

# Natural Language Processing

Christian Rossi

Academic Year 2024-2025

## **Abstract**

This course introduces students to the challenges and methodologies related to the analysis and production of natural language sentences, both written and spoken. The course explores the current role of stochastic models and Deep Learning, as well as new opportunities to combine traditional, formally based models with stochastic models. Topics covered include morphology, syntax, semantics, pragmatics, voice, prosody, discourse, dialogue, and sentiment analysis.

The course includes practical exercises where students can test the models and techniques presented in the lectures. Applications explored during the course will include: human-machine and human-human interaction analysis based on language; linguistic and prosodic analysis and generation for rehabilitation; pattern recognition and research for sentiment analysis in critical interactions; complexity analysis in text and overall spoken communication; and the development of user profiles that account for expression preferences in forensic, educational, and clinical contexts.

---

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Natural language . . . . .	1
1.1.1	Natural Language Processing . . . . .	1
1.1.2	History . . . . .	2
1.2	Text analysis . . . . .	2
1.2.1	Text mining . . . . .	3
1.2.2	Characters encoding . . . . .	3
1.2.3	Tokens . . . . .	3
1.2.4	Text normalization . . . . .	4
1.2.5	Morphology and lemmatization . . . . .	4
1.3	Regular expressions . . . . .	5
1.3.1	Regular expressions in text mining . . . . .	5
<b>2</b>	<b>Text classification</b>	<b>7</b>
2.1	Supervised Learning . . . . .	7
2.1.1	Overfitting . . . . .	8
2.2	Text classification . . . . .	8
2.2.1	Feature extraction . . . . .	8
2.2.2	Word frequencies . . . . .	9
2.2.3	Preprocessing . . . . .	10
2.3	Linear classifier . . . . .	10
2.3.1	Multinomial Naïve Bayes . . . . .	11
2.3.2	Logistic regression . . . . .	12
2.3.3	Support Vector Machines . . . . .	12
2.4	Model Evaluation . . . . .	13
2.4.1	Multi-class classifiers . . . . .	14
<b>3</b>	<b>Text search and clustering</b>	<b>15</b>
3.1	Introduction . . . . .	15
3.2	Term weighting . . . . .	16
3.2.1	Inverse Document Frequency . . . . .	16
3.2.2	Term Frequency Inverse Document Frequency . . . . .	16
3.3	Text normalization . . . . .	17
3.3.1	Document length normalization . . . . .	17
3.3.2	Cosine similarity normalization . . . . .	17
3.3.3	Alternative length normalization . . . . .	18
3.4	Indices . . . . .	19

---

3.4.1	Inverted indices . . . . .	19
3.4.2	Positional indices . . . . .	19
3.5	Ranking . . . . .	20
3.5.1	Reranking . . . . .	20
3.5.2	Ranking metrics . . . . .	20
3.5.3	Regression reranking . . . . .	21
3.6	Text clustering . . . . .	21
3.6.1	Clustering algorithms . . . . .	21
3.6.2	Topic modelling . . . . .	23
<b>4</b>	<b>Language models</b>	<b>25</b>
4.1	Introduction . . . . .	25
4.1.1	Markov models . . . . .	25
4.1.2	Language model evaluation . . . . .	26
4.1.3	N-gram limitations . . . . .	27
4.2	Word embeddings . . . . .	27
4.2.1	Training . . . . .	27
4.2.2	Properties . . . . .	28
4.2.3	Word2Vec . . . . .	28
4.2.4	FastText . . . . .	29
4.2.5	Usage . . . . .	30

# CHAPTER 1

---

## Introduction

---

### 1.1 Natural language

The origins of spoken language are widely debated. Estimates range from as early as 2.5 million years ago to as recent as 60,000 years ago, depending on how one defines human language.

The development of written language, however, is more clearly documented. The first known writing systems emerged in Mesopotamia (modern-day Iraq) around 3500 BCE. Initially, these were simple pictograms representing objects, but over time, they evolved into abstract symbols representing sounds, paving the way for more complex communication.

The characteristics of human language are as follows:

- *Compositional*: language allows us to form sentences with subjects, verbs, and objects, providing an infinite capacity for expressing new ideas.
- *Referential*: we can describe objects, their locations, and their actions with precision.
- *Temporal*: language enables us to convey time, distinguishing between past, present, and future events.
- *Diverse*: thousands of languages are spoken worldwide, each with unique structures and expressions.

#### 1.1.1 Natural Language Processing

One reason to care about Natural Language Processing is the sheer volume of human knowledge now available in machine-readable text. With the rise of conversational agents, human-computer interactions increasingly rely on language understanding. Furthermore, much of our daily communication is now mediated by digital platforms, making Natural Language Processing more relevant than ever.

However, Natural Language Processing is a challenging field. Human language is highly expressive, allowing people to articulate virtually anything—including ambiguous or nonsensical statements. Resolving this ambiguity is one of the core difficulties in computational linguistics. Moreover, meaning can be influenced by pronunciation, emphasis, and context, making interpretation even more complex. Fortunately, language is often redundant, allowing for error correction and inference even when mistakes occur.

### 1.1.2 History

The field of Natural Language Processing has its roots in linguistics, computer science, speech recognition, and psychology. Over time, it has evolved through various paradigms, driven by advancements in formal language theory, probabilistic models, and Machine Learning.

During World War II, early work in Natural Language Processing was influenced by information theory, probabilistic algorithms for speech, and the development of finite state automata.

Between 1957 and 1970, two primary approaches emerged. The symbolic approach, based on formal language theory and AI logic theories, focused on rule-based processing. Meanwhile, the stochastic approach leveraged Bayesian methods, leading to the development of early Optical Character Recognition systems.

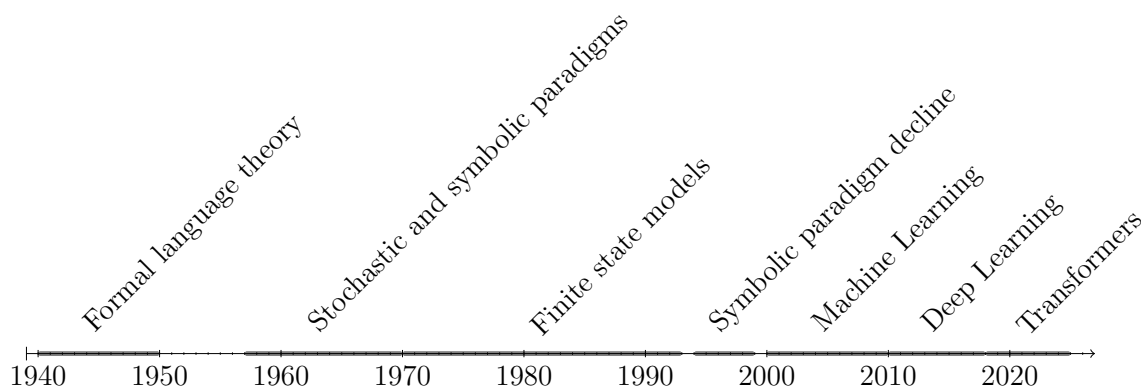
From 1970 to 1993, the focus shifted toward empirical methods and finite-state models. Researchers worked on understanding semantics, discourse modeling, and structural analysis of language.

By the mid-1990s, symbolic approaches began to decline, and the late 1990s saw a surge in data-driven methods, fueled by the rise of the internet and new application areas.

The 2000s marked a deep integration of Machine Learning into Natural Language Processing. The increasing availability of annotated datasets, collaboration with Machine Learning and high-performance computing communities, and the rise of unsupervised systems solidified empiricism as the dominant paradigm.

From 2010 to 2018, Machine Learning became ubiquitous in Natural Language Processing, with Neural Networks driving major advances in conversational agents, sentiment analysis, and language understanding.

Since 2018, the field has been revolutionized by Transformer architectures. Pretrained language models, such as BERT and GPT, have enabled transfer learning at an unprecedented scale, leading to the rise of massive online language models.



