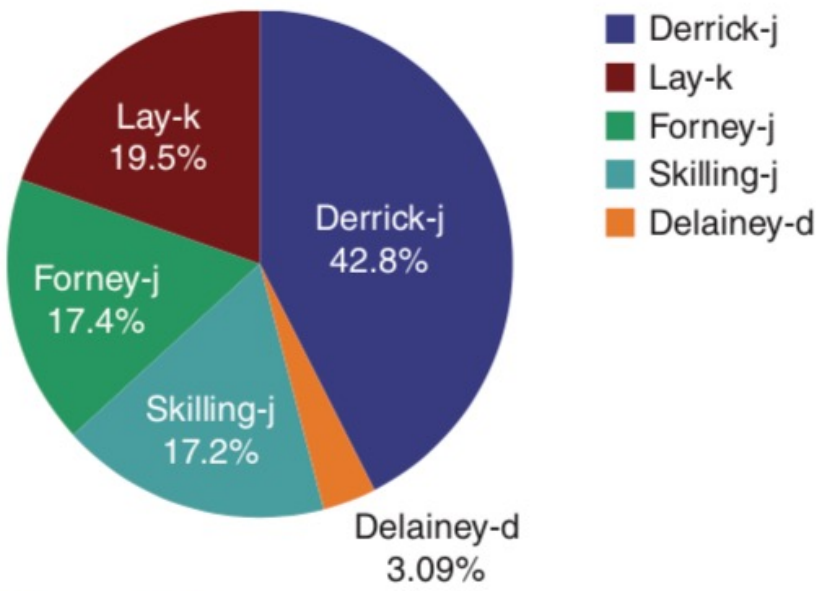


Figure 4.37 Proportion of emails by user (Pie chart).

- **Visualize:** We will create a simple pie chart or donut chart to check for proportions.
- Refer to Figs. 4.37.
- **Inference:** James Derrick, followed by Kenneth Lay, John Forney, Jeffery Skilling, and David Delainey, is the sequence of people with the highest proportion of emails.

Visualizing Information Flow

Query: What is the timeline of the emails that are sent or received by each user?

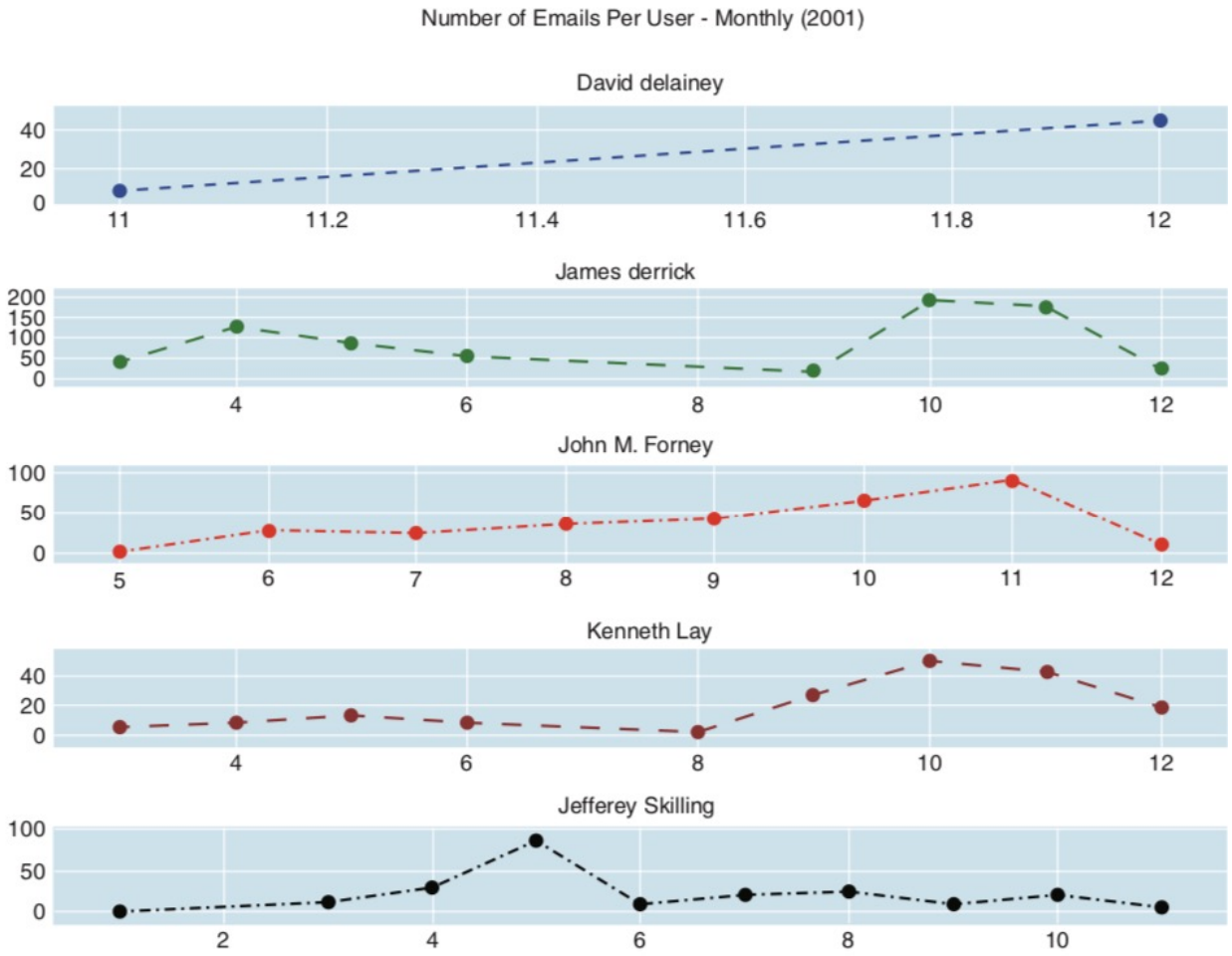


Figure 4.39 Email counts by month and user.

- **Visualize:** Refer to Fig. 4.39, line chart.
- **Inference:**
- Maximum number of communications happened in the later part of the year i.e., October–December 2001.
- This is understandable as Enron Corp was going through a crisis during that period.

Query: What is the timeline of the emails that are sent or received by each user?

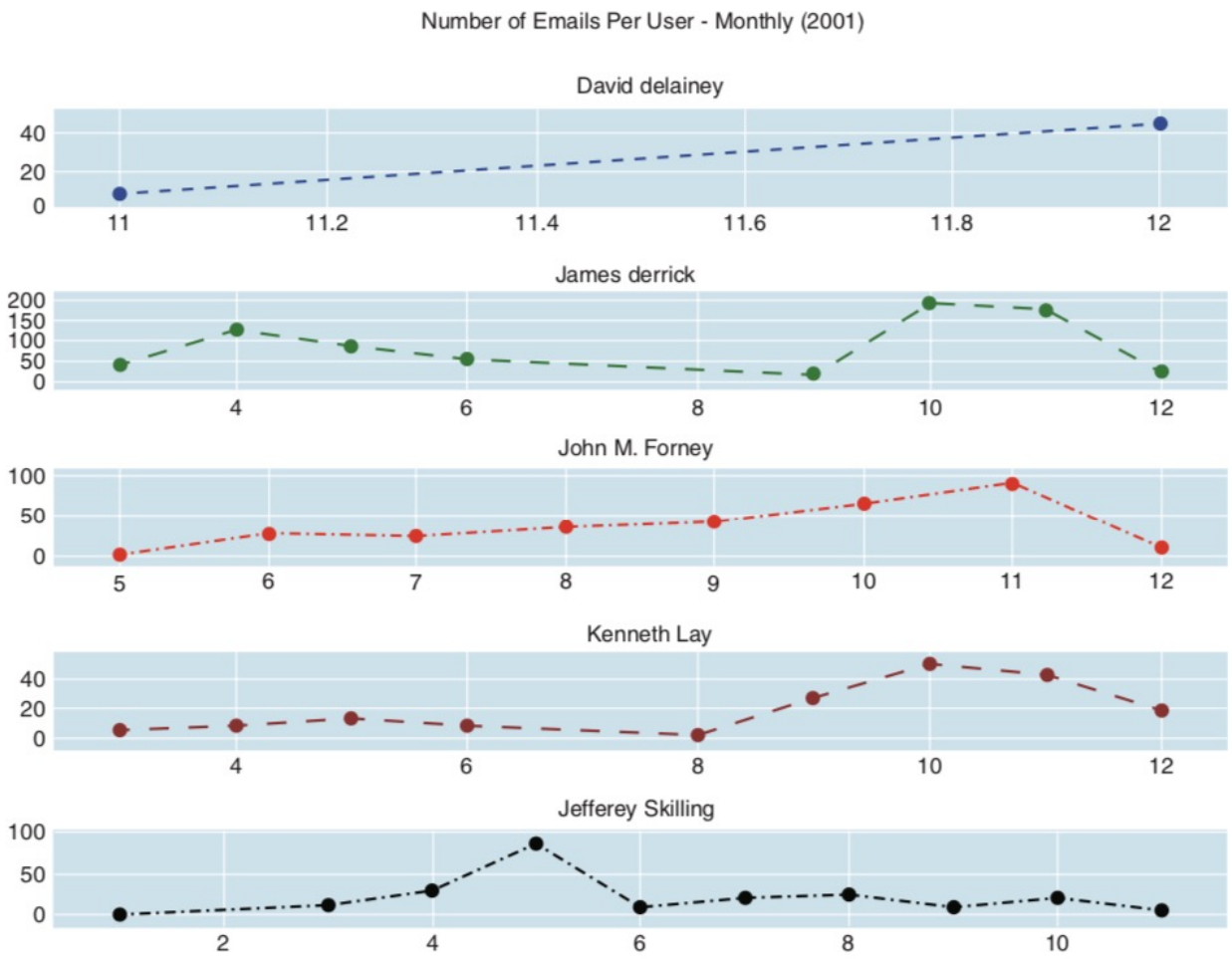


Figure 4.39 Email counts by month and user.

- Interestingly Jeffery Skilling was a consistent sender/ recipient till 2001, but there is an increase in communication involving him in May-2001.
- The reason being he resigned from Enron Corp. as CEO in August-2001, hence, the period before that must be a crucial time for him.
- He must be communicating more than usual during that period. This ensures the quality of the data.
- David Delainey was active only in the months November-2001 and December-2001.

