TA*ll*: a new Shiny app of Text Analysis for *All*

Massimo Aria[1,4], Corrado Cuccurullo[2,4], Luca D'Aniello[1,4], Michelangelo Misuraca[3,4] and Maria Spano[1,4]

[1] *University of Naples Federico II, 80126 Napoli, Italy*
[2] *University of Campania "Luigi Vanvitelli", 81043 Capua (CE), Italy*
[3] *University of Calabria, 87036 Arcavacata di Rende (CS), Italy*
[4] *K-Synth spin-off, University of Naples Federico II, 80126 Napoli, Italy*

Abstract

The rapid technological advancements in recent years allowed to process different kinds of data to study several real-world phenomena. Within this context, textual data has emerged as a crucial resource in numerous research domains, opening avenues for new research questions and insights. However, many researchers lack the necessary programming skills to effectively analyze textual data, creating a demand for user-friendly text analysis tools. While languages such as R and python provide powerful capabilities, researchers often face constraints in terms of time and resources required to become proficient in these languages.

This paper introduces TA*ll* - Text Analysis for All, an R Shiny app that includes a wide set of methodologies specifically tailored for various text analysis tasks. It aims to address the needs of researchers without extensive programming skills, providing a versatile and general-purpose tool for analyzing textual data. With TA*ll*, researchers can leverage a wide range of text analysis techniques without the burden of extensive programming knowledge, enabling them to extract valuable insights from textual data in a more efficient and accessible manner.

Keywords

text analysis, shiny app, web app

1. Introduction

In the era of big data, researchers across various disciplines are increasingly faced with the challenge of analyzing vast amounts of textual data.

Textual data, such as research articles, social media posts, customer reviews, and survey responses, hold valuable insights that can contribute to the advancement of knowledge in fields ranging from social sciences to healthcare and beyond.

Researchers seek to analyze textual data to uncover patterns, identify trends, extract meaningful information, and gain deeper insights into various phenomena. By employing advanced natural language processing (NLP) techniques and machine learning algorithms, researchers can explore the semantic and syntactic structures of texts, perform topic detection, polarity detection, and text summarization among other analyses. Moreover, the advent of digital platforms and the proliferation of online content have generated vast amounts of textual data that were previously inaccessible or challenging to obtain. Researchers can tap into these resources to explore new research questions, validate existing theories, and generate novel insights.

By harnessing the power of computational tools and techniques, researchers can efficiently process and analyze large volumes of text, significantly reducing the time and effort required compared to manual analysis. Moreover, there is a growing recognition of the need for text analysis tools that cater to individuals who may not possess extensive programming skills. While programming languages like R and python provide powerful capabilities for data analysis, not all researchers have the time or resources to acquire proficiency in these languages.

This paper presents the first version of TA*ll* - Text analysis for All - a new R Shiny app that brings together all the major advancements in text analysis developed in recent years. For researchers who lack programming skills, TA*ll* offers a viable solution, providing an intuitive interface that allow researchers

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

maria.spano@unina.it (M. Spano); massimo.aria@unina.it (M. Aria); corrado.cuccurullo@unicampania.it (C. Cuccurullo); luca.daniello@unina.it (L. D'Aniello); michelangelo.misuraca@unical.it (M. Misuraca)

0000-0002-3103-2342 (M. Spano); **0000-0002-8517-9411** (M. Aria); **0000-0002-7401-8575** (C. Cuccurullo); **0000-0003-1019-9212** (L. D'Aniello); 0000-0002-8794-966X (M. Misuraca)

to interact with data and perform analyses without the need for extensive programming knowledge.

TA*ll* offers a comprehensive workflow for data cleaning, pre-processing, statistical analysis, and visualization of textual data, by combining state-of-the-art text analysis techniques into an R Shiny app.

2. Discovering TA*ll* workflow

First TA*ll* combines the functionality of a set of R packages developed for NLP tasks (see: https://cran.r-project.org/web/views/NaturalLanguageProcessing.html) with the ease of use of web apps using the Shiny package environment. TA*ll* workflow aims to facilitate the discovery and analysis of text data by systematically processing and exploring the content.

Figure 1 shows the three main steps of a TA*ll* workflow.

