Natural Language Processing (NLP)

Fundamentals of Data Science

# Introduction to Text Analytics

Why text analytics and why now? This is driven by a “perfect storm” of two factors:

1. Over 80% of all digitized information is stored in unstructured textual format.
2. Recent advances in data science and deep learning have opened up new possibilities in computer vision and natural language processing.

Natural language processing helps data scientists unlock insights from free-form text (usually large amounts of it). There are numerous applications of NLP including the following examples:

* Analyzing consumer feedback or customer experience by capturing information from publicly available social media sources, blogs, comments, and Twitter to understand and improve brand image
* Analyzing customer surveys and feedback cards to capture the level of customer satisfaction with a product or service (e.g., links to the surveys we frequently see printed at the bottom of cash register checkout receipts)
* Capturing and summarizing doctors’ notes and patient medical records in preparation for office visits or telemedicine appointments
* Mining contracts, which are written in free-form text (and frequently in multiple languages) and contain information ripe for analytics (e.g., dollar amounts, dates, and terms-and-conditions clauses)
* Researching new breakthrough technologies by mining information from peer-reviewed journals and scientific publications to identify new, emerging technologies that can reveal candidate companies for investment or mergers and acquisition efforts
* Carrying out human resources (HR) tasks such as evaluating employee engagement surveys (with the goal of improving productivity and reducing turnover) and parsing job applicants’ résumés and cover letters.
* Parsing through large volumes of legal documents to determine applicable law and identify relevant cases (thereby reducing legal fees and providing unbiased, more consistent results across multiple languages).
* In elections and political campaigns, identifying areas of concern for voters by looking at a large volume of news, blogs, and posts (with the goal of creating messages that resonate with a target electorate).

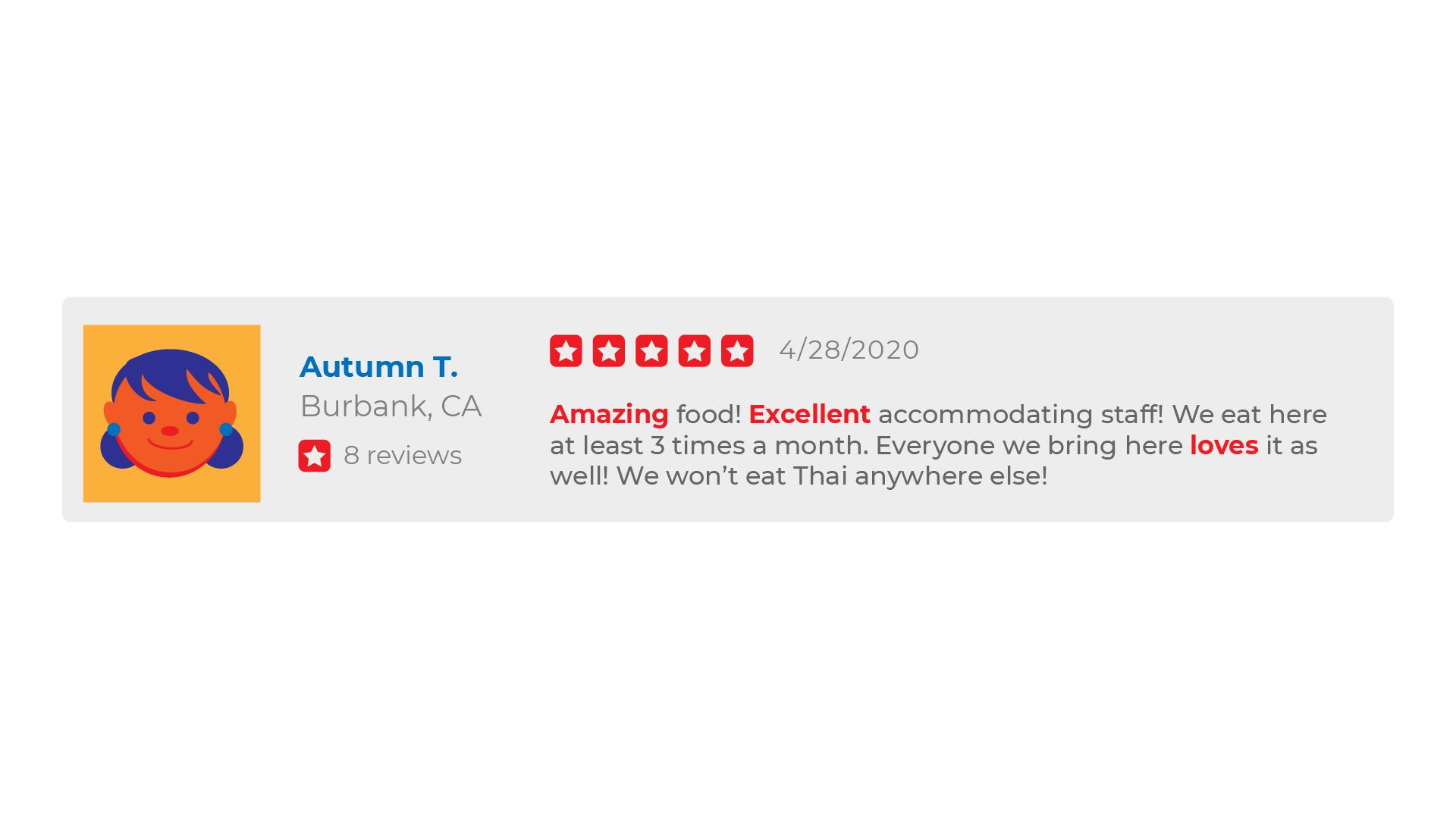
In traditional data mining applications, you can either train the model on target variables (supervised modeling) or let the model find natural patterns in the data (i.e. unsupervised clustering). The same concepts apply to NLP problems, to which we can apply both supervised and unsupervised machine learning approaches. You can build NLP classifiers where the target variable can be topic, sentiment, etc., or you can build NLP clusters such as topic modeling, document/sentence clustering, etc.

