INTRODUCTION TO TEXT MINING

Marcus Birkenkrahe

January 6, 2023

README

You will learn:

- ☐ The basic definition of practical text mining
- ☐ Why text mining is important to the modern enterprise
- ☐ Examples of text mining used in enterprise
- ☐ The challenges facing text mining
- ☐ Example workflow for processing natural language
- ☐ A simple text mining example
- ☐ When text mining is appropriate

Source: Kwartler, 2019, chapter 1

What is "text mining"?

> "Text mining is the process of distilling actionable insights from text." Kwartler, 2019

☐ What does "distilling actionable insights" mean?[1]

> "Text [data] mining is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights." IBM, 2023

☐ What does "structured" and "unstructured data" mean? What are "meaningful patterns"?[2]

☐ What about the text on this page: structured or not?[3]

What is text mining in practice (= business)?

- Identify useful social media posts for customer services
- Measure campaign success for marketing purposes
- Match job descriptions to resumes for human resources

Why should you care about text mining?

> " To be truly customer-centric in a hyper-competitive environment, an organization should be listening to their constituents whenever possible. Yet the amount of textual information from these interactions can be immense, so text mining offers a way to extract insights quickly." Kwartler, 2019

Concrete examples:

[1] Distillation is a process of extracting an essence (a wanted substance) and getting rid of unwanted substances. Actionable insights are insights that one can use to make decisions (action in business is usually accompanied by decision-making).

[2] (Source) Structured data are data in tabular format with specific data types for digital processing. Unstructured data do not have a specific data format.

[3] The Org-mode file is semi-structured! Semi-structured data carry meta information in the form of markup - e.g. HTML, XML, JSON, or Org-mode: the header information at the top of the file structures the data, as does the Org-mode format itself, which comes with a markup language.