## 1.2 Text analysis

Before applying any NLP algorithm, it is important to standardize and clean the text. Preprocessing ensures consistency and improves the accuracy of downstream tasks. Common cleaning steps include:

- Before tokenization, removing non-content information such as HTML tags, converting text to lowercase, and eliminating punctuation.
- After tokenization, filtering out stopwords (extremely common words that add little meaning), removing low-frequency words, and applying stemming or lemmatization to reduce vocabulary size.

### 1.2.1 Text mining

Text may need to be extracted from various sources, each with its own challenges:

- Textual documents, HTML pages, and emails often contain formatting elements that should be removed.
- Binary documents are more complex to process. In PDFs, text may be laid out in multiple columns, requiring careful reconstruction. If all PDFs follow a consistent format, handwritten rules may suffice; otherwise, Machine Learning techniques may be needed.
- Scanned documents require Optical Character Recognition, which relies on Deep Learning to convert images to text. However, Optical Character Recognition is not flawless and can introduce recognition errors.

### 1.2.2 Characters encoding

When storing and processing text, different character encodings must be considered.

ASCII encoding represents only 128 characters, mapping letters and symbols to numerical values. While sufficient for basic English text, it cannot handle many linguistic symbols.

UTF-8 encoding supports over 149,000 Unicode characters, covering more than 160 languages. Unicode is essential for processing texts that use non-Latin scripts. It also preserves special characters, such as diacritical marks in Italian and English.

### 1.2.3 Tokens

In many languages, spaces serve as natural boundaries between words, making tokenization straightforward. However, if spaces weren't available, we would need alternative methods to segment text. Since they do exist in most languages, they are commonly used for tokenization.

Despite this, tokenizing text isn't always simple. Hyphenated words can pose challenges, as some languages construct long, compound words that may need to be split for effective processing. In other cases, meaningful units are spread across multiple non-hyphenated words in multiword expressions. Additionally, punctuation cannot always be blindly removed, as certain clitics (words that don't stand alone) depend on them for meaning.

For languages like Chinese, tokenization is even more complex since it does not use spaces to separate words. Deciding what constitutes a word is non-trivial, and a common approach is to treat each character as an individual token. Other languages, such as Thai and Japanese, require sophisticated segmentation techniques beyond simple whitespace or character-based tokenization.

A more advanced method, sub-word tokenization, can be useful for handling long words and capturing morphological patterns within a language. Instead of relying purely on spaces, data-driven techniques determine the optimal way to segment text. This is particularly important for Machine Learning applications, where models benefit from explicit knowledge of a language's structure. A common approach is byte-pair encoding.

In some tasks, text must be split into sentences rather than just words. Sentence segmentation often relies on punctuation marks, which typically indicate sentence boundaries. However, periods are more ambiguous, as they also appear in abbreviations, numbers, and initials. A common approach is to tokenize first and then use rule-based or Machine Learning models to classify periods as either part of a word or a sentence boundary.

### 1.2.4 Text normalization

In many applications, such as web search, all letters are converted to lowercase. This process significantly reduces the vocabulary size and improves recall by ensuring that variations in capitalization do not affect search results. Since users often type queries in lowercase, this normalization helps retrieve more relevant documents.

For classification tasks, removing case can simplify the learning process by reducing the number of distinct tokens. With fewer parameters to learn, models can generalize better even with limited training data.

However, case folding is not always beneficial. In some contexts, capitalization carries meaningful information. Machine translation and information extraction may also benefit from preserving case distinctions.

Beyond case folding, word normalization involves converting words or tokens into a standard format, ensuring consistency in text processing. This step is particularly crucial in applications like web search, where variations in word forms should not hinder retrieval performance.

**Stopwords** Stopwords are the most frequently occurring words in a language. They typically have extremely high document frequency scores but carry little discriminative power, meaning they do not contribute much to understanding the main topic of a text. Removing stopwords can sometimes improve the performance of retrieval and classification models, mainly by reducing computational and memory overhead. Eliminating common words can also speed up indexing by preventing the creation of excessively long posting lists. However, stopword removal is not always beneficial. In some cases, stopwords play an important role in understanding meaning and context.

### 1.2.5 Morphology and lemmatization

Morphology, a fundamental concept in linguistics, refers to the analysis of word structure. At its core, it involves breaking words down into their smallest meaningful units, known as morphemes.

**Definition** (*Morpheme*). A morpheme is the smallest linguistic unit that carries meaning.

A morpheme can be a root (base form) or an affix, which can appear as a prefix, infix, or suffix.

**Definition** (*Lexeme*). A lexeme is unit of lexical meaning that exists regardless of inflectional endings or variations.

**Definition** (*Lemma*). A lemma is the canonical form of a lexeme.

**Definition** (*Lexicon*). A lexicon is the set of all lexemes in a language.

**Definition** (*Word*). A word is an inflected form of a lexeme.

**Lemmatization** Lemmatization is the process of reducing words to their lemma, or base form. By normalizing words to a common root, it helps deal with complex morphology, which is essential for many languages with rich inflectional systems.



**Stemming** Stemming is a simpler approach that removes affixes based on predefined rules, often without considering the actual meaning or structure of the word. Unlike lemmatization, stemming does not require a lexicon.

Porter stemming algorithm (1980) is one of the most widely used stemming algorithms, it applies a set of rewriting rules to reduce words to their stems. While computationally efficient, stemming can introduce errors such as collisions (different words may be reduced to the same stem) and over-stemming (some words may be shortened excessively, losing meaning).

While stemming is computationally cheaper, lemmatization provides more linguistically accurate results, making it preferable for tasks requiring precise language understanding.

## 1.3 Regular expressions

Text documents are fundamentally just sequences of characters. Regular expressions provide a powerful way to search within these sequences by defining patterns that match specific character sequences. Regular expressions are useful for:

- *Pattern detection*: determine whether a specific pattern exists within a document.
- *Information extraction*: locate and extract relevant information from a document whenever the pattern appears.

Name	Formula	Description
Exact match	aaa	Matches the exact sequence aaa
Sequence choice	(aaa bbb)	Matches either aaa or bbb
Wildcard	.	Matches any single character except for a newline
Character choice	[]	Matches any one character inside the square brackets
Newline	\n	Represents a newline character
Tab	\t	Represents a tab character
Whitespace	\s	Matches any whitespace character
Non-whitespace	\S	Matches any non-whitespace character
Digit	\d	Matches any digit ([0-9]).
Word character	\w	Matches any word character ([a-zA-Z0-9_])
Zero or more times	*	Matches the preceding character zero or more times
One or more times	+	Matches the preceding character one or more times
Zero or one time	?	Matches the preceding character zero or one time
Exactly $n$ times	{ $n$ }	Matches $n$ occurrences of the preceding character
From $n$ to $m$ times	{ $n,m$ }	Matches between $n$ and $m$ occurrences of the preceding character

### 1.3.1 Regular expressions in text mining

Regular expressions offer a powerful way to define patterns that can extract specific content from text documents. This allows for highly customizable and efficient text processing.

The advantages of regular expression based text extraction are:

- *Simplicity*: regular expressions are a straightforward way to specify patterns.
- *Precision*: extraction rules can be finely tuned to target specific patterns, which reduces false positives.

The limitations of regular expression based text extraction are:

- *Manual rule creation*: writing extraction rules usually requires manual effort, which can be time-consuming and complex.
- *False positives*: regular expressions can still yield false positives, where the pattern matches unintended content.
- *False negatives*: false negatives occur when the rules are not broad enough to capture all valid cases.
- *Lack of context awareness*: regular expressions typically work on isolated patterns, without understanding the context in which the pattern appears.

## CHAPTER 2

---

### Text classification

---

#### 2.1 Supervised Learning

Machine Learning involves techniques that help machines become more intelligent by learning from past data to predict future outcomes.

**Definition** (*Machine Learning*). A computer program is considered to learn from experience  $E$  when it improves its performance on a specific task  $T$  based on that experience, as measured by a performance metric  $P$ .