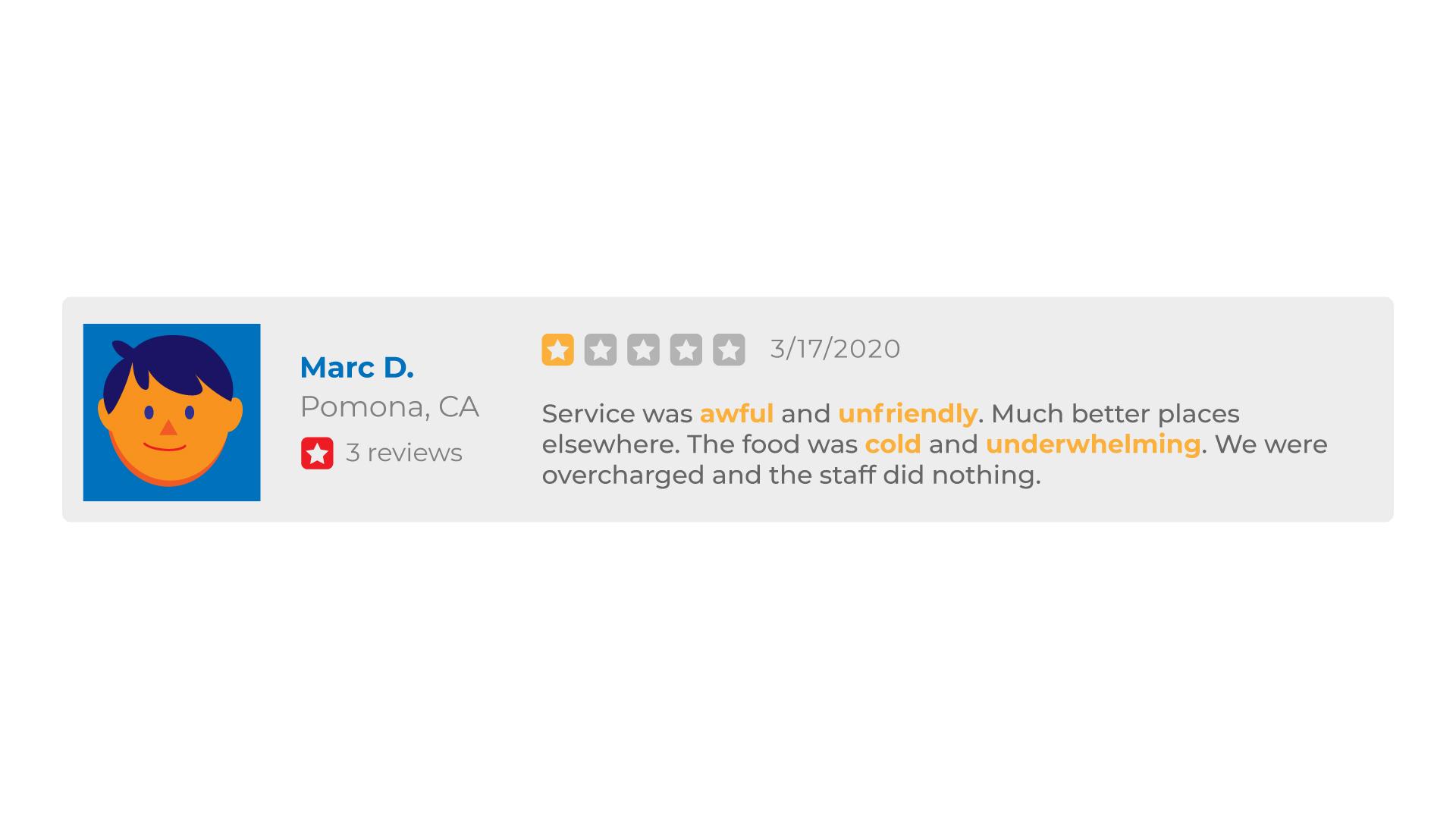
The most common NLP algorithms are:

* **Sentiment analysis:** Identify either positive or negative attitudes toward a product, organization, or person. This can be both *supervised* and *unsupervised*.
* **Text summarization:** Reduce large amounts of text to a few sentences/paragraphs keeping the most salient information. This is usually *unsupervised*.
* **Topic modeling and keyword extraction:** Extract the most important keywords/topics and measure their relevance within the context of the text. This is usually *unsupervised*
* **Text classification:** Assign topics or genres to subject categories; detect spam; identify authorship; identify language; etc. This is the broadest category of NLP and is usually *supervised*.

Sentiment Analysis

The most common and perhaps the most popular application of NLP is sentiment analysis, which is used to determine whether textual content (such as a movie review, customer comment, or even a tweet) is positive or negative. The results of sentiment analysis models generally correspond well to typical “human” reviews (Yelp, Tripadvisor, Google reviews, Amazon product reviews, IMDb movies reviews, etc.) where a one-star review is typically negative (strong negative sentiment) and a five-star review is positive (strong positive sentiment). Three-star ratings correspond to neutral sentiment.





There are two major types of sentiment models: *dictionary-based sentiment* and *supervised-learning-based sentiment*, each having its pros and cons. *Dictionary models* are based on the positive or negative sentiment values associated with each word or token. One of the most popular dictionary-based models (for English) is based on the [Loughran-McDonald Sentiment Word Lists](https://sraf.nd.edu/textual-analysis/resources/), where the list of sentiment words is arranged by category (negative, positive, uncertain, litigious, strong modal, weak modal, and constraining). This model is easy to apply and understand, and works relatively well in financial industry sentiment analysis (i.e. tracking the sentiment in financial markets). However, these models do not understand words associated with negotiation, sarcasm, or even many basic contexts. “I loved this phone… not” is an example of an expression in which a dictionary-based analysis will certainly fail.

*Supervised models*require large volumes of labeled training data (at least 10,000 examples to train a stable model) and can be trained to recognize context and long-term dependencies between the words. While trained classifiers (e.g., trained on IMDB movie reviews) are readily available and work very well within their domains, these models are not generalizable, which means you cannot take a sentiment model trained on movies and apply it to a banking customer satisfaction survey.