Visualizing Data Hierarchy

Query: Who sent out or received the maximum number of emails, and when?

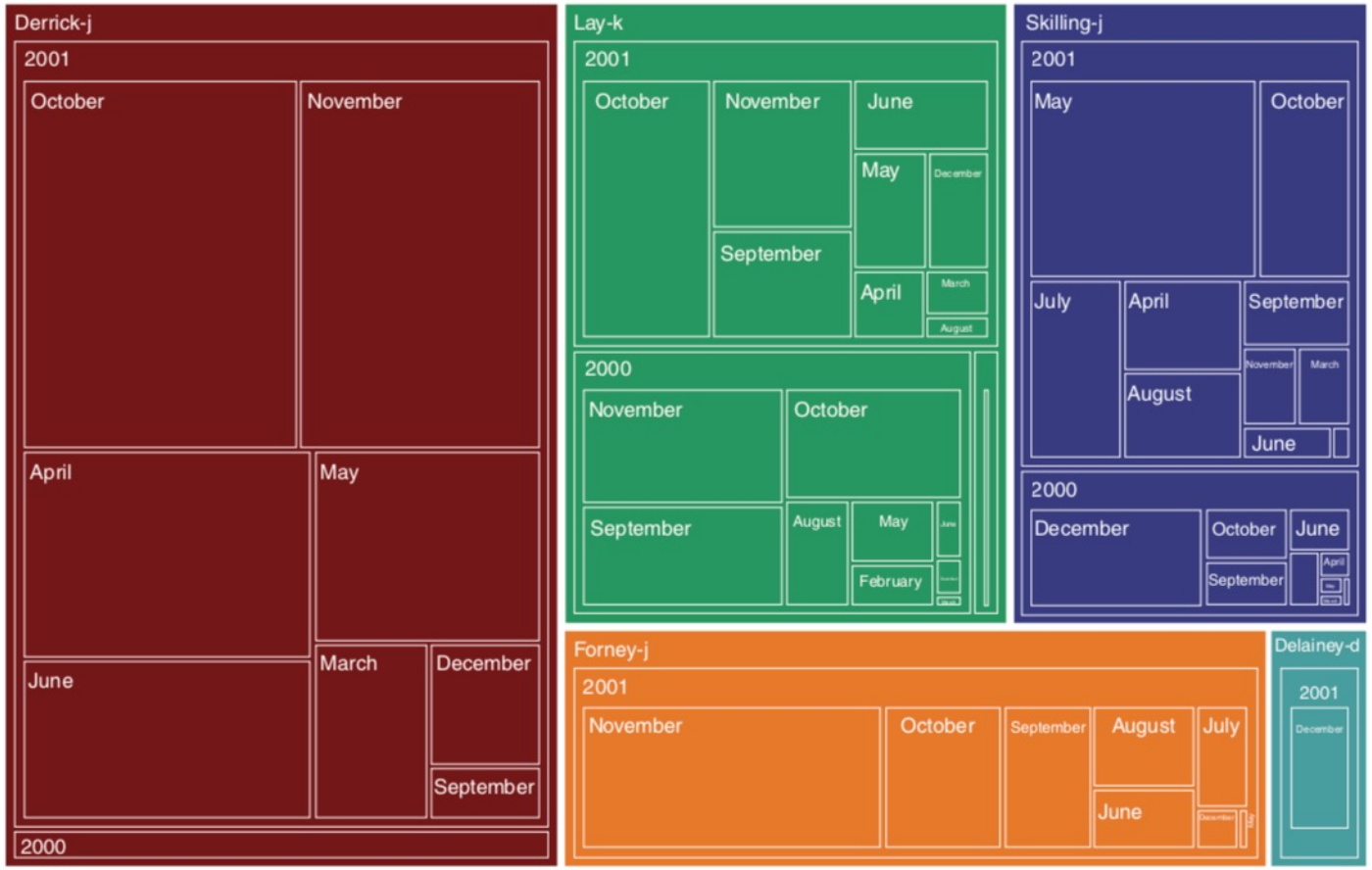


Figure 4.40 Tree map- email counts by user-year-month.

- **Visualize:**
- Tree maps are to visualize this data hierarchy.
- It helps get an overall sense of the hierarchical data as a recursive set of nested rectangles.
- The bigger the value of the parameter, the larger is the size of the rectangle.
- The size and colour of the rectangles vary based on the different categorical/numerical values in data.

Query: Who sent out or received the maximum number of emails, and when?

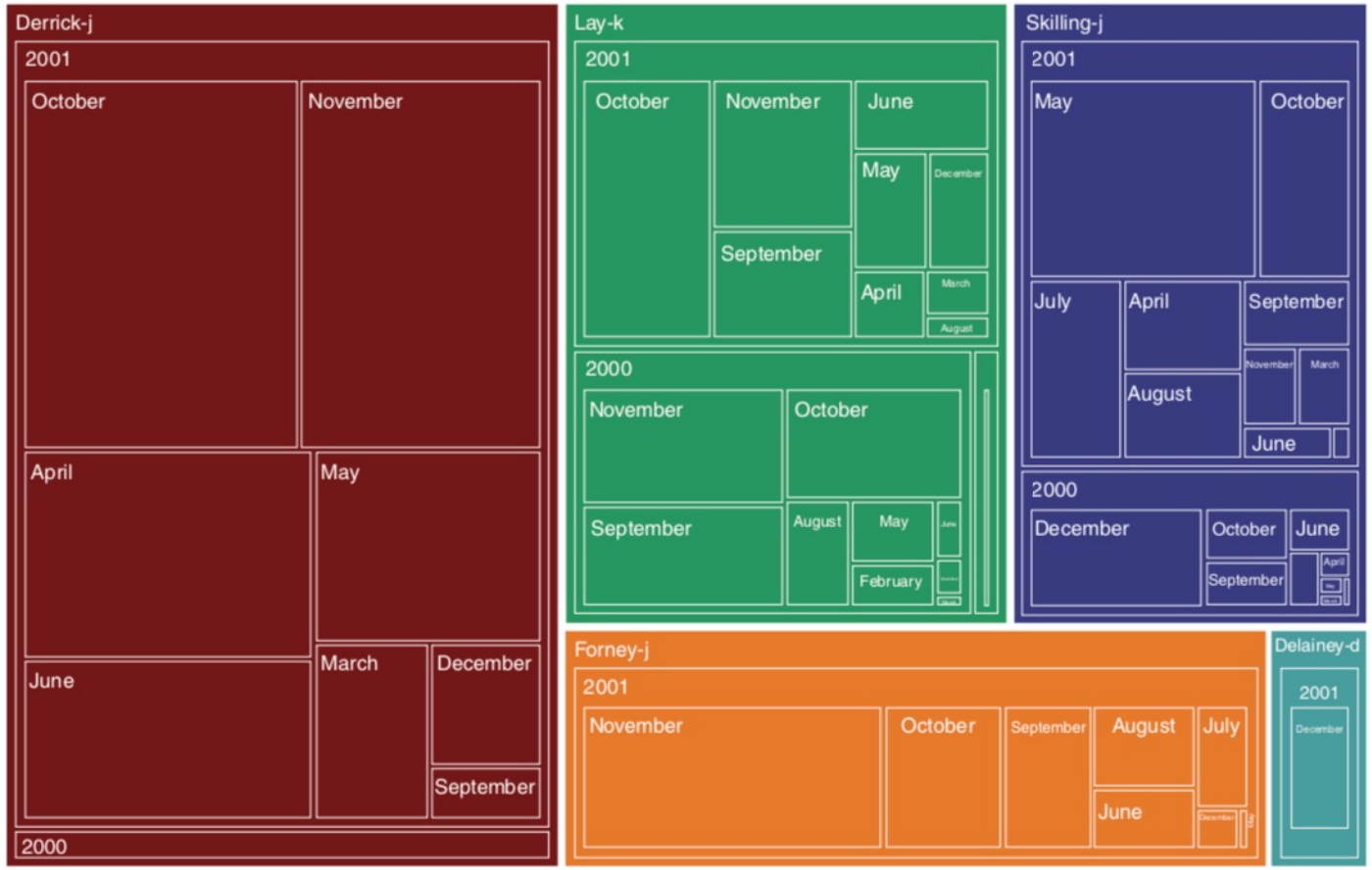


Figure 4.40 Tree map- email counts by user-year-month.

- The focus is to visualize the bigger and smaller components in the data and how they are categorized.
- Although they are not recommended when you want to make exact comparisons (Refer Fig. 4.40).

Query: Who sent out or received the maximum number of emails, and when?

- **Inference:**
- October, November, and December are the months when most of the conversations happened by each user.
- Another interesting observation is that Jeffery Skilling had the maximum number of conversations in May 2001.
- To explain this, the summer of 2001 was when Enron was starting to crumble.

Query: Who sent out or received the maximum number of emails, and when?

- According to Kevin Hannon, then chief operating officer of the company's flailing broadband unit, at a May 2001 meeting of Enron executives, Jeffery Skilling said, "They're on to us", which was an indication that the investment community was able to figure out how Enron was making money.[9]
- This must have caused panic and hence the increased number of emails from Jeffery Skilling in May 2001 as he tried to hide the various operations as well as financial losses of trading businesses using mark-to-market accounting.[10]

[9]Source: Alleged Skilling Quote Shocks Courtroom - https://www.cfo.com

[10]Source: U.S. Securities and Exchange Commission. "Complaint: Jeffrey K. Skilling, Richard A. Causey." Accessed Jan. 19, 2021.

Visualizing Networks

- Network graph is an important tool in data visualization which represents interconnected entities.
- A network is made up of two important components:
    1. **node-** which represents an entity
    2. **edge-** which represents the connection between any two nodes.
- For example, in the case of email data, if we want to visualize “who is communicating with whom?”,
- In network graph, users will be the nodes and if an email is sent out to one user from another, it creates an edge between them.
- The focus of a network graph visualization is in discovering the relationship between two entities in the form of a network.

Visualizing Networks

- NetworkX, Pyvis and visdcc are a few python packages to visualize network graphs.
- For a smaller chunk of data, NetworkX gives good net- work visualizations.
- But if we are dealing with a relatively larger dataset, Pyvis gives us a better output and provides manual interactions such as zooming in and out and dragging the nodes from one place to another for better visualization.
- For visualizing the network, we have filtered the emails of the five persons of interest (POIs) for the timeline from January 2001 to December 2001.
- It is further divided into two groups as described below:

1. **Group 1**- consists of emails for the duration January 2001 to June 2001.
2. **Group 2**- consists of emails for the duration July 2001 to December 2001.

