Deep Analysis of the Geothermal Literature Using Natural Language Processing

Mohammad J. Aljubran[1], Alaa S. Alahmed[2], Ahmed S. Alkhalifah[3], Matt Hall[4], Robert Leckenby[4]

[1]Stanford University, School of Earth, Energy, and Environmental Sciences, Stanford, CA

[2]Saudi Aramco, Energy Systems Engineering, Dhahran, Saudi Arabia

[3]King Fahad Medical City, As Sulimaniyah, Riyadh, Saudi Arabia

[4]Agile Scientific, Mahone Bay NS B0J 2E0, Canada

aljubrmj@stanford.edu

Keywords: Natural Language Processing, Deep Learning, Literature

ABSTRACT

With the globally growing volume of geothermal literature, data analysis has become useful to advance professional and academic research and development efforts. Furthermore, it is essential to leverage state-of-the-art algorithms to develop useful tools based on existing databases. This work utilized statistical and deep learning techniques to draw insights based on the geothermal literature. We scraped the International Geothermal Association (IGA) database using the Stanford University search engine. We gathered and preprocessed all 18,873 publications archived in this database, where headers included publication title, authors, year, keywords, abstract, language, conference, and session type.

Analysis shows that the three geothermal events with the largest volume of publications historically are the Geothermal Resources Council Transactions, World Geothermal Congress, and Stanford Geothermal Workshop. Using natural language processing (NLP) techniques, we “geoparsed” each abstract to figure out what location in geographical coordinates it is concerned about. This allowed for developing an interactive world heatmap showing the focus of geothermal research efforts historically. Latent Dirichlet Allocation (LDA) was used to cluster the geothermal literature into a total of nine topics. we also developed a geothermal literature intelligent search engine using term frequency -- inverse document frequency (TF-IDF) and cosine similarity. Preprocessing the “authors” data, we developed a coauthorship graphical network encompassing researchers within the geothermal community and reflecting the level of collaboration between them. Finally, a deep learning model was developed to perform text generation and auto-completion using the state-of-the-art generative pretrained transformers (GPT-2) fine-tuned to the geothermal literature.

We conclude this paper by introducing an open-source application programming interface (API) demonstrating and offering these insights and tools for public use. This live API is designed to continuously read from the IGA Stanford University search engine to ensure up-to-date results. You may access this API at http://steaming-geothermal-analytics.info.

INTRODUCTION

Global warming has led to many concerning changes, e.g. shifting wind patterns, rising sea levels, variable plant blooming seasons, amongst others. Nowadays, limiting greenhouse gas emissions represents an important objective in the pursuit of reversing the rising temperature levels. Replacing fossil fuels with clean renewable energy alternatives has become primary means to reach net-zero carbon dioxide emissions. Amongst the different resources of renewables, geothermal energy stands out with the ability to provide continuous baseload, and even dispatchable, power to the electricity grid to support its seasonally variable alternatives. With the increase of global interest and investment, the yearly volume of geothermal literature has noticeably grown as a result. Precise and continuous analysis of this literature represents a major task to facilitate access of the relevant literature data and information to the academic and professional geothermal community. This paper describes an analysis of the entire geothermal literature using a variety of state-of-the-art natural language processing (NLP) techniques. The International Geothermal Association (IGA) geothermal conference papers library was used as the source of geothermal papers and articles. This work resulted in three major NLP and deep learning tools: 1) intelligent search, 2) author network, and 3) text generation and auto-completion.

2. TRENDS

A Python module was first programmed to collect all 18,873 publications, as of 23 December 2020, stored in the IGA library. Collected data included publication title, authors, year, keywords, abstract, language, conference name, and session type. It is important to note that the different geothermal conferences follow various data collection and reporting formats, hence careful preprocessing was required. Fig. 1 demonstrates the yearly count of publications historically, where we observe a spike every five years coinciding with the World Geothermal Congress occurrences. Moreover, there is an increasing trend of publications over time which indicates growing interest and research and development efforts across the geothermal community. Fig. 2 shows ranking based on publication count of the different geothermal events globally. The top three events are the Geothermal Resources Council Transactions, World Geothermal Congress, and Stanford Geothermal Workshop.