# Social Media Analytics

Social media analytics (SMA) is a very popular application of NLP. While there is no distinction between the application of NLP algorithms to SMA or to survey mining, there *are* differences based on the sources of the data. While data scientists can always create their own custom web crawlers, the need to bring data in from more than one source (such as a number of different social media sites) complicates the process. This is where commercial social media aggregators such as Gnip or Sprinklr can provide quick access to ready-to-analyze data (for a fee).

Typically, SMA measures a combination of volume and sentiment associated with customer discussions. The volume of social media posts is usually referred to as “buzz” and data scientists are interested in both the measure and share of buzz:

* **Measure of buzz (overall number of messages):** How much or frequently customers are talking about a particular product or brand
* **Share of buzz (number of messages for your product/brand vs. the competition):** How much of the discussion is about your product vs. your competitor’s product

A successful SMA implementation is not limited to just effective text processing or the ability to structure customer feedback—or even to obtaining accurate sentiment measurements. Augmenting your data with other sources of information—such as market research, sales, traffic, weather patterns, etc.—further amplifies SMA’s effectiveness. SMA also must be overlaid with a variety of traditional, legacy sources of information, including customer sales data, call center activity, and traditional market research. SMA is most effective when used over time to gain insights into how the “buzz” is changing in response to what is happening in the marketplace along with the results of competitive actions.

Another dimension of SMA that deals with the interconnectedness of people on social media is called *social network analysis*. The goal is to automatically discover linkages among social network users along with who are the leaders, followers, and central authorities.

# Topic Modeling

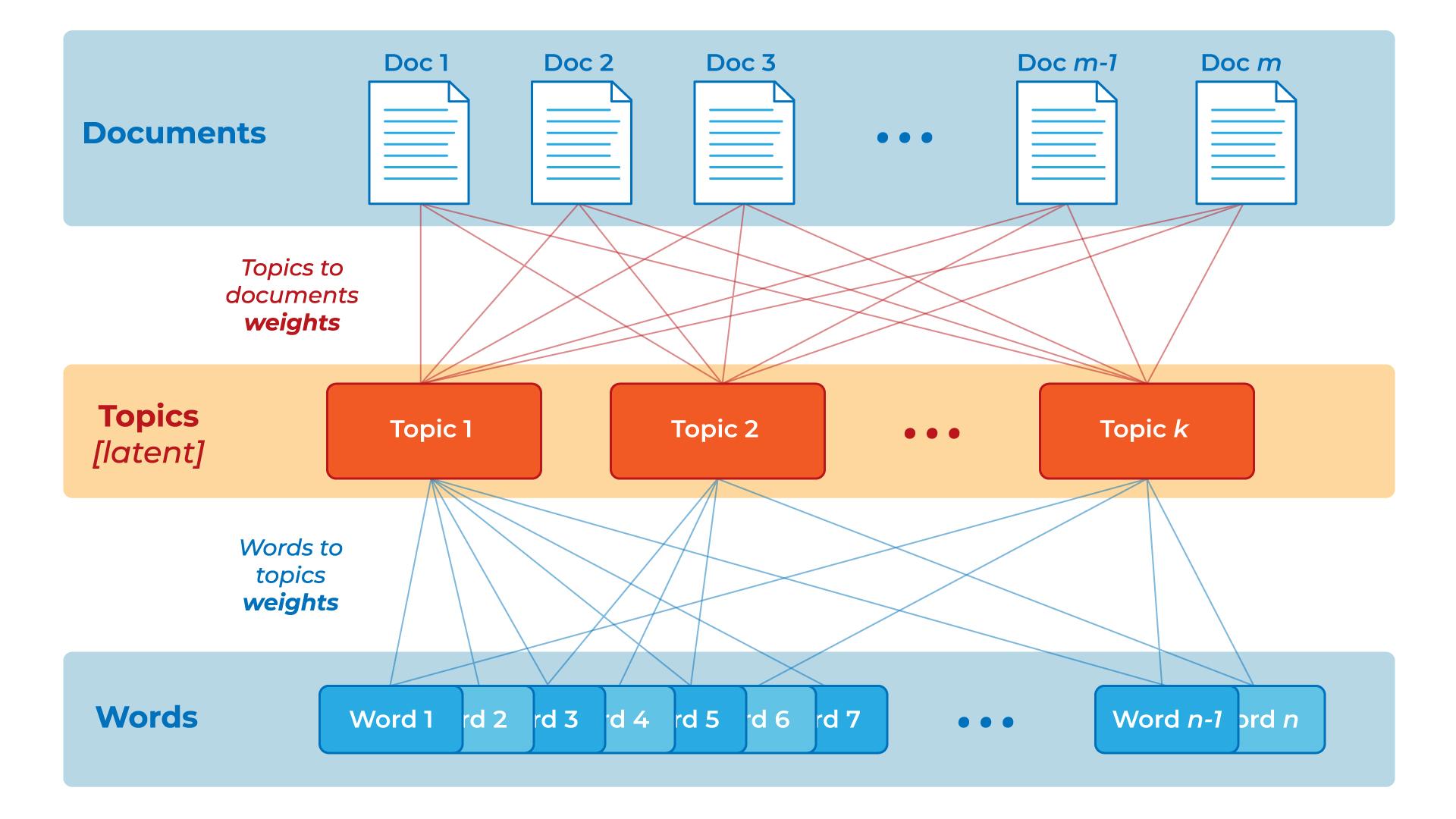
*Topic modeling*, also known as *topic detection*, is a suite of probabilistic machine learning algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. The common assumptions behind all topic modeling methods are that (1) each document consists of a mix of topics and (2) each topic consists of a collection of words/terms. The topics are “hidden” or “latent” constructs in between documents and words. The goal of topic modeling is to discover these latent variables (i.e., topics) that shape the meaning and semantics in the document collection.