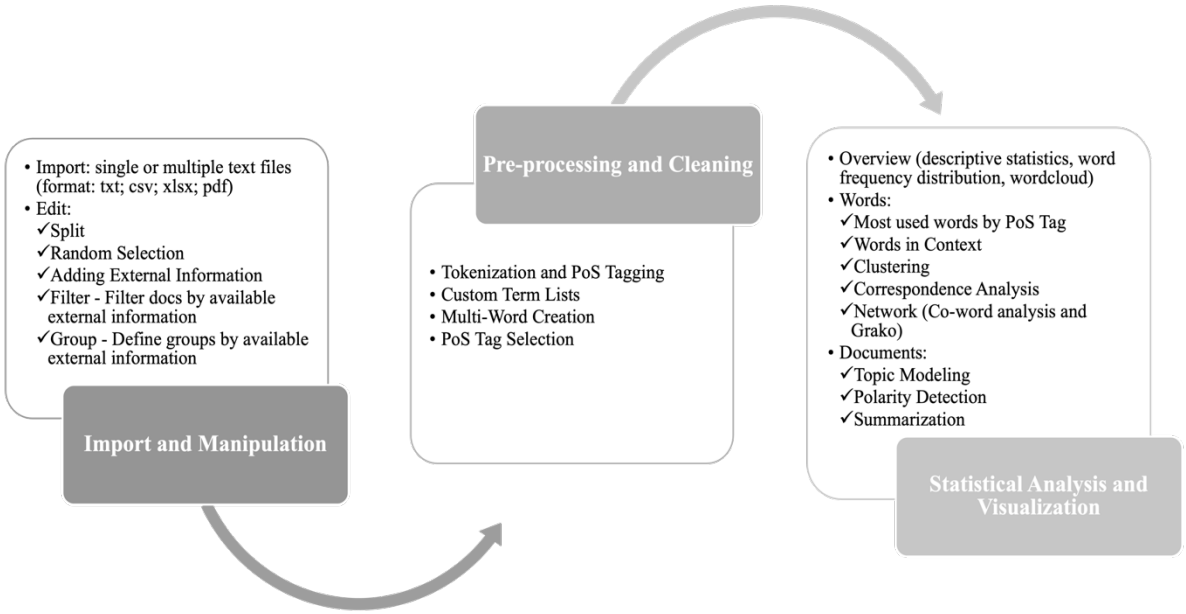

Figure 1: Discovering TA*ll* tabs, methods and workflow

The first step **Import and Manipulation** involves importing one or multiple text files in various formats, such as txt, csv, xlsx, and pdf, allowing easy loading of a diverse range of textual data. Subsequently, texts could be subjected to several editing actions, including the division into smaller segments, such as chapters or paragraphs, or the selection of texts' subsets for sampling or random analysis purposes. Users can supplement the imported texts with additional external information (e.g., author, publication date, rating) attached to the texts or added during the analysis. Concerning both the aim of analysis and the availability of external variables, texts could be filtered, enabling to focus on specific subsets or grouped for comparison purposes.

Before beginning the **Pre-processing and Cleaning** step, a language model was necessary for the annotation process (i.e., tokenization, PoS tagging, and lemmatization).

TA*ll* utilizes pre-trained models provided by Universal Dependencies Treebanks. Universal Dependencies (https://universaldependencies.org) is a framework for consistent annotation of grammar (Part-of-Speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 500 contributors producing over 200 treebanks in over 100 languages. By using these models TA*ll* supports the analysis of texts written in 60 different languages.

Each text is parsed into individual tokens (words) and tagged with its respective Part-of-Speech (PoS) label to better understand word usage patterns. All the subsequent statistical analyses could be performed alternatively on tokens or lemmas. Moreover, users can define and load custom lists of words for various research purposes (e.g., to substitute synonyms, remove domain stopwords, and semantically tag specialized lexicons'terms).

A crucial aspect when we deal with text analysis is to identify and handle multi-word expressions and collocations. To face this issue, TA*ll* performs *Rapid Automatic Keyword Extraction* (RAKE) algorithm [1] that uses a delimiter-based approach to identify candidate keywords and scores them using word co-occurrences that appear in the candidate keywords.

At the end of pre-processing and cleaning phase, users can select specific PoS tags to focus their analyses considering only content-bearing words (e.g., nouns, adjectives, verbs, collocations).

Statistical analysis and Visualization step opens the opportunity of exploring cleaned texts by performing one or more approaches as listed in Figure 1. Descriptive statistics (e.g., number of tokens, types, sentences, lexical measures), word frequency distribution, and wordclouds provide an initial overview of the text corpus. TA*ll* tabs are then organized by considering two levels of analysis: words and documents.