In supervised learning, each training example is represented as a vector in a feature space. These training examples are labeled with the correct class. The task is to divide the feature space in such a way that the model can make accurate predictions for new, unseen data points.

In practice, however, data is rarely perfectly clean or neatly separable. Often, different classes of data overlap, meaning they may not be linearly separable. Additionally, instances are described by many features, and not all features are equally useful for distinguishing between classes. Some features might provide more meaningful information, while others might be less relevant.

To address this, classifiers divide the feature space into regions, and the boundaries between these regions can either be linear or non-linear. Linear models use simple decision boundaries to separate classes. On the other hand, non-linear models are capable of creating more complex decision boundaries to better fit the data.

**Training** The process of training a model involves finding a formula that can predict the correct labels for new instances. The learning algorithm takes in the training data and their corresponding labels, and then searches for the best parameters for the model. These parameters are adjusted in order to minimize prediction loss on the training data. The learning algorithm operates based on its own settings, called hyperparameters, which control aspects of the learning process.

**Hyperparameters** Hyperparameters play a crucial role in determining the model's behavior. These are parameters that govern the learning algorithm itself, and they can influence the complexity of the model.

### 2.1.1 Overfitting

Overfitting is a common challenge in Machine Learning. As the model becomes more complex, the error on the training data tends to decrease. However, at some point, the model may begin to memorize the training data rather than learning generalizable patterns, leading to poor performance on unseen data. This is known as overfitting. The goal is to find the model that strikes the right balance, one that generalizes well to new, unseen data, rather than simply fitting the training data perfectly.

To prevent overfitting, we need to carefully select hyperparameters that control the model's complexity. Since the training data alone doesn't tell us how well the model will generalize, and the test data should be reserved for final evaluation, we use a separate validation dataset. This dataset is a portion of the training data held out during the training process. The model is trained multiple times with different hyperparameter settings, and its performance is evaluated on the validation set. By comparing how well each configuration generalizes, we can choose the hyperparameters that lead to the best performance.

## 2.2 Text classification

Text classification involves training a model to assign documents to specific categories. It's a widely-used task across various fields. Classification problems can take different forms, including:

- *Binary classification*: where the output is either one of two possible labels.
- *Ordinal regression*: where the output is an ordered value, representing categories with a natural rank.
- *Multi-class classification*: where the output corresponds to one category from a set of predefined options.
- *Multi-label classification*: where the output is a set of categories that can overlap or be chosen simultaneously.

### 2.2.1 Feature extraction

Text can be arbitrarily long, so it can't be fed directly into a model. To make text usable for machine learning, we first need to extract meaningful features. Features are the useful signals in the document that help predict its category. To do this, we need to convert text data into a vector of features that a classifier can process.

When training data is limited (with only a few documents available), some approaches to feature extraction include:

- *Syntax-based features*: such as the number of capitalized words.
- *Part-of-speech features*: like the count of verbs versus proper nouns.
- *Reading difficulty features*: such as average word or sentence length

However, the most common and effective features to extract are the words themselves. The vocabulary of the document provides significant signals. The frequency of word occurrences offers additional context.

One popular method is the Bag-of-Words (BoW) model. This model represents documents as vectors of word counts. It results in a sparse representation (long vectors with many zero entries).

**One-hot encoding** One-hot encoding could be an option to create fixed-dimensional feature vectors. However, there are practical limitations: represents each word as a binary feature, creating a vector where each dimension corresponds to a word in the vocabulary. To solve this, we often sum the one-hot encodings. This reduces the number of features to the size of the vocabulary. While this discards word order, it retains the critical information about the vocabulary and word occurrences.

## 2.2.2 Word frequencies

In text data, certain statistical laws describe how term frequencies behave across documents and collections:

- *Heap's law*: this law states that the vocabulary size grows with the square root of the document or collection length:

$$V(l) \propto l^\beta$$

Here,  $\beta \approx 0.5$ . This means the number of unique words in a document or collection increases slowly as the length of the document or the size of the collection grows.

- *Zipf's law*: this law describes the frequency of a token being inversely proportional to its rank:

$$\text{ctf}_t \propto \frac{1}{\text{rank}(t)^s}$$

Here,  $s \approx 1$ . In simple terms, a small number of words (the most frequent ones) appear very often, while a large number of words appear very rarely.

Heap's Law is derived from Zipf's Law and can be understood through models like random typing, showing how the vocabulary of a document or collection grows more slowly compared to its length.

**Bag of Words** The Bag of Words model represents a document as a collection of its terms, ignoring grammar and word order but keeping track of the frequency of each word. In this model:

- The vocabulary of a document is much smaller than the vocabulary of the entire collection, so the terms in the document generally give a good representation of its content.
- The BoW representation typically includes the count of occurrences of each term, although a binary representation (indicating presence or absence of words) can also be used with minimal loss of information.

However, the BoW model has limitations. It produces a sparse representation of the text, meaning most of the values in the vector are zero. The model completely ignores word order, which means it cannot capture the sequence or context in which words appear.

To improve this, we can extend BoW to include  $n$ -grams, which capture sequences of words. This can enhance performance, but it significantly increases the number of features, requiring more data to avoid overfitting.

### 2.2.3 Preprocessing

In natural language processing, preprocessing is a critical step to prepare the text data for machine learning. Common preprocessing tasks include tokenization, spelling correction, and other forms of cleaning the text.

**Tokenization** The default tokenizer in the scikit-learn toolkit for machine learning in Python is simple and efficient. However, for more complex tokenization, the NLTK (Natural Language Toolkit) offers a tokenizer that uses advanced regular expressions to identify different types of tokens, such as words, numbers, punctuation.

**Spelling correction** When dealing with text data, misspellings can often occur, especially with user-generated content. Correcting these misspellings can improve model performance. One way to approach spelling correction is by using a probabilistic model.

If we had an enormous corpus of misspellings and their correct versions, we could estimate the probability of a misspelling being corrected in a certain way by using string edit distance, which measures how much one string differs from another. The edit distance counts the minimum number of operations (insertions, deletions, substitutions, or transpositions) needed to convert one string into another.

To apply this to spelling correction, we use Bayes' Rule to reverse the conditional probability:

- The likelihood of a misspelling being corrected to a particular word can be computed using the edit distance between the misspelled word and the candidate correction.
- The prior  $\Pr(\text{correct})$  represents the popularity or frequency of the word in a large corpus. This can be estimated by counting how often the candidate correction appears in a corpus.
- The likelihood  $\Pr(\text{observed} \mid \text{correct})$  represents the probability of observing the misspelled word given the correct word.

In practical terms, the prior shows how often a word appears in a large corpus, while the likelihood shows how likely it is that the observed misspelling corresponds to a particular correction, based on string edit distance. Since the denominator in Bayes' Rule is the same for all candidate corrections, we can ignore it during the calculation and normalize probabilities later.

To improve spelling correction, we can incorporate contextual information. This can be done by considering bigrams (pairs of consecutive words) instead of just individual words (unigrams). By calculating the bigram probabilities  $\Pr(\text{bigram})$ , the model can use both the misspelled word and the previous word in the sentence as features. This approach turns the spelling correction process into a Naïve Bayes model with two features: the misspelled word and the previous word in the sentence.

## 2.3 Linear classifier

In text classification, where documents are often represented with a bag-of-words model, linear classifiers are commonly used due to the high dimensionality of the feature space. Linear models work by assigning a parameter to each word in the vocabulary, making them highly interpretable. This approach allows us to understand which terms have the greatest impact on the prediction and the extent of their influence.

**Decision boundaries** Linear classifiers create decision boundaries that are represented as hyperplanes in an  $n$ -dimensional vector space. The model includes a weight vector, which is the same size as the feature vector, along with a bias term.

### 2.3.1 Multinomial Naïve Bayes

Naïve Bayes is one of the oldest and simplest text classification algorithms. It is called naïve because it makes a simplifying assumption: that word occurrences are statistically independent of each other given the class label.