## Latent Dirichlet Allocation

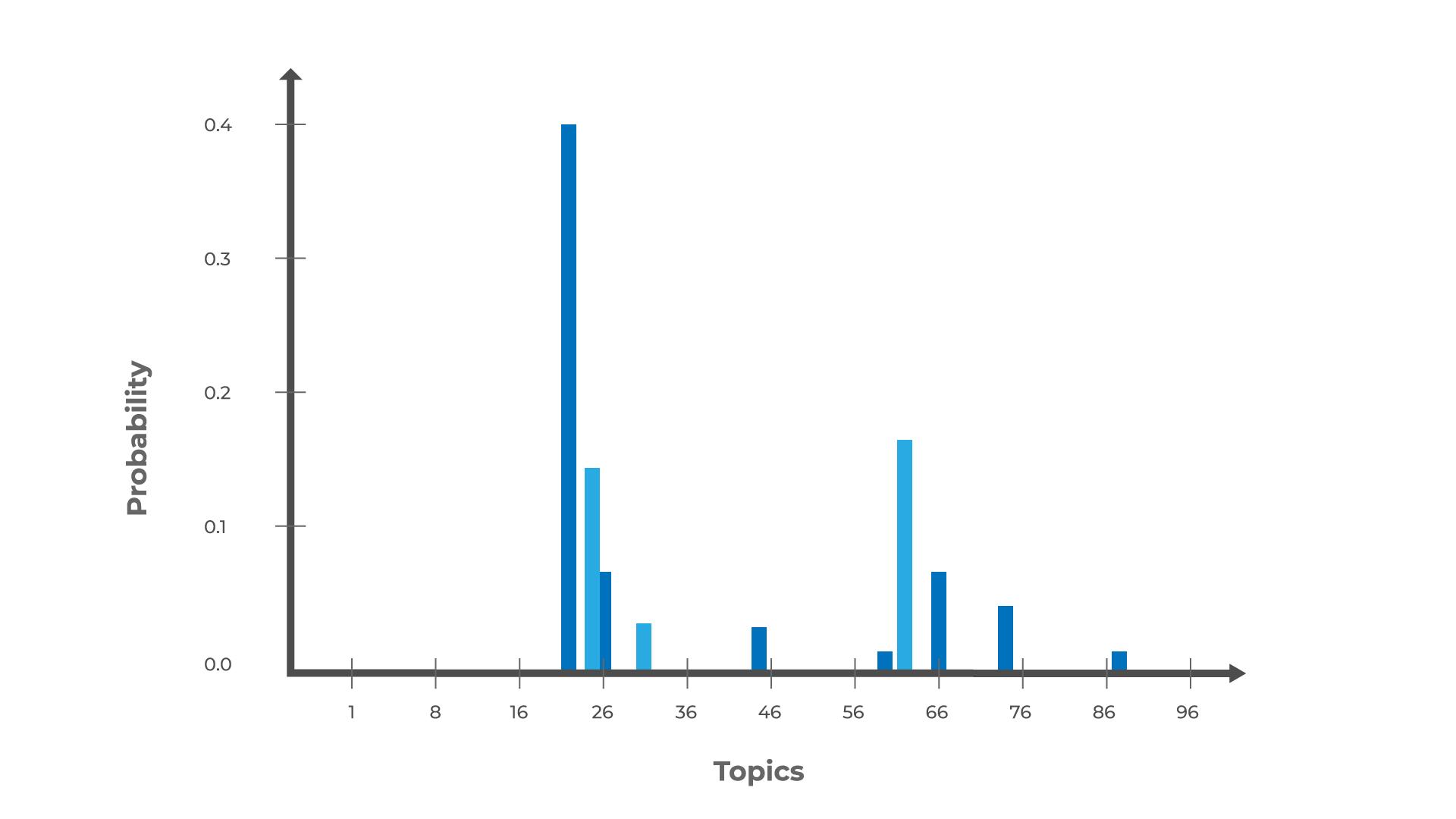
Latent Dirichlet allocation (LDA) is the most popular and perhaps the most effective topic detection technique currently in use. While there have been other recent attempts to develop newer models that can detect latent topics more effectively, the vast majority of current practices involve LDA. (Alternative models involve deep learning—RNN/LSTM-type neural network architectures—or WordEmbedding/Word2Vec-type approaches.)

LDA uses Dirichlet priors/distributions for the document-to-topic and topic-to-term associations/allocations. Dirichlet distribution, the core function of LDA, is a family of continuous multivariate probability distributions parameterized by a vector α of positive reals. It is, in fact, a multivariate generalization of the beta distribution, and hence it is also called multivariate beta distribution.

LDA is a generative probabilistic model that employs an unsupervised learning process—given a set of training data, it aims to identify the underlying distribution by generating samples from the same distribution. At the highest level, LDA portrays a three-level hierarchical probabilistic allocation model, in which each document is modeled as a weighted mixture of the underlying set of topics and each topic is, in turn, modeled as a weighted mixture of the underlying set of terms. The following figure illustrates the hierarchical structure of LDA.



