Applying Lexical Link Analysis to Discover Insights from Public Information on COVID-19

Ying Zhao[a,1,2] **and Charles C. Zhou**[b,1,2]

[a]Information Sciences Department, Naval Postgraduate School, Monterey, CA 93943, USA; [b]Quantum Intelligence, Inc., Monterey, CA 93943, USA

This manuscript was compiled on April 10, 2020

SARS-Cov-2, the deadly and novel virus , which has caused a worldwide pandemic and drastic loss of human lives and economic activities. An open data set called the COVID-19 Open Research Dataset or CORD-19 contains large set full text scientific literature on SARS-CoV-2. The Next Strain consists of a database of SARS-CoV-2 viral genomes from since 12/3/2019. We applied an unique information mining method named lexical link analysis (LLA) to answer the call to action and help the science community answer high-priority scientific questions related to SARS-CoV-2. We first text-mined the CORD-19. We also data-mined the next strain database. Finally, we linked two databases. The linked databases and information can be used to discover the insights and help the research community to address high-priority questions related to the SARS-CoV-2's genetics, tests, and prevention.

information mining | text mining | data mining | coronavirus | SARS-CoV-2 |COVID-19| clade | mutation | genetics | tests | prevention

This paper is to answer the call to action to the Nation's data scientists and AI experts to mobilize and develop new text and data mining techniques that can help the science community answer high-priority scientific questions related to SARS-Cov-2, the deadly and novel virus, which has caused a worldwide pandemic and drastic loss of human lives and economic activities.

Recently, an open data set, namely the COVID-19 Open Research Dataset or CORD-19 (2) containing more than 40,000 full text scientific literature on SARS-CoV-2, was released (1, 2) for researchers across technology, academia, and the government. It is hopeful that machine learning and AI community might employ advanced data sciences to surface unique insights in the body of data and help answer with high-priority questions related to genetics, incubation, treatment, symptoms, and prevention. The corpus is updated weekly as new research is published in peer-reviewed publications and archival services like bioRxiv, medRxiv, and others. In this paper, we show how to apply an unique information mining method lexical link analysis (LLA) to link unstrurctured and structured data address the research questions below:

1. How to extract themes and topics that address the high-priority questions of genetics, incubation, treatment, symptoms, and prevention of COVID-19?

2. How to extract valued information such as information with authority, insights and innovation for combating SARS-Cov-2? What is the authoritative information and insightful information?

3. What are the timelines of themes and topics across all the research literature?

We show LLA that integrates text and data mining into a single platform so that insights from data can be visualized, and validated to answer the high-priority questions.

Materials and Methods

Overview of Lexical Link Analysis. Traditionally in social networks, the importance of a network node is a form of high-value information. For example, the leadership role in a social network (7, 8) is measured according to centrality measures (9). Among various centrality measures, a common practice is to sort and rank information based on authority. Current automated methods, such as graph-based ranking used in many search engines (10), require established hyperlinks, citation networks, social networks, or other forms of crowd-sourced collective intelligence. However, these methods are not applicable to situations where there exist no pre-established relationships among network nodes such as the data set SARS-Cov-2. This makes traditional methods difficult to apply. Furthermore, current methods mainly score popular information. High-value information can be totally different depending on specific applications. Popular and authoritative information identified by the current methods can be useful for marketing applications or crowdsourcing applications. Emerging and anomalous information is important for looking for insights and innovation. In paper, we show how to apply game-theoretic framework of lexical link analysis (LLA) to discover and rank high-value information from unstructured and structured data from SARS-Cov-2.

In LLA, a complex system can be expressed in a list of attributes or word features with specific vocabularies or lexicon terms to describe its characteristics. LLA is first a data-driven text analysis method. Fig. 1 shows an example of extracting and learning word pairs, or bi-grams as lexical terms, from text data. Words from a text document are represented as nodes, which form word pairs or bi-grams, via the links between any two nodes. For instance, the node "anitiviral" in Fig. 1 is formed with "chain-terminating," "broadspectrum," etc. as bi-gram word pairs (word features). In contrast to human-annotated word networks, such as WordNet (13), LLA automatically discovers word pairs, divides them into clusters and themes, and displays them as word networks. Fig. 1 shows an example of LLA.

