Do You Know Where Your Research Is Being Used?An Exploration of scientific literature using Natural Language Processing

Theodore J. LaGrow, Jacob Bieker, Boyana Norris

University of Oregon

Author Note

Theodore J. LaGrow and Jacob Bieker are undergraduate researchers for Boyana Norris in the High-Performance Computing Laboratory at the University of Oregon.

Correspondence concerning this article should be addressed to Theodore J. LaGrow, Jacob Bieker, Boyana Norris, Computer and Information Science, University of Oregon, Eugene, OR 97403.

Contact: tlagrow@uoregon.edu, jbieker@uoregon.edu, norris@cs.uoregon.edu

Abstract

In such a complex and dynamic field as computer science, it is of interest to understand what resources are available, how much the resources are used, and for what the resources are used. We demonstrated the feasibility of automatically identifying resource names on a large-scale from scientific literature in arXiv’s database and showed that the generated data can be used for exploration of software and topics. While scholarly literature surveys can provide some insights, large-scale computer-based approaches to identify mentions of technology and methods from primary literature is needed to automate systematic cataloguing and facilitate the monitoring of usage in a more effective method. We developed a software using Natural Language Processing to evaluate if the article relates to the technology and methods in question. We were then able to evaluate a trend of technology and methods used in each specific areas of science. As we continue to expand this software, we will analyze the researchers’ sentiment about the technology and methods.

*Keywords*: natural language processing, scientific literature, database, computer software

Do You Know Where Your Research Is Being Used?

An Exploration of scientific literature using Natural Language Processing

With expanding databases of scientific articles, there is exceedingly greater access to publishing’s of specific scientific topics. Hucka and Grahams (2016) suggest in their article *Software search is not a science, even among scientists,* “When searching for ready-to-run software, the top five approaches overall are: (i) search the Web with general-purpose search engines, (ii) ask colleagues, (iii) look in the scientific literature.” These dated technology search methods can be painstaking and arduous. We were curious if there was a method to finding trends of technology usage utilizing large data from these databases.

Recently, linguistic machine learning has been implemented to inference across large data sets. (Bird et al, 2009) Scientific databases can be incorporated into large sets of collections from a given number of articles using various methods of text extraction and filtering. The use of linguistic machine learning can be used to understand greater meaning with this type of large data. We decided to use Natural Language Processing to explore and infer about what types of technologies and methods are being used in varying disciplines of science.

**Natural Language Process Overview**

Bird et al. (2009) describes Natural Language Processing as the ability of a computer program to understand human speech as it is spoken. Natural Language Processing is a field of artificial intelligence and computational linguistics concerned with the interactions between computers and natural languages. The inference usually consists of human language. Modern NLP is based on machine learning, especially statistical machine learning. The programing paradigm of machine learning is different from that of most prior attempts at language processing. Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. (Jones, 2001) Starting in the late 1980s, however, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. This was due to the steady increase in computational power over time. (Jones, 2001) Machine learning calls for using general learning algorithms, often, grounded in statistical inference. This is to automatically learn such rules through the analysis of large corpora of typical real-world examples. A corpus is a set of documents (or sometimes, individual sentences or strings) that have been hand-annotated with the correct values to be learned.

**Method**

To create the corpus of data, we first parsed through arXiv.org search results for our topics of interest. arXiv.org is a major hub of researchers to publish their articles as their papers get peer-reviewed. The four topics we were curious about were: galaxy evolution, Hawkes Processes, t-cell receptor genomes, and Natural Language Process itself. We then went through the downloaded PDFs, extracting text using PDFMiner and Python. We decided to only extract the first 100 articles from the topic searches because of computing capabilities of our computers. Once we converted the PDFs to text, we applied filters to the text to remove non-alphanumeric characters, and removed any lines that were less than seven characters, to clean up the text documents. Once the documents were cleaned, we used the Natural Language Toolkit to parse the text, giving us the parts of speech of each word, a frequency distribution of n-grams containing predefined interesting words, and finding words similar to the user-defined interesting words. n-grams take an interesting word and use it as a center point in the string of the ‘n’ given length. We decided to use n-grams of 15 because the average length of a sentence is 6-7 words giving us roughly the sentence on either side of the interesting word. Once that was done, we went trough the collection of n-grams, only taking the noun phrases from the n-grams and counting the occurrences of each noun phrase. The counted noun phrases became the basis for the generated word clouds, which visualize the hierarchical significance of the word to the corpus of data related to the discipline being examined.

