Hickman pre-processing recommendations

Recent advancements in text mining have opened up new opportunities for organizations to analyze and extract insights from the vast amounts of natural language text data available to them. However, the preprocessing of text data, which involves the decisions made before analyzing the data, plays a crucial role in determining the quality and validity of the results obtained from text mining. Existing studies have provided varying recommendations for text preprocessing techniques, and there is a lack of consistency in the methods used in primary studies. To address this, the authors of this paper conducted two complementary reviews of computational linguistics and organizational text mining research to develop evidence-based recommendations for text preprocessing. These recommendations take into account the type of text mining being conducted, the research question, and the characteristics of the data set. The authors also suggest guidelines for reporting text mining studies to ensure transparency and reproducibility.

**Preprocessing techniques in text mining.**

This section of the document discusses various preprocessing techniques used in text mining and their effects on text mining analyses. It also provides recommendations for when and whether to apply these techniques in closed and open vocabulary text mining.

The section covers the following preprocessing techniques:

1. Lowercase conversion: Converting all letters in a corpus to lowercase. It is generally recommended to use lowercase conversion during open vocabulary text mining as it decreases data dimensionality and usually does not reduce validity.
2. Handling negation: Negations play an important role in language as they alter the meaning of subsequent words. It is recommended to handle negation by appending "not\_" to each word it precedes, creating a new distinct unigram. This captures more semantic information, reduces errors, and increases validity in both open and closed vocabulary text mining.
3. Correcting spelling errors and expanding contractions and abbreviations: Correcting spelling errors and expanding contractions and abbreviations can reduce data dimensionality and improve textual analysis. However, caution should be exercised in certain cases where errors and idiosyncrasies carry meaningful information, such as in the analysis of individual differences.
4. Removing non-alphabetic characters: Removing punctuation, symbols, and numbers from text is commonly done to focus on words and phrases. However, exceptions may apply in cases where punctuation and non-alphabetic characters are relevant to the research question, such as in the analysis of linguistic complexity or sentiment.
5. Removing stop words: Stop words are commonly occurring words in a corpus that may become uninformative. It is recommended to remove stop words during open vocabulary text mining preprocessing, but they should be retained in closed vocabulary text mining. Exceptions may apply in cases where stop words are relevant to individual differences.

The author concludes by discussing feature extraction in text mining, including closed and open vocabulary text mining. Closed vocabulary text mining uses dictionaries to count relevant words, while open vocabulary text mining extracts n-grams from text using a tokenizer. Tokenization helps improve validity by capturing semantic information.

**Text mining preprocessing: effects, techniques, accuracy**

This section of the document discusses the different preprocessing techniques used in text mining and their effects on the accuracy of machine learning predictions. It emphasizes that certain techniques such as lowercase conversion, handling negation, spelling corrections, and expanding contractions and abbreviations tend to improve system accuracy. On the other hand, techniques such as nonalphabetic character removal, emoticon recognition, stop word removal, and stemming are found to be less beneficial. The section also reviews previous studies in computational linguistics that compare the effects of different preprocessing decisions and provides a summary of their findings. In addition, the section presents a review of preprocessing practices in organizational research that utilizes text mining, noting that only a small percentage of articles reported using preprocessing techniques.

**Preprocessing in text mining.**

This section of the document discusses the preprocessing techniques used in text mining research, specifically in the context of open vocabulary and closed vocabulary text mining. The section notes that multiple preprocessing techniques are often used in open vocabulary studies, while closed vocabulary studies rarely engage in extensive preprocessing. The section also mentions the importance of tokenization and handling negation in text mining. Additionally, the section emphasizes the need for researchers to report the software used and the rationale behind their preprocessing decisions to ensure transparency and replicability of research methods. Recommendations for preprocessing in both open and closed vocabulary text mining are provided, along with best practices for reporting text mining research.

**Limitations in unstructured data analysis and future work.**

This section of the document discusses the limitations and future work of analyzing unstructured data in organizational research. It suggests that researchers can pair open vocabulary text mining approaches with strong theoretical reasoning to derive grounded and cross-validated insights. The section also discusses the removal of sparse or frequent terms in open vocabulary preprocessing, noting that there is no consensus on the appropriate threshold for removal. It recommends that the effect of these practices should be empirically assessed. Additionally, the section mentions the use of neural network topic models in computational linguistics, which require less preprocessing. The section highlights the potential for extracting other textual features that require different preprocessing techniques and suggests exploring these features in future research. It emphasizes the importance of theoretical rationale in guiding preprocessing decisions, feature extraction, and analysis. Lastly, it states that the recommendations provided aim to ensure that preprocessing decisions are empirically and theoretically justified, transparent, and replicable.

What is a word problems of tokenization

Tokenization is a crucial step in natural language processing, but it poses challenges due to the complexity of determining what should be considered a token. This document explores the methods and choices involved in tokenizing text, including recognizing punctuation, numbers, dates, and abbreviations. Regular expressions are suggested as a tool for recognizing and classifying tokens, with examples given for numbers, dollar values, and dates. The use of a lexicon is also discussed for recognizing abbreviations and their meanings. However, despite these approaches, tokenization remains a difficult problem with no perfect solutions.

* NLP: The section explains that tokenization is an important step in computational lexicography, as it prepares the text for further automated processing. The authors discuss methods for tokenizing text, including recognizing punctuation, numbers, dates, acronyms, and abbreviations using regular expressions. The authors also mention the challenges and choices involved in tokenization, such as dealing with hyphenation and recognizing word and sentence boundaries. The section concludes by acknowledging that tokenization can be a non-trivial problem and that the choices made during tokenization can have implications for subsequent linguistic analysis.
* Recognition, classification and abreviations: In this section of the document, the author discusses the use of regular expressions to recognize and classify different types of tokens, specifically numbers, dollar values, and dates. The author provides examples of regular expressions for each type of token. They also discuss the importance of recognizing abbreviations and the challenges associated with it. The author suggests using the structure of abbreviations to identify and classify them. They propose three classes of abbreviations based on their structure. Additionally, the author explores the use of the corpus itself as a filter to identify abbreviations and validate their recognition. The section concludes by mentioning the remaining unrecognized abbreviations and the percentage of correctly recognized sentence boundaries.
* Techniques: This section of the document discusses the use of a lexicon to recognize abbreviations and their meanings. The section presents a filter sequence that can be applied to determine whether a string terminated by a period is an abbreviation or a sentence-ending word. The section also mentions related work on sentence boundary recognition, such as using neural nets or regression analysis. Additionally, the section discusses the importance of tokenization and the challenges it presents, such as deciding whether a token corresponds to a single class or a sequence of classes. The section concludes by stating that tokenization in automated text processing systems raises difficult questions with imperfect answers.

Preprocessing

**Text Preprocessing Importance**

This section introduces the topic of text preprocessing for unsupervised learning in political science. It explains the importance of preprocessing decisions and their impact on the results of models. The authors argue that substantive theory is often too vague to guide feature selection and that advice from the supervised literature may not be applicable. To address this, they propose a statistical procedure and software to examine the sensitivity of findings under different preprocessing regimes. The section also provides information on the word count, availability of preText software, and the contact details of the authors. Additionally, it mentions the availability of replication data for the paper.

**Importance of n-grams**

This section discusses the importance of including n-grams in text analysis, particularly in cases where individual words have ambiguous meanings. It mentions that n-grams are contiguous sequences of tokens of length n and can improve the interpretability of bag-of-terms statistical analyses of text. The section also discusses the practice of removing infrequently used terms from the corpus preprocessing, as they may not contribute much information about document similarity and can reduce the size of the vocabulary. The section highlights that there are 128 different possible combinations of preprocessing choices, which can impact the final document-term matrix. Additionally, the section discusses the differences between supervised and unsupervised approaches to feature selection in text analysis, emphasizing that decisions on preprocessing steps should be guided by substantive knowledge. Finally, the section provides a brief description of the datasets used in the analyses presented in the rest of the paper.

**Preprocessing choices matter.**

This section of the document provides two examples of how the choices made during preprocessing can significantly impact the results and inferences made in unsupervised analysis.

In the first example, the Wordfish model is used to analyze the ideological positions of UK Conservative and Labour manifestos from 1983 to 1997. The author demonstrates that depending on the preprocessing steps taken, the model may yield different rank orders of the manifestos, leading to different conclusions about the ideological positions of the parties. The author points out that these different specifications can lead to significantly different interpretations of the data, potentially allowing a researcher to support a wide range of hypotheses.

In the second example, the author applies Latent Dirichlet Allocation (LDA) to analyze the Congressional Press Releases corpus. By varying the preprocessing steps, the author shows that the optimal number of topics and the interpretation of those topics can vary widely. This means that researchers may draw different substantive conclusions based on their choices in preprocessing.

Overall, the section highlights the importance of considering the effects of preprocessing choices on the results and inferences made in unsupervised analysis. It also suggests that there is little concrete guidance for researchers in making these choices, emphasizing the need for further research in this area.

**Preprocessing decisions' impact**

This section of the document discusses the importance of preprocessing decisions in topic modeling and how they can impact the results and conclusions drawn from the analysis. The researchers conducted an analysis to assess the variation in the proportion of topics that contain key terms across different preprocessing specifications. They found that there was a great deal of variation in the presence of these key terms in the topic top-20 terms. This variation could lead to different substantive conclusions about the topics being discussed depending on the preprocessing specification chosen. The researchers also noted that some key terms, such as "insurance," did not exhibit as much variation in their prevalence across preprocessing specifications. They concluded that researchers should be aware of the potential impact of preprocessing decisions on their results and conclusions. The section also introduces a method called preText for assessing sensitivity to preprocessing decisions.

**preText score analysis**

The author discusses the mean movement of the preText score, which is calculated based on different preprocessing steps applied to a corpus. The section explains how regression analysis was performed for each corpus, and the results are presented in Figure 5. The interpretation of the regression results is discussed, with a negative parameter estimate indicating that a preprocessing step reduces the preText score, while a positive parameter estimate indicates that it increases the preText score.

An illustration of the approach is provided using the Wordfish example from a previous section. The results show that the theoretically selected preprocessing specification and the averaged results produce different ordering for the Conservative manifestos, highlighting the importance of assessing the robustness of results by considering different preprocessing choices.

**Preprocessing's impact on findings**

This section of the document discusses the importance of preprocessing choices in text analysis and the potential impact on research findings. The authors argue that researchers should not arbitrarily choose preprocessing specifications and instead should be systematic and transparent in their approach. They highlight the potential for "forking paths" in interpretation of results, and suggest that researchers should consider multiple preprocessing options and compare the robustness of their findings. The section also references related work in other fields, such as psychology, that have advocated for similar approaches. The section concludes by acknowledging that their method is not exhaustive and that further research is needed to explore other preprocessing steps.

**Topic modeling evaluation.**

This section of the document includes information about topic modeling and evaluation metrics. It discusses how the log-likelihood of held out test data can be used to compare different models, with higher log-likelihood indicating a better model. The perplexity of a test set, which is a normalization of the log-likelihood, is also mentioned as a commonly used metric for evaluating topic models. The section also includes a replication of topic model results and a comparison of different preprocessing specifications in a Wordfish analysis of Irish parliamentary budget debate speeches. The results suggest that different preprocessing choices can affect the interpretation of the Wordfish scores.