The Journal of Open Source Software

ConTEXT Explorer: a web-based text analysis tool for exploring and visualizing concepts across time

Ziying Yang[1], **Gosia Mikolajczak**[1], **and Andrew Turpin**[1]

1 University of Melbourne

DOI: 10.21105/joss.03347

Software

- Review
- Repository
- Archive

Editor: Fabian-Robert Stöter

Reviewers:

- @sara-02
- @baileythegreen

Submitted: 03 May 2021
Published: 23 June 2021

Summary

ConTEXT Explorer is an open Web-based system that assists in exploring the context of concepts (combinations of co-occurring words and phrases) over time in text documents. It provides a user-friendly interface to the analysis of user-provided text data and integrates functionalities of the Whoosh search engine, Spacy, Gensim and Plotly Python libraries. By providing suggestions for query expansion and producing interactive plots, `ConTEXT Explorer` facilitates exploratory data analysis, which can serve as the basis for subsequent text classification.

Statement of need

With the availability of digital sources of data and associated tools, automated text analysis is becoming increasingly popular in humanities and social sciences. While for very large coprora, unsupervised text mining methods like topic modelling (Blei et al., 2003) can provide some useful summaries of data, many social science and humanities applications require analysis of data in context. That is, simple "bags of words" automatically mined and presented in isolation from the orignial text are often not meaningful for complex questions involving human behaviour and society. Inevitably, human interpretation is required to make sense of such patterns. For corpora with more than several hundred documents, there is a need for computational tools that can assist researchers in exploring the context in which "bags of words" (we will call them concepts from now on) occur.

Similarly, there is a need for tools that assist in the manual construction of concepts from text coropora. In particular, manual intervention to judge the semantic intent of words (e.g., word sense disambiguation) is usualy needed to filter keywords to add to concepts that might be generated by automatic methods such as query expansion (Buckley et al., 1994) or comparison of word embeddings (Mikolov et al., 2013). For example, if a researcher is interesed in finding articles about same sex marriage, they might start with "same_sex marriage"[1] as a concept. Automated methods processing a corpus of news articles might suggest related words like 'matrimony,' 'union,' 'erosion,' and 'puzzlement'[2]. Depending on the research question and the context of these words, some might be relevant to the concept and should be included, while others are not. It requires complex human judgement to make the distinction. `ConTEXT Explorer` is a tool to assist construction of such concepts in context.

Most existing computational methods underlying automated text processing require at least a working knowledge of relevant methods and programming languages (such as R or Python).

[1] We use underscore to join mulitple words into a single phrase.

[2] This is an example where we apply `ConTEXT Explorer` in the Australian Research Council Discovery Project (DP180101711) "Understanding Political Debate and Policy Decisions Using Big Data."

ConTEXT Explorer is designed to lower these barriers to entry, by allowing an application of information retrieval and machine learning methods to text analysis without programming knowledge.

ConTEXT Explorer is developed using **Dash** (Plotly Technologies Inc., 2015) in Python, and integrates the following packages.

- **Spacy** pipeline (Honnibal et al., 2020) - for pre-processing the text corpora uploaded by users.
- **Whoosh** (Whoosh, 2021) - for building a search engine, which allows ranking of sentences relevant to the given concepts, and word frequency analysis at the sentence and document level.
- **Gensim** (Řehůřek & Sojka, 2010) - for training a word2vec (Mikolov et al., 2013) model for the uploaded corpus, which allows the user to find words related to other words for expanding concepts.
- **Plotly** (Plotly Technologies Inc., 2015) - for visualizing results in interactive graphs, which can be customized and saved as PNG files.

Current text analysis tools require either previous knowledge of programming (e.g., R, Python), or are commercial products (e.g., **RapidMiner** (RapidMiner, 2021), **Google Cloud Natural Language API** (Google, 2021)). One exception that we are aware of is **Voyant Tools** (Sinclair & Voyant Tools Team, 2012), which is an open-source web-based text analysis tool built in Java. It allows the users to explore their data using some basic text analysis techniques such as word frequencies (at the document level), word cloud, and word context (words appearing around a chosen term). ConTEXT Explorer provides several functionalities that give a user a deeper understanding of the text, which are currently not available in Voyant Tools, such as: concept suggestions, sentence ranking and concept grouping. It includes models allowing discovery of similar terms, and a search engine allowing retrieval of sentences relevant to concept terms, which can be used for concept expansion, and visualization of concepts over time. One key feature is that a caoncept can be either a a conjunction or disjunction of bags of words.

Compared to commercial text analysis systems such as **RapidMiner** (RapidMiner, 2021), which include some complex analysis techniques, ConTEXT Explorer is open-source (free) and easy to install. It enables researchers to browse text interactively for concepts (bags of words) in their corpus before mining the text in machine learning driven systems.

ConTEXT Explorer is designed to help users interested in defining concepts, and exploring their trends over time. This could be a particularly helpful as an input for some popular text analysis systems such as **MonkeyLearn** (MonkeyLearn, 2021), which enable text classification, tagging, and training AI machine learning models, but require prior knowledge of the data.

ConTEXT Explorer is engineered to allow for integration of other Python packages into the analysis process. It can be easily combined with other Python APIs (such as MonkeyLearn), once the concept groups are defined.

Key features

ConTEXT Explorer has been tested locally under macos and as a server running under Ubuntu.

Build a corpus

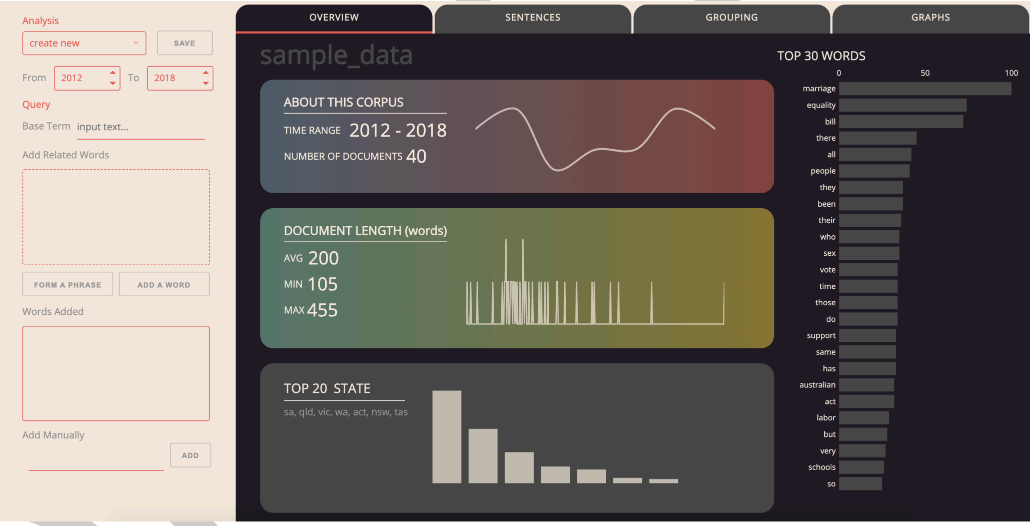
Users are asked to format their text documents as a CSV file (with each document saved in a separate row), before uploading it into ConTEXT Explorer. At minimum, users are asked to provide document text and publication year. Users can also upload columns with additional document information (such as document author, title, and so on).

`ConTEXT Explorer` processes the submitted file in the following steps.

1. Sentencize and tokenize English text using Spacy (Honnibal et al., 2020). This allows ranking of documents and speeds up document search.
2. Index the documents, and build a search engine for the corpus using Whoosh (Whoosh, 2021) and the Okapi BM25F (Robertson & Zaragoza, 2009) ranking function.
3. Remove stop words, lemmatize remaining words, and generate a word2vec (Mikolov et al., 2013) model for the corpus using Gensim (Řehůřek & Sojka, 2010).

For each corpus, users can create a new analysis, or load a pre-saved analysis to the dashboard.

Dashboard


Analysis: create new

From 2012 To 2018

Base Term: input text...

sample_data

ABOUT THIS CORPUS

TIME RANGE 2012 - 2018

NUMBER OF DOCUMENTS 40

DOCUMENT LENGTH (words)

AVG 200

MIN 105

MAX 455

TOP 20 STATE

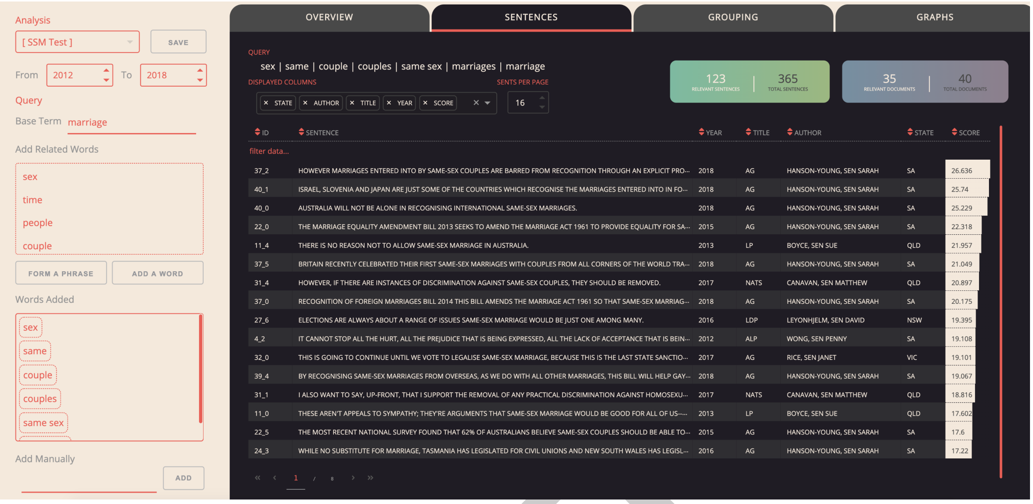
sa, qld, vic, wa, act, nsw, tas

Figure 1: The starting dashboard.

As shown in Figure 1, the dashboard interface has two panes. On the left-hand side, users can:

- select the year range of documents to be displayed in search results;
- add or delete query terms (single words or phrases) to create a concept;
- save the current query as a new analysis; and
- download the subset of the corpus filtered by the query terms.

Overview. The overview tab summarizes the corpus information such as the total number of documents, year range, document length, most frequent words in the corpus, and most frequent values for selected metadata.



QUERY: sex | same | couple | couples | same sex | marriages | marriage

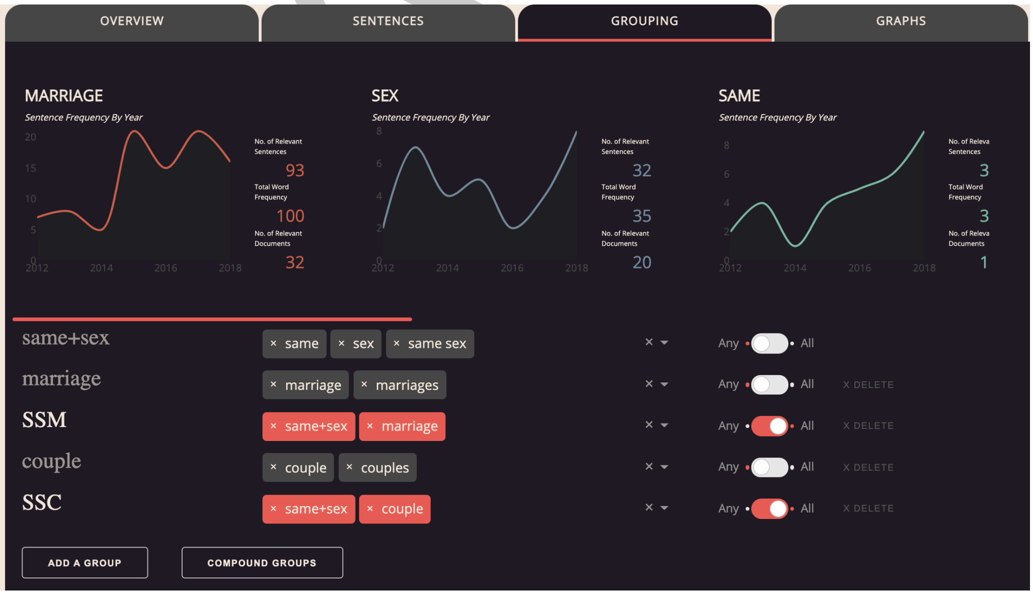
123 RELEVANT SENTENCES | 365 TOTAL SENTENCES