This assumption means that each word contributes independent information about the class. It simplifies the process of calculating the model's parameters, making the algorithm easy to implement. However, in practice, this assumption doesn't hold since words are often correlated with each other. Despite this, Naïve Bayes still produces effective predictions, though the assumption can make the model seem overly confident in certain cases.

If all instances of a word appear exclusively in one class, we can have issues with probability estimation. To avoid this problem, we use smoothing, which consists in adding a small pseudo-count  $\alpha$  for each feature. This helps prevent zero probabilities for unseen word-class combinations. The value of  $\alpha$  can be selected to optimize performance or set to a default value (if  $\alpha = 1$ , this technique is known as Laplace smoothing).

**Independence assumption** The independence assumption is not necessarily a big problem. the assumption simplifies both model estimation and prediction, which, in turn, makes the process more efficient. While this theoretically reduces the model's accuracy slightly, in practice, Naïve Bayes often works well for some tasks.

Advantages	
Speed	Naïve Bayes is incredibly fast to train, requiring just one pass over the training data. No need for complex optimization routines like gradient descent
Stability	It's a reliable model even with limited data. If the conditional independence assumption holds, it provides the best possible performance
Disadvantages	
Scalability	Naïve Bayes doesn't perform as well on large datasets compared to other classifiers because redundant features are counted multiple times
Calibrating probabilities	The predicted probabilities are not well calibrated, meaning they can be less reliable for certain applications

### 2.3.2 Logistic regression

The farther a point is from the decision boundary, the more confident we are in our prediction. The signed distance of a point from the hyperplane is given by:

$$s(\mathbf{x}) = \boldsymbol{\theta}\mathbf{x} - b$$

To convert the signed distance  $s(x)$  into a probability, we need a function that maps the entire range of real values  $\mathbb{R}$  to the probability range  $[0, 1]$ . The standard function used for this purpose is the logistic curve (also known as the sigmoid function):

$$\sigma(s) = \frac{1}{1 + e^{-s}}$$

This function outputs a probability of 0.5 at the decision boundary ( $s = 0$ ). The slope, or the speed of probability change, depends on the magnitude of  $\boldsymbol{\theta}$

Advantages	
Well-calibrated probabilities	Logistic regression produces well-calibrated probability estimates
Scalability	It can be trained efficiently and scales well to large numbers of features
Interpretability	The model is explainable, since each feature's contribution to the final score is additive
Disadvantages	
Linearity assumption	Logistic regression assumes feature values are linearly related to log-odds
Sensitivity to assumptions	If the linearity assumption is strongly violated, the model will perform poorly

### 2.3.3 Support Vector Machines

Imagine a dataset where two classes are clearly separable into two groups. There are many possible positions for the linear decision boundary. We want to select a boundary that generalizes well to new, unseen data, avoiding overfitting.

The SVM approach finds the maximum margin hyperplane that separates the two classes. The margin, denoted  $\gamma$ , is the distance from the hyperplane to the closest points on either side. These closest points are called support vectors. Support vectors are the points that lie exactly on the margin. They prevent the margin from expanding and thus help define the location of the boundary. In a  $d$ -dimensional space, you need at least  $d + 1$  support vectors to define the hyperplane.

In contrast to logistic regression, where the position of the hyperplane depends on the entire dataset, the SVM hyperplane's position is determined only by the closest points. The convex hull of the data points helps define the boundary, and moving internal points does not affect the hyperplane.



**Hard margin SVM** A basic SVM is also a linear classifier that finds a hyperplane in feature space that best separates the two classes. While logistic regression and Naïve Bayes also find linear decision boundaries, the difference lies in the loss function used to find the model parameters:

- Logistic regression uses negative log-likelihood, which penalizes points based on the probability of incorrect predictions, even if they are correctly classified.
- SVM uses hinge loss, which only penalizes points that are on the wrong side of the margin (or very close to it).

Mathematically, the loss function for SVM is:

$$\mathcal{L}(\mathbf{w}) = \sum_i w_i^2 + \sum_j \varepsilon_j$$

Here,  $\varepsilon_j$  is the error for a prediction  $(x_j, y_j)$ , and it is defined as:

$$\varepsilon_j = \max(0, 1 - y_j \mathbf{w} x_j)$$

This formulation reflects the hinge loss, which penalizes misclassified points or those close to the margin.

**Soft margin SVM** In the case of non linearly separable dataset, SVMs still aim to separate the classes by penalizing points that are on the wrong side of the margin, based on their distance from the hyperplane. Support vectors are now the points that are either misclassified or very close to the margin, contributing a non-zero amount to the loss function.

The objective function to minimize remains similar to the hard margin case, but it includes a penalty for misclassified points:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_i w_i^2 + C \sum_j \varepsilon_j$$

Here,  $\varepsilon_j$  is the distance from the  $j$ -th support vector to the margin and  $C$  is a hyperparameter that controls the trade-off between minimizing the margin size and penalizing errors. A large  $C$  places more emphasis on minimizing misclassifications, while a smaller  $C$  allows a larger margin even at the cost of more misclassifications.

**SVM and logistic regression** Both SVM and logistic regression are linear classifiers, but they differ in the loss functions they use: Logistic regression uses log-likelihood, which penalizes points based on the probability of incorrect predictions, including those correctly classified. SVM uses hinge loss, which only penalizes points on the wrong side of the margin or those very close to it.

## 2.4 Model Evaluation

When evaluating a classification model, a common starting point is the confusion matrix, which summarizes the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Several metrics are derived from the confusion matrix to assess the performance of a model.

**Accuracy** Accuracy is the proportion of correct predictions, and it can be calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

**Precision** Precision measures the proportion of positive predictions that were actually correct. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

A high precision indicates that when the model predicts positive, it is likely to be correct.

**Recall** Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives that were correctly identified by the model. It is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

High recall means the model is good at identifying positive instances, though it might also include more false positives.

**F-measure** The F1 score combines precision and recall into a single metric by taking the harmonic mean of the two:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is particularly useful when you need a balance between precision and recall, especially when the class distribution is imbalanced.

**Area under the ROC curve** The Area under the ROC Curve (AuC) is another important metric, often used when dealing with different thresholds for making predictions. It measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various thresholds. A higher area under the curve indicates better model performance, as it signifies that the model is better at distinguishing between classes regardless of the threshold.

### 2.4.1 Multi-class classifiers

When working with multi-class classifiers, the confusion matrix expands to an  $n \times n$  matrix, where  $n$  is the number of classes. In this case, precision and recall are calculated for each class, treating each class as the positive class in a one-vs-all fashion.

For a given class  $i$ , precision is the ratio of true positives for that class to the sum of true positives and false positives. Similarly, recall for class  $i$  is the ratio of true positives for that class to the sum of true positives and false negatives.

To combine the precision and recall across all classes, two methods are commonly used:

- *Macro-average*: takes the average of the precision and recall across all classes, treating each class equally, regardless of how many instances belong to that class. This method gives equal importance to all classes, which can be useful when the classes are imbalanced.
- *Micro-average*: aggregates the true positives, false positives, and false negatives across all classes before calculating precision and recall. This method gives more weight to classes with more data points. The micro-average is useful when the number of data points varies significantly between classes.

## CHAPTER 3

---

### Text search and clustering

---

#### 3.1 Introduction

Information retrieval is the task of finding relevant content that match a user’s information need. To achieve this, we typically extract keywords from the user’s query and search for documents containing those keywords.

**Vocabulary matching** A simple but effective approach in text retrieval is vocabulary matching. If the vocabulary is well-distributed across documents, we can use fast indexing techniques to quickly locate relevant documents. However, this method has limitations. Some queries may not have an exact match in the document collection. Many documents might contain all the query terms, making it difficult to rank them effectively. To address these challenges, we can:

- Assign scores to keywords based on how discriminative they are.
- Expand document representations by incorporating additional signals, such as page importance.
- Train machine learning models to improve retrieval performance.

