DMIF, University of Udine

Text Mining

Andrea Brunello
andrea.brunello@uniud.it

May, 2020

Introduction

Text mining can be defined as:

The process of extracting new, previously unknown information from different written resources.

Written sources can be, for instance, websites, books, e-mails, reviews, articles.

Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluating and interpreting of the output.

In this scope, *text* is defined as a kind of unstructured data consisting of a series of paragraphs, each composed of one or more sentences, each made by strings called words.

Sentences start with a capital letter and end with a full stop. A sentence may have one or several *clauses*, that can be joined to one another by conjunctions or by relative pronouns.

A *paragraph* is an ordered list of sentences consistent with a particular subtopic. Paragraphs starts with an indentation and end with a carriage return.

A *word* is considered as the basic text unit. Nevertheless, sometimes words *go together* , for instance: "as well as", "in contrast", the phrasal verbs, . . .

Text may be stored and processed using different formats, which increases the difficulty of information mining tasks:

- MS Word
- MS Powepoint
- PDF Adobe
- XML
- HTML
- Plain text

Text Preprocessing

Typical text mining tasks include:

- Visualization

- Classification

- Clustering

- Tagging/annotation

- Concept/entity extraction

- Sentiment analysis

- Document summarization

However, typically, before any of these tasks, some deal of preprocessing has to be performed over the text.

Since text is a kind of unstructured data, it is difficult to analyze it in its original form.

To apply classical analysis approaches, we have to describe it by means of a fixed number of attributes.

In this sense, the first step is that of performing *text indexing*, that is, converting texts into a list of words. Then, we have to devise a way to assign an importance score to each word for each text the words occurs, a task that is named *term weighting*.
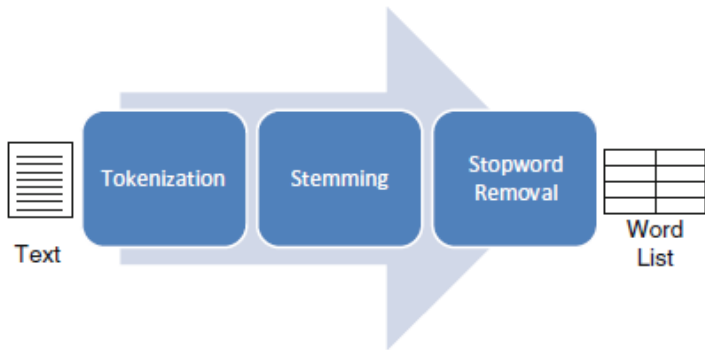
The tokenization process involves segmenting text into tokens.

Tokens are identified by looking at white spaces, punctuation and special characters.

Words with one ore more special characters may be discarded entirely, or just the special characters can be removed (e.g., 25%).

Words are transformed to lowercase, so to have a unique representation for each word.

Text categorization refers to the process of assign a category or some categories among predefined ones to each document, automatically. Text categorization is a pattern classification task for text mining and necessary for efficient management of textual information systems.

Tokenization

Tokens
text
categorization
refers
to
the
process
of
assign
a
category
....

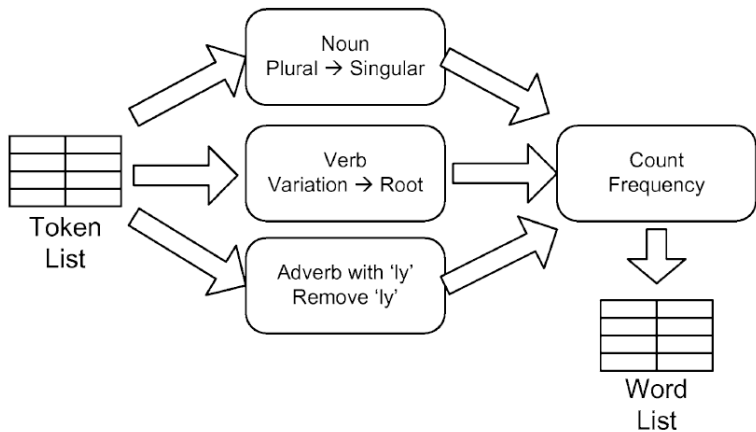Stemming is the process of mapping each token obtained from the previous step into its root form.