35 RELEVANT DOCUMENTS | 40 TOTAL DOCUMENTS

ID	SENTENCE	YEAR	TITLE	AUTHOR	STATE	SCORE
37_2	HOWEVER MARRIAGES ENTERED INTO BY SAME-SEX COUPLES ARE BARRED FROM RECOGNITION THROUGH AN EXPLICIT PRO...	2018	AG	HANSON-YOUNG, SEN SARAH	SA	26.636
40_1	ISRAEL, SLOVENIA AND JAPAN ARE JUST SOME OF THE COUNTRIES WHICH RECOGNISE THE MARRIAGES ENTERED INTO IN FO...	2018	AG	HANSON-YOUNG, SEN SARAH	SA	25.74
40_0	AUSTRALIA WILL NOT BE ALONE IN RECOGNISING INTERNATIONAL SAME-SEX MARRIAGES.	2018	AG	HANSON-YOUNG, SEN SARAH	SA	25.229
22_0	THE MARRIAGE EQUALITY AMENDMENT BILL 2013 SEEKS TO AMEND THE MARRIAGE ACT 1961 TO PROVIDE EQUALITY FOR SA...	2015	AG	HANSON-YOUNG, SEN SARAH	SA	22.318
11_4	THERE IS NO REASON NOT TO ALLOW SAME-SEX MARRIAGE IN AUSTRALIA.	2013	LP	BOYCE, SEN SUE	QLD	21.957
37_5	BRITAIN RECENTLY CELEBRATED THEIR FIRST SAME-SEX MARRIAGES WITH COUPLES FROM ALL CORNERS OF THE WORLD TRA...	2018	AG	HANSON-YOUNG, SEN SARAH	SA	21.049
31_4	HOWEVER, IF THERE ARE INSTANCES OF DISCRIMINATION AGAINST SAME-SEX COUPLES, THEY SHOULD BE REMOVED.	2017	NATS	CANAVAN, SEN MATTHEW	QLD	20.897
37_0	RECOGNITION OF FOREIGN MARRIAGES BILL 2014 THIS BILL AMENDS THE MARRIAGE ACT 1961 SO THAT SAME-SEX MARRIAG...	2018	AG	HANSON-YOUNG, SEN SARAH	SA	20.175
27_6	ELECTIONS ARE ALWAYS ABOUT A RANGE OF ISSUES SAME-SEX MARRIAGE WOULD BE JUST ONE AMONG MANY.	2016	LDP	LEYONHJELM, SEN DAVID	NSW	19.395
4_2	IT CANNOT STOP ALL THE HURT, ALL THE PREJUDICE THAT IS BEING EXPRESSED, ALL THE LACK OF ACCEPTANCE THAT IS BEIN...	2012	ALP	WONG, SEN PENNY	SA	19.108
32_0	THIS IS GOING TO CONTINUE UNTIL WE VOTE TO LEGALISE SAME-SEX MARRIAGE, BECAUSE THIS IS THE LAST STATE SANCTIO...	2017	AG	RICE, SEN JANET	VIC	19.101
39_4	BY RECOGNISING SAME-SEX MARRIAGES FROM OVERSEAS, AS WE DO WITH ALL OTHER MARRIAGES, THIS BILL WILL HELP GAY...	2018	AG	HANSON-YOUNG, SEN SARAH	SA	19.067
31_1	I ALSO WANT TO SAY, UP-FRONT, THAT I SUPPORT THE REMOVAL OF ANY PRACTICAL DISCRIMINATION AGAINST HOMOSEXU...	2017	NATS	CANAVAN, SEN MATTHEW	QLD	18.816
11_0	THESE AREN'T APPEALS TO SYMPATHY; THEY'RE ARGUMENTS THAT SAME-SEX MARRIAGE WOULD BE GOOD FOR ALL OF US--...	2013	LP	BOYCE, SEN SUE	QLD	17.602
22_5	THE MOST RECENT NATIONAL SURVEY FOUND THAT 62% OF AUSTRALIANS BELIEVE SAME-SEX COUPLES SHOULD BE ABLE TO...	2015	AG	HANSON-YOUNG, SEN SARAH	SA	17.6
24_3	WHILE NO SUBSTITUTE FOR MARRIAGE, TASMANIA HAS LEGISLATED FOR CIVIL UNIONS AND NEW SOUTH WALES HAS LEGISL...	2016	AG	HANSON-YOUNG, SEN SARAH	SA	17.22

Figure 2: The sentences tab of the dashboard, with some query terms shown in the left pane.

Sentences. This tab shows the ranking of relevant sentences based on query terms defined in the left pane. Sentences are ranked by the Okapi BM25F ranking function, and the computed similarity score for each sentence is shown in the "SCORE" column. The table can be sorted and filtered by column values. Users can click on each sentence to see its full content in a pop-up window, which also allows checking of the frequency of individual terms, and to add them to the query.



MARRIAGE — No. of Relevant Sentences 93; Total Word Frequency 100; No. of Relevant Documents 32

SEX — No. of Relevant Sentences 32; Total Word Frequency 35; No. of Relevant Documents 20

SAME — No. of Releva Sentences 3; Total Word Frequency 3; No. of Releva Documents 1

same+sex	× same × sex × same sex	Any ○ All	
marriage	× marriage × marriages	Any ○ All	X DELETE
SSM	× same+sex × marriage	Any ● All	X DELETE
couple	× couple × couples	Any ○ All	X DELETE
SSC	× same+sex × couple	Any ● All	X DELETE

Figure 3: The grouping tab of the dashboard, showing the term frequency of the added terms across time (top), and some examples of query groups (bottom).