**Classifier** An alternative approach is to train a classifier that directly predicts the relevance of a document to a query. This involves representing both the query and the document as a combined feature vector and predicting the probability that a user finds the document relevant to the query. However, this approach presents several challenges:

- A simple linear classifier would fail to capture complex interactions between query and document terms.
- A non-linear model that accounts for pairwise term interactions would be extremely large—potentially requiring billions of parameters for a vocabulary of 100,000 words.
- Training such a model would require massive labeled datasets of (query, document, relevance) pairs.
- Retrieval speed would suffer if we had to score every document individually.

## 3.2 Term weighting

In information retrieval, term weighting plays a crucial role in determining the relevance of documents to a query. The idea is to identify a subset of query terms that are most likely to return the most relevant documents. Generally, the smaller the subset of terms, the more specific and on-topic the resulting document set is likely to be. Thus, we rank documents based on how small the set of relevant documents is for a given query term subset.

One way to estimate how many documents a query term subset would return is by calculating the probability that a random document contains those terms. Documents can then be ranked based on how unlikely it is to see so many query terms in them.

Assuming the terms in the query are independent, we can express the probability of a document containing all the terms in the query as:

$$\Pr(q \subseteq d') = \prod_{t \in q} \Pr(t \in d') = \prod_{t \in q} \frac{\text{df}_t}{N}$$

Here,  $\text{df}_t$  is the document frequency, or the number of documents containing term  $t$  and  $N$  is the total number of documents in the corpus.

### 3.2.1 Inverse Document Frequency

To rank documents, we can score them based on how unlikely it is for them to contain all the query terms. This gives rise to the inverse document frequency weighting:

$$\text{score}(d) = -\log \prod_{t \in q \cap d} \Pr(t \in d') = \sum_{t \in q \cap d} \log \frac{N}{\text{df}_t}$$

IDF is a measure that assigns weights to terms based on how rare they are across the entire document collection. This is a standard measure from information theory that quantifies the information gained from observing a term. Essentially, the IDF measures the surprise or information content of encountering a term in a document. This same concept is used in text compression algorithms to determine how many bits should be used to represent a word.

A common variation of IDF uses the odds of observing a term rather than the probability, which results in the following document score:

$$\text{score}(d) = \sum_{t \in q} \frac{N - \text{df}_t + 0.5}{\text{df}_t + 0.5}$$

Here, a smoothing factor of 0.5 is added to all counts to prevent terms with very low frequencies from disproportionately affecting the ranking. This smoothing helps to handle rare terms without letting them dominate the results.

### 3.2.2 Term Frequency Inverse Document Frequency

While IDF helps to weight terms by their rarity, there's more to a document than just its vocabulary. Some documents may contain the same query term multiple times, making them more likely to be relevant to the query. To account for this, we introduce term frequency (the number of times a term appears in a document). The simplest way to include term frequency is to multiply the IDF score by the term frequency:

$$\text{score}(q, d) = \sum_{t \in q} \text{tf}_{t,d} \log \frac{N}{\text{df}_t}$$

Here,  $\text{tf}_{t,d}$  is the number of occurrences of term  $t$  in document  $d$ .

This can be motivated as follows: instead of calculating the probability that a random document contains the term, we calculate the probability that a document contains the term exactly  $k$  times:

$$\Pr(t, k) \cong \Pr(\text{next token} = t)^k$$

The next token probability is estimated using the term occurrences across the entire collection:

$$\Pr(\text{next token} = t) = \frac{\text{ctf}_t}{\sum_{t'} \text{ctf}_{t'}}$$

Here,  $\text{ctf}_t$  is the collection term frequency for term  $t$  and  $\sum_{t'} \text{ctf}_{t'}$  is the total term frequency across all terms. The score can then be expressed as:

$$\text{score}(q, d) = - \sum_{t \in q} \text{tf}_{t,d} \log \frac{\text{ctf}_t}{\sum_{t'} \text{ctf}_{t'}}$$

Although this formulation is slightly different, it's conceptually similar to TF-IDF. Using document frequency instead of collection frequency doesn't drastically change the outcome, and in some cases, may even make the formula more robust.

**Variations** In practice, TF-IDF has been found to perform well, but it assumes a linear relationship between term frequency and document relevance. In most cases doubling the occurrences of a term in a document should not double the document's score. The score should improve with more occurrences of the term, but not linearly.

As a result, common alternatives include using a logarithmic scale for term frequency:

$$\log(1 + \text{tf}_{t,d}) \quad \max(0.1 + \log(\text{tf}_{t,d}))$$

## 3.3 Text normalization

When ranking documents, it's important to normalize for document length. Longer documents tend to have a larger vocabulary, which makes it more likely they will contain the query terms. However, this doesn't necessarily mean they are more relevant to the user's search. In fact, shorter documents with the same term count should often be preferred.

### 3.3.1 Document length normalization

One simple way to normalize for document length is to divide the term frequency by the document length. However, the most common method of normalization uses the L2 norm (also called the Euclidean norm) instead of the L1 norm (which is simply dividing by the document length).

### 3.3.2 Cosine similarity normalization

In the Vector Space Model, each document is represented as a vector of term frequencies weighted by their inverse document frequency. The vector for a document might look like this:

$$\mathbf{d} = (\text{tf}_{1,d} - \text{idf}_1, \dots, \text{tf}_{n,d} - \text{idf}_n)$$

To compute the similarity between a query and a document, we measure the cosine of the angle between their vectors. The cosine similarity formula is:

$$\text{similarity}(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\|\mathbf{d}_1\| \|\mathbf{d}_2\|}$$

Here  $\mathbf{d}_1$  represents the query vector, and  $\mathbf{d}_2$  represents the document vector. The cosine of the angle is used because it produces a similarity value in the range  $[0, 1]$ . To calculate the cosine similarity, the vectors are normalized by their Euclidean (L2) norm, rather than the length of the document in terms of tokens (which would be an L1 norm).

### 3.3.3 Alternative length normalization

There have been many studies into alternative methods of length normalization.

**Pivoted Length Normalization** One notable method is Pivoted Length Normalization, which aims to retain the beneficial information from longer documents while preventing them from being unfairly favored. The idea behind Pivoted Length Normalization is that longer documents generally contain more information, but simple length normalization could lose valuable length information. Instead, PLN adjusts the normalization to account for both the document length and the average document length in the corpus:

$$\frac{\text{tf}_{t,d}}{L_d} \rightarrow \frac{\text{tf}_{t,d}}{bL_d + (1-b)L_{\text{avg}}}$$

Here,  $L_d = \sum_t \text{tf}_{t,d}$  is the length of the document,  $L_{\text{avg}} = \frac{1}{N} \sum_d L_d$  is the average document length across the corpus, and  $b$  is a parameter that controls the balance between the document length and the average length, with  $0 \leq b \leq 1$ .

**Best March 25** Another widely used length normalization method is BM25 (Best Match 25), a ranking function that builds upon the ideas of TF-IDF and length normalization. The formula for BM25 is as follows:

$$\text{RSV}_d = \sum_{t \in q} \log \frac{N}{\text{df}_t} \frac{(k_1 + 1) \text{tf}_{t,d}}{k_1 \left( (1-b) + b \cdot \frac{L_d}{L_{\text{avg}}} \right) + \text{tf}_{t,d}}$$

Here,  $k_1$  and  $b$  are parameters that control how much weight is given to term frequency and document length,  $N$  is the total number of documents in the corpus,  $\text{df}_t$  is the document frequency of term,  $\text{tf}_{t,d}$  is the term frequency of  $t$  in document  $d$ , and  $L_d$  is the length of the document. Some of the reasons for its lasting effectiveness include:

- *Asymptotic term importance*: the influence of a term on the document's score decreases as its frequency increases. This means documents with extremely high term frequency won't always be ranked at the top.
- *Parameter control*: the  $k_1$  and  $b$  parameters allow fine-tuning based on the corpus, and default values typically work well. However, these parameters can be improved through validation on a specific dataset.