Bi-gram also allows LLA to be extended to structured data (15) including meta-data such as the ones for CORD-19, where a word

Significance Statement

In this paper, we show how to apply an unique information mining method lexical link analysis (LLA) to link unstructured (CORD-19) and structured (Next Strain) data sets to relevant publications, integrate text and data mining into a single platform to discover the insights that can be visualized, and validated to answer the high-priority questions of genetics, incubation, treatment, symptoms, and prevention of COVID-19.

Y.Z. and C.C.Z. designed and performed research, and wrote the paper.

The authors declare no conflict of interest.

[1] Y.Z. and C.C.Z contributed equally to this work.

[2] Email correspondence: yzhao@nps.edu or charles.zhou@quantumii.com

is an attribute combined with its possible values. LLA is related to but significantly different from the methods such as bag-of-words (BOW) methods, Automap (7), Latent Dirichlet Allocation (LDA) (14), Latent Semantic Analysis (LSA) (16), Probabilistic Latent Semantic Analysis (PLSA) (17) and can be jointly used with NEE (20, 21), POS methods (25). LLA is related to unsupervised learning algorithms such as k-means, principal component analysis (PCA), and spectral clustering (18). In a social network, the most connected nodes are typically considered the most important nodes. However, the uniqueness of LLA is that we consider anomalous information (word features) might be more interesting. Community detection algorithms have been illustrated by Newman (11, 12) in terms of a quality function as the "modularity" measure for a community (cluster) and optimized using a dendrogram-like greedy algorithm. The uniqueness of LLA includes new value metrics to identify high-value information that are not presented in the other methods.The new value metrics consider a game-theoretic framework (19, 35).

LLA Applied to CORD-19. LLA automatically discovers popular (P) themes, emerging (E) themes , and anomalous (A) themes (as defined below) as follows:

- Popular (P) themes: These themes resemble themes generated from the eigenvalue centrality measure in the network sciences. The themes represent the main topics in a data set. They can be insightful information in two folds: 1) These word pairs are more likely to be shared or cross-validated across multiple diversified domains, so they are considered authoritative; 2) These themes could be less interesting because they are already in public consensus and awareness and can be considered as popular.
- Emerging (E) themes: These themes tend to become popular or authoritative over time. An emerging theme has the intermediate number of inter-connected word pairs. They are emerging and important themes and can be high-value for further investigation.
- Anomalous (A) themes: These themes may not seem to belong to the data domain when compared to other themes. They are interesting outliers and can be high-value for further investigation.

Fig. 2 shows an example of extracted themes from a text data set where popular themes, e.g. 517(P), emerging theme, e.g. 42(E), and anomalous theme e.g. 478(A) among others are listed. Fig. 8 shows a drill-down visualization for theme 517(P) labeled "infection coronavirus, global coronavirus" The labeled nodes in Fig. 8 are the words with the most connections with other words (via bi-gram measures in LLA).

The Next Strain Database.

Data. Understanding the spread and evolution of a virus such as SARS-Cov-2 is important for effective public health measures and surveillance. A global strain tracking tool from nextstrain.org (4, 6), i.e., which strain evolves from which other strain. Nextstrain consists of a database of viral genomes, a bioinformatics pipeline for phylodynamics analysis of many novel virues such as Zika, Ebola, and SARS-Cov-2 with an interactive visualization platform. The visualization integrates sequence data with other data types such as geographic information, serology, or host species. Nextstrain compiles the current understanding of phylogenetic analysis into a single accessible location, open to health professionals, epidemiologists, virologists and the public alike (5).

Insights. Fig. 3 shows the phylogenetic tree of the SARS-Cov-2's similarity and development around the globe from 12/3/2019 to 3/25/2020 where the 2649 cases of patients' genomic data uploaded to the nextstrain.org website. Visually, Fig. 3 shows the strain in North America and Europe mutated to a different branch of the tree around 2/25/2020 and is more contagious. Fig. 4 shows the same data set colored by clades. Tab. 1 shows a split of the 2649 cases to three time periods: before 2/25/2020, from 2/25/2020 to 3/16/2020, after 3/16/2020. We computed the average case per day for Clade A2, A2a (the new mutation) and Clade A1a and other types. Clade A2, A2a is the mutated strain because it only has the data starting from 1/28/2020. Clade A1a, etc. and A2, A2a are statistically different ($p < 0.0001$) measured via average cases per day for each time period, i.e., 6.29 vs 0.24 (before 2/25/2020), 38.81 vs 46.86 (between 2/25/2020 and 3/16/2020), and 17.89 vs 31.33 (after 3/16/2020) . A2a almost started on 2/25/2020 since only 11 cases of A2, A2a happened before 2/25/2020. Clade A2,A2a is more contagious than Clade A1a, and other since the average cases per day is much higher, 46.86 and 31.33 between 2/25/2020 and 3/16/2020. After 3/17/2020, many countries in North America and Europe implemented "shelter-in-place," the average cases per day decreased for both clades, however, the mutated clade decreased statistically significantly slower (31.33) than A1a, B1, B2, etc. together (17.89).

Fig. 5 shows how the first case of Clade A2a linked to earlier cases of Clade A2. There are two insights as follows:

- Among the total 1277 cases of Clade A2, A2a, there were only 34 cases in Asia, including four cases in China and 18 cases in Japan. There were 955 cases in Europe, 216 cases in North

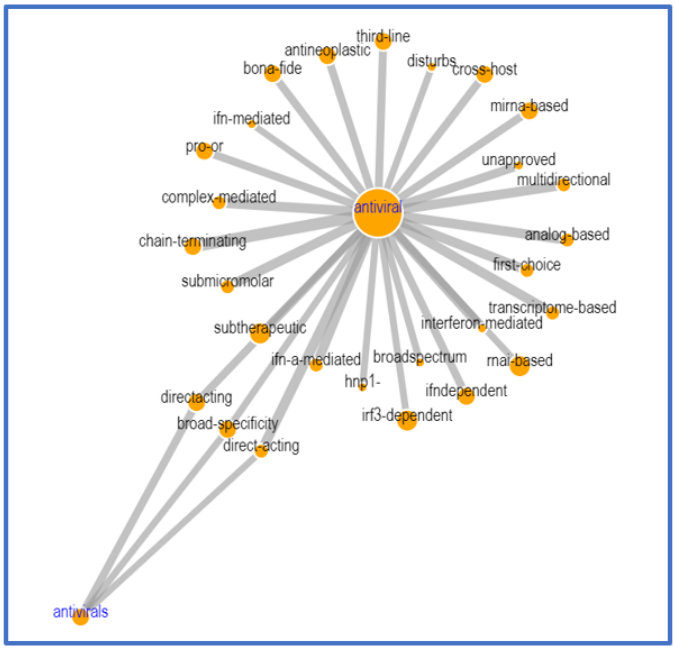

Fig. 1. LLA Example

ID	Theme Keywords
142(P)	di il,la il,il di
517(P)	infection coronavirus,global coronavirus
42(E)	antibody limited,antibody antibodies
374(E)	patients group,group canine,patients canine
233(E)	complete support,main
423(E)	protein mutant,gene mutant,gene protein,protein gene
126(E)	data functional,data primer
50(E)	assay rapid,database
478(A)	distinct membranes,membranes distinct,structures
114(A)	peptide inhibitor,peptide prevent,inhibitor peptide
451(A)	sequence rna,rna final,sequence final
480(A)	national increasing,vivo
318(A)	unique similarly,mm
152(A)	vector recombinant,plasmid vector,vector plasmid,recombinant vector,plasmid recombinant
100(A)	usa university,university usa,short
464(A)	detect,signaling,receptors
123(A)	chain,medicine,media
204(A)	reduce,weight,staining
515(A)	expressing strain,kit
471(A)	program software,numerous

Fig. 2. LLA themes

Table 1. Cases

A1a, B1, B2, etc	Cases	Days	Cases/Per Day
12/24/2019-2/24/2020	396	63	6.29
2/25/2020-3/16/202	815	21	38.81
3/17/2020-3/25/2020	161	9	17.89
A2, A2a			
1/28/2020-2/24/2020	11	46	0.24
2/25/2020-3/16/202	984	21	46.86
3/17/2020-3/25/2020	282	9	31.33

America, and 15 cases in Australia.

- Among the four cases of Clade A2, A2a in China, three were collected from 1/28/2020 to 2/6/2020, they were submitted around 3/20/2020. The fourth case was collected on 3/23/2020 and submitted at the same time.

At least the data shows the Clade A2, A2a in Europe and North America are more contagious and virulent.

Linking the Next Strain Data Set to the CORD-19 Data Set. The next strain data can be linked to the CORD-19 data set using LLA. We first extracted another document data set with documents pertinent to the next strain database. We then fused the two data sources based on themes in Fig. 2.

Fig. 6 shows a match matrix of number of matched word pairs from two document sources (the next strain and CORD-19), there are 2012 matched word pairs. Fig. 7 shows the examples of matched word pairs. These matched word pairs (concepts) are also grouped based on these themes. Fig. 8 shows a popular theme of word pair appeared in both data sources. Fig. 9, Fig. 10, and Fig. 11 show examples of emerging themes appeared in both data sources. Fig. 12 shows an example of an anomalous theme appeared in both data sources. Emerging themes are interesting topics for researchers to drill down and discover information in CORD-19 pertinent to high-priority questions of the SARS-Cov-2 of genetics (Group 423(E)), tests (Group 50(E)), and prevention (Group 42(E)).

Conclusion and Challenge to Future. We applied an unique information mining method lexical link analysis to conduct a preliminary study to the call to action and help the science

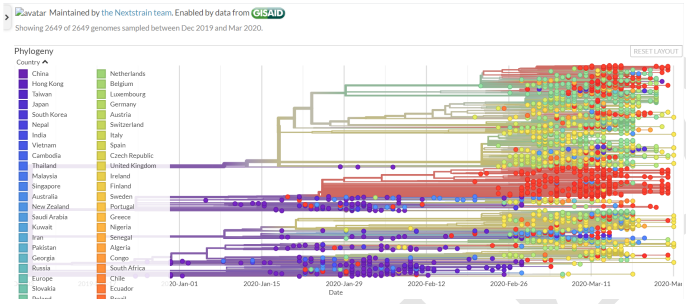

Fig. 3. Next strain data by country from 12/3/2019 to 3/25/2020. Source: nextstrain.org

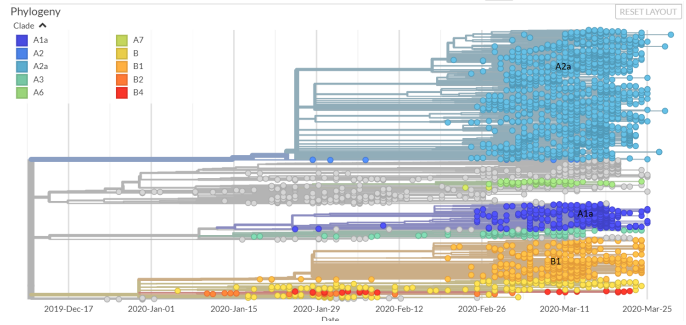

Fig. 4. Next strain data by clade from 12/3/2019 to 3/25/2020. Source: nextstrain.org

Strain	Admin Divis	gisaid_epi_l	url	Country	Age	Sex	Submission Date	Originating	Region	Clade	Collection Date	Author	genbank_acces	Location	Exposure	History
Germany/BavPat1/2020	Bavaria	EPI_ISL_406862		Germany		Male	Older	Charite Uni	Europe	A2	1/28/2020	Corman et al		Starnberg		
Shanghai/SH0014/2020	Shanghai	EPI_ISL_416327		China	47	Female	One week ago	Shanghai Pu	Asia	A2	1/28/2020	Wang et al				
Shanghai/SH0086/2020	Shanghai	EPI_ISL_416386		China	38	Male	One week ago	Shanghai Pu	Asia	A2	1/31/2020	Wang et al				
Shanghai/SH0025/2020	Shanghai	EPI_ISL_416334		China	81	Male	One week ago	Shanghai Pu	Asia	A2	2/6/2020	Wang et al				
Italy/CDG1/2020	Lombardy	EPI_ISL_412973		Italy	38	Male	Older	Department	Europe	A2a	2/20/2020	Stefanelli et al				
France/IHF1465/2020	Hauts de Fra	EPI_ISL_418218		France	56	Male	3-7 days ago	Centre Hosp	Europe	A2a	2/21/2020	Albert et al		CompiÃ¨gne		
Italy/UniMI02/2020	Lombardy	EPI_ISL_417446		Italy	76	Male	3-7 days ago	Laboratory i	Europe	A2a	2/24/2020	Zehender et al		Milan		
Italy/UniMI01/2020	Lombardy	EPI_ISL_417445		Italy	71	Male	3-7 days ago	Laboratory i	Europe	A2a	2/24/2020	Zehender et al		Milan		
Netherlands/Bericum_1363564/2020	North Braba	EPI_ISL_413565		Netherlands			One month ago	Foundation	Europe	A2a	2/24/2020	Nieuwenhuijse et al		Bericum		
Switzerland/TI9486/2020	Ticino	EPI_ISL_413996		Switzerland	70	Male	One month ago	Laboratoire	Europe	A2a	2/24/2020	LAUBSCHER Florian et al, et al				
Italy/UniMI03/2020	Lombardy	EPI_ISL_417447		Italy	80	Male	3-7 days ago	Laboratory i	Europe	A2a	2/24/2020	Zehender et al		Milan		
Finland/FIN-25/2020	Helsinki	EPI_ISL_412971		Finland	24	Female	Older	HUS Diagno	Europe	A2a	2/25/2020	Smura et al			Italy	
Germany/Baden-Wuerttemberg-1/2020	Baden-Wue	EPI_ISL_412912		Germany			Older	State Health	Europe	A2a	2/25/2020	Corman et al			Italy	
Spain/Madrid201105/2020	Madrid	EPI_ISL_418255		Spain	24	Male	3-7 days ago	HOSPITAL U	Europe	A2a	2/25/2020	Iglesias-Caballero et al				
Brazil/SPBR-01/2020	Sao Paulo	EPI_ISL_412964		Brazil	61	Male	Older	Hospital Isra	South Ameri	A2a	2/25/2020	Jaqueline Goes de Jesus et al		Sao Paulo	Italy	
Denmark/SSI-01/2020	Copenhager	EPI_ISL_416142		Denmark	45	Male	One week ago	Department	Europe	A2a	2/26/2020	Rasmussen et al				
Switzerland/GE3805/2020	Geneva	EPI_ISL_413997		Switzerland	28	Male	One month ago	Laboratoire	Europe	A2a	2/26/2020	LAUBSCHER Florian et al, et al			Italy	

Fig. 5. The first case of Clade A2a. Source: nextstrain.org

		Match Score	cord_19	nextstrain	Uniqueness Score
1	cord_19	2102.00	_	2102.00	777591.00
2	nextstrain	2102.00	2102.00	_	6490.00

Fig. 6. LLA match matrix

[517]diseases infectious
[517]tract respiratory
[517]respiratory acute
[517]disease infectious
[517]viruses respiratory
[517]syndrome respiratory
[517]virus influenza
[517]pcr real-time
[517]health global
[517]viruses influenza
[517]control disease
[517]size sample
[517]infections respiratory
[517]symptoms respiratory
[517]rate mortality
[517]disease respiratory
[517]severity disease
[517]products pcr
[517]failure respiratory
[517]pathogens respiratory
[517]infections tract
[517]infection tract
[517]illness respiratory
[517]influenza seasonal

Fig. 7. LLA matched list

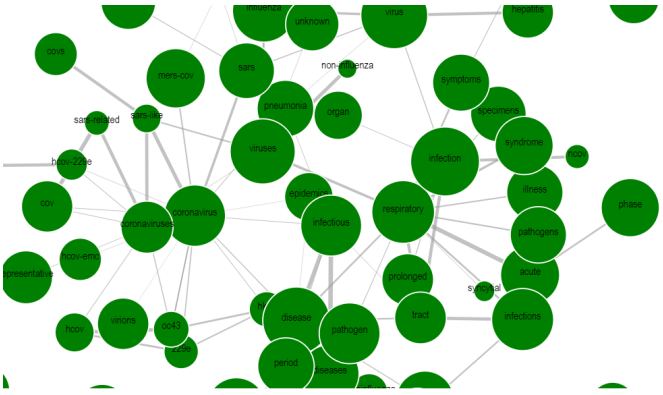

Fig. 8. LLA group 517(P)

community answer high-priority scientific questions related to SARS-Cov-2. We first text-mined an unstructured database, i.e. the COVID-19 Open Research Dataset or CORD-19. We also data-mined a structured database, i.e., the next strain database, covering the period from 12/3/2019 to 3/25/2020. Finally, we linked two databases and certain publications, and discovered the insights and methodologies. The linked documents from CORD-19 can help address high-priority questions related to SARS-COV-2's genetics, tests, and prevention.

There are some un-answered questions we need to ponder:
1. If certain data and publications before 12/3/2019 need to be mined and analysed, can one estimate the outcome of COVID-19 outbreak after 12/3/2019?
2. Can we use the data and publications up to today to estimate what will happen one month later?

1. COVID-19 Open Research Dataset (CORD-19). 2020. Version 2020-03-13. Retrieved from https://pages.semanticscholar.org/coronavirus-research. doi:10.5281/zenodo.3715506
2. https://techcrunch.com/2020/03/16/coronavirus-machine-learning-cord-19-chan-zuckerberg-ostp
3. https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/
4. Nextstrain https://nextstrain.org/ncov
5. github.com/nextstrain
6. Hadfield et al. (2018), Nextstrain: real-time tracking of pathogen evolution, Bioinformatics (2018)
7. Center for Computational Analysis of Social and Organizational Systems (CASOS). (2009) AutoMap: extract, analyze and represent relational data from texts. Retrieved from http://www.casos.cs.cmu.edu.
8. Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences," USA, 99(12), 7821–7826.
9. Freeman, L.C. (1979). Centrality in social networks I: conceptual clarification. Social Networks, 1: 215-239.
10. Brin, S. and Page, L. (1998). The Anatomy of a large-scale hypertextual web search Engine. Computer Networks and ISDN Systems, 30:107-117.

Fig. 9. LLA group 42(E)

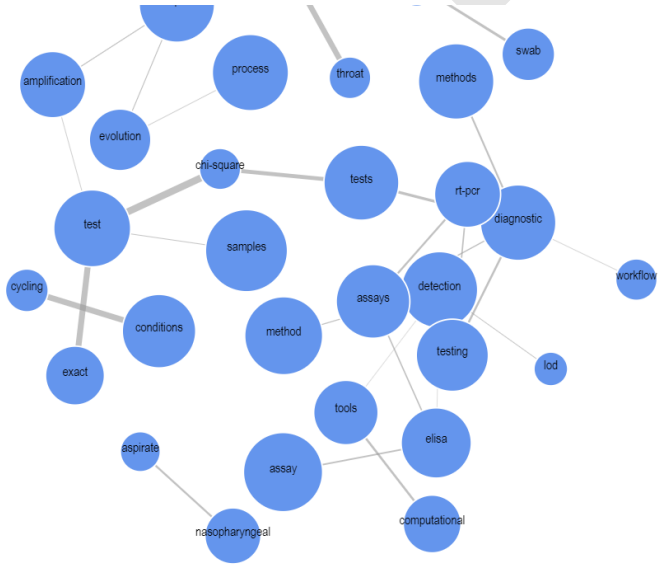

Fig. 10. LLA group 50(E)

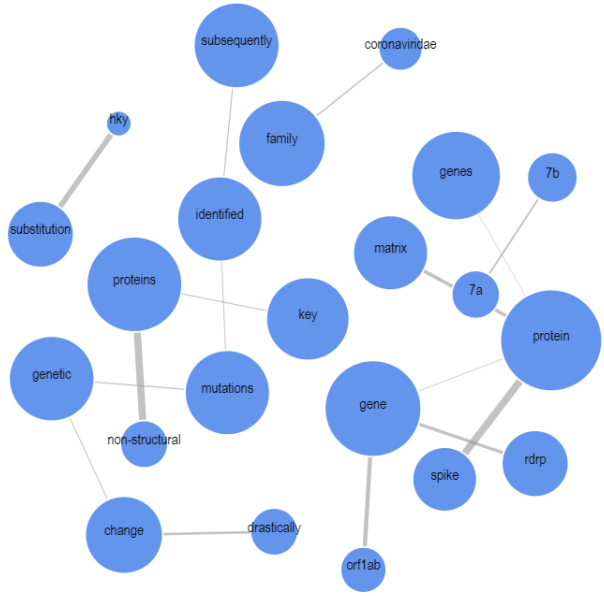

Fig. 11. LLA group 423(E)