The algorithmic details of LDA can be found in Blei et al. (2003), which is considered the seminal work for the LDA algorithm. In order to further explain the internal structure and superior applicability of LDA, the lead author of this article, David M. Blei, a professor at UC Berkeley, presented several illustrative examples in his CACM review article (Blei, 2012). In one of those examples, Blei fit a 100-topic LDA model to 17,000 articles from the popular journal, *Science*. The following figure shows the results of this study, where the inferred topic proportions for the example is shown on the left while the top 15 frequent (i.e., weighted) words from the most prevalent four latent topics are given in the table below.



| **Genetics** | **Evolution** | **Disease** | **Computers** |
| --- | --- | --- | --- |
| human | evolution | disease | computer |
| genome | evolutionary | host | models |
| dna | species | bacteria | information |
| genetic | organisms | diseases | data |
| genes | life | resistance | computers |
| sequence | origin | bacterial | system |
| gene | biology | new | network |
| molecular | groups | strains | systems |
| sequencing | phylogenetic | control | model |
| map | living | infectious | parallel |
| information | diversity | malaria | methods |
| genetics | group | parasite | networks |
| mapping | new | parasites | software |
| project | two | united | new |
| sequences | common | tuberculosis | simulations |

# Natural Language Processing (NLP) Functions

The first step in most NLP applications is called *information extraction* (IE) or *numerisizing* the textual content. (Textual content is also called a *corpus.*) This is where data scientists convert unstructured text into structured data that can be put into a table with columns and rows, and used for pattern identification and knowledge discovery tasks.