## 3.4 Indices

Search engines are designed to deliver results as quickly as possible, since any delay can impact user experience and attention. Responses must be returned within tenths of a second, so search engines are optimized for speed.

At the heart of this efficiency is the inverted index, which is the core data structure used by search engines to retrieve documents.

### 3.4.1 Inverted indices

An inverted index consists of posting lists, which map term IDs to document IDs. The basic idea is to create a mapping between terms and the documents that contain them, so that when a user searches for a term, the system can quickly find all the relevant documents.

To optimize for speed and reduce storage space, inverted indices often use integer compression algorithms, allowing for quick decompression and reducing the overall size of the index.

When calculating a retrieval function, the process typically involves joining posting lists. The documents within these lists are sorted by term frequency. This sorting allows for early termination of the results list computation, so irrelevant documents are discarded quickly.

### 3.4.2 Positional indices

In many cases, it's not just about whether a term appears in a document, but where it appears. To capture this, some indices maintain positional information, recording the exact locations of terms within documents. This allows for the calculation of proximity between query terms, which can be a useful indicator of relevance.

Moreover, the location of words within a webpage can influence their importance. In addition, certain statistically significant bigrams and trigrams are often identified and indexed separately. These are usually discovered using a technique like pointwise mutual information, which measures the association between terms. These bigrams or trigrams often have their own posting lists, as they can provide more context to queries.

#### 3.4.2.1 Crawlers

To populate the index, web crawlers scour the web, following hyperlinks to discover and add new pages to the search engine's database. Effective crawling involves two main challenges:

- *Prioritizing URLs*: the crawler must decide which URLs to visit first based on factors like relevance and likelihood of finding fresh content.
- *Re-visiting websites*: determining how often to revisit a website to check for updates is critical to ensuring that the index remains fresh and up to date.

At the scale of the web, crawlers must also be robust enough to handle different types of content, including dynamically generated pages. Additionally, web crawlers must detect and manage duplicate content. Many different URLs may lead to the same content, and the crawler needs to ensure that it doesn't index the same page multiple times.

To manage these challenges, a distributed crawler architecture is typically used, with a centralized URL list to keep track of the pages the crawler needs to visit. Crawlers also respect robots.txt files, which are placed in the root directory of websites. These files tell crawlers which pages or sections of the site they are allowed to crawl and index, helping website owners manage how their content is indexed.

## 3.5 Ranking

In web search, search engines rely on a variety of indicative features to determine the most relevant results for a user query. Search engines combine hundreds of signals to generate the most relevant search results. To do this efficiently, rank learning offers an automated and coherent method of combining diverse signals into a single retrieval score. It optimizes this score based on metrics that are important to users.

### 3.5.1 Reranking

The reranking process follows these steps:

1. *Start with a query*: the user enters a search query.
2. *Generate initial ranking*: use keyword-based ranking to retrieve an initial set of results.
3. *Truncate ranking*: limit the ranking to a manageable number of candidates for further evaluation.
4. *Calculate feature values*: compute relevant features for each candidate document.
5. *Normalize features*: normalize each feature at the query level to make the comparison between documents more consistent.
6. *Training*: use ground truth relevance labels to train the model.

### 3.5.2 Ranking metrics

There are several metrics used to evaluate the performance of search engine rankings. Here's a breakdown:

- *Precision at depth  $k$* : the percentage of relevant documents in the top  $k$  results:

$$\text{precision} = \frac{\text{number of relevant documents in top } k}{k}$$

- *Recall at depth  $k$* : the percentage of all relevant documents that are found in the top  $k$  results.

$$\text{recall} = \frac{\text{number of relevant documents in top } k}{\text{total relevant documents}}$$

- *F-measure at depth  $k$* : a combination of precision and recall, providing a single score that balances both:

$$\text{F1} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- *Mean Average Precision*: this is the average of the precision values at each rank position where a relevant document appears. It estimates the area under the precision-recall curve:

$$\text{MAP} = \frac{\sum_{k=1}^n (\text{precision}(k) \cdot \text{rel}(k))}{\text{total number of relevant documents}}$$



- *Normalized Discounted Cumulative Gain*: this metric is more faithful to the user experience, as it gives lower ranks less importance, aligning with the natural way users interact with search results. It's normalized at the query level:

$$\text{NDCG}(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^k \frac{2^{R(j,m)} - 1}{\log_2(1 + m)}$$

### 3.5.3 Regression reranking

Reranking can be treated as a regression problem, where the goal is to predict the relevance of a document based on various features. The model can then use standard regression techniques to combine these features and predict relevance scores. During training, the loss function can be defined in three different ways:

- *Pointwise*: this approach calculates the loss based on individual query-document pairs (Mean Squared Error).
- *Pairwise*: this method considers the relative ordering of document pairs (number of incorrectly ordered pairs).
- *Listwise*: this considers the entire ranking list (optimizing NDCG at the query level).

## 3.6 Text clustering

Text clustering is the process of grouping documents into coherent subgroups based on similarities in their content. These subgroups can be based on various criteria.

The key challenge in text clustering is to accurately measure the similarity between documents. This similarity is typically determined by analyzing the content of the documents. Traditionally, documents are represented as tf-idf vectors, which are scaled bag-of-words representations. In this approach:

- *Sparse representation*: most of the values in the vector are zero, meaning that each document is represented by a vector with many empty or zero entries.
- *Similarity measure*: the similarity between documents is usually based on the number of shared words, with the importance of each word being scaled by its rarity.

### 3.6.1 Clustering algorithms

There are several popular clustering algorithms used for grouping documents based on similarity.

#### 3.6.1.1 k-means

$k$ -means searches for exactly  $k$  clusters, each represented by a centroid, in the following way:

1. Randomly initialize  $k$  centroids.
2. Assign each data point to the nearest centroid.
3. Recompute the centroids by averaging the data points in each cluster.

4. repeat steps 2 and 3 until convergence.

It scales well to large datasets and does not need pairwise distance calculations needed, which makes it faster. However, it assumes clusters are globular, which may not be ideal for text data, relies on the Euclidean distance metric, which might not be the best choice for all types of data.

The number of clusters must be specified in advance. Choosing the right value of  $k$  can be tricky, though the elbow method can help. The algorithm can converge on a local minimum; running it multiple times can help mitigate this issue. Scaling of the data affects clustering results, especially for text, where tf-idf weighting and document length normalization are important.

### 3.6.1.2 k-medoids

$k$ -medoids is similar to  $k$ -means, but it uses medoids instead of centroids. A medoid is a point from the dataset itself that is closest to all other points in the cluster. In each iteration, the algorithm reassigns data points to the cluster with the closest medoid and then recomputes the medoids.

It is flexible because it works with various distance metrics. Since a medoid is an actual data point (not a mean), it provides a more realistic representation of the cluster. However, it has higher computational complexity compared to  $k$ -means because it requires calculating distances between pairs of points, which is an  $\mathcal{O}(n^2)$  operations.

### 3.6.1.3 Agglomerative hierarchical clustering

Agglomerative hierarchical clustering is a hierarchical clustering builds a hierarchy of clusters, known as a dendrogram. Agglomerative hierarchical clustering works by merging smaller clusters bottom-up in the following way:

1. Assign each document to its own group.
2. Merge the two most similar groups.
3. Repeat until only one group remains.

To merge clusters, hierarchical clustering computes distances between them using various linkage criteria:

- *Complete-linkage*: maximum distance between points in two groups.
- *Single-linkage*: minimum distance across groups.
- *Average-linkage*: average distance across groups.

The choice of linkage criteria affects the shape and tightness of the clusters. Complete or Average-linkage tends to create tight, globular clusters. Single-linkage can lead to long, thin clusters.

One advantage is that it works with any distance metric or similarity function. Thus, the dendrogram provides useful insight into the structure of the data. However, its high time complexity makes it unsuitable for large datasets.

### 3.6.1.4 Density-Based Spatial Clustering of Applications with Noise

DBScan is a density-based clustering algorithm that does not require the number of clusters to be specified in advance. It has two key parameters, namely  $\varepsilon$  (radius of the neighborhood around each point) and `minPoints` (minimum number of points required to form a cluster). The algorithm classifies points as:

- *Core points*: points with at least `minPoints` within their  $\varepsilon$ -neighborhood.
- *Border points*: points that are not core points but lie within the  $\varepsilon$ -neighborhood of a core point.
- *Noise points*: points that are neither core nor border points.

In this algorithm we do not need to specify the number of clusters upfront, it is robust to noise and outliers, and can find arbitrary-shaped clusters, making it highly flexible. However, the performance depends on the chosen parameters and may not work well when clusters have different densities.

## 3.6.2 Topic modelling

Topic modeling is a soft clustering technique for documents, meaning that each document can belong to multiple clusters (or topics) to varying degrees. It is a way to extract the underlying themes or topics from a collection of documents.

Topic modeling reduces the dimensionality of documents by representing them in a lower-dimensional space compared to the high-dimensional vocabulary space. Each topic is described by a probability distribution over words, where the terms should ideally reflect a single theme. Similarly, documents are represented as probability distributions over topics.

**Matrix decomposition** Topic modeling is often approached as a form of matrix decomposition. The general idea is to decompose a term-document matrix (which represents word counts or tf-idf scores for each term in each document) into smaller matrices representing terms  $\cdot$  topics and topics  $\cdot$  documents. Instead of dealing with a large  $V \times D$  matrix (where  $V$  is the vocabulary size and  $D$  is the number of documents), we decompose it into two much smaller matrices:

$$V \times T \quad (\text{terms} \cdot \text{topics}) \quad T \times D \quad (\text{topics} \cdot \text{documents})$$

Here,  $V$  is the vocabulary size,  $T$  the number of topics, and  $D$  the documents count. This decomposition reduces the number of parameters that need to be estimated, making the topic modeling process more efficient. The most used modeling techniques are:

- *Latent Dirichlet Allocation*: LDA is the most famous technique in topic modeling. It uses a Dirichlet prior to estimate the parameters. In LDA, each document is assumed to be a mixture of topics, and each topic is a mixture of words.
- *Non-negative Matrix Factorization*: a related technique to LDA, which factorizes the document-term matrix into two non-negative matrices representing term-to-topic and topic-to-document relations.
- *Latent Semantic Indexing*: LSI applies Singular Value Decomposition to a TF-IDF matrix to uncover latent semantic structure.

**Applications** Topic modeling helps address several challenges:

- *Polysemy*: it can disambiguate words with multiple meanings.
- *Synonymy*: it identifies synonyms that might be used interchangeably.
- *Short documents*: it improves the representation of documents that may have limited vocabulary due to their short length.

In addition to its utility in improving document representation, topic modeling is also useful for dimensionality reduction. After modeling topics, further dimensionality reduction techniques can help visualize collections of documents in a more interpretable way.

### 3.6.2.1 Generative model

The Generative Model for Latent Dirichlet Allocation is a probabilistic process that describes how the words in a document are generated from topics. Here's how it works:

1. *Choose word proportions for each topic*: each topic is associated with a probability distribution over words.
2. *Choose topic proportions for each document*: each document is modeled as a distribution over topics.
3. *For each word in a document*: choose a topic based on the document's topic proportions and choose a word based on the topic's word distribution.

Estimating the parameters of the topic model involves updating the topic and word distributions iteratively. This is typically done using Bayesian priors to avoid overfitting, ensuring that the model doesn't just memorize the data but rather generalizes well. Gibbs sampling or other sampling techniques are used to avoid local maxima during parameter optimization.

The hyperparameters of this method are:

- $\alpha$ : the prior on the topic distribution for each document (controls how concentrated the topic distribution is).
- $\beta$ : the prior on the word distribution for each topic (controls how concentrated the word distribution is).

By iteratively updating these parameters and sampling from the distribution, the model learns the underlying structure of topics in a collection of documents.

---

## Language models

---

### 4.1 Introduction

**Definition** (*Statistical language model*). A statistical language model is a probability distribution over sequences of words.

Given this distribution over sequences, we can condition the next word based on previous words and generate new sequences by sampling from it. In essence, language models serve as general-purpose text generators.

Language models identify statistical patterns in text and leverage these patterns to predict the next word in a sequence. By predicting each subsequent word with increasing accuracy, language models can generate entire sentences or even longer passages.

#### 4.1.1 Markov models

Natural language utterances can be of arbitrary length, but we only have a finite set of parameters to model them. One way to define a probability distribution over such variable-length sequences is to predict the next word based on a fixed number of previous words.

The simplest models for this are  $n$ -gram models, which count sequences of  $n$  words in a large corpus. Using longer  $n$ -grams provides better predictions, though the model can become more sparse and complex as  $n$  increases.

Several techniques improve the performance of Markov models:

- *Smoothing* (regularization): this technique adds a small constant to all counts before estimating probabilities, which helps avoid assigning zero probability to unseen  $n$ -grams. The smoothed probability can be computed as:

$$\Pr(w_n \mid w_{n-1}, w_{n-2}) = \frac{\text{count}(w_{n-2}w_{n-1}w_n) + \alpha}{\text{count}(w_{n-2}w_{n-1}) + V\alpha}$$

Here,  $\alpha$  is a pseudocount (often set to  $\alpha = 1$ ), and  $V$  is the size of the vocabulary. This ensures that all possible  $n$ -grams have a non-zero probability, even if they haven't been seen in the training data.

- *Backoff*: instead of inventing values for unseen  $n$ -grams, the backoff technique uses lower-order models when an  $n$ -gram is not found in the training data. This helps maintain the flow of predictions while keeping the model manageable.

- *Interpolation*: this technique combines higher-order and lower-order models by blending their probabilities. To determine the weight of each model, interpolation parameters (lambdas) are chosen to maximize the likelihood on a held-out development set.

**Generative Markov model** A generative Markov model can be used to estimate the probability of the next word and generate text in various ways:

- *Greedy*: this approach chooses the most probable term:

$$w^* = \operatorname{argmax}_t \Pr(w_n = t \mid w_{n-k}, \dots, w_{n-1})$$

- *Random sampling*: in this method, a term is sampled according to its probability:

$$w^* \sim \Pr(w_n \mid w_{n-k}, \dots, w_{n-1})$$

- *Top-k sampling*: sampling is restricted to the top  $k$  most likely terms:

$$w^* \sim \Pr(w_n \mid w_{n-k}, \dots, w_{n-1}) \mathbf{1}(w_n \in \text{top-}k)$$

- *Temperature sampling*: sampling is limited to likely terms by raising the probabilities to a power:

$$w^* \sim \Pr(w_n \mid w_{n-k}, \dots, w_{n-1})^{\frac{1}{T}}$$

Here,  $T$  is the temperature (higher temperature means a more uniform sampling).

- *Beam search*: this technique searches forward one step at a time for the most likely sequence  $(w_n, w_{n+1}, w_{n+2}, \dots)$  while limiting the search space to a maximum set of  $k$  candidate sequences.

Greedy techniques always produce the same text, while sampling generates different text each time. Output from lower-order  $n$ -gram language models may produce text that is non-sensical but potentially grammatical.

### 4.1.2 Language model evaluation

To determine if one language model is better than another, we use two main evaluation approaches:

- *Extrinsic*: the model is used in a downstream task, and the performance is evaluated based on the task's outcomes.
- *Intrinsic*: the model's parameters are trained on a dataset, and its performance is evaluated on a held-out dataset. The likelihood of the model producing the observed data is used to assess how well it is performing.