Stemming is usually applicable to nouns, verbs, and adjectives.

The output of this process is a list of words in their root forms.

Language-dependent stemming rules are required for this process (e.g., Porter algorithm).

Varied Form	Root Form
better	good
best	good
simpler	simple
simplest	simple
assigning	assign
assigned	assign
assignment	assign
complexity	complex
analysis	analyze
categorization	categorize
categorizing	categorize
categorizes	categorize

Stop word removal is the process of removing stop words from the list of tokens.

Stop words are the most common words in a language, though there is no single universal consensus on them.

Some examples are prepositions, such as 'in', 'on', 'to', and so on. Conjunctions such as 'and', 'or', 'but', and 'however'.

Intuitively, one may want to remove stop words since they do not bring any information, being far too common.

Caveat: in some cases, stop words may actually be important. For instance, consider the English rock band *The Who*. Or, the phrase *I told you that she was not happy*, which may be left with just ['told', 'happy'].

Dealing with single words has a major drawback: it is not possible to keep into account the word order and context.

For instance, consider the two phrases, which may look that same if just the set of single words is considered:

- *It seems you were right not inviting him*
- *It seems you were not right inviting him*

An *n-gram* is a contiguous sequence of *n* items from a given sample of text or speech.

By means of n-grams, it is possible to keep better track of words contexts, or to consider to phrasal verbs, such as 'get around', 'pass out', 'carry on'.

Full sentence	It does not, however, control whether an exaction is within Congress's power to tax.
Unigrams	"It"; "does"; "not,"; "however,"; "control"; "whether"; "an"; "exaction"; "is"; "within"; "Congress's"; "power"; "to"; "tax."
Bigrams	"It does"; "does not,"; "not, however,"; "however, control"; "control whether"; "whether an"; "an exaction"; "exaction is"; "is within"; "within Congress's"; "Congress's power"; "power to"; "to tax."
Trigrams	"It does not"; "does not, however"; "not, however, control"; "however, control whether"; "control whether an"; "whether an exaction"; "an exaction is"; "exaction is within"; "is within Congress's"; "within Congress's power"; "Congress's power to"; "power to tax."

Term weighting refers to the process of calculating and assigning a weight to each word (or, in general, n-gram) for each text in which the words occurs, reflecting its importance degree.

Such an importance is typically bound to the term frequency, which can be declined in two different ways:

- *absolute term frequency*: the number of occurrences of the given word in a text

- *relative term frequency*: the ratio of word occurrences with respect to the maximal ones or the total number of terms in the text

Just relying on absolute/relative word frequencies is not enough.

Given a set of documents from the same domain, some words are naturally more frequent than others (e.g., *money* in a financial domain).

Despite their higher frequency, they might not be very informative, because they are common (think of them as domain-related stop words).

TF-IDF statistic tries to cope with such an issue.

Term frequency–inverse document frequency (TFIDF) is a numerical statistic that is intended to reflect how important a word is to a document in a collection.

Its value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word.

Given a word t in the document d, its weight w_{td} is given by:

$$W_{td} = log(\frac{N}{DF_t})TF_{td}$$

where N is the total number of documents in the collection, DF_t is the number of documents which include the word t, and TF_{td} is the number of occurrences of the word t in the document d.