Visualizing Networks – Why Year 2001?

- The year 2001 was chosen to analyse the network as it was the most critical year for Enron Corp.
- From listing as ‘the most innovative company’ to going down to bankruptcy at the end of year, many events happened in 2001.[11]
- Enron was crowned ‘the most innovative company’ by *Fortune* for six consecutive years: 1996–2001 till 2001.16
- In February 2001, Jeffrey Skilling was appointed Chief Executive Officer of Enron, replacing Kenneth Lay.
- In August 2001, Jeffrey Skilling resigned from the post stating personal reasons and Kenneth Lay was again appointed CEO. He said there were “no issues” with Jeffrey Skilling’s resignation.

[11]Source: Timeline: Enron - https://www.theguardian.com/business/2006/ jan/30/corporatefraud.enron

Visualizing Networks - Why Year 2001?

- In October 2001, a White House team concluded that a collapse of Enron would pose little risk to the US economy.
- In October 2001, Enron reported a $618m loss, its first quarterly loss, and disclosed a $1.2bn reduction in shareholder equity.
- In October 2001, Enron shares lost a fifth of their value.
- In November 2001, Enron agreed to be acquired by Dynegy for
- $9bn. This offer was later cut by Dynegy within a month.
- In December 2001, Enron filed for bankruptcy, the biggest in US history.

Visualizing Networks

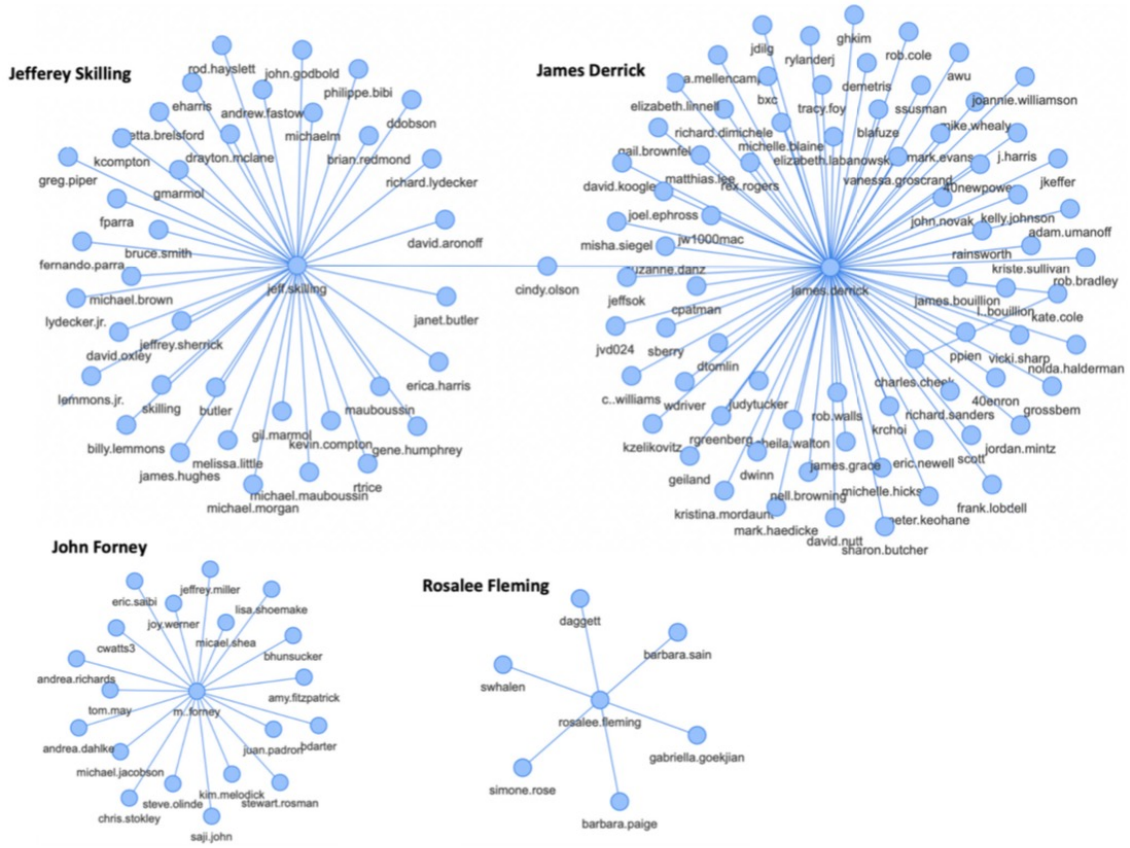


Figure 4.41 Network graph- [Jan 2001–June 2001].

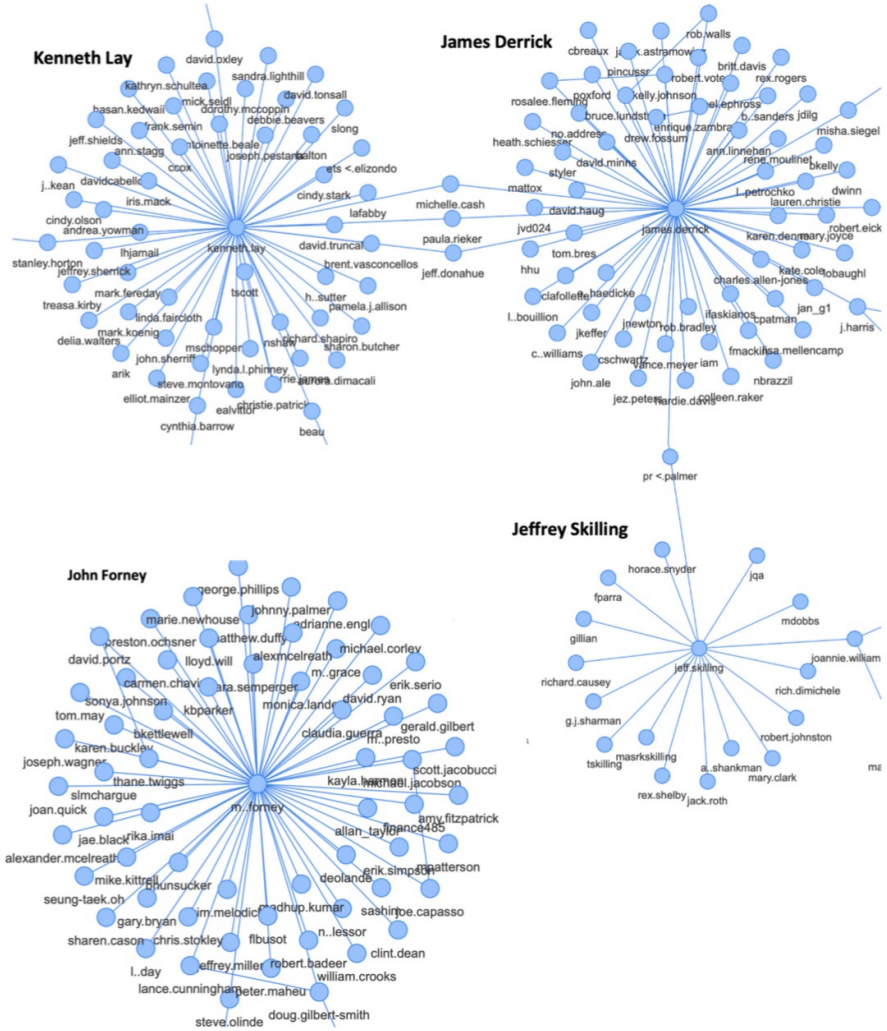


Figure 4.42 Network graph- [July 2001–December 2001].

Visualizing Networks

- **Visualize:** Figure 4.41 is the network graph of the five POIs for the timeline January 2001 to June 2001.
- The network graph has four major clusters corresponding to Jefferey Skilling, John Forney, James Derrick and Rosalee Fleming.
- Figure 4.42 is the network graph of the five POIs for the timeline July 2001 to December 2001.
- The network graph has four major clusters corresponding to Jefferey Skilling, John Forney, James Derrick and Kenneth Lay.

Visualizing Networks

- **Inference:**
- In the first half of the year (Fig. 4.41), Kenneth Lay did not have many conversations, but in the second half of the year (Fig. 4.42), his email traffic increased a lot as we can see a cluster which corresponds to Kenneth Lay's emails.
- This justifies that after Jeffery Skilling resigned in August 2001 and Kenneth Lay was again appointed CEO, the email traffic shifted to Kenneth Lay.
- There are very few emails of Jeffery Skilling in the second half of the year (Fig. 4.42), which indicates his absence in the second half due to his resignation.
- As expected, there is no cluster corresponds to David Delainey, as he has very few email conversations in year 2001, hence it did not contribute much to the network graph.

Visualizing Networks

- **Inference:**
- Interestingly, in Fig.4.41,we do not see any cluster corresponding to Kenneth Lay, rather a small cluster is shown corresponds to “Rosalee Fleming”.
- On further investigation we found out that she was Kenneth lay’s assistant, and it appears that at many instances she used to send out emails as ‘on behalf of Kenneth Lay’.[12]
- Overall, the email traffic increased a lot in the second half, which can be justified as Enron was going through a tough phase during that period.

[12]Source:https://www.salon.com/2003/10/14/enron_22/

Visualizing Networks

- **Inference:**
- Another interesting thing to note is that while everyone else had their own niche the conversations between Kenneth Lay and James Derrick had increased in the second half of the year (Fig. 4.42).
- Our investigation of what might be happening at Enron around that time revealed that James Derrick played an important and crucial role in 2001.
- Sherron Watkins, a former Vice President at Enron, discovered in mid-2001 that Enron manipulates their financial statements among various rumours about the company's collapse.
- She informed Kenneth Lay about it and warned him that Enron "will implode in a wave of accounting scandals" and not to rely on Vinson & Elkins, citing a conflict of interest.[13]

[13]Source: https://money.cnn.com/2002/01/16/companies/enron_lawyers/index.htm

Visualizing Networks

- **Inference:**
- Kenneth Lay then turned to Enron's general counsel James Derrick.
- Both Kenneth Lay and James Derrick asked law firm Vinson & Elkins to investigate the matter despite being warned not to trust them.
- James Derrick was a former Vinson & Elkins partner and imposed two restraints on employees conducting the probe: do not second guess its accountants from Andersen and do not analyse every transaction, the paper reported.[14]
- This makes Derrick an interesting person in Enron's collapse as he was the connecting link between Enron and the law firm Vinson & Elkins, and he was in close contact with Enron's CEO Kenneth Lay during the investigation.
- It, however, remains unanswered as to how much he was aware of what was happening inside Enron.

[14]Source: https://money.cnn.com/2002/01/16/companies/enron_lawyers/index.htm

Visualizing Content
Word Embedding

- Word embedding is a method to represent text data in a numerical manner i.e., representing words as a vector.
- Word2vec is a neural network structure to generate word embeddings that can also measure a semantic relationship between words.
- In other words, the word2vec model is based on the **hypothesis** that words will have a similar semantic meaning if they occur in similar linguistic contexts.
- It maps words having similar meaning, geometrically close to each other in their numerical representation.

Word Embedding

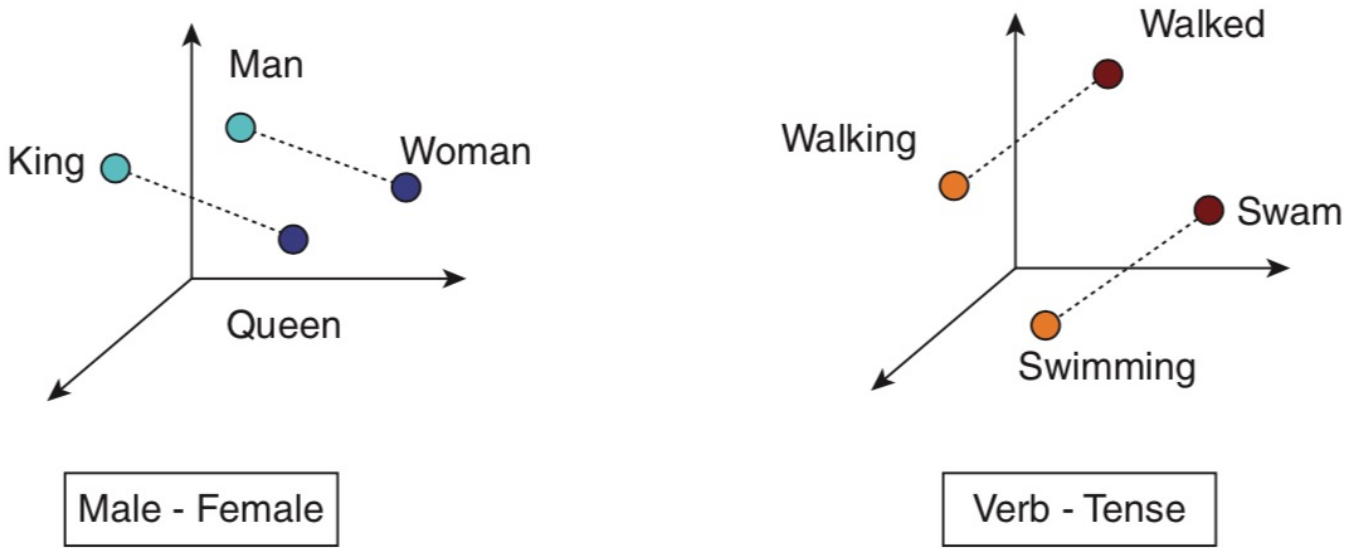


Figure 4.43 Word embedding- vector relationship among words.

- For example, it learns that what a man is to woman is similar to what a king is to queen.
- Its vector representation will calculate it as “Queen = king – man + woman”.
- Gensim[15] is an open-source python library that allows us to develop word embedding model using word2vec algorithm.

[15]Source: https://anaconda.org/anaconda/gensim

Word Embedding

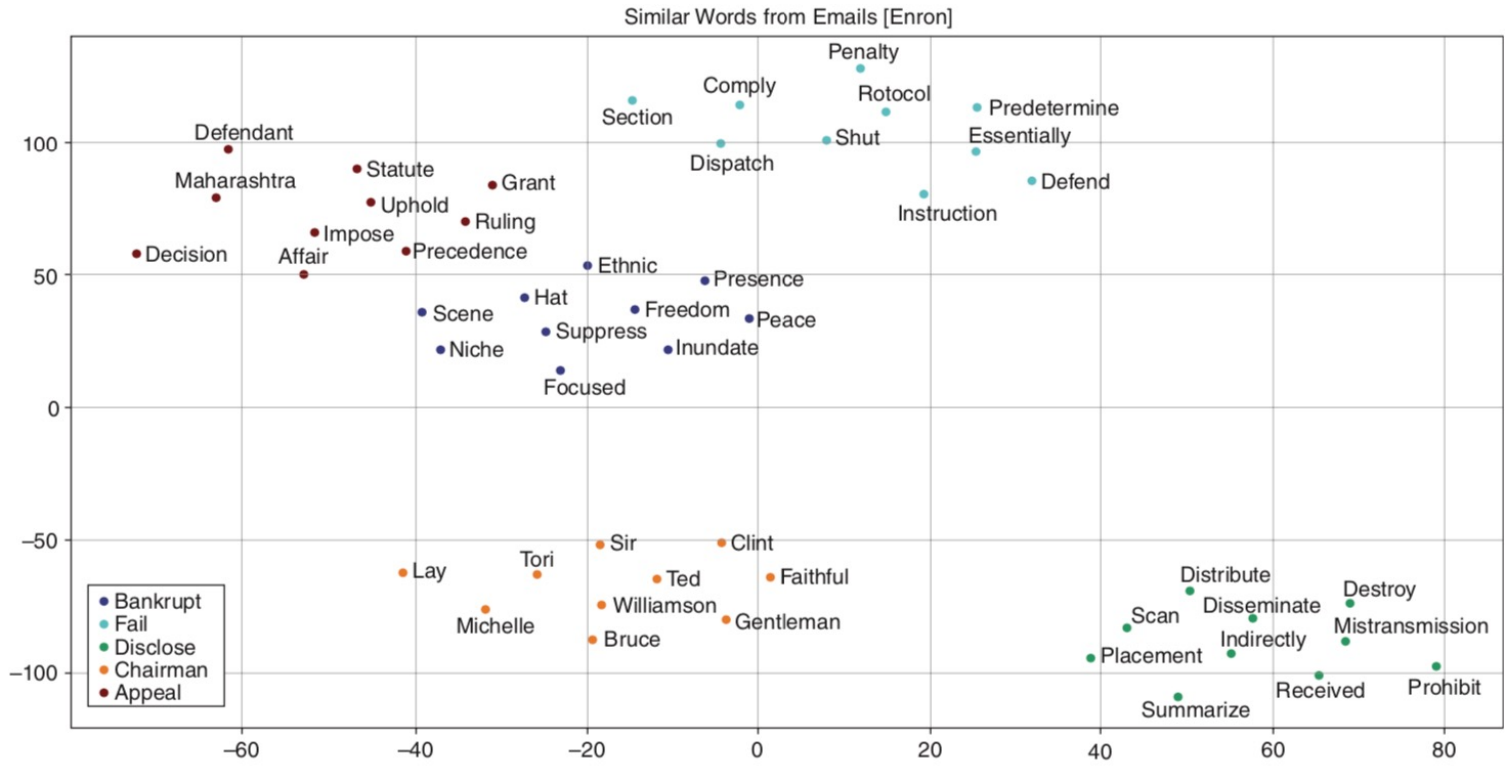


Figure 4.44 Word embedding [Similar words from Enron emails].

- **Inference:** In Fig. 4.44, we can visualize the most similar words corresponding to the list of words chosen in the algorithm (bankrupt, fail, disclose, chairman, and appeal).

Word Embedding

- Genism library was used to build this word2vec model and matplotlib was used for the visualization.
- Using this visualization, one can look for the most similar words corresponding to a particular word to include as many data points as possible, which may help in providing a wider perspective in a scenario.
- As we can see that the most similar words for the keyword “chairman” are “lay”, “tori”, “sir”, “gentleman”, “Williamson”, etc. one can look out for possible reasons of similarity of these words with “chairman”.
- As it is easy to interpret that Kenneth lay was the CEO and chairman of Enron, hence similarity occurs and ‘sir’, and ‘gentleman’ might be used to address him respectfully.

Word Embedding

- Joannie Williamson, used to work in the chairman's office and was former assistant to both Kenneth lay and Jefferey skilling.
- Hence, 'Williamson' has high similarity with 'chairman'.
- Another interesting observation that can be made from the chart is that "appeal" word shows high similarity with the word "Maharashtra".

Word Embedding

- On further investigation, we found out that Enron (as majority share- holder) along with GE, and Bechtel (minority shareholder) built a Power Plant in Dabhol, Maharashtra.
- From 1992 to 2001, the construction and operation of the plant was mired in controversies related to disruption in Enron and at the highest political levels in India and the United States.[16]
- By May 2001, operations of the project were shut and after the infamous bankruptcy of Enron, the majority shares owned by Enron were purchased by Bechtel and GE.[17]

[16]Source: https://web.archive.org/web/20161025181723/http://www.finance-mba. com/Dabhol_fact_sheet.pdf
[17]Source: https://en.wikipedia.org/wiki/Dabhol_Power_Station

Topic Modelling

- Topic modelling is an unsupervised machine learning algorithm method for extracting topics that exist in a dataset.
- In other words, topic models are clusters of words over the document collection where each cluster can be represented as a topic.
- Topic model helps group a collection of words that form a topic.
- Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are a few algorithms in topic modelling.

Topic Modelling
Latent Dirichlet Allocation (LDA)

- LDA works on the basic assumption that a document consists of a set of topics and each topic consists of a set of words.
- Hence, LDA algorithm then tries to reverse engineer this concept.
- It inputs a collection of documents and extracts set of topics from them.
- Gensim library is used to build the model and PyLDAvis is used for the visualization. PyLDAvis is an interactive chart.
- It plots the topic clusters and shows the proportion of words in each topic.

Latent Dirichlet Allocation (LDA)

The visualization consists of two parts:

1. **Bubble chart** is called Intertopic Distance Map, where each circle represents different topics and the distance between them represents the semantic relationship between two topics.

- The closer the circle, the similar they are to each other; farther the circle, more is the dissimilarity in the topics.
- The area of the circle is directly proportional to the number of words in that topic.

Latent Dirichlet Allocation (LDA)

2. **Horizontal bar chart** represents the frequency of the words.