**Definition** (*Perplexity*). Perplexity is a measure of how well a language model predicts new text.

It quantifies the level of surprise or confusion when encountering new data and reflects how unlikely the observed data is under the model. The perplexity is computed through the following steps:

1. Compute the probability of the observed sequence under the model.
2. Normalize the probability for the length of the text sequence.
3. Invert the probability to calculate uncertainty. Minimizing perplexity is equivalent to maximizing probability, so a lower perplexity indicates a better model.

Perplexity is closely related to other metrics used in training and evaluating predictive models:

- *Negative Log Likelihood* (nLL): the negative logarithm of the probability of the sequence. When divided by the sequence length, this gives the per-word nLL. Perplexity can be derived as:

$$\text{perplexity} = 2^{nLL}$$

- *Crossentropy*: the expected log surprise under the model. It represents the number of bits needed to quantify the surprise or uncertainty of a sequence.

### 4.1.3 N-gram limitations

As the value of  $n$  increases in  $n$ -gram models, the probability of encountering a specific sequence in the training corpus decreases exponentially. This results in sparse data, where many possible  $n$ -grams are never seen during training. When the model backs off to shorter  $n$ -grams, this significantly limits its ability to make accurate predictions.

Moreover, to generate reasonable and coherent language, we need to model long-distance dependencies, where the relationship between words may span across many tokens. Unfortunately, as  $n$  grows, both memory and data requirements scale exponentially with the length of these dependencies, making traditional Markov models impractical for capturing such long-range relationships.

## 4.2 Word embeddings

Word embeddings, which emerged around 2013, significantly improved performance on nearly every natural language processing task. They are dense vector representations of words in a high-dimensional space, typically with between 100 and 1000 dimensions. These embeddings are much more compact compared to the sparse one-hot encoding of terms, which typically requires a vector the size of the entire vocabulary. Given that document collections often have vocabularies ranging from 100,000 to 1 million tokens, word embeddings provide a much more efficient way to represent words.

Similar to one-hot encodings, word embeddings can also be aggregated to represent larger units of text. This aggregation allows for the capture of semantic meaning in a more computationally efficient manner.

### 4.2.1 Training

Word embeddings are produced using supervised machine learning models, specifically models that are trained to predict a missing word based on the surrounding context. The context can either include only previous words (causal models) or both previous and future words (non-causal models). In this setup, the training process can be described as follows:

- *Features*: the words in the current context.

- *Target*: the missing word that we aim to predict from the sequence.

This is essentially a multi-class classification problem, where the model needs to estimate the probability for every word in the vocabulary being the missing word.

**Challenges** A key challenge is that even a simple linear classifier for this task requires a large number of parameters. For example, if we use a multi-class linear classifier to predict the missing word, with a bag-of-words feature vector (ignoring word order), the model will require a parameter vector of the size of the vocabulary for each vocabulary term. Therefore, the total number of parameters will scale quadratically with the size of the vocabulary.

This quadratic growth in parameters was a significant issue before deep learning techniques emerged, as the vocabulary size in traditional NLP tasks was often very large. However, with modern deep learning techniques, we can overcome this limitation more efficiently.

### 4.2.2 Properties

Word embeddings possess several intriguing and sometimes surprising properties that make them powerful tools for natural language processing tasks:

- *Semantic clustering*: neighbors in the embedding space are often semantically related. Words that share similar meanings or appear in similar contexts are typically located near each other in the embedding space.
- *Dense and distributed representation*: word embeddings use multiple dimensions to capture semantic relationships. However, individual dimensions of the embedding vector are generally not directly interpretable.
- *Meaningful translations*: despite the individual dimensions being hard to interpret, translations in the embedding space have meaningful semantic implications.
- *Additive semantics*: certain semantic relationships can be represented as additive vectors.
- *Analogies*: word embeddings can encode analogies between words.
- *Discovering relationships*: embeddings can uncover various relationships between words, such as synonyms, antonyms, and other semantic connections, purely based on how words co-occur in text. These relationships are often encoded in the geometry of the embedding space.

The low-dimensional nature of word embeddings means that semantically similar terms tend to have similar representations in the vector space. This allows the model to generalize better from semantically related examples, improving its ability to handle unseen data. Moreover, embeddings place similar concepts close together in the vector space, making them useful for discovering implied (but unknown) properties of concepts.

### 4.2.3 Word2Vec

Word2Vec, developed in 2013 by Mikolov et al., is one of the most popular and influential word embedding models. It builds on early work by Bengio et al. in 2003 and was later followed by GloVe (2014) by Pennington et al. Word2Vec solved the problem of the large parameter space in traditional models by using a bag-of-words representation.

There are two main versions of Word2Vec:



- *Continuous Bag Of Words*: this version is trained to predict an observed word based on the surrounding context words. The context consists of all terms occurring within a symmetric window around the target word. The model predicts the target word  $c$  given the observed words  $w$  by applying a softmax function over the dot product of the word embeddings. Directly optimizing this model is computationally expensive because it requires summing over all possible target words. To address this, Mikolov et al. introduced negative sampling, where they sample some negative examples (words that were not observed) and turn the problem into a binary classification task. This greatly reduces the computational load. Although on modern GPUs, optimizing softmax directly is no longer a problem.
- *Skip-gram*: in this version, the model is trained to predict the observed words given a single context word. It works in exactly the same way as CBOW, but here the context is the sum of the word vectors around the target word.

In summary, Skip-gram is a 1-to-1 prediction model, while CBOW is a many-to-1 prediction model.

Word embeddings from Word2Vec can be viewed as a form of matrix decomposition. The model uses a square co-occurrence matrix, where each cell in the matrix represents the co-occurrence of a pair of words within a fixed-size context window. By factorizing this matrix, Word2Vec generalizes the information from these co-occurrence windows to produce the word embeddings.

**GloVe** GloVe (Global Vectors for Word Representation) is another model for generating word embeddings. It offers a probabilistic interpretation for the translation of words in the embedding space. The training objective of GloVe is formulated as fitting an objective function that is approximated by minimizing a weighted least squares objective, with several tricks required to ensure convergence. GloVe improves upon Word2Vec by focusing on global co-occurrence information, rather than just local context windows, making it a more globally-informed method for generating word embeddings.

**Word2Vec and GloVe** Word2Vec is typically faster to train, has lower memory requirements, and produces more accurate models, especially with the skip-gram approach. GloVe focuses more on capturing global word co-occurrences, making it different from the local context modeling of Word2Vec. According to Levy et al., skip-gram tends to outperform CBOW, except for when using FastText (a variant of Word2Vec).

#### 4.2.4 FastText

Word embeddings work well when the vocabulary is fixed, but they face challenges when dealing with new or unseen words in the test set. If a word is not present in the trained model's vocabulary, we don't have an embedding for it and, traditionally, would have to ignore it. This is problematic, especially since we can often infer the meaning of the word from the letters or characters contained within it.

FastText, introduced by Bojanowski et al. in 2016, provides an elegant solution to this problem. Instead of learning embeddings for entire words directly, FastText splits words into smaller fixed-length character sequences (subword units), specifically character  $n$ -grams. This allows FastText to learn embeddings for these subword units, and then combine these embeddings to

form the representation of the entire word. FastText, therefore, is a powerful extension of traditional word embeddings, as it can deal with the dynamic nature of language and is particularly well-suited to languages with rich morphology.

### 4.2.5 Usage

In causal models, the context words are restricted to those that occur before the missing word, which makes these models suitable for language modeling. In this setup, the model predicts the next word in a sequence based on the preceding words, allowing it to capture longer dependencies than traditional  $n$ -gram models. This capability enables a more sophisticated understanding of word sequences and improves the quality of text generation and prediction tasks.

In non-causal models, the context can include both previous and future words. These models are often used to generate word embeddings that can serve as additional feature vectors to represent words. The embeddings can significantly improve performance across a wide range of tasks, such as:

- Training classifiers for sentiment analysis or other classification tasks.
- Machine translation, where embeddings are used to translate text from one language to another by leveraging the semantic meaning of words.

By incorporating semantic knowledge from the context, these models benefit from the rich relationships between words, improving the quality of predictions and translations.