Frequency of term in a large set of documents — Frequency of term on a single page

Common stop words. Low TF-IDF → The, and, because

Less frequent terms earn higher TF-IDF with increased usage → car, remained

Terms with the highest TF-IDF may indicate importance → Auto repair / auto repair
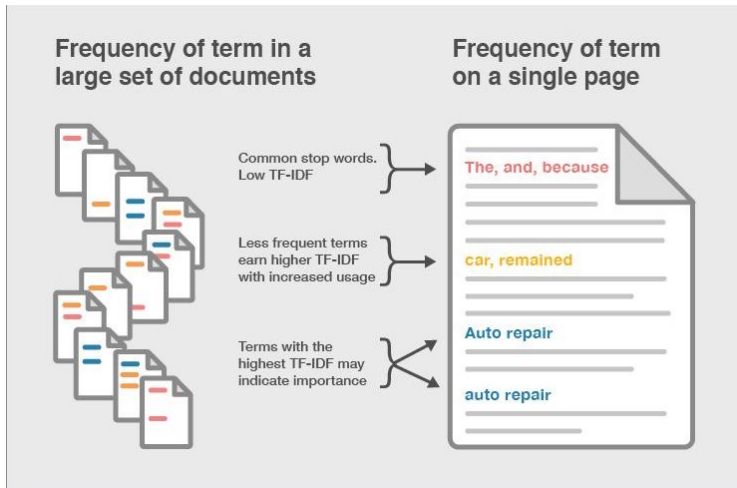
Text Mining Tasks

Text visualization is an essential task to perform *exploratory data analysis*, i.e., getting to know your data before delving into more complex analysis tasks.

Also, visualizing text you can intuitively, by glancing at them:

- understand the overall content of a document
- group documents
- compare documents

Some kinds of visualization techniques are rather intuitive (e.g., plot the histogram of word weights), some other are more complex (clouds, trees, . . .).

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We wer losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of yo

Word clouds show the text word count in an intuitive way.

Bill Clinton 1993

Barack Obama 2009

Andrea Brunello

They show all phrases starting with a given sequence of words.

12
hits

let up until those americans who seek jobs can find them - - (applause) - - until th

back down on the basic principle that if americans can't find affordable coverage , w

a plan that adds one dime to our deficits - - either now or in the future

sign

it if it adds one dime to the deficit , now or in the future , period .

not

make that same mistake with health care .

waste time with those who have made the calculation that it's better politics to kill th

- - and i will not accept the status quo as a solution .

accept the status quo as a solution .

i will

make sure that no government bureaucrat or insurance company bureaucrat gets between you and

protect medicare .

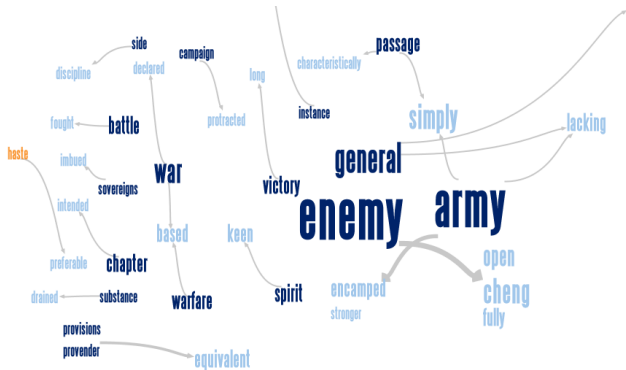**continue** to seek common ground in the weeks ahead .

be there to listen .

It may also be interesting to consider pairs of words that are connected through some other word. For instance, this is the result of looking at words connected by the term *is* in *Sun Tzu's The Art of War*.

Classification / Regression

Classification is defined as the process of assigning one or more categories among the predefined ones to each data item.

In *regression* tasks, instead, we assign a number to each data item (the domain is continuous).

Considering text, one may want to assign a category to each document, for instance, the list of topics discussed in it; or, we may want to determine the text's overall sentiment score.

A labelled training dataset is typically needed to train suitable classification or regression models.

If such a dataset is not available, one may try to rely on clustering first.

Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

Clustering involves the definition of a suitable metric to evaluate the distance between the objects.

There are two main categories of clustering:

- *Hard clustering*, where each item is assigned to only one cluster
- *Soft clustering*, here each item may belong to more than one cluster (e.g., degree of membership)

Taeho Jo Text Mining Concepts, Implementation, and Big Data Challenge, Volume 45, Springer