- When we hover over a specific bubble, the words contained in that topic are highlighted.
- The length of the bar chart indicates how significant the word is in identifying the topic corresponding to the circle.
- This also shows the importance of the word for the given topic.

Latent Dirichlet Allocation (LDA)

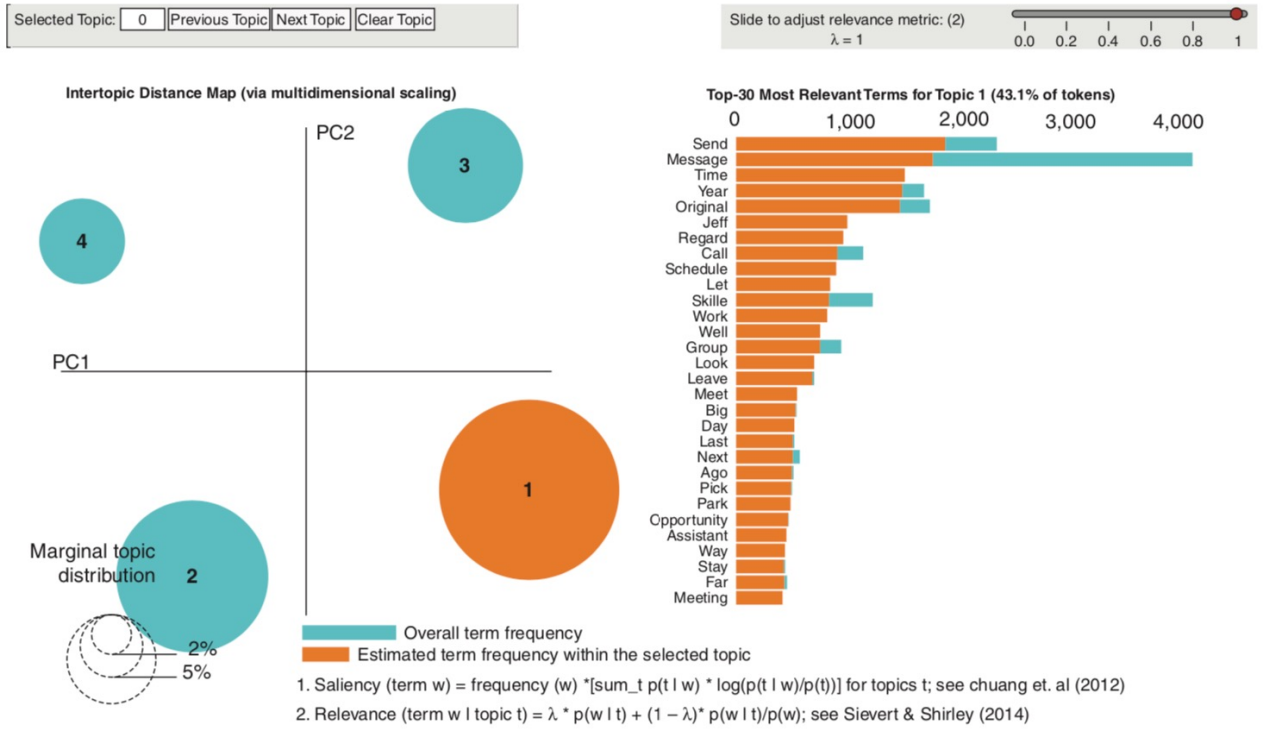


Figure 4.46 Topic modelling- top 30 words in topic 1.

- **Inference:**
- **Topic 1-** contains a lot of meeting-related words, such as “message”, “meeting”, “schedule”, “send”, “email” etc.
- The interpretation is that they may be from emails where meeting scheduling or meeting-related discussions happened.

Latent Dirichlet Allocation (LDA)

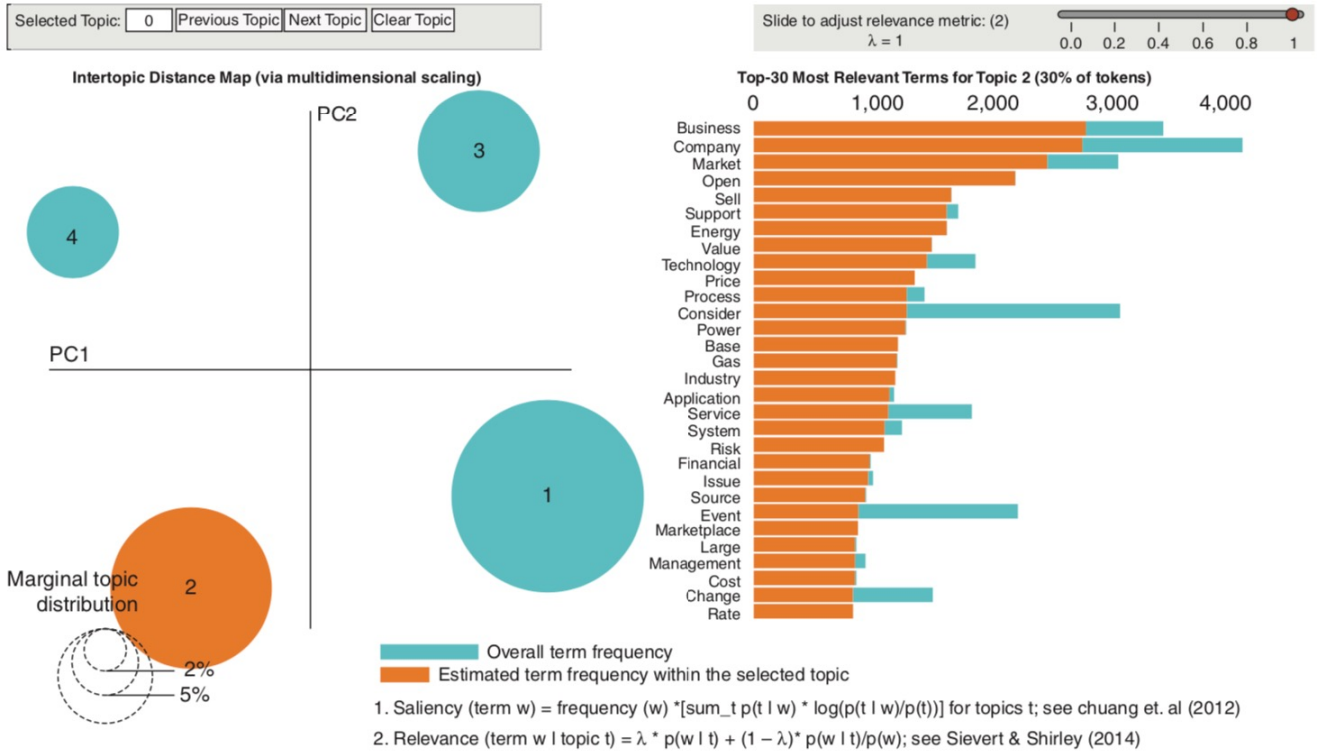


Figure 4.47 Topic modelling- top 30 words in topic 2.

- **Topic 2-** contains terms like “business”, “market”, “power”, “energy”, “gas”, which are directly related to the core business of Enron.
- This topic may be from emails where core business communications occurred.

Latent Dirichlet Allocation (LDA)

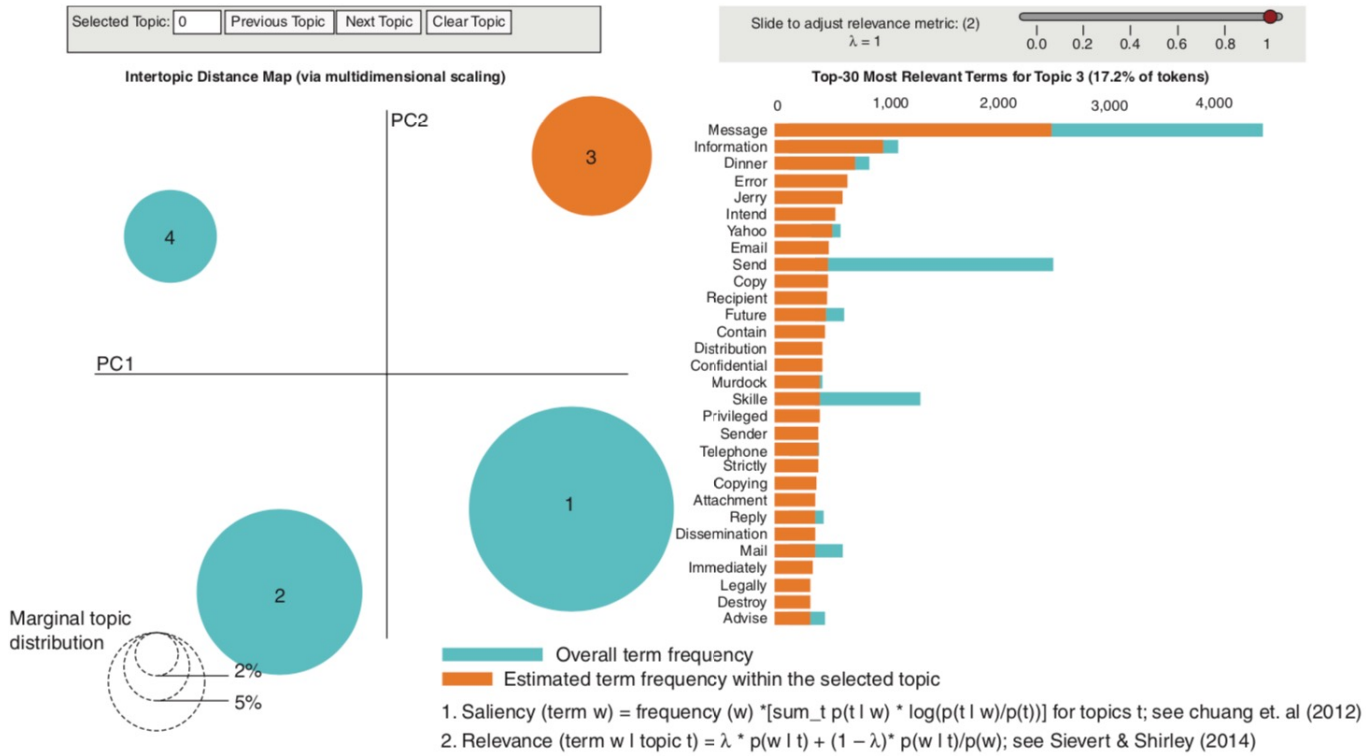


Figure 4.48 Topic modelling- top 30 words in topic 3.