**Dictionary of Interesting Words**

simulation, software, code, analysis, using, program, analyzed, scripted, automated, description, implements, function, modifies, operated, pipeline, helps, allows, manipulate, processed

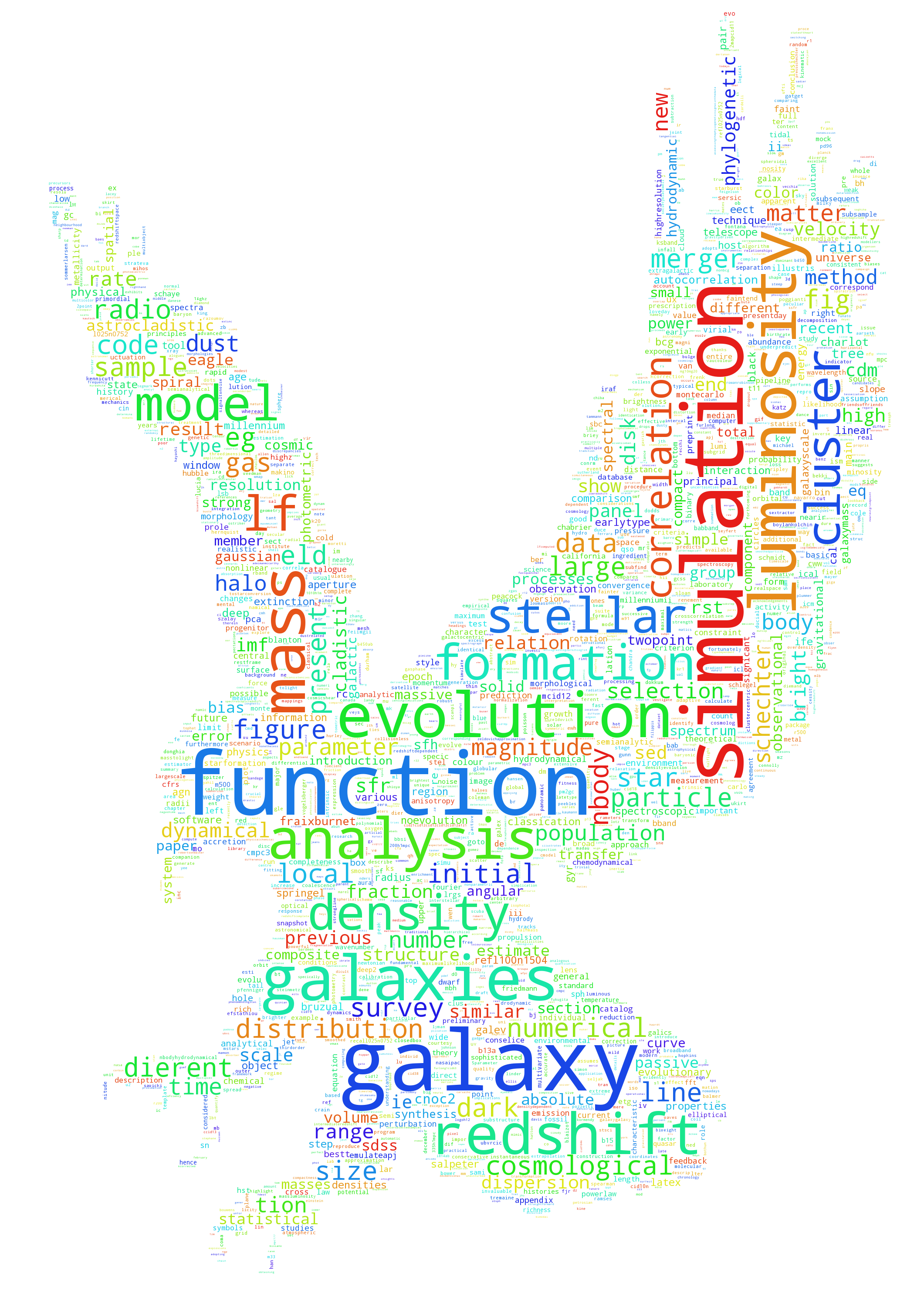
**Results**

We found that each data set outputted a variety of similar words. A few similar words included: function, method, and analysis. These words had relatively high frequencies compared to the more unique words related to the data sets. We suspect since these words are in our interesting word dictionary, they typically occur close to the other interesting words in our corpus. Interesting results found include: Gadget (a galaxy imaging technology), Velvet (an assembly program), and morphological (a method dealing with the structure of things).

**Output Frequencies**

Search keyword: Hawkes Process, limit top 30 frequencies after analysis

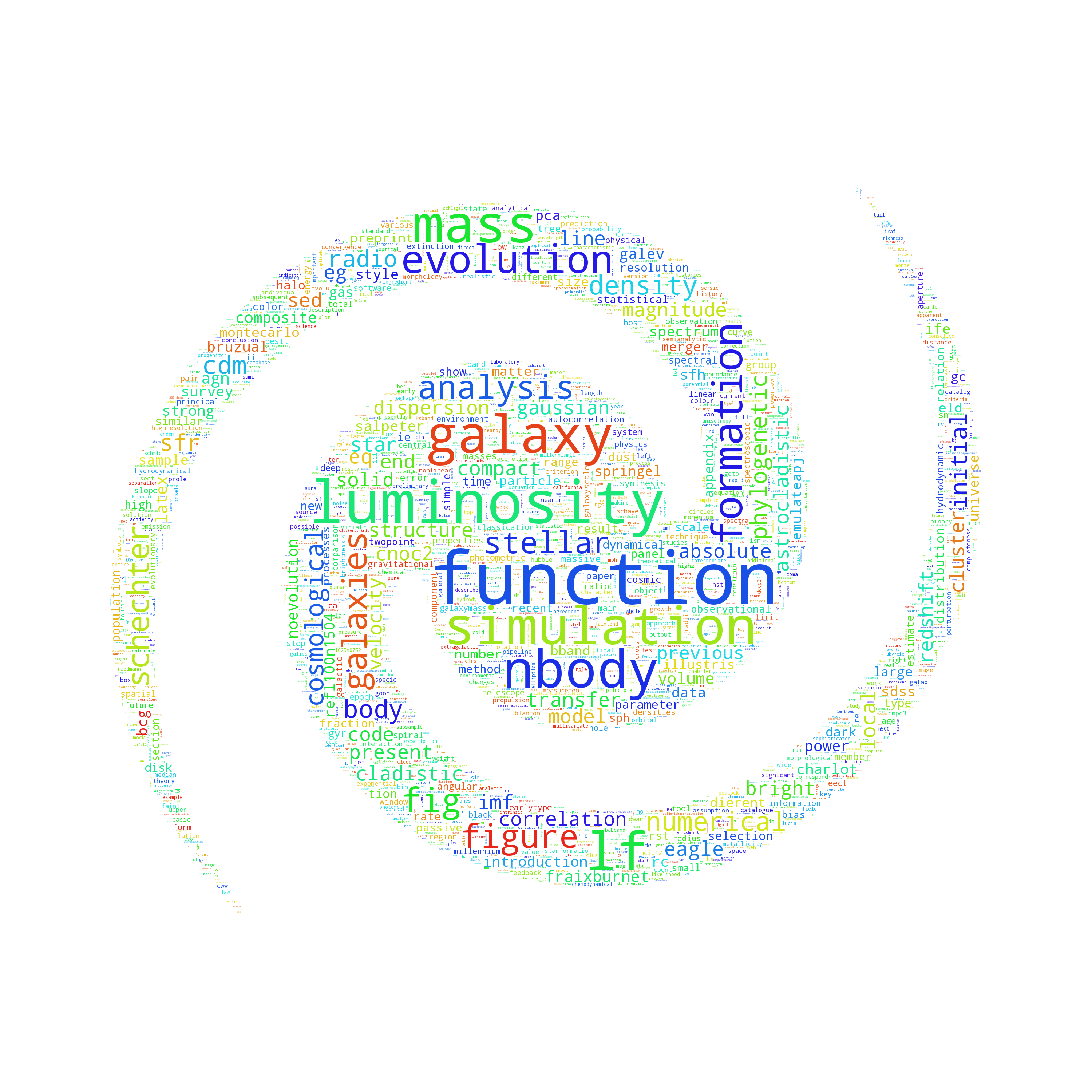
{'hawkes': 815, 'rate function': 349, 'large deviation principle': 113, 'lemma': 109, 'exciting function': 107, 'point processes': 99, 'eq': 95, 'theorem': 90, 'poisson': 82, 'fig': 78, 'residual analysis': 78, 'hawking': 74, 'ix': 70, 'black hole': 67, 'intensity function': 58, 'correlation function': 54, 'conditional intensity function': 54, 'contrast function': 51, 'excitement function': 51, 'consider': 49, 'genome analysis': 48, 'numerical simulations': 44, 'simulation study': 44, 'morphological’ : 42, 'partition function': 42, 'exponential function': 40, 'distribution function': 39, 'cost function': 39, 'kernel function': 38, 'wienerhopf': 38, 'fourier': 37}



*Figure 1: Output distribution word cloud of the search keyword: Hawkes Process*

Search keyword: galaxy evolutions, limit top 30 frequencies after analysis

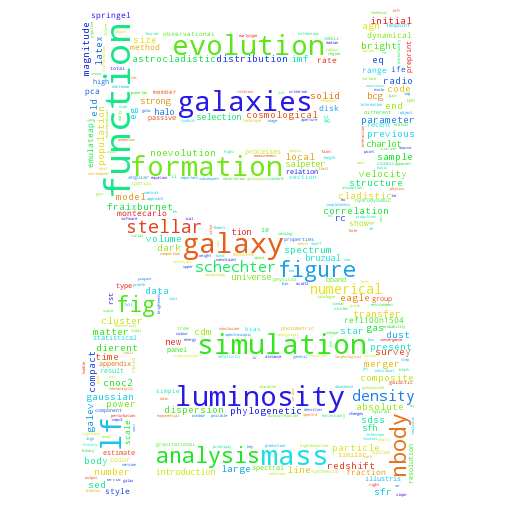
{'luminosity function': 332, 'nbody': 145, 'fig': 128, 'figure': 113, 'schechter': 101, 'galaxy luminosity function': 72, 'galaxy evolution': 71, 'galaxy formation': 70, 'cdm': 67, 'phylogenetic analysis': 65, 'body simulations': 63, 'numerical simulations': 60, 'initial mass function': 54, 'cosmological simulations': 53, 'astrocladistics': 52, 'mass function': 49, 'stellar mass': 45, 'transfer': 45, 'eagle': 45, 'local density': 39, 'compact galaxies': 39, 'gaussian': 37, 'cladistic analysis': 36, ‘gadget-3’: 36, 'radio galaxy luminosity function': 36, 'star formation': 36, 'cluster galaxies': 33, 'correlation function': 33, 'bright end': 33}