Detailed analysis of words includes a set of statistical methods mainly devoted to topic detection. The most intuitive approach is to identify and visualize through dynamic plots the most frequently used words for each PoS tag, looking obviously at their absolute frequency but also considering more complex weighting schemes as *term frequency/inverse document frequency* (TF-IDF) [2] to uncover words with the highest discriminative power. Despite the simplicity, often analyzing the frequency distribution of words gives a general idea of the contents in the text collection, but it is not enough to identify topics. A topic can be represented as a set of meaningful words with syntagmatic relatedness [3].
Following this definition, the three methods most widely shared in the literature [4] are implemented in TA*ll*:

- *Clustering* [5, 6] to group similar words based on their usage patterns or context;
- *Correspondence Analysis* [7, 8] to explore semantic relationships among words, identifying the latent structure of the text collection;
- *Network* (Co-word analysis and Grako) [9] to analyze co-occurrence patterns of words within texts, highlighting subsets of words strictly related through community detection algorithms [10].

The documents tab includes a set of statistical methods to cope with specific tasks where the focus is properly on the entire documents:

- *Topic Modeling* to identify both prominent topics and their distribution within documents using the well-known Latent Dirichlet Allocation (LDA) algorithm [11]. Moreover, TA*ll* estimates the number of topics automatically through the measures proposed in [12, 13, 14, 15], but users can also explore different solutions by setting the number of the desired topics;
- *Polarity Detection* to determine the polarity (positive, negative, neutral) of documents by choosing among different lexicons (i.e., Hu & Liu [16], Loughran & McDonald [17], nrc [18]);
- Summarization to concisely summarize each text to capture key insights rapidly. TA*ll* performs TextRank algorithm [19], based on applying Google's PageRank [20] to the network of sentences for extracting the most relevant ones.

This comprehensive workflow provides users with the statistical methods to process texts efficiently and share their results and workflows with collaborators by downloading plots and reports from TA*ll*, facilitating and speeding up all analysis steps. paragraph in every section does not have first-line indent. Use only styles embedded in the document.

3. Conclusion and remarks

This paper presented a brief overview of the first version of TA*ll,* a new shiny app for importing, pre-processing, and analyzing textual data.
Our idea stems from the now growing need to analyze textual content to today's ever-increasing number, offering the opportunity to explore it quickly and efficiently, even for those without programming skills. Using a user-friendly text analysis tool, researchers can focus more on their domain expertise and research questions rather than spend significant time learning programming languages or writing complex code. Tools like TA*ll* democratize text analysis, making it accessible to a broader audience and promoting interdisciplinary collaboration.
Moreover, general-purpose software can be used in every research field and encourages reproducibility and transparency in research. paragraph in every section does not have first-line indent. Use only styles embedded in the document.

References

[1] S. Rose, D. Engel, N. Cramer, W. Cowley, 2010. Automatic keyword extraction from individual documents, pages 1–20. Wiley Online Library.

[2] G. Salton, C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

[3] H. Schütze, 1993. A vector model for syntagmatic and paradigmatic relatedness. In Making sense of words: 9th annual conference of the UW Centre for the New OED and Text Research.

[4] M. Misuraca, M. Spano, 2020. *Unsupervised Analytic Strategies to Explore Large Document Collections*, pp. 17–28. Heidelberg: SPRINGER, 06.

[5] A. K. Jain, M. N. Murty, P. J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

[6] D. Xu and Y. Tian, 2015. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2:165–193.

[7] J. P. Benzécri, 1982. *Histoire et préhistoire de l'analyse des données*. Dunod, Paris.

[8] L. Lebart, A. Salem, L. Berry, 1997. *Exploring textual data*, volume 4. Springer Science & Business Media.

[9] M. Callon, J.-P. Courtial, W. A. Turner, S. Bauin, 1983. From translations to problematic networks: An introduction to co-word analysis. *Social science information*, 22(2):191–235.

[10] S. Fortunato, D. Hric, 2016. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, nov.

[11] D. M. Blei, A. Y. Ng, M. I Jordan, 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993– 1022.

[12] T. L. Griffiths, M. Steyvers, 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

[13] R. Deveaud, E. Sanjuan, P. Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17:61–84, 06.

[14] J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, 2009. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7):1775–1781. Advances in Machine Learning and Computational Intelligence.

[15] R. Arun, V. Suresh, C. E. Veni Madhavan, M. N. Narasimha Murthy, 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 391–402, Berlin, Heidelberg. Springer Berlin Heidelberg.

[16] M. Hu, B. Liu, 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177, New York, NY, USA. Association for Computing Machinery.

[17] T. Loughran, B. McDonald, 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

[18] S. Mohammad, P. Turney, 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34, Los Angeles, CA, June. Association for Computational Linguistics.*

[19] R. Mihalcea, P. Tarau, 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

[20] L. Page, S. Brin, R. Motwani, T. Winograd, 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.