Grouping. The top part of this tab shows the number of sentences containing each query term within the user-defined year range. In the bottom part, users can group the query terms using "Any" or "All" operators. Groups can be further combined into more complex groups.

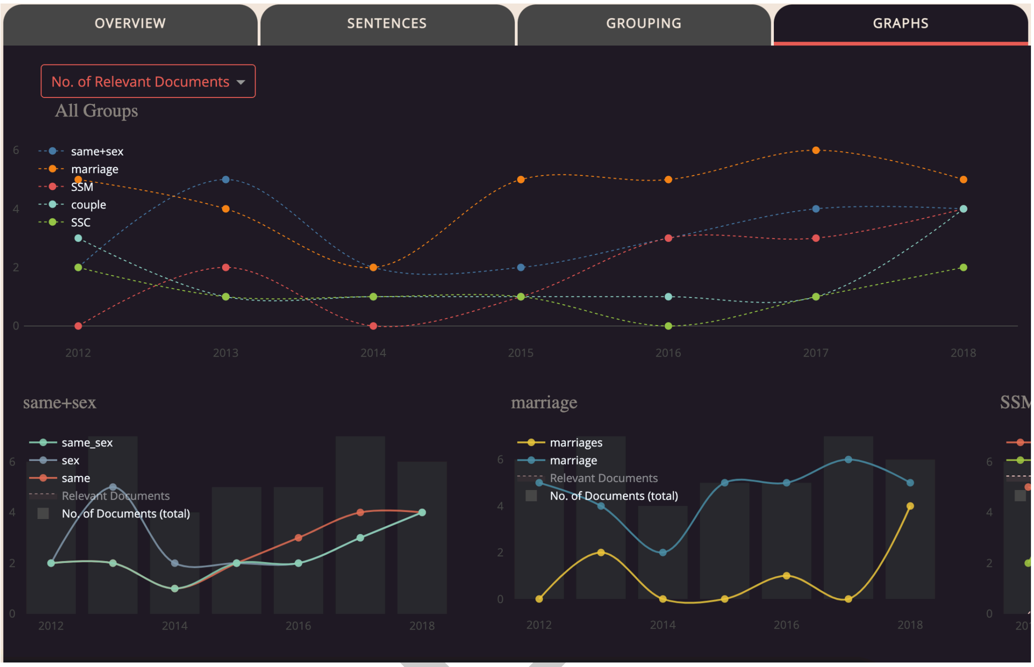


Figure 4: 'Graphs' tab, showing the aggregated graph for all groups (top) and individual graph for each group (bottom)

Graphs. Based on the query groups generated in the previous tab, this page displays aggregated and individual plots, which allow to compare groups (top) and individual terms within each group (bottom). Users can choose the number of relevant documents, number of sentences, or proportion of document as the y-axis of the graphs. All graphs are plotted by Plotly (Plotly Technologies Inc., 2015) which allows users to interact with every trace in the graphs.

Save and reload an analysis

As mentioned in Section 3.2, users are able to save the details of their analysis (including added terms and generated groups), and reload it to view all of the ranking, groups and graphs from the index page.

Acknowledgements

Development of `ConTEXT Explorer` has been supported by the Australian Research Council Discovery Project (DP180101711) "Understanding Political Debate and Policy Decisions Using Big Data" awarded to the third author.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Buckley, C., Salton, G., & Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. *SIGIR '94*, 292–300. https://doi.org/10.1007/978-1-4471-2099-5_30

Google. (2021). *Cloud natural language – google cloud*. https://cloud.google.com/natural-language/

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

MonkeyLearn. (2021). *How MonkeyLearn works*. https://monkeylearn.com/how-it-works/

Plotly Technologies Inc. (2015). *Collaborative data science*. Plotly Technologies Inc. https://plot.ly

RapidMiner. (2021). *Best data science & machine learning platform*. https://rapidminer.com/

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Sinclair, G. R., Stéfan, & Voyant Tools Team, the. (2012). *Voyant tools (web application)*. https://voyant-tools.org/

Whoosh. (2021). *Whoosh, a pure python search engine library*. https://whoosh.readthedocs.io/en/latest/intro.html