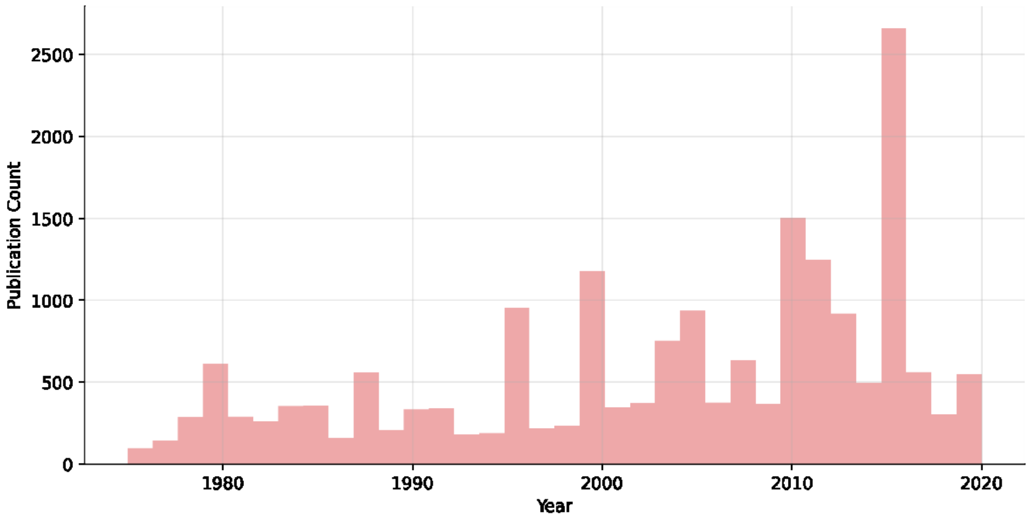


Figure 1: Vertical histogram demonstrating the geothermal literature publication count over 1975-2020. The publication count spikes every five years which coincides with the World Geothermal Congress occurrences.

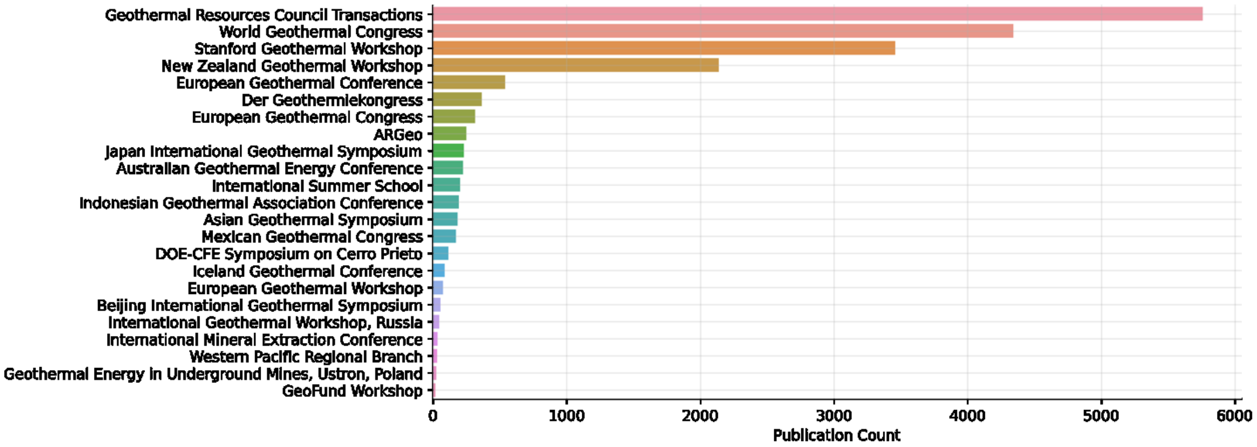


Figure 2: Horizontal histogram of the geothermal literature publication count across the different events and conferences. The top three events are the Geothermal Resources Council Transactions, World Geothermal Congress, and Stanford Geothermal Workshop.

To visualize the growing interest in geothermal resources historically, we analyzed the geothermal literature to find out what locations are discussed most frequently in abstracts. This was done using “geoparsing” algorithms which correlate geographical coordinates to specific words or combinations of words. The project API demonstrates an animation of the most frequently referenced countries/states in the geothermal literature over the span of 1975-2020. Figs. 3 and 4 show 2020 snippets of the most frequently referenced countries and U.S. states, respectively. These geographical heatmaps reflect the academic and professional research and development activity in the geothermal arena. For instance, Fig. 3 shows that locations across the U.S. dominate the global geothermal literature with nearly 75% of the total mentions; meanwhile, Indonesia comes second. These observations naturally coincide with the fact that the U.S. and Indonesia have the highest geothermal production rates and also confirm their active role in the geothermal literature. Similarly, Fig. 4 shows that the western U.S. states have the lion’s share, i.e. over 60% mentions come from literature focused on locations across California and Nevada.

