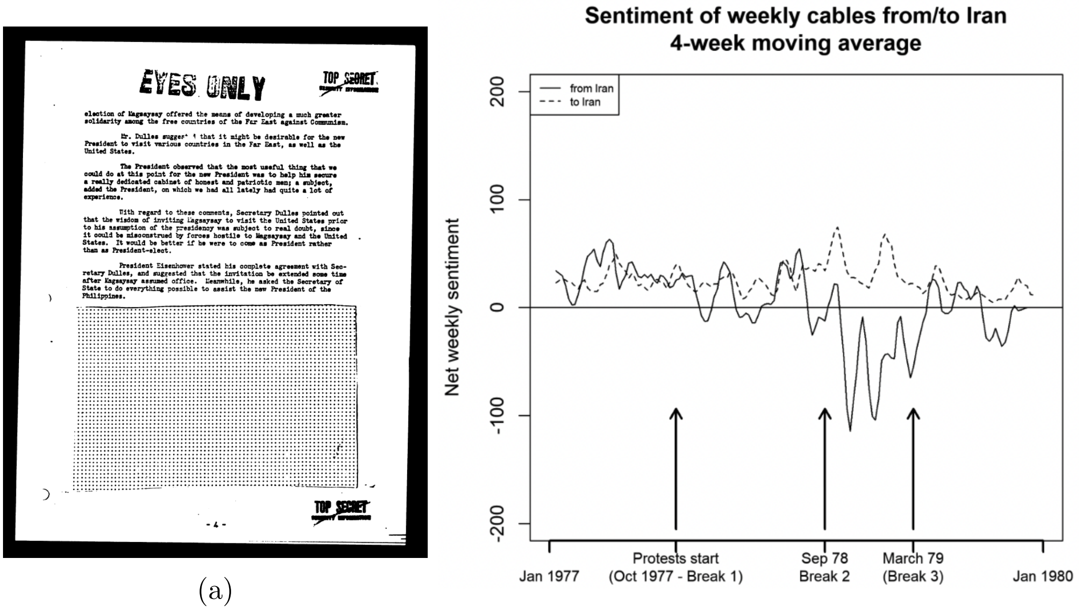
# Deciphering Historical Records through NLP and Graph data science - DeHiR NLP

Luis Salamanca, Fernando Pérez-Cruz, Matthew Connelly and Raymond P. Hicks

August 2022

## 1 Project description

The digitization wave has changed the way we view and handle archival data. Records and proceedings are scanned at large volumes and made public via online portals. These documents are a promising data source to address a magnitude of research questions, in particular in the social sciences. Henceforth, these corpora have opened enormous opportunities to historians, who can now easily access many records with extensive and invaluable information. Nevertheless, only through the analysis of large corpora we can get a further understanding on how, for many historical events, complex relations between different actors, and seemingly irrelevant affairs, may have been the initial spark. Therefore, the access to these larger corpora opens a world of new possibilities, but at some cost: now the scale does not allow a manual processing of the documents.



EYES ONLY

TOP SECRET
SECURITY INFORMATION

election of Magsaysay offered the means of developing a much greater solidarity among the free countries of the Far East against Communism.

Mr. Dulles sugges⁺ ⁱ that it might be desirable for the new President to visit various countries in the Far East, as well as the United States.

The President observed that the most useful thing that we could do at this point for the new President was to help him secure a really dedicated cabinet of honest and patriotic men; a subject, added the President, on which we had all lately had quite a lot of experience.

With regard to these comments, Secretary Dulles pointed out that the wisdom of inviting Magsaysay to visit the United States prior to his assumption of the presidency was subject to real doubt, since it could be misconstrued by forces hostile to Magsaysay and the United States. It would be better if he were to come as President rather than as President-elect.

President Eisenhower stated his complete agreement with Secretary Dulles, and suggested that the invitation be extended some time after Magsaysay assumed office. Meanwhile, he asked the Secretary of State to do everything possible to assist the new President of the Philippines.

[redacted]

- 4 -

TOP SECRET
SECURITY INFORMATION

(a)

(b)

Figure 1: In (a), example of confidential declassified cable. In (b) some results of using sentiment analysis for analysing US cables during the Iranian revolution [1].

