Iran.

C. Queries

We now turn to a more heuristic evaluation of the search engine. A set of compound plain text string queries were executed using the frontend interface of the application in an effort to simulate a potential use case scenario of a non-technical user interested in crime informatics.

More specifically, one plausible scenario involves an non-technical user (investigator) who explores the *modus operandi* of criminal groups using expert's opinions from open media sources. To retrieve relevant information, the user can search for "drugs + NORP", that will return results linking the string "drugs" with a series of ethnic groups, nationalities and religious or political groups. A second search might focus on identifying what places/infrastructure co-occur with keywords such as trafficking, cocaine or heroin. Desired results can be retrieved using the FAC tag along with any of the aforementioned search terms (e.g. heroin + FAC). In addition to key places and infrastructure, users can also explore the geographical location (LOC) or administrative jurisdictions associated with drug-related crime. Finally, using the NER tags PERCENT and MONEY, a user can find hits that discuss drug-related activities in conjunction with monetary values, quantitative measurements, and percentages. This search can provide longitudinal estimations of the street value of the recovered drugs.

V. Conclusions and Future work

The paper presents a high-level overview of a search engine that was build scratch, using a large corpus of news quotes provided by the GDELT project [6]. The search engine enables users to quickly drill through news articles and to identify quotations provided by experts and public figures. The added value of the proposed solution is based on the seamless integration of spaCY's NER model with Elasticsearch, allowing non-technical users to access novel information using a combination of strings and things. Furthermore, we have shown that aggregations and analytics can bring about interesting pathways for exploring fundamental sociological questions such as the reproduction of gender bias and stereotypes in the media. From the implementation side, there are numerous bottlenecks that need to be addresses. The search engine crawls raw data that have already been pre-processed and the quotations identified. While the work of the GDELT project has been extremely valuable, there is still room for improving the accuracy of the mechanism used in identifying quotes. In addition, the current version of the search engine implements the default NER model from the sapCy. The model has been reported to have an accuracy of about 80% when tested on a general web corpus [7]. Future iterations of the search engine will explore domain specific models using both static and contextualized word embeddings [28] in downstream tasks, that will increase the overall accuracy and quality of the user-defined queries.

References

[1] R. Keyes, *The quote verifier: Who said what, where, and when*. St. Martin's Griffin, 2007.

[2] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of artificial intelligence research*, vol. 11, pp. 95–130, 1999.

[3] M. Mohamed and M. Oussalah, "Identifying and extracting named entities from wikipedia database using entity infoboxes," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 7, 2014.

[4] J. Turcotte, C. York, J. Irving, R. M. Scholl, and R. J. Pingree, "News recommendations from social media opinion leaders: Effects on media trust and information seeking," *Journal of Computer-Mediated Communication*, vol. 20, no. 5, pp. 520–535, 2015.

[5] K. Niklewicz, "Weeding out fake news: an approach to social media regulation," *European View*, vol. 16, no. 2, pp. 335–335, 2017.

[6] GDELT, "The global quotation graph." [Online]. Available: https://blog.gdeltproject.org/announcing-the-global-quotation-graph/

[7] SpaCy, "Named entity recognition." [Online]. Available: https://spacy.io/usage/linguistic-featuresnamed-entities

[8] Elasticsearch. [Online]. Available: https://www.elastic.co/elastic-stack

[9] Flask. [Online]. Available: https://flask.palletsprojects.com/en/1.1.x/

[10] NGINX. [Online]. Available: https://www.nginx.com/

[11] R. Blanco, M. P. Matthews, and P. Mika, "Quote-based search," Oct. 21 2014, uS Patent 8,868,558.

[12] E. Segalis, G. Chechik, Y. Matias, Y. Leviathan, and Y. Tzur , "Systems and methods for searching quotes of entities using a database," Aug. 8 2017, uS Patent 9,727,617.

[13] E. Segalis, G. Chechik, Y. Matias, Y. Leviathan, and Y. Tzur, "Systems and methods for searching quotes of entities using a database," Feb. 5 2019, uS Patent 10,198,508.

[14] Elasticsearch, "Elasticsearch mapping." [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping.html

[15] Elasticsearch, "Mapper annotated text plugin." [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/plugins/current/mapper-annotated-text.html

[16] Elasticsearch, "Terms aggregation." [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-terms-aggregation.html

[17] Elasticsearch, "Significant text aggregation." [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/master/search-aggregations-bucket-significanttext-aggregation.html

[18] CSC. It center for science, pouta cloud services. [Online]. Available: https://research.csc.fi/cloud-computing

[19] Prometheus. [Online]. Available: https://prometheus.io/

[20] Python elasticsearch client. [Online]. Available: https://elasticsearch-py.readthedocs.io/en/master/

[21] SpaCY, "Annotation specifications." [Online]. Available: https://spacy.io/api/annotation

[22] Elasticsearch, "How to find and remove duplicate documents in elasticsearch." [Online]. Available: https://www.elastic.co/blog/how-to-find-and-remove-duplicate-documents-in-elasticsearch

[23] OpenSSL. [Online]. Available: https://www.openssl.org/

[24] "Demo of the prototype entity oriented search engine." [Online]. Available: https://www.humcomp.ml

[25] Elasticsearch, "Simple query string query." [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-simple-query-string-query.html

[26] Lucene. Class tfidfsimilarity. [Online]. Available: https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

[27] D. Petrelli, M. Beaulieu, M. Sanderson, G. Demetriou, P. Herring, and P. Hansen, "Observing users, designing clarity: A case study on the user-centered design of a cross-language information retrieval system," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 10, pp. 923–934, 2004.

[28] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proceedings of the 2019 EMNLP and the 9th IJCNLP*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65. [Online]. Available: https://www.aclweb.org/anthology/D19-1006