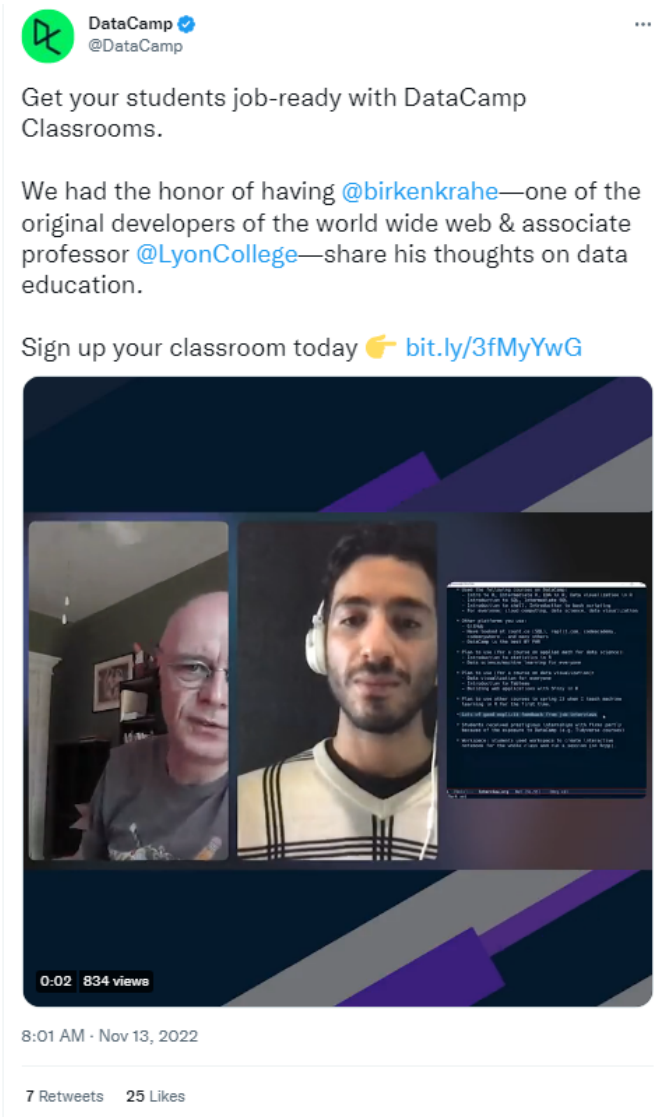


Figure 1: twitter.com/DataCamp featuring interview with me

Figure 2: twitter.com/birkenkrahe campaigning for fsf.org

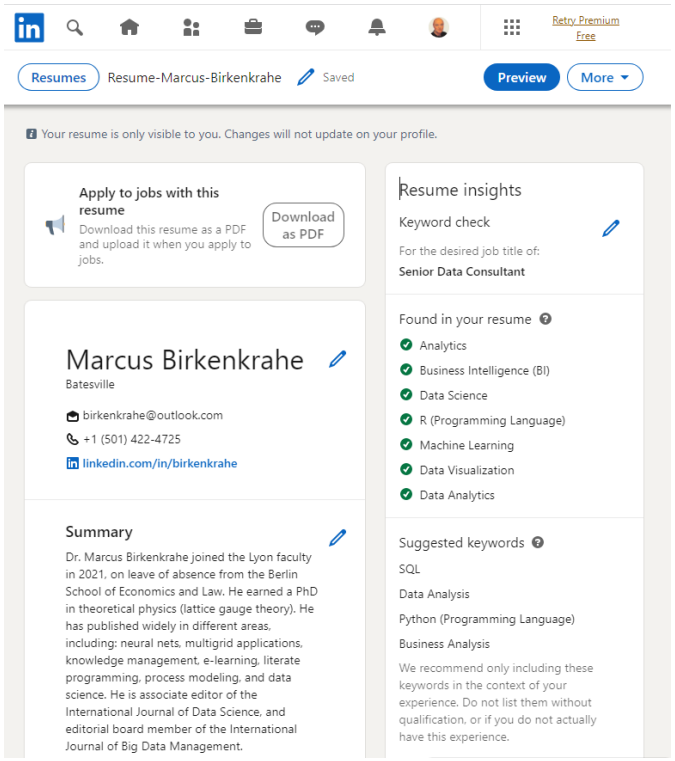


Figure 3: automatic resume built by LinkedIn.com

- Relevance of social media for public opinion (e.g. Twitter)
- Growth of online content from an organization (e.g. blogs)
- Digitization of paper records (e.g. healthcare)
- Automatic translation of natural language (e.g. Google Translate)
- Augmentation of human work through chatbots (e.g. ChatGPT)

Vox populi - the "wisdom of crowds"

In 1907, Francis Galton published a paper exploring whether a crowd could be intelligent. His analysis used entries submitted to a contest held at an agricultural fair in Plymouth, England [1]. The participants, many of whom were butchers and farmers, competed at judging how much a "fat ox" would weigh after it had been slaughtered and dressed. They wrote their estimates on cards, which were later loaned to Galton for analysis. There were 787 eligible entries. The median value of these estimates (what Galton called the "middlemost estimate") was 1207 pounds, amazingly close to the actual weight of the dressed ox: 1198 pounds. Although the individual estimates varied widely, they had a central tendency very close to the true value. Even though the contestants made errors, some guessing too high and others guessing too low, the "vox populi" seemed to produce an estimate that was better than what could be expected from any individual expert.

Figure 4: Source: Patten, 2015

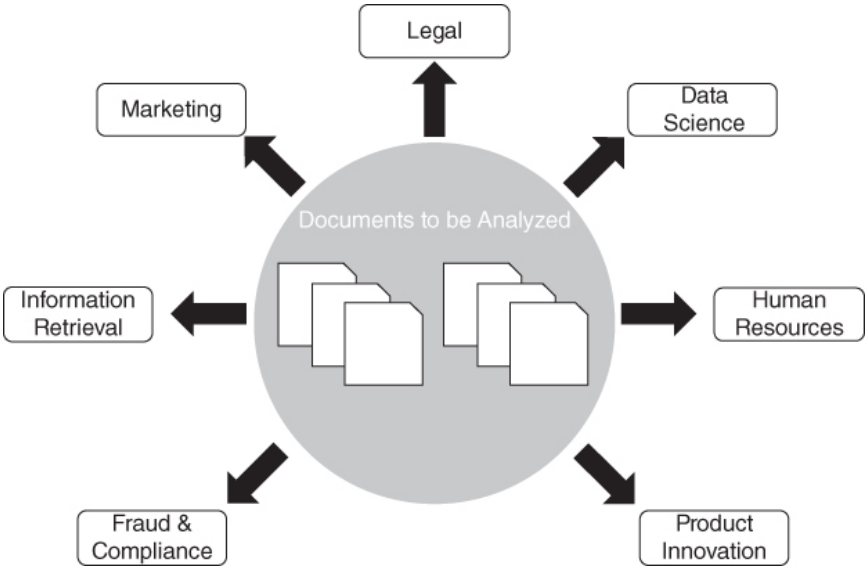
> "Under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them." Surowiecki, 2005

The "right circumstances": no assessment bias =

1. Assessors need to exercise *independent* judgements
2. Assessors need to possess *diverse* information understanding
3. Assessors need to rely on *decentralized*, *local* knowledge.
4. There has to be a way to *aggregate* or tabulate the results.

5. ☐ How about Amazon.com reviews - do they meet these conditions?[4]

Beneficiaries and benefits of text mining



- Benefits include:
  1. Trust among stakeholders because little to no *sampling* is needed to extract information (all available text sources can be used).
  2. The methodologies can be applied quickly (text processes fast).
  3. Using R allows for *auditable* and *repeatable* methods.
  4. Text mining identifies novel *insights* or reinforces existing perceptions based on all relevant information.
- The "opinion" of ChatGPT looks comprehensive as always - does this chatbot represent "vox populi"? Are all criteria fulfilled?[5]

[4](1) reviews may not be independent since reviewers have access to old reviews, which may influence them (it's harder to have a different opinion from everyone else). (2) Diversity is hard to measure but in the case of Amazon.com, a national audience can be seen as highly diverse (there are nearly 150 mio subscribers of Amazon Prime in the US alone). (3) Local here means "not only at a distance" - only "verified purchase" reviews fulfil this condition in principle. (4) Tabulation of the reviews relies on text mining, and hence - unlike in the case of Galton - not on recording simple numbers. Stochastic procedures (probability distributions) are involved.

[5]ChatGPT is source from a very large number of textual documents but it is impossible to ascertain any of the criteria when identfying the chatbot as the "assessor".

☐ Whom would you trust more - the expert author or the chatbot?[6]

When to use and when not to use text mining

Example use case	Recommendation
Survey texts	Explore topics using various methods to gain a respondent's perspective.
Reviewing a small number of documents	Don't perform text mining on an extremely small corpus, as the results and conclusion can be skewed.
Human resource documents	Tread carefully; text mining may yield insights, but the data and legal barriers may make the analysis inappropriate.
Social media	Use text mining to collect (when allowed) from online sources and then apply preprocessing steps to extract information.
Data science predictive modeling	Text mining can yield structured inputs that could be useful in machine learning efforts.
Product/service reviews	Use text mining if the number of reviews is large.
Legal proceeding	Use text mining to identify individuals and specific information.

- "Use case"

TODO Summary

TODO TM Glossary

TERM	MEANING
Text mining	Identify patterns in text
Structured data	Tabular data (rows and columns)
Semi-structured data	Markup with meta data

References

- IBM (2023). What is text mining? URL: ibm.com/topics/text-mining.
- Kwartler, T (2019). Text Mining in Practice with R. Wiley.
- Patten, S B (2015). The Wisdom of Crowds (Vox Populi) and Antidepressant Use. Clin Pract Epidemiol Ment Health (11):1-3. URL: doi.org/10.2174%2F1745017901510011001

[6] For me personally, knowledge about a source increases trust in believing that source while lack of knowledge decreases the trust. In the case of ChatGPT, I asked the bot about its sources but its answer was redundant and not overly satisfying (see for yourself).

- Surowiecki J (ed) (2005). The wisdom of crowds. New York First Anchor Books. crowds.