*Figure 2: Output distribution word cloud of the search keyword: galaxy evolution*

Search keywords: T-cell receptor genome, limit top 30 frequencies after analysis

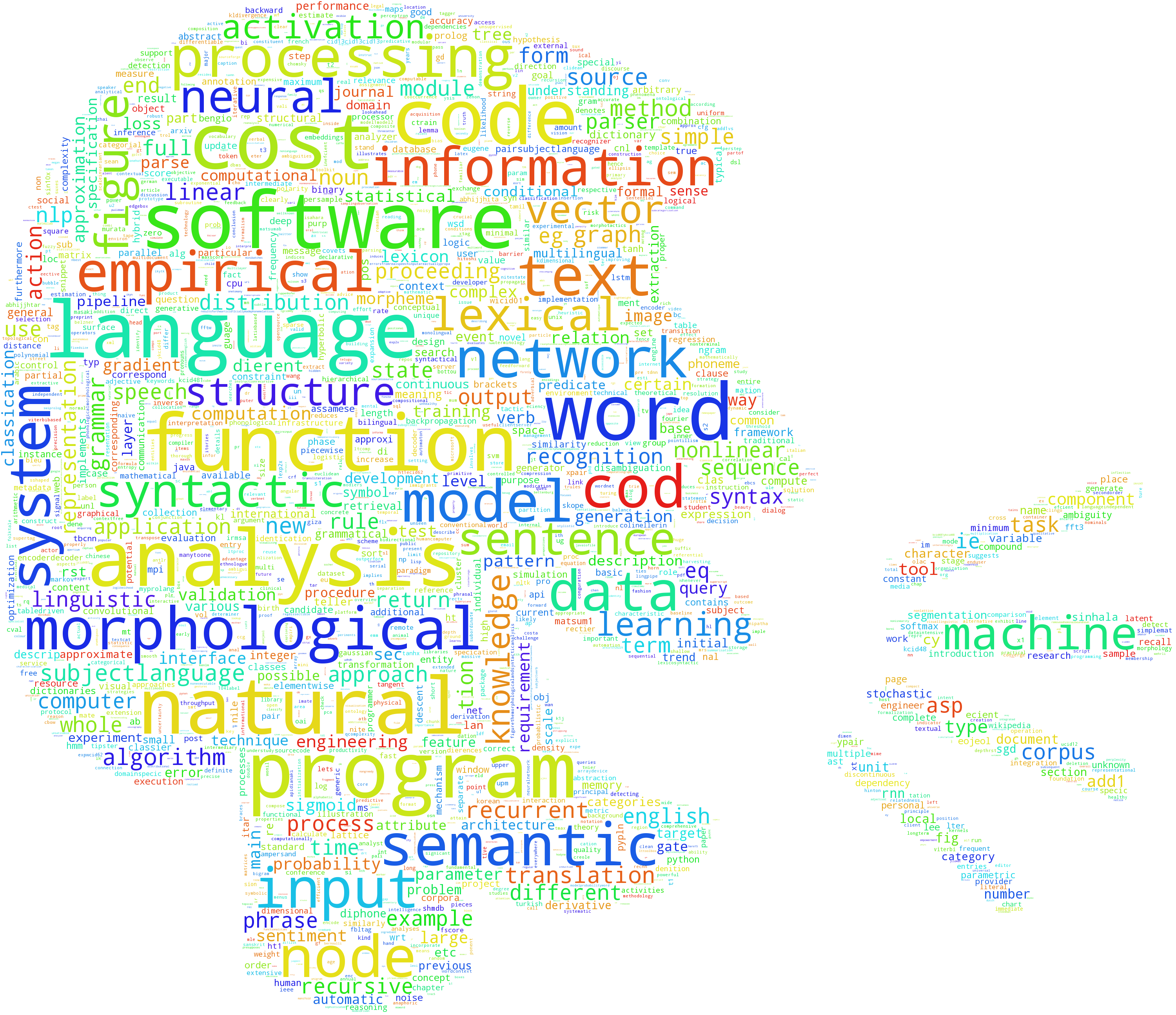
{'monte carlo': 244, 'eq': 244, 'fig': 241, 'figure': 119, 'tcr': 82, 'dna': 71, 'rna': 68, 'snps': 59, 'chipseq': 59, 'numerical simulations': 58, 'partition function': 54, 'ligand concentration': 53, 'methods': 52, 'correlation function': 52, 'mc': 50, 'gillespie': 46, 'rnaseq': 44, 'microarray analysis': 43, 'maximum likelihood': 41, 'bayesian': 39, 'simulation study': 39, 'velvet': 39, 'data analysis': 38, 'snp': 37, 'stochastic simulation': 36, 'cluster size': 36, 'covariance function': 35, 'dierent values': 30, 'greens': 28, 'phylogenetic analysis': 27, 'quantitative analysis': 27}