- **Topic 3-** we can see few interesting terms such as “confidential”, “information”, “strictly”, “legally” “destroy”, etc.
- These may be from the set of emails where we can find some information regarding what went wrong which led to the Enron collapse.

Latent Dirichlet Allocation (LDA)

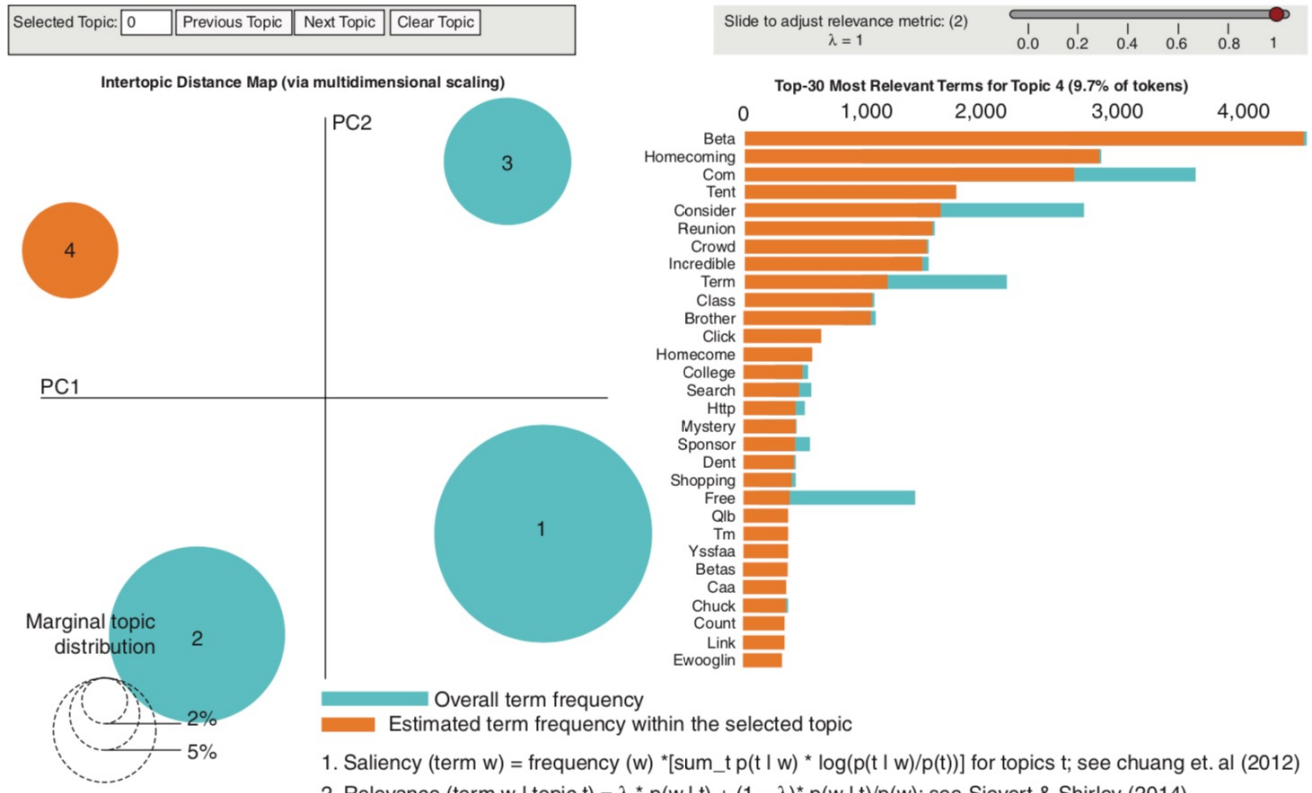


Figure 4.49 Topic modelling- top 30 words in topic 4.

- **Topic4-**contains terms like “reunion”, “homecoming”, “brother”, “shopping”, which are more related to personal communications.

Latent Dirichlet Allocation (LDA)

- Topic analysis can help us extract those emails where Enron crisis related communications happened.
- Further analysis of those emails can give us a better understanding and insight into the most infamous bankruptcy in the history of the United States.

t-SNE Clustering

- The t-SNE (t-distributed Stochastic Neighbourhood Embedding) clusters of all topics in four different colours corresponding to the four topics.
- t-SNE is a manifold learning algorithm, which is used to explore and visualize high dimensional data.
- For data visualization of high dimensional data, it is one of the most popular and efficient **dimensionality reduction** technique.
- **Dimensionality reduction** is the process of trans- forming a high- dimensional data in a low-dimensional space.

t-SNE Clustering

- It constructs two probability distributions.
- One over the original dataset and another in a lower-dimensional dataspace.
- Then the algorithm tries to minimize the distance between both the distributions.
- It is used to visualize the high dimensional data in a 2-D scatter plot.
- The plot helps us in figuring out how close or scattered the clusters are in 2-dimensional space.
- It constructs two probability distributions.
- One over the original dataset and another in a lower-dimensional dataspace.
- Then the algorithm tries to minimize the distance between both the distributions.
- It is used to visualize the high dimensional data in a 2-D scatter plot.
- The plot helps us in figuring out how close or scattered the clusters are in 2-dimensional space.

t-SNE Clustering

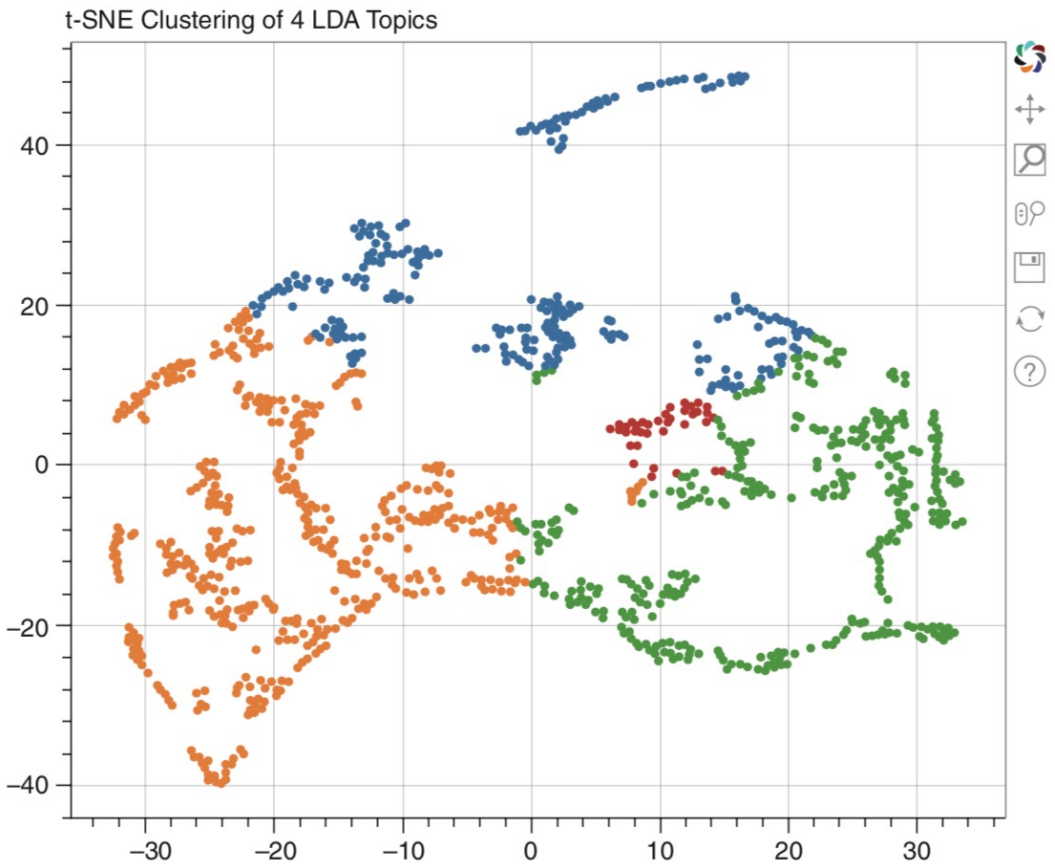


Figure 4.50 Topic clusters of 4 LDA topics.

- The four topics are coloured in four colours- red, orange, green and blue.
- Each rectangle represents a document, and the colour of the rectangle represents the topic it belongs to.
- Each word in a document corresponds to a topic based on its colour.
- Based on the number of words in the topic and the corresponding weightage of the word, a document has been assigned to a topic.

Sentence Chart

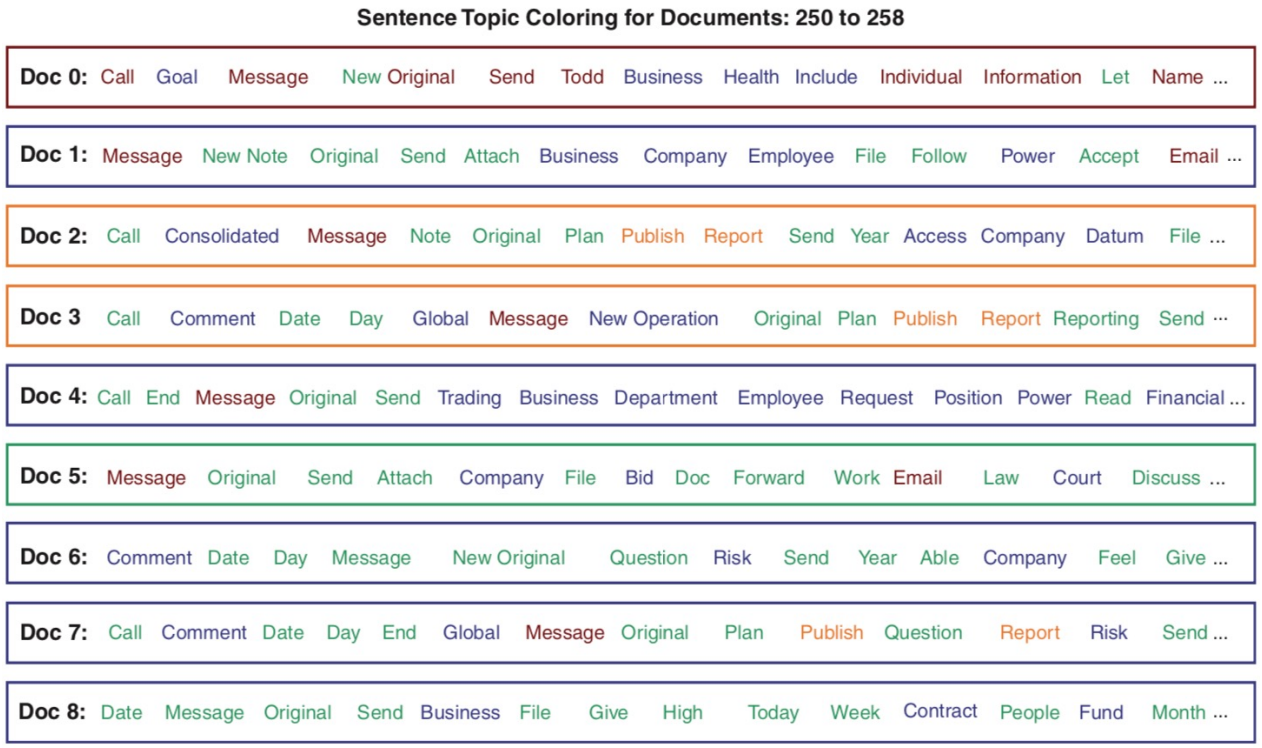


Figure 4.51 Sentence topic coloring for documents.