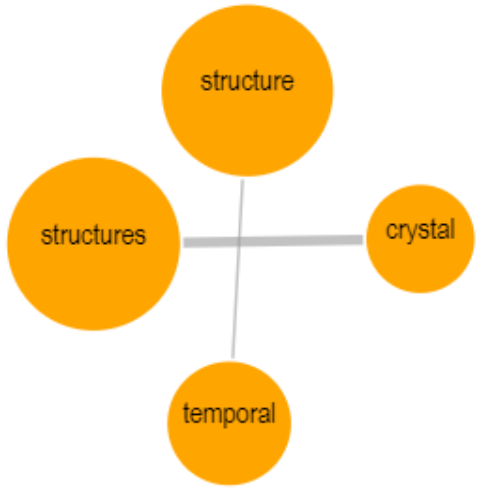

Fig. 12. LLA group 478(A)

11. Newman, M. E. J. (2003). "Fast algorithm for detecting community structure in networks," 2003. Retrieved from http://arxiv.org/pdf/cond-mat/0309508.pdf
12. Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E, vol. 74, no. 3.
13. Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11).
14. Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, 2003. Retrieved from http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf.
15. US patent 8,903,756. (2014). System and method for knowledge pattern search from networked agents. 2014. Retrieved from https://www.google.com/patents/US8903756
16. Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, 281-285.
17. Hofmann, T. (1999). Probabilistic latent semantic analysis," In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden.
18. Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14 (pp. 849 – 856), (2002). MIT Press. Retrieved from http://ai.stanford.edu/ ang/papers/nips01-spectral.pdf
19. Zhao, Y., Zhou, C., and Huang, S. (2019). Theory and use case of game-theoretic lexical link analysis. In the proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).
20. InXight 1997. Retrieved from http://en.wikipedia.org/wiki/Inxight
21. MUC-7: PAPERS: SYSTEMS - Named Entity Tasks Retrieved from http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.htmlnamed
22. Yatsko, V. A. and Vishnyakov, T. N. 2007. A method for evaluating modern systems of automatic text summarization. Automatic Documentation and Mathematical Linguistics, 41(3):93-103.
23. Soergel, D. 1985. Organizing information: Principles of data base and retrieval systems. Orlando, FL: Academic Press.
24. Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the Association for Computational Linguistics, pp. 417–424.
25. Kristina Toutanova and Christopher Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-ofspeech tagger. In EMNLP/VLC 1999, pages 63–71.
26. C.J. van Rijsbergen, S.E. Robertson and M.F. Porter, 1980. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587).
27. Salton G and Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (5): 513-523.Self-Organization, SO, wiki, retrieved from http://en.wikipedia.org/wiki/Self-organization
28. Jiampojamarn S and Cercone N (2005) Biological named entity recognition using n-grams and classification methods. In Proceedings of PACLING
29. Bekkerman R and Allan J (2003) Using Bigrams in Text Categorization. Retrieved from http://people.cs.umass.edu/ ronb/papers/bigrams.pdf
30. N-gram, http://en.wikipedia.org/wiki/N-gram
31. sciSpaCy https://allenai.github.io/scispacy/
32. sciBert https://github.com/allenai/scibert
33. Beltagy, I., Lo, K., Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. https://arxiv.org/abs/1903.10676
34. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.
35. Zhao, Y. Zhou C. (2018). A Game-Theoretic Lexical Link Analysis for Discovering High-Value Information from Big Data. In the proceedings the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Barcelona, Spain, 28-31 Aug. 2018 (ASONAM 2018), page 621 – 625. Retrieved from https://ieeexplore.ieee.org/document/8508317.