*Figure 3: Output distribution word cloud of the search keyword: T-cell receptor genome*

Search keywords: Natural Language Processing, limit top 30 frequencies after analysis

{'cost function': 260, 'figure': 159, 'morphological analysis': 118, 'empirical cost function': 114, 'nlp': 102, 'proceedings': 101, 'asp': 88, 'function f': 79, 'syntactic analysis': 76, 'english': 75, 'eq': 74, 'language': 71, 'function node': 70, 'x language': 58, 'fig': 56, 'sec': 54, 'cost function c': 52, 'y subjectlanguage': 47, 'sigmoid function': 47, 'lexical analysis graph': 46, 'function approximation': 44, 'empirical cost function c': 42, 'sentiment analysis': 41, 'semantic analysis': 41, 'morphological': 40, 'activation function': 39, 'pairsubjectlanguage code': 37, 'recursive function': 36, 'machine learning': 35, 'teller machine': 34}



*Figure 4: Output distribution word cloud of the search keyword: Natural Language Processing*

**Limitations**

We only used 100 articles for each scientific topic. Each search varied in amount of PDFs, however we wanted to make sure we were able to keep around the same corpus size for each analysis. The data sets ranged to around 600,000 strings and 29,000,000 characters after being parsed with n-grams. These are not terribly large files of text, however to iterate over each string can take some time. The program took around twenty minutes to run the corpus creation where we downloaded each PDF and extracted and filtered the text, then another half hour to run our analysis program. The PDF parser program we developed is not the most efficient we could have used. Most of the time the parser worked, however at time when a PDF was older than a certain data, or had too many pictures, or was too short, the text would come out at all one text block or ASCII characters and we would have to through that document away. In the future, we will seek proprietary means of extracting text from PDFs.

**Conclusion**

The results of our analysis demonstrate we can evaluate trends of technology and methods in various disciplines. This information lays the groundwork for building a network of software used by various researchers to evaluate the effectiveness of National Science Foundation funding of different software projects. From these initial results, we are planning on continuing improving the software to extract common methods and tools used in research in any given discipline from the literature, with the hope of connecting researchers to tools that they might not know about, or informing the development of future software packages to better tailor them to their users.

References

Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.

Hucka, M., & Graham, M. J. (2016, May 8). ArXiv.org cs arXiv:1605.02265. Retrieved May 11, 2016, from https://arxiv.org/abs/1605.02265

Jones, Karen S. "Natural Language Processing: A Historical Review". Language (2001): n. pag. Print.

Noam Chomsky's Theories on Language. (n.d.). Retrieved May 13, 2016, from http://study.com/academy/lesson/noam-chomsky-on-language-theories-lesson-quiz.html

What is natural language processing (NLP)? - Definition from WhatIs.com. (n.d.). Retrieved May 13, 2016, from http://searchcontentmanagement.techtarget.com/definition/natural-language-processing-NLP