- In Fig. 4.51:
- The four topics are coloured in four colours- red, orange, green and blue.
- Each rectangle represents a document, and the colour of the rectangle represents the topic it belongs to.
- Each word in a document corresponds to a topic based on its colour.

Sentence Chart

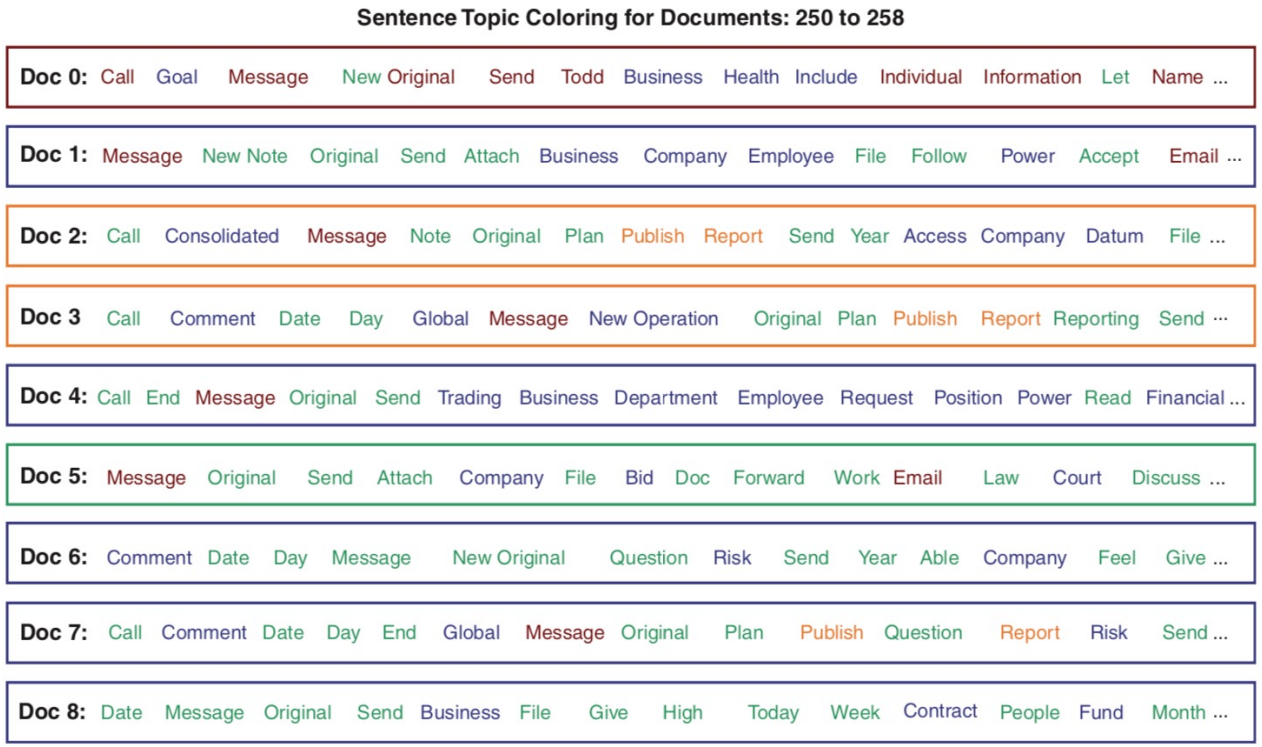


Figure 4.51 Sentence topic coloring for documents.

- **Inference:**
- We can see that words belong to topic coloured ‘blue’ and ‘green’ are more as compared to the words belong to topic coloured ‘orange’ and ‘red’.
- It is interesting to see that in document 2, and document 3, though most of the words correspond to the topics coloured ‘green’ and ‘blue’, the overall document corresponds to the topic-coloured ‘orange’.
- This implies that the weightage/ importance of the words “publish”, and “report” is higher than other words in the documents.

Sentence Chart

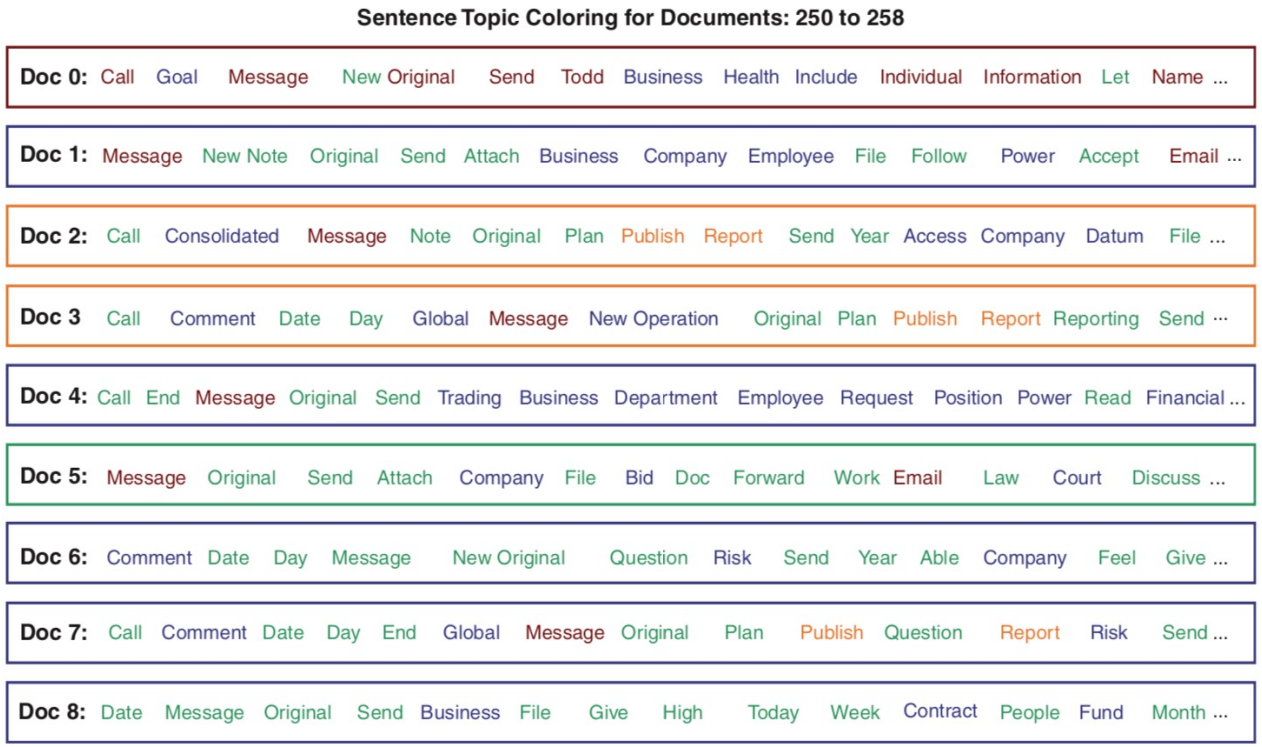


Figure 4.51 Sentence topic coloring for documents.

- **Inference:**
- Contrary to that, in document 7, we have words “publish”, and “report” but still the document does not belong to the topic coloured ‘orange’.
- We can infer that the weightage/importance of the word “comment”, “global”, and “risk” must be higher than that of “publish”, and “report” and hence, document 7 belongs to topic coloured ‘blue’.

Bullet Graph

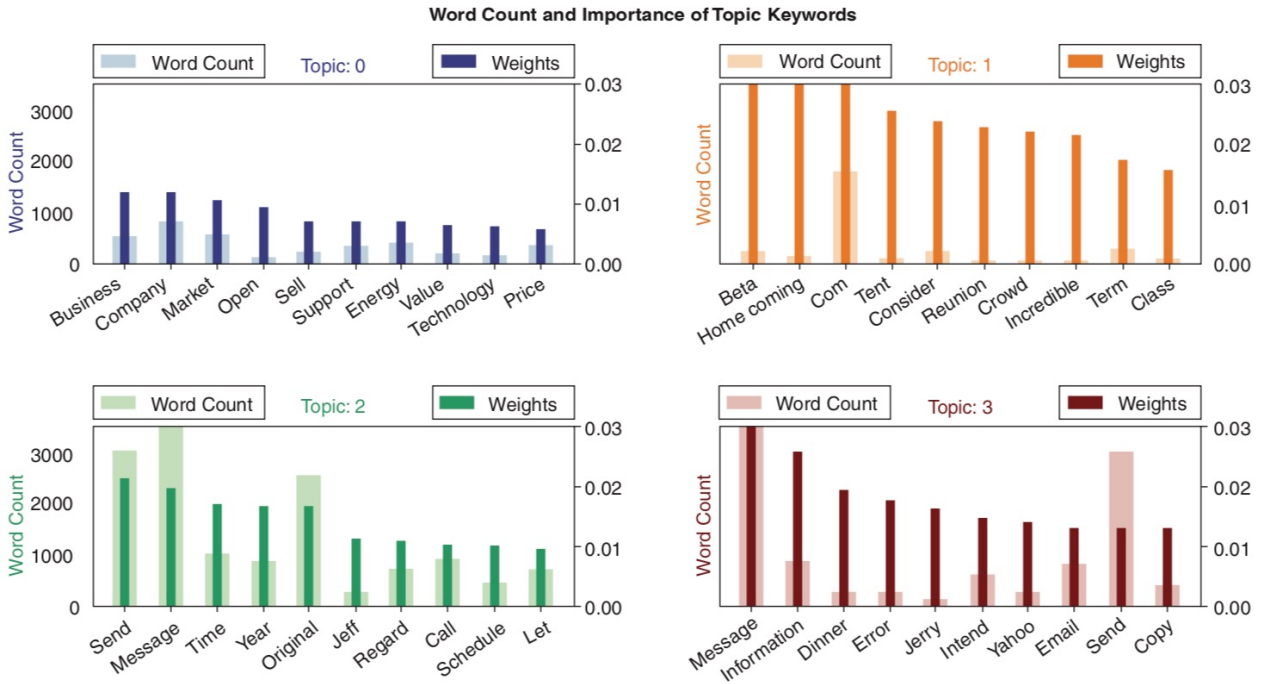


Figure 4.52 Word count and importance of topic keywords.

