Module 4: Exploratory Text Mining

What is Text Mining?

Text mining is the process of extracting meaningful information and patterns from unstructured text data. It combines techniques from linguistics, statistics, machine learning, and information retrieval to transform raw text into structured knowledge that can be analyzed.

The Process of Text Mining

1. **Text Collection:** Gathering documents, emails, reviews, reports, etc.
2. **Text Preprocessing:** Cleaning and preparing data (lowercasing, stopword removal, stemming, lemmatization, handling numbers/dates/URLs).
3. **Feature Extraction:** Converting text into structured forms (Bag-of-Words, TF-IDF, word embeddings).
4. **Pattern Discovery:** Applying algorithms to find associations, clusters, sentiments, or trends.
5. **Knowledge Representation:** Presenting results as summaries, dashboards, or decision-support insights.

How is Text Mining Used in Business?

- **Customer Sentiment Analysis:** Used for understanding reviews and feedback from customers.
- **Fraud Detection:** Used for finding suspicious terms in financial/textual records.
- **Market Intelligence Analysis:** Used to discover trends in news and reports.
- **Email Filtering:** Used for classifying spam vs. non-spam.
- **Contractual Analysis:** Used to extract legal terms, risks, obligations, etc.

A Warm-Up Example

Scenario: Suppose a company collects $10\ 000$ customer reviews. One of the raw reviews is: "The delivery was late and the support team was unhelpful."
Task: What can we gather from this review, based on text mining?

- Upon preprocessing, the following keywords standout: ["delivery", "late", "support", "team", "unhelpful"].
- Frequency analysis shows words like "late", and "unhelpful" are common among many other reviews as well.
- This signals service issues.
- Based on only this one review, we can gather that the customer was not very happy. However, what happens when you are a multinational conglomerate like Amazon or Reliance Industries, it is impractical to read each review one-by-one, hence, we need to do sentiment analysis.

Word Counts

- **Word counts** are the total number of times a word appears in a text or a collection of documents.
- The purpose of word counts is to help identify common words or potential keywords in a corpus.
- Example: A customer posts the following review: "The product is good. The product is cheap."
- The word counts analysis is

Word	Count
"the"	2
"product"	2
"is"	2
"good"	1
"cheap"	1

Term Frequencies

- The **term frequency** is the frequency of a term in a document, usually normalized by the total number of words.
- Mathematically, it is given by

$$TF(t,d) = \frac{\text{Number of times the term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

- It is used to measure how important a word is in a document, relative to its length.
- Example: In the previous example, the term frequency for the word "product" is

$$TF(\text{"product"}) = \frac{2}{8} = 0.25.$$

n-Grams

- **n-grams** are sequences of n consecutive words in a text.
- They are used to capture context and word patterns, not just individual words.
- **Unigrams:** These are **single** words: ["The", "product", "is"].
- **Bigrams:** These are **2-word** sequences: ["The product", "product is", "is good"].
- **Trigrams:** These are **3-word** sequences: ["The product is", "product is good"].
- Applications in business:
  - Used to detect common phrases in customer reviews ("fast delivery", "poor customer service").
  - Used to improve text classification by including phrases instead of just single words.

Zipf's Law

- **Zipf's law** is an empirical law about the frequency of words in natural language. It states that:

> *In a given corpus, the frequency of any word is inversely proportional to its rank in the frequency table.*

- Mathematically, Zipf's law is given by

$$f(r) \propto \frac{1}{r^s}$$

where $f(r)$ is the frequency of the word, r is the rank of the word when words are sorted by frequency (most frequent $= 1$), $s \approx 1$ for natural languages (close to 1).

- Equivalently, Zipf's law can be expressed as

$$\boxed{f(r) = \frac{C}{r}}$$

  where C is a constant that depends on the corpus size.
- The key insights that we gauge from Zipf's law are
  - There are a few words which are extremely common, e.g., "the", "and", "of".
  - Most words are rare, appearing only once or twice in a large corpus.
  - This creates a long-tail distribution: The top 10 words might cover 20–30% of all word occurrences.

Applications of Zipf's Law in Business

- **Stopword Identification:** Words that appear extremely often (like "the", "is") can be safely removed since they carry little semantic information.
- **Keyword Extraction:** Focuses on words in the middle frequency range, which often represent important topics.
- **Data Compression/Storage:** Rare words appear less often. This implies that dictionaries and storage in NLP pipelines are optimized.
- **Language Modeling:** Helps in predicting word probabilities and smoothing techniques.

Example Application of Zipf's Law

Situation: While reading an executive report, we have the following word counts in the corpus:

Word	Count	Rank
"the"	1000	1
"product"	500	2
"is"	400	3
"excellent"	50	4
"fast"	25	5

- Checking Zipf's law
  - $f(1)/f(2) = 1000/500 = 2 = 2/1.$
  - $f(2)/f(3) = 500/400 = 1.25 \approx 1.5 = 3/2.$
- The law is approximate, but captures the heavy head and long tail behavior.

Keywords in Context (KWIC)

- **Keywords in Context (KWIC)** is a concordance tool widely used to study how words are used in real texts.
- It is a way of displaying all occurrences of a given keyword in a text (or a corpus) together with their surrounding words (the context).
- Instead of just listing word counts, KWIC shows the meaning, usage, and nuance of how words appear in sentences.

Why is KWIC Important?

- Helps understand word semantics and usage.
- Reveals collocations (words that often appear together).
- Useful for sentiment analysis, market research, business intelligence, and lexical studies.
- Often used in search engines, chatbots, and digital assistants.

Co-occurrences

- A **co-occurrence** simply means that two words appear near each other in a corpus.
- It is a more general concept than collocation.
- Any two words appearing within a defined window (say, ± 5 words, or the same sentence, or the same document) are considered co-occurring.
- Example: "AI" and "innovation" co-occur within the same sentence. This suggests a semantic link.

Collocations

- A **collocation** is a pair or group of words that tend to occur together more often than expected by chance.
- They form a kind of "natural combination" in a language.
- Example: We say "strong tea", not "powerful tea".
- Example: We say "make a decision", not "do a decision".

The key difference between collocation and co-occurrence is:

- Collocations = statistically strong word pairs (like "customer satisfaction").
- Co-occurrences = any words appearing together (like "AI" and "innovation" in the same sentence).

- Collocations are detected using association measures:
  - **Pointwise Mutual Information (PMI):** This measures how much more often x and y occur together than if they were independent. Mathematically,

$$\boxed{PMI(x,y) = \log\left[\frac{P(x,y)}{P(x)\cdot P(y)}\right]}$$

    where $P(x,y)$ is the joint probability distribution, and $P(x)$ and $P(y)$ are the probability distributions of x and y, respectively.
  - **Chi-Square Test (χ^2):** This tests whether co-occurrence frequency is significantly higher than random chance. Mathematically,

$$\boxed{t(x,y) = \frac{O-E}{\sqrt{O}}}$$

    where O and E are the observed and expected co-occurrences, respectively.

Word Clouds, Bar Charts, and Co-occurrence Networks

- **Word clouds** give a visual representation of word frequency. Words that appear more often are displayed larger and/or bolder. They are used for exploratory data analysis (EDA) and summarization.
- **Bar charts** A quantitative visualization of word counts, term frequencies, or TF-IDF scores (which we will see later on). Unlike word clouds, they show exact values and comparisons. They are used for the top k-terms, sentiment terms, and for n-grams.
- **Co-occurrence networks** are graphs where the nodes are the words, and the edges are the co-occurrences (how often the words appear together). The edges can be weighted by frequency or PMI scores. They are used for relationship discovery, cluster analysis, and trend analysis.

Lab 4

In this lab session, we will use all that we learned in exploratory text mining and apply it to **real-world business problems**.