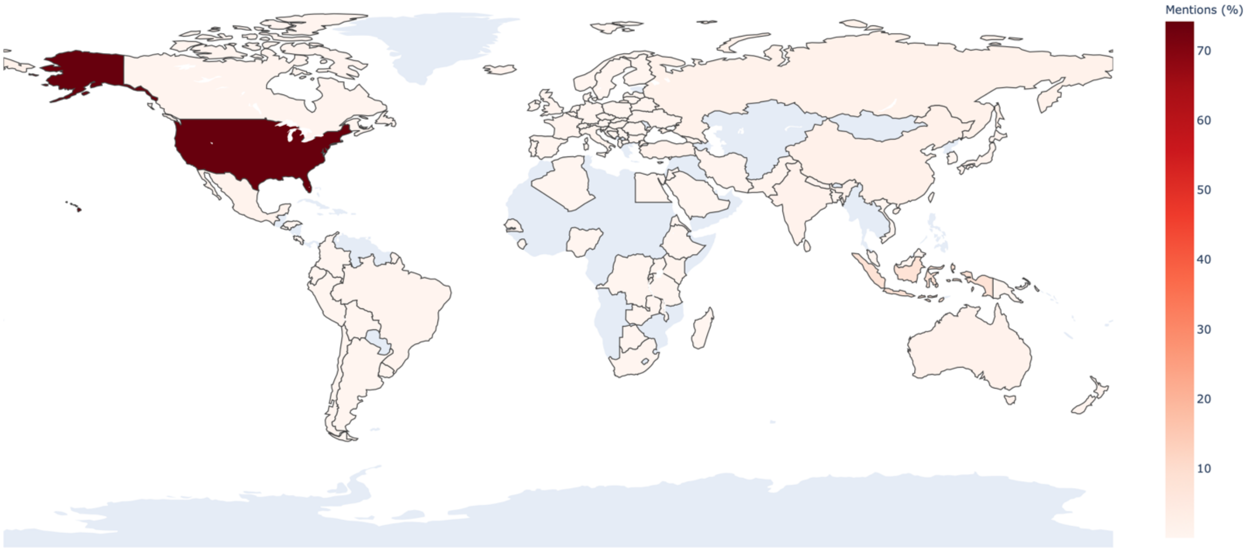


Figure 3: World heatmap showing the relative frequency of mentioning a country (or locations within it) in the 2020 publications of the geothermal literature. It is evident that the U.S. is taking the lead with nearly 75% of the total contribution. An interactive version of this heatmap is accessible at http://steaming-geothermal-analytics.info.

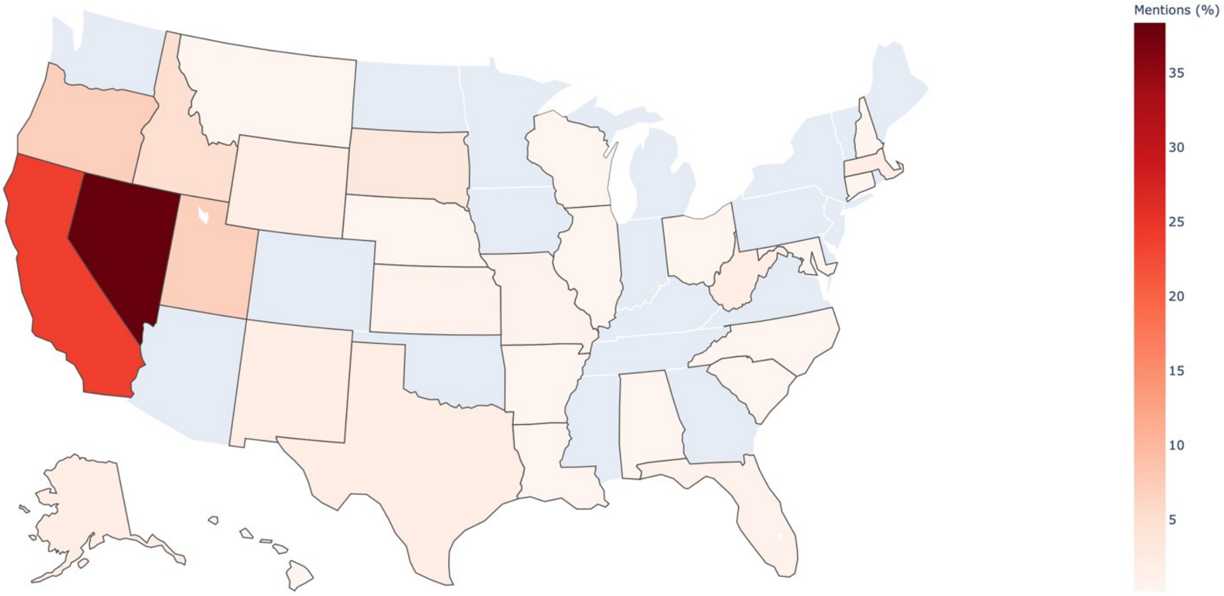


Figure 4: U.S. heatmap showing the relative frequency of mentioning a state (or locations within it) in the 2020 publications of the geothermal literature. Note that Nevada and California alone contribute with over 60% of mentions across the U.S. An interactive version of this heatmap is accessible at http://steaming-geothermal-analytics.info.

3. TOPIC CLUSTERING

Topic clustering offers an unsupervised means for automatically organizing, searching, and summarizing literature. Topic clustering helps with discovering hidden themes within large volumes of literature and classifying publications stochastically under one or more themes. In this task, we aimed to group the different geothermal articles into clusters based on their corresponding abstract content using NLP clustering techniques. Each abstract is first preprocessed through a pipeline of cleansing steps before clustering. This pipeline consisted of tokenization, removal of stopwords (e.g. “the”, “a”, “an”, “in”, etc.), creation of bigrams (sequences of two adjacent words), stemming, and lemmatization of words to only keep nouns, adjectives, verbs, and adverbs (Denny and Spirling 2017).

The resulting geothermal literature corpus was then clustered using the Latent Dirichlet Allocation (LDA) algorithm (Blei et al. 2003, Falush et al. 2003, Pritchard et al. 2000). LDA is a popular generative unsupervised model that aims to cluster documents (paper abstracts in this case) into unobserved groups or topics. It assumes that documents with similar topics contain similar words. While other document clustering techniques rely on deterministic word-frequency clustering, LDA represents documents as probability distributions of topics and topics as probability distributions of words. Sparse Dirichlet priors are used to model the document-topic and topic-word distributions, which captures the natural skewness found in topics within documents and words within topics. To compute the posterior, LDA also utilizes multinomial distributions which describe the likelihood of a topic appearing in a given document and the likelihood of a word appearing in a given topic.

In LDA modelling, the number of topics must be assumed *a priori*. The goal is to produce coherent topic clusters while minimizing model complexity. Hence, we conducted an iterative process where we performed LDA topic clustering for different numbers of topics and measured topic coherence (Röder et al. 2015). As seen in Fig. 5, the choice of nine topics was found optimal to simultaneously minimize model complexity and maximize topic coherence. To enable visualization of topics, principal component analysis (PCA) was used to project the topic vectors in two dimensions. The project API webpage hosts an interactive visualization of the resulting LDA topics along with a ranked list of their corresponding words (Sievert and Shirley 2014, Chuang et al. 2012). Note that this insightful visualization allows for adjusting a relevance metric (λ), which defines the relevance of a word to a topic. This adjustable metric allows for normalizing a word's frequency in a particular topic using its frequency in the whole corpus of documents. Using λ = 1 means that we only consider the probability of a word in a topic while using λ = 0 means that we consider the pure specificity of a word to a topic. For instance, choosing λ = 1 in this work would result in the word "geothermal" to appear in a high rank across most topics due to lack of relevance normalization, which is misleading. It is generally recommended to use λ values in the range of 0.3 to 0.6 (Sievert and Shirley 2014). Fig. 6 demonstrates a snippet of the LDA interactive plot with λ = 0.3 where topic seven is highlighted in the horizontal histogram. Given the drilling-focused terms (e.g. "drilling", "well", "depth", "borehole", "bit", "mud", etc.), it is evident that LDA dedicated topic seven to drilling publications in the geothermal literature. Table 1 shows common words of the nine LDA topics which clearly demonstrates the representative and independent focus of each individual topic.

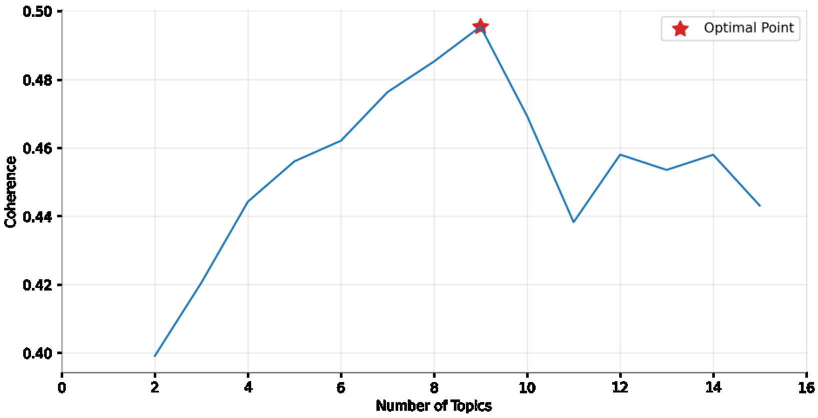
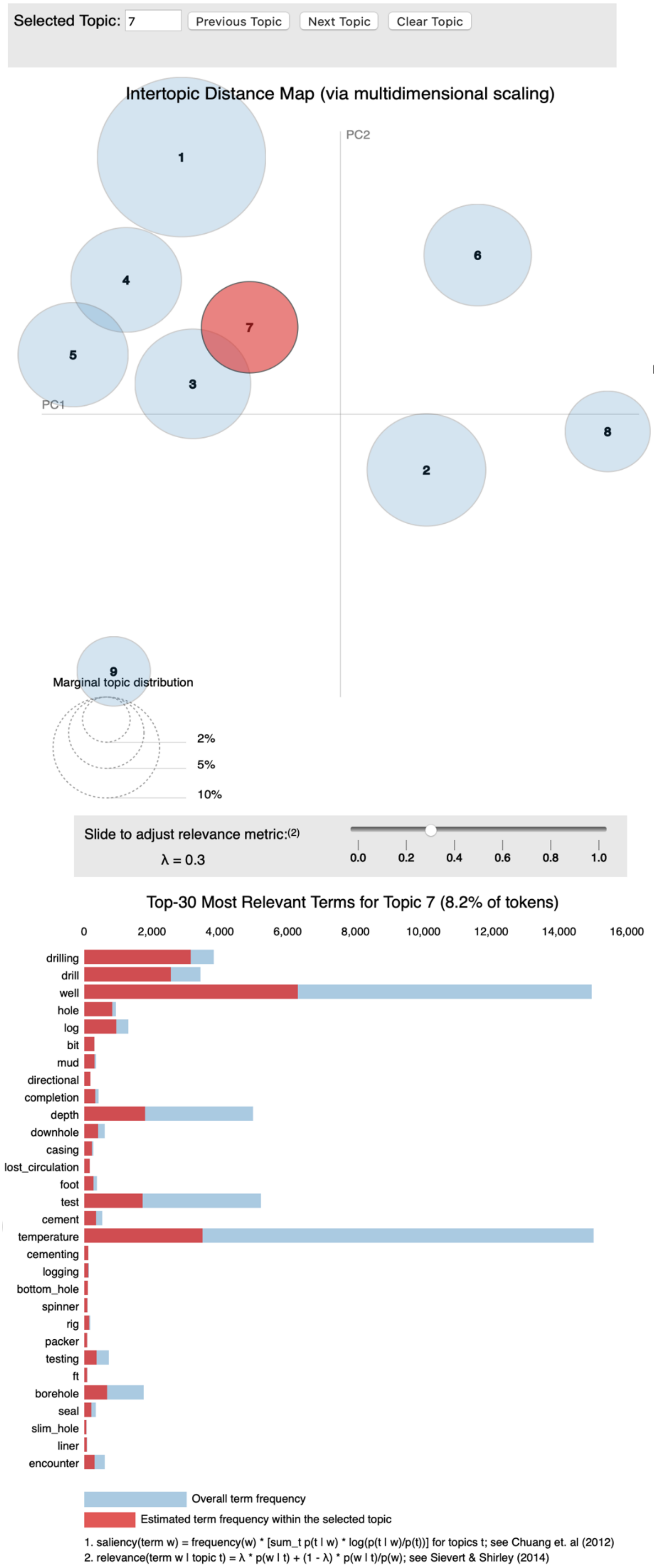


Figure 5: Effect of number of topics on the coherence measure across LDA topics. Nine is chosen to be the optimal number of topics that maximizes coherence and minimizes model perplexity based on this iterative process.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6: Demonstration of the LDA topic results with PCA plot (top) and topic seven term frequency histogram (bottom) based on relevance factor of λ = 0.3. An interactive version of this map is accessible at http://steaming-geothermal-analytics.info.

Table 1: LDA Topics and Common Words

LDA Topic	Sample of Most Common Words
1	Geothermal; energy; resource; project; development; power
2	Water; fluid; spring; hot; temperature; thermal
3	Fracture; flow; heat; permeability; hydraulic; stress
4	Model; reservoir; simulation; numerical; parameter; analysis
5	Steam; production; plant; power; reinjection; turbine
6	Seismic; fault; area; survey; map; earthquake
7	Drilling; well; hole; log; depth; temperature
8	Resistivity; alteration; rock; hydrothermal; basement; age
9	Corrosion; silica; scale; concentration; brine; tracer

4. INTELLIGENT SEARCH

In this section, we introduced an improved search engine to facilitate the allocation of the most relevant publications to a search query. Unlike the traditional keyword-based search engines, this approach projects the publication abstracts to a latent hyperspace using a deterministic vector space model called term frequency—inverse document frequency (TF-IDF). TF-IDF measures the importance of a word to a document with respect to the corpus (Shi et al. 2009). Abstracts were first preprocessed through a pipeline of steps before clustering. This included tokenization, removal of stopwords (e.g. “the”, “a”, “an”, “in”, etc.), creation of bigrams (sequences of two adjacent words), stemming, and lemmatization of words to only keep nouns, adjectives, verbs, and adverbs (Denny and Spirling 2017). To compute the TF matrix, we construct a counting matrix whose rows and columns represent abstracts and corpus terms, respectively. In other words, the element at the i^{th} row and j^{th} column indicates the count of the j^{th} word in the i^{th} abstract. Meanwhile, IDF adds a normalization factor where words which are generally frequent in the corpus (e.g. “geothermal”, “project”, “study”, etc.) have smaller IDF factors, hence their importance diminishes with respect to an abstract in particular.

Subsequently, each abstract is embedded into a vector whose length is equal to the total number of unique terms in the corpus. When a new search query is made by the user, the query is transformed into a TF-IDF vector and compared to all abstracts in the geothermal literature using cosine similarity. Because we only have 18,873 publications in total, this process is computationally fast. In the case where the literature is significantly larger, the computation can be reduced in different ways, e.g. performing PCA on the TF-IDF vectors, only comparing the query TF-IDF vector to those publications which belong to the same LDA topic, etc. Table 2 shows the top five publications returned by this tool for an example search query: “Future of geothermal in New Zealand”. This tool is made available on the project API with additional features, where the user can filter by year, author, conference, amongst others.

Table 2: Ranked Search Results for the Query “Future of geothermal in New Zealand”

Title	IGA Database Link (Stanford Mirror Site)
2015 New Zealand Country Update	https://pangea.stanford.edu/ERE/db/IGAstandard/record_detail.php?id=23259
Geothermal: The Next Generation	https://pangea.stanford.edu/ERE/db/IGAstandard/record_detail.php?id=29182
New Zealand’s Supercritical Opportunity: Moving from Potential Resource to Deployed Technology	https://pangea.stanford.edu/ERE/db/IGAstandard/record_detail.php?id=29485
The Rise and Rise of Geothermal Heat Pumps in New Zealand	https://pangea.stanford.edu/ERE/db/IGAstandard/record_detail.php?id=19232
2020 New Zealand Country Update	https://pangea.stanford.edu/ERE/db/IGAstandard/record_detail.php?id=29354

5. AUTHOR NETWORK

In this section, we performed network analysis on the geothermal community based on the published literature. This involves calculating insightful statistics on the individual author contribution to the geothermal literature and also the connectivity of authors as a network. An undirected graph of the author network is first constructed, i.e. authors are nodes which are connected by edges based on coauthorship (Hagberg et al. 2008). Fig. 7 shows a Circos plot representation of the undirected graphical model of the geothermal community based on literature coauthorship. In this plot, the top 50 authors in publication count are arranged on a circle with connections in between. While Roland Horne has the highest contribution with 158 publications in total, author nodes are oriented such that the number of publications decreases clockwise, i.e. top three authors in order are Roland Horne, Joseph Moore, and Karsten Pruess. Meanwhile, the width of the grey edges represents the frequency of coauthorship between two authors.

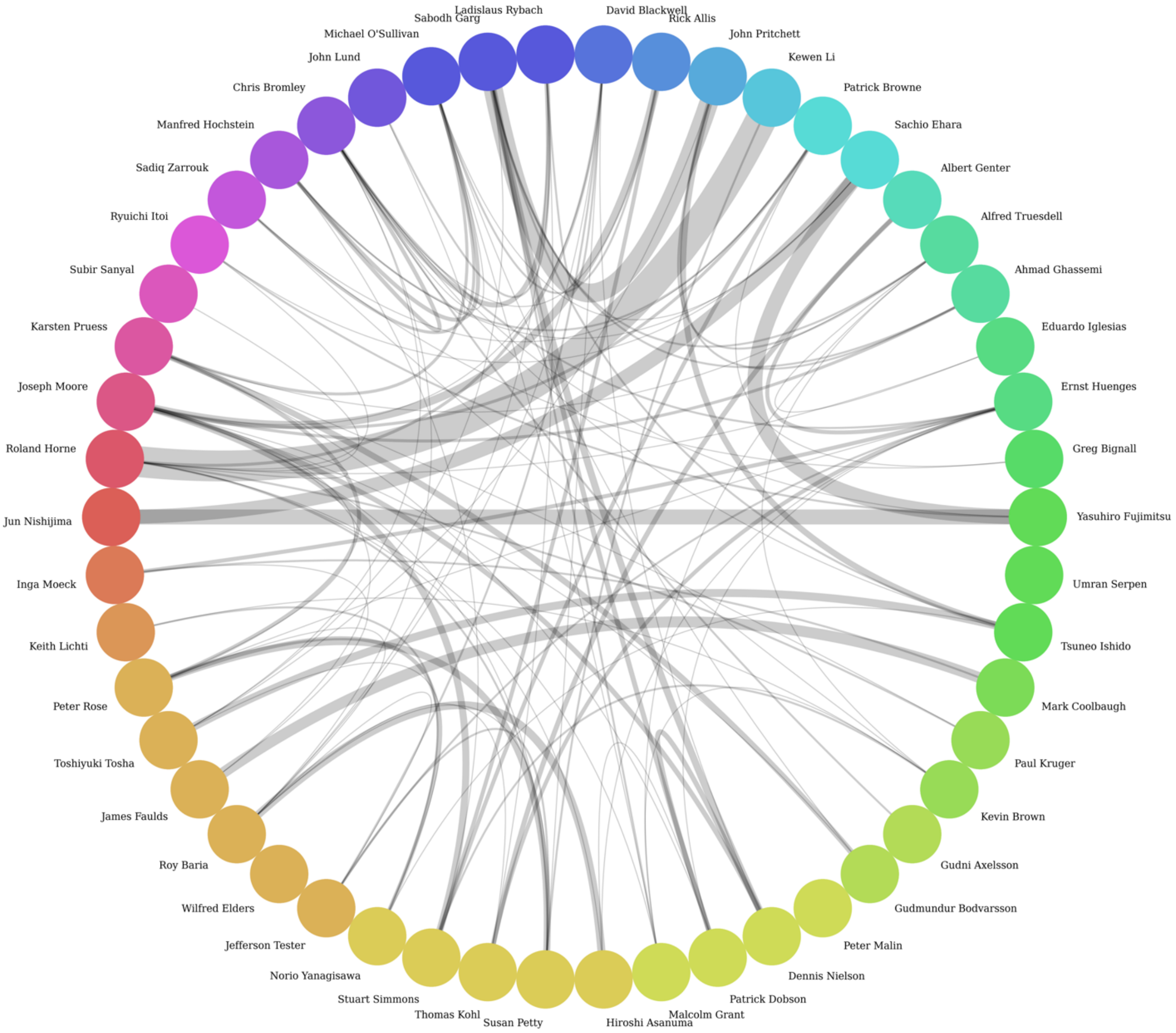


Figure 7: Circos plot of the top 50 authors in the geothermal literature where authors are arranged based on publication count starting at the nine o'clock position and decreasing clockwise.

While publication count is a primary metric, another important measure is author centrality (Bavelas 1948). We focused on two centrality measures: degree centrality and betweenness centrality. Degree centrality is simply the total number of edges connected to an author node (Freeman 1978). Degree centrality is a measure of "collaboration", where higher scores indicate that an author is relatively more collaborative. Meanwhile, betweenness centrality measures node vorticity or the number of times an author node contributes into the shortest path between two other author nodes (Freeman 1977, Brandes 2001). Betweenness centrality reflects how essential an author is in connecting to other authors within the network. Fig. 8 shows degree centrality and betweenness centrality scores of the top ten authors in terms of publication count. Roland Horne, Joseph Moore, and Chris Bromley are the top three authors with respect to these centrality measures. In fact, centrality measures of the author network graph provide other practical applications. They allow for finding the shortest path/s between any two authors, which provides users with personalized shortest paths to connect with any author within the geothermal community based on their personal connections.

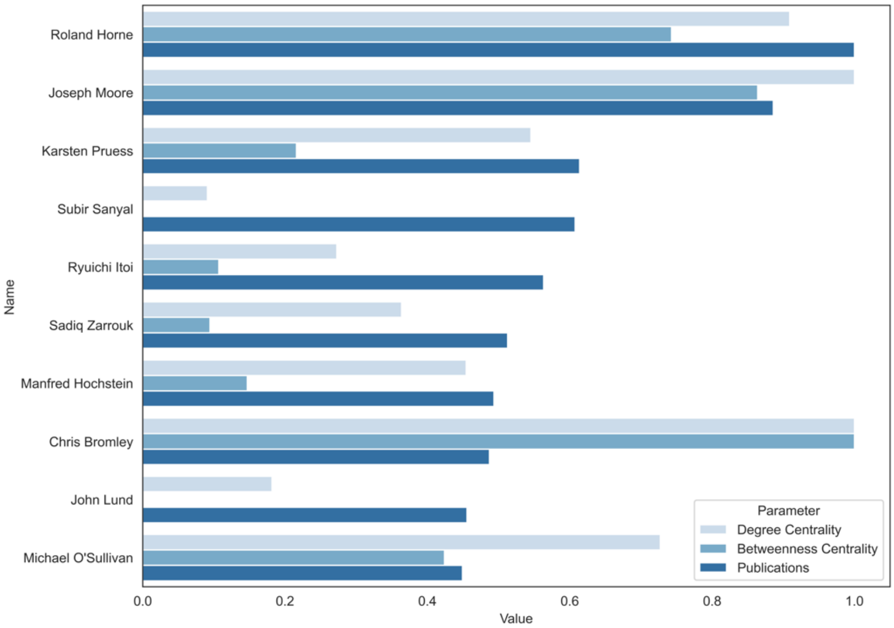


Figure 8: Ranking of top ten authors based on publication count. Degree centrality and betweenness centrality scores are also included.

6. TEXT GENERATION AND AUTO-COMPLETION

Deep learning models proved to have many applications in the NLP arena. One such application is text generation and sentence/paragraph auto-completion. This application is similar to the popular smartphone language models which predict the next word based on what you have just typed. The majority of language modelling tasks (e.g. machine translation, question answering, sentiment analysis, etc.) are approached using supervised techniques and algorithms (e.g. recurrent neural networks, temporal convolutional neural networks, encoder-decoder architectures, etc.). Yet, these models are task-specific (Kirkpatrick et al. 2017) and sensitive to changes in the data distribution (Recht et al. 2018). Instead, unsupervised multitask models were introduced to tackle multiple tasks simultaneously and produce generalizable models (Caruana 1997). In this work, we utilized a version of the OpenAI's state-of-the-art generative pretrained transformer (GPT-2) which is a decoder that autoregressively generates the next token (term) in a sequence of tokens (Radford et al. 2019). While the model only outputs the probability of the next token in a sequence, beam search was used to choose the next token by maximizing the average log-likelihood (Boulanger-Lewandowski et al. 2013).

The number of trainable GPT-2 weights (parameters) depends on the number of transformer blocks. To develop this geothermal text generation tool in particular, we utilized GPT-2 with 124 million weights pretrained on WebText, a large corpus of English text. Using a pretrained model ensures that the transformer blocks can already construct grammatically and structurally correct English sentences. However, the model does not understand the geothermal content in scientific details as it is trained on text from various domains and based on a variety of writing styles. Hence, we retrained GPT-2 on the geothermal literature abstracts to learn how to generate or auto-complete an abstract in the domain of geothermal energy. This model is made accessible on the project API for users to use. The model can be used to stochastically generate geothermal text, or auto-complete an input text. Table 3 shows examples of this tool output based on selected inputs.

Table 3: Example Input/Output of Text Generation and Auto-Completion

Input	Model Output
empty	The experimental data are presented as a quadrangle diagram and show that the modal distributions of fracture aperture, aperture-to-depth, post-stimulation fluid velocity, and surface area are strongly correlated with the reservoir temperature. The results of the field experiments are compared with the analytical solutions. A method is presented to determine the rate of water injection into a reservoir using two different methods: The first one is a reservoir simulation based on the conservation of mass and energy (enthal …
The future of the geothermal industry	The future of the geothermal industry will depend on the future development of geothermal technologies in the countries. In the cases of the energy technology of a particular type of geothermal resource, the next demand is the high temperature geothermal resources of the world. For this purpose, the government of Indonesia has taken the lead in the development of geothermal resources in Indonesia. The development of a geothermal resource and of geothermal energy resources has been prioritized by the government. The geothermal resource in Indonesia has been classified as "medium …

Enhanced geothermal systems	Enhanced geothermal systems and in particular the EGS Collab project at the Sanford Underground Research Facility in South Dakota with the aim of identifying barriers to further development of the geothermal industry. The fluid flows in the sample dominated reservoir from the reservoir through a network of fractures within the rock matrix to the fluid during circulation. In this paper a model of fluid flow within a fault to explain the flow in the reservoir is presented. The model is based on a mathematical model for the strongly fractured region where the …
Tracer tests	Tracer tests in the Los Azufres-Hueco geothermal field, Mexico, are important to the geothermal development. A new geochemical model is developed based on chemical and isotopic data from a geothermal fluid which is mixing with a cold meteoric water. Results show that the recharge of the meteoric water is derived from the discharge of the deep recharge of the volcanic zone. The meteoric water has a high concentration of boron causing it to enter the geothermal reservoir …
Fracture modelling	Fracture modelling in the Pannonian Basin is presented. The modelling of the Pannonian Basin is based on the geological and structural work of the region's most prominent geothermal areas, geochemical and geophysical methods. In the Pannonian Basin, the known geothermal resources are stored in the Tertiary basement, which is not easily accessible. The new two-dimensional (2D) models of the Pannonian basin reveal …

CONCLUSION

This work performed a deep analysis of the geothermal literature, and introduced a live programming interface API where such insights are demonstrated and updated. The work further introduced three tools to facilitate personalized search and manipulation of the geothermal literature. These tools are: 1) intelligent search, 2) author network, and 3) text generation and auto-completion. Careful preprocessing and cleansing of the geothermal literature had to be performed at first to ensure consistent and high-quality data before launching into analytics and development. Statistics, trends, and tools were developed using a variety of state-of-the-art NLP and deep learning algorithms. These visualizations and tools have been made available publicly, and are hosted on a live and open-source API which can be accessed at http://steaming-geothermal-analytics.info.

REFERENCES

Bavelas, Alex. 1948. A mathematical model for group structures. *Human organization* **7** (3): 16-30.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* **3** (Jan): 993-1022.

Boulanger-Lewandowski, Nicolas, Bengio, Yoshua, and Vincent, Pascal. 2013. Audio Chord Recognition with Recurrent Neural Networks. *Proc.,* ISMIR335-340.

Brandes, Ulrik. 2001. A faster algorithm for betweenness centrality. *Journal of mathematical sociology* **25** (2): 163-177.

Caruana, Rich. 1997. Multitask learning. *Machine learning* **28** (1): 41-75.

Chuang, Jason, Manning, Christopher D, and Heer, Jeffrey. 2012. Termite: Visualization techniques for assessing textual topic models. *Proc.,* Proceedings of the international working conference on advanced visual interfaces74-77.

Denny, Matthew and Spirling, Arthur. 2017. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *When It Misleads, and What to Do about It (September 27, 2017).*

Falush, Daniel, Stephens, Matthew, and Pritchard, Jonathan K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164** (4): 1567-1587.

Freeman, Linton C. 1977. A set of measures of centrality based on betweenness. *Sociometry*: 35-41.

Freeman, Linton C. 1978. Centrality in social networks conceptual clarification. *Social networks* **1** (3): 215-239.

Hagberg, Aric, Swart, Pieter, and S Chult, Daniel. 2008. Exploring network structure, dynamics, and function using NetworkX, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* **114** (13): 3521-3526.

Pritchard, Jonathan K, Stephens, Matthew, and Donnelly, Peter. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155** (2): 945-959.

Radford, Alec, Wu, Jeffrey, Child, Rewon et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* **1** (8): 9.

Recht, Benjamin, Roelofs, Rebecca, Schmidt, Ludwig et al. 2018. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:180600451.*

Röder, Michael, Both, Andreas, and Hinneburg, Alexander. 2015. Exploring the space of topic coherence measures. *Proc.,* Proceedings of the eighth ACM international conference on Web search and data mining399-408.

Shi, Congying, Xu, Chaojun, and Yang, Xiaojiang. 2009. Study of TFIDF algorithm. *Journal of Computer Applications* **29** (6): 167-170: 180.

Sievert, Carson and Shirley, Kenneth. 2014. LDAvis: A method for visualizing and interpreting topics. *Proc.,* Proceedings of the workshop on interactive language learning, visualization, and interfaces63-70.