Natural language processing (NLP) is at the forefront of the revolution that the resurgence of neural networks brought to many different fields, advancing problems that were thought unmanageable. The countless availability of written data, and self-supervised training methods, has enabled the development of massive learning models (LM) that can capture subtle semantic and grammatical information [2, 3]. By computing embeddings for the input text, these LMs enable solving many downstream tasks that can be of great interest for the analysis of massive records: topic modelling (TM) to understand the main issues discussed; name entity recognition (NER) to extract keywords of relevance; semantic role labelling to summarize a text in its main constituents, etc. Besides, the similarity between different texts, as captured in the computed embeddings, can be leveraged to pose new predictions problems. These approaches have been widely used already in different social science fields, but their usage is still relevant, and newly generated corpora can benefit from them.

However, currently most of the results acquired using these techniques are mostly analysed and exploited in an independent fashion, without trying to further integrate them. This leads to all possible relations between different entities, that might hold important information, being neglected in the analysis. For example, we can obtain insightful results from TM (e.g. distribution of topics throughout the years related to specific historical events) and NER (e.g. high presence of specific organizations mentioned in some given documents), but if we would not proceed to integrate both, we could be missing subtler information/patterns such as: given the discussion of some specific topic throughout some years, we find a higher prevalence of a given organization, and the frequent relation of it with an specific actor. Similar networks of interest and interaction may hold quite insightful information, but enabling such analyses require novel representations beyond standard relational databases, and compartmentalized analyses.

Nowadays, knowledge graphs and graph databases provide the required flexibility to enable such representation. They are built around nodes (entities of interest) connected through relations, holding both different properties that characterize them. Not only they are optimized to store millions of nodes and relations, but provide also utter flexibility, i.e. new types of nodes and relations can be added or deleted on-the-fly, as well as properties. Beyond, current graph databases technologies, such as Neo4J, provide optimized tools for graph analysis. These enable running a plethora of graph data science (GDS) methods such as community detection, node embedding, ranking, etc., on the vastness of the information held by the graph, therefore yielding patterns that could not have been found otherwise.

In the present project, we aim at harnessing the possibilities of NLP and GDS and their application onto a novel historical corpus: the Foreign Relations of the United States (FRUS), a collection of more than 300,000 diplomatic documents of the United States Government, hand-selected by historians. This collection is consistently growing and now covers documents from 1620 through the 1980s. Given the eminent exploratory nature of the current project, we will collaborate with Matthew Connelly, professor of History at Columbia University and his History Lab team who apply data science and natural language processing techniques to a large collection of declassified government documents (as can be seen in Fig. 1a, together with some results in Fig. 1b). The first objective of the project is to run Named Entity Recognition and Named Entity Linking on the FRUS corpus which will allow us to unify the various aliases of entities into a standardized name. This will enable then to parse relevant information for each entity from outside sources, and use it to construct a knowledge graph. This information, together with features derived from the FRUS text, will allow to create embeddings for the different entities, that will be exploited to unveil similarities between them. While many of these similarities will be related to known historical events, hence validating the method, others will pose new explanations and hypothesis. We expect these will shed some light into different phenomena of modern times, such as the ebb and flow of different countries in US diplomatic relations and how the United States viewed, and used, different international organizations in the post-World War II world. Nonetheless, the aims of the project will evolve as we dig deeper into the data, and further understand it and the different possibilities it holds.

The present project will allow the student gaining a greater familiarity with the use of different NLP and GDS methodologies, as well as the use of knowledge graphs to perform a more holistic encoding and analysis of the information contained in written records. It is important to highlight that the project is eminently applied, so we foresee mostly the application of up-to-date methods rather than the advancement of the state-of-the-art. Still, in the framework of the present interdisciplinary collaboration, we expect to unveil many interesting insights from the data that will be published in conferences and/or journals on the domain field.

## 2 Additional information

- **Difficulty of the project:** From moderate to very challenging.
- **What will you learn?** natural language processing, graph data science, language models.
- **Requirements:** Good python level, some knowledge of either Keras or PyTorch, experience with git, machine learning fundamentals, creative thinking, **interest in history**.
- **Supervisors:** Luis Salamanca (luis.salamanca@sdsc.ethz.ch), Fernando Pérez-Cruz (fernando.perezcruz@sdsc.ethz.ch), Matthew Connelly (mjc96@columbia.edu) and Raymond P. Hicks (rh2883@columbia.edu).

## External references

- **History lab**: dedicated to using data science to recover and repair the fabric of the past, beginning with declassified documents, some of the earliest examples of electronic records. http://history-lab.org/
- **Analysis for the Washington Post**: research results from the History Lab. URL to article

## Bibliography

## References

[1] Matthew Connelly, Raymond Hicks, Robert Jervis, and Arthur Spirling. New evidence and new methods for analyzing the iranian revolution as an intelligence failure. *Intelligence and National Security*, 36(6):781–806, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. URL: `https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.