- Figure 4.52 shows the visualization of the importance of topic keywords along with the word count in the document for all four LDA topics.
- **Inference:**
- It is interesting to note that frequency of a keyword appears in a document is not necessarily associated with the importance (weightage) of the keyword in the document.

Bullet Graph

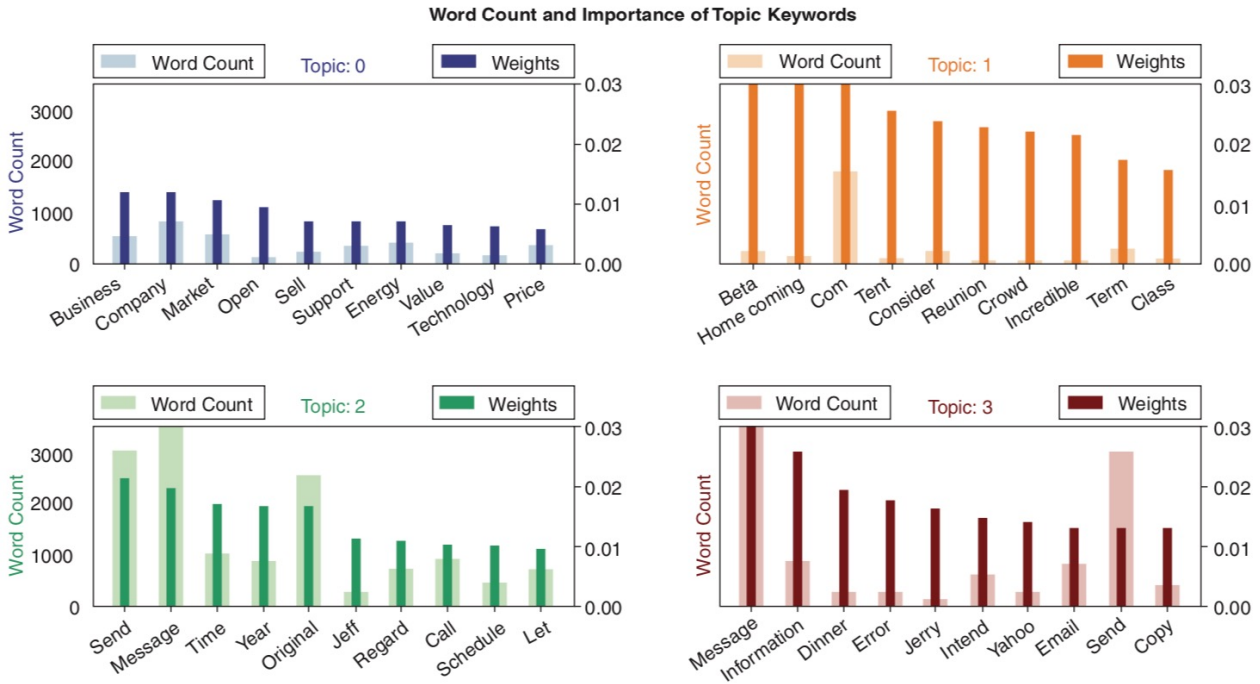


Figure 4.52 Word count and importance of topic keywords.

- In Topic 1 keywords such as “beta”, “homecoming”, “reunion” is not very frequent, but they are among the most important key-words for the topic.
- In Topic 2 keywords “original” is around two times more frequent than words such as “time” and “year”, but all these key- words have almost equal importance.
- Keyword “send” is frequently used in Topic 2 and Topic 3, but it has more weightage in Topic 2 compared to Topic 3.

Bullet Graph

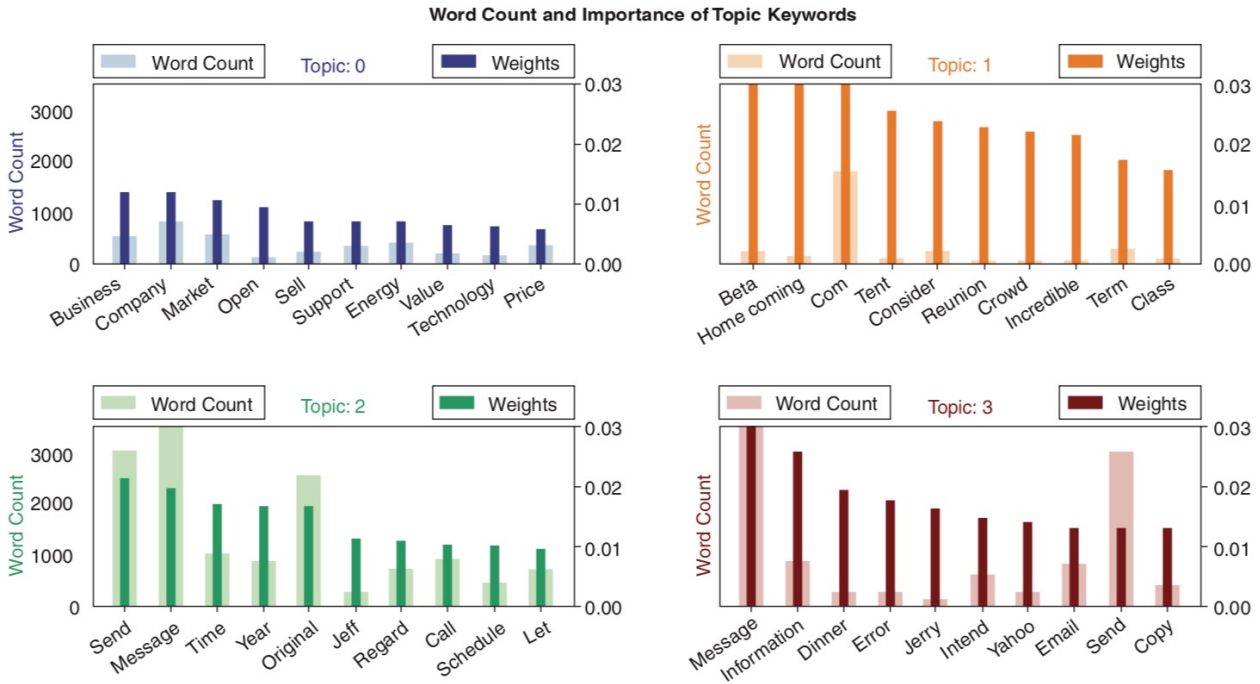


Figure 4.52 Word count and importance of topic keywords.

- This implies the keyword “send” is more representative of Topic 2 compared to Topic 3.
- Similarly, “message” is frequently used in Topic 2 and Topic 3, but it has more weightage in Topic 3 compared to Topic 2.
- This implies the keyword “message” is more representative of Topic 3 compared to Topic 2.

Summary on the packages used for text visualizations

Table 4.5 Summary on the packages used for text visualizations

Chart	Package	Comment
Word cloud [Figures 4.6–4.11]	WordCloud	Tools available: Wordart, TagCrowd, wordsClouds.com, Wordcloud by Jason Davies
Word Tree [Figures 4.15–4.16]		Tools available: Wordtree by Jason Davies
Word Frequency Visualization [Figures 4.26–4.27]	Scattertext, empath	
Tree Map [Figure 4.41]	Plotly.express	
Network Graph [Figures 4.42–4.43]	networkx	
Word Embedding [Figure 4.45]	Matplotlib.pyplot, Matplotlib.cm	
Topic Model [Figures 4.47–4.50]	PyLDAvis.gensim_models	
Sentence Chart [Figure 4.52]	Matplotlib.patches.Rectangle	
Topic clusters [Figure 4.51]	Bokeh.plotting, Bokeh.models, Bokeh.io	

References:

- [Anon, 2017] – Anon (2017), Market Research Report, Text Analytics Market to 2025 – *global Analysis and Forecast by Deployment Type, Technology, Application and Verticals*. Technology, Media and Telecommunications, The Insight Partners, TIPTE100000198, July 2017.
- [Anon, 2020] – Anon (2020), Social Media Sentiment Analysis helped a Sports Technology Firm to drive Customer Satisfaction rates by 3X, Business wire, February 2020, available at https:// www.businesswire.com/news/home/20200225005830/en/Social- Media-Sentiment-Analysis-Helped-a-Sports-Technology-Firm-to- Drive-Customer-Satisfaction-Rates-by-3x-Head-to-Quantzig's-New- Success-Story-for-Detailed-Insights, last accessed February 15, 2021.
- [Pedrazzoli, 2020] – C. Pedrazzoli (2020) – Harry Potter Spells, Chiara Pedrazzoli, Tableau Public, Published in August 2020, available at https://public.tableau.com/app/profile/chiara. pedrazzoli/viz/HarryPotterSpells_15986333032660/HPSpells Overview, last accessed February 15, 2021.
- [Rampell,2009]–Rampell(2009)–Obamain'09vs.Clintonin'93 by Catherine Rampell, Economix: *Exploring the science of everyday life, The New York Times*, September 2009, available at https://webcache. googleusercontent.com/search?q=cache:skCkL0Nb9SIJ:https:// economix.blogs.nytimes.com/2009/09/09/obama-in-09- vs-clinton-in-93/+&cd=1&hl=en&ct=clnk&gl=in&client=safari, last accessed February 15, 2021.
- [Zwolenski, 2014] – M. Zwolenski and L. Weatherill (2014), "The Digital Universe Rich Data and the Increasing Value of the Internet of Things," *Australian Journal of Telecommunications and the Digital Economy*, vol. 2, no. 3, 2014.

Thank You!