Suppose that the city of Pleasantville is running a campaign to reduce drunk driving. In order to deploy proactive policing methods, city leaders would like to identify where and when most DUI arrests have taken place by mining free-form text appearing in police reports. They must look for:

* Type of offense (to identify DUIs)
* Place and time of the offense (to deploy squad cars proactively)
* Names, addresses, and vehicle makes/models (to more effectively profile the offenders)

In order to extract this information, data scientists need to apply multiple information retrieval techniques:

* Segment and tokenize the text (into sentences, words, etc.).
* Apply parts-of-speech tagging to identify the linguistic roles of the words and the relationships between the words/tokens.
* Apply stemming or lemmatization to bring the tokens into their simplest word forms.
* Conduct *named entity recognition* (NER) to identify geographical locations, people, organizations, dates, times, etc.
* Run *semantic similarity analytics* to identify common synonyms and related tokens.
* Run *relation extraction* to understand relations between specific entities such as located-in, founded-by, cures, etc.

Text mining has its own language with many technical terms and acronyms. Following is a list of some of the terms and concepts that are commonly used in text mining:

## Corpus

In linguistics, a corpus (plural corpora) is a large, structured set of texts (now usually stored and processed electronically), prepared for the purpose of conducting knowledge discovery.

## Stemming

Stemming is the process of reducing inflected words to their stem (or base or root) forms. For instance, *stemmer*, *stemming*, and *stemmed* are all based on the root *stem*.

## Lemmatization

Similar to stemming, lemmatization is also used to convert inflected words to their stem/root forms. While stemming does this syntactically, lemmatization does it semantically using a term dictionary.

## Stop Words

Stop words (or noise words) are words that are filtered out prior to or after processing of natural language data (i.e., text). Even though there is no universally accepted list of stop words, most natural language processing tools use a list that includes articles (a, am, the, of, etc.), auxiliary verbs (is, are, was, were, etc.), and context-specific words that are deemed not to have differentiating value.

## Tokenizing

A token is a categorized block of text in a sentence. The block of text corresponding to the token is categorized according to the function it performs. This assignment of meaning to blocks of text is known as tokenizing. A token can look like anything; it just needs to be a useful part of the structured text.

N-grams

Sometimes a single word may not convey the actual meaning. Instead, its meaning is embedded in a multi-word combination. For instance, the United States or Federal Bureau of Investigation are only meaningful if they are used in combination of two and three words, respectively. In natural language processing, frequently occurring n-word combinations are identified and then these n-grams are treated as unique terms for further processing.

## Part-of-Speech Tagging

This is the process of marking up the words (or n-grams) in a text that correspond to a particular part of speech (e.g., nouns, verbs, adjectives, adverbs), based on each word’s definition and the context in which it is used.

## Morphology

Morphology is a branch of the field of linguistics and a part of natural language processing that studies the internal structure of words (patterns of word formation within a language or across languages).

## Term-by-Document Matrix (Occurrence Matrix)

This type of matrix is a common representation schema for the frequency-based relationship between terms and documents in tabular format, where terms are listed in rows, documents are listed in columns, and the frequency between the terms and documents is listed in cells as integer values.

## Singular-Value Decomposition

This dimensionality reduction method is used to transform a term-by-document matrix to a manageable size by generating an intermediate representation of the frequencies using a matrix manipulation method similar to principal-component analysis.

# References

Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Blei, D., Ng, A., & Jordan, M. (2003). [Latent Dirichlet Allocation](https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf).