SAS® GLOBAL FORUM 2021

Paper 1055-2021

Natural Language Processing—An Introduction

Colleen M. Farrelly, Staticlysm/Datasembly

ABSTRACT

Natural Language Processing (or NLP) is a growing field of machine learning concerned with understanding text data and integrating insights from text data with other data sources. Often, insights contained in text data aren't captured in relational data or other sources. Through processing and analyzing text data, these insights can be integrating wit relational data within a single model, dashboard, or other analytics tool.

INTRODUCTION

Many fields of analytics, including healthcare and law, can benefit tremendously from the integration of text data with traditional data sources, such as data captured and stored in relational databases. For instance, consider numeric data related to a patient's health. One might find blood pressure readings, height, weight, and medications prescribed for recorded conditions. However, subjective assessments by a particular physician may provide more insight—such as how the patient appears in the ER, what the patient had been doing when symptoms presented… This information can enhance predictive models by providing a context outside the patient's vitals and the boxes clicked on a patient history.

This paper elucidates some of the tools used to process text data and wrangle it into something compatible with relational databases. Some of these tools include parsing text data into pieces, tagging important features in the text, and analyzing specific word choices.

PARSING DATA

Parsing text data into individual sentences or words is often a necessary first step to wrangling and analyzing text data. Sometimes, text documents can be as long as novels, and collections of text documents can include billions of documents of that size. Big data tools, such as Spark, often provide parsing tools to help wrangle that volume of data; however, smaller volumes can often be done on a laptop with Python or R.

Tokens represent the basic building blocks of text, including words and punctuation. For instance, in the previous sentence, tokens would include each word, the comma, and the period. Punctuation styles often vary between writers, and analyzing these types of tokens can provide insight into the individual generating the text data. Words typically feed into later stages of NLP pipelines and can provide insight into topics in the document, key word usage, and parts of speech usage patterns.

Some words in text data aren't particularly important to an analytics goal, and it's possible to define stop words, which are words to strip from a list of tokens. For instance, if one is interested in certain diseases of interest, most tokens can be stripped from the token list, as they won't be relevant to a list of words associated with a disease of interest.

Because some words can be turned into different parts of speech, some applications will require text to be parsed into root words. For instance, "go" and "going" are different forms of the same verb. Some applications need to treat these as the same word rather than two separate words. Humans do this naturally, but computers need explicit instructions to associate these correctly. Stemming and lemmatizing algorithms provide these instruction sets, albeit in slightly different algorithms.

TAGGING FEATURES

Tagging features is another important processing step in NLP applications. Some applications may wish to quantify features within a document. Others may wish to recognize features occurring within a document. Others might want to parse sentences into their key components.

Many feature taggers exist. Parts of speech tagging is often important in grammar and personality analysis applications. Tokens within a document are tagged as nouns, verbs, adjectives, interjections, or other parts of speech; some taggers are specific enough to tag parts of speech as more specific parts of speech, such as parsing nouns into proper nouns or common nouns.

Clause identification and grammatical relations are common tasks within speech data, which can require filling in missing data that was not translated or transcribed. Many options exist for this type of tagging, as well. Often, these taggers involve deep learning models that infer relationships within the text.

Entity recognition is another important application within feature tagging. For instance, legal documents might contain the name of a defendant, and tagging each document as containing that name or not allows an algorithm to discard documents that a paralegal or lawyer would not need to examine if only documents with the defendant named are important to the upcoming court proceedings. Discarding half of the documents automatically alleviates the human's work load and allows cases to proceed more quickly.

DERIVING SENTIMENT

Sentiment analysis is another important NLP task, which is language-dependent and can involve simple positive/negative poles or more complicated emotional detection within the text document. Many English-based sentiment tools exist within NLP packages, including the afinn sentiment analyzer in Python's NLTK package. Simple sentiment analysis matches each word of a document to lists of positive, negative, and neutral words while counting the instances of each type of match. Tallies from the positive and negative contribute to the overall sentiment score of the document.

More complicated dictionaries that match words to anger, happiness, surprise, and other emotion word lists exist, and it's fairly straight-forward to create custom matching algorithms with open-source lists that associate words with emotions. If needed, one can create a pilot study to tag words with a desired emotion or personality trait for custom sentiment analysis algorithms.

INTEGRATING DATA

Once text data is processed, it's often necessary to integrate the output with a relational database or to create a new relational database to hold the insights, as many dashboard applications and machine learning models assume this format as input. There are a few useful approaches to ease this integration after processing text data.

Vectorization and word frequency matrices usually involves document summaries of the token data. This is useful when the analysis of words is the core problem one is trying to solve with the text analytics. For instance, one might create a deep learning model based on patient outcomes given words used in medical notes.

Summary tables of sentiment, parts of speech numbers, and other document attributes can also be created out of NLP pipelines. For instance, one might want to create speech pattern classifiers to understand the likelihood of someone giving a false statement to police. Summarizing documents based on parts of speech usage, specific word frequencies, or sentiment might be useful for that sort of problem, as these often load onto measures of personality and future risk behaviors.

EXAMPLES OF NLP IN ANALYTICS

Let’s consider a few applications of NLP that are common in analytics today to see how text data can help in real-world situations. Let’s first consider a problem involving customer feedback on a product. Sentiment analysis can provide valuable insight into customer satisfaction over time. Applying sentiment analysis to each comment coming into a platform in real-time, which can feed into a dashboard platform like Tableau or PowerBI, allows marketers to track customer satisfaction beyond survey choices over time to see if customers are growing dissatisfied.

Entity recognition, in which specific words are tagged within documents, can be applied to medical notes and integrated with relational data on a patient to create predictive models related to success on a given treatment or medication compliance or risk of hospital readmission. For instance, psychosocial factors in the patient history taken might provide insight into relapse risk in patients discharged from a substance abuse program, such that the counselor or doctor can better guide the patient to resources that mitigate relapse risk.

One of the more unique analyses of text data is within the field of psychometrics, where personality traits and behavioral attributes are derived from text data. Traits such as extraversion often show through in a person’s writing. Truthfulness or state of mind when a legal statement is written shines through in text data, as well. It’s even possible to determine authorship of a particular text or author’s mood when a text was written from analyzing word choice, grammatical features, and ratios of parts of speech.

OTHER USES OF NLP

Many other uses of NLP exist, including chatbots, virtual personal assistants, translation services, and sentence completion. Translation services and autonomous agents are a relatively fruitful field of research at the moment, with many contributions by authors in the developing world and many open problems related to languages without large training corpuses. Readers who are interested in these applications are encouraged to look in the published literature, as well as within the developing world’s start-up sectors, which often include these types of NLP applications in their technology platforms.

CONCLUSION

NLP tools provide a way to squeeze meaning out of text data, which often provide another view of a particular problem than relational data alone. Myriad applications exist, and it’s likely more problems will be solved with the help of text data in the future, particularly within the legal and medical fields. This paper provides a quick overview of a subset of tools that currently exist. More are being developed daily.

While NLP can provide many pieces of information related to a given problem, it’s important to wrangle the data into a useable format before feeding it into dashboards, machine learning/statistical models, or databases. As more tools are developed to analyze text data (and other unstructured data, such as video or image data), integration with other relational data sources will become even more important to comprehensive analytics.

Interested readers are encouraged to do their own literature and software reviews so that they can choose the best tools for the problem they’re wrangling. However, it’s important to note that NLP-derived insight is only as good as the text documents being analyzed; low-quality data will produce low-quality results.

REFERENCES

Reference examples:

Dunnmon, J. A., Ratner, A. J., Saab, K., Khandwala, N., Markert, M., Sagreiya, H., ... & Ré, C. 2020. "Cross-modal data programming enables rapid medical machine learning." *Patterns*, 100019.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. 2011. "Learning word vectors for sentiment analysis." In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142-150.

Pennebaker, J. W. 2011. "The secret life of pronouns." *New Scientist*, 211, 2828, 42-45.

Polsley, S., Jhunjhunwala, P., & Huang, R. 2016. "Casesummarizer: a system for automated summarization of legal texts." In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, 258-262.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., ... & Chapman, W. 2018. "Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances." *Journal of biomedical informatics*, 88, 11-19.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Colleen M. Farrelly
Staticlysm, LLC
cfarrelly@med.miami.edu