**Grant-to-Author recommender system**

This project demonstrates the usage of natural language processing and machine learning techniques for the goal of streamlining the process of identifying potential grant call applicants. I created a recommender system that matches the content of grant calls with the research abstracts of authors by calculating the distance between their neural embeddings.

The grant calls’ contents were scraped from NIH’s “Grants & Funding” website. For this purpose, I used hand-crafted heuristics and pattern matching that were refined over several iterations. Eventually I was able to find the “Purpose” and/or “Background” sections (or an equivalent section) of most grant calls and store them. As the variance of the grant pages’ contents in NIH’s website is very large (especially the html formatting), scraping the purpose and/or background of all grants with 100% accuracy was not feasible. For some entries, the sections were partially scraped, and in others, lines from other sections were scraped as well. As the deviations were not high, the resulting embeddings should still possess good representation capabilities. About a quarter of the grant calls did not fit into any of the patterns of the rest, and therefore were filtered out of the database. Eventually, around 12,000 grant calls were scraped and stored.

The authors along with their abstracts were collected using Semantic Scholar’s API. Papers were retrieved in batches of 1,000. Some of the papers did not include an “abstract” section or information about the authors due to privacy, and those were filtered out. As NIH’s grant calls are mostly related to health, medicine and biology, only papers in the fields of medicine and/or biology were collected. As an additional filtering step with the purpose of collecting only high-quality and relevant papers, only papers with at least 10 citations and from 2019 and onwards were collected. Also, as ~12,000 grant calls were scraped, I chose to collect around 20,000 papers so database sizes are on the same scale (code allows it to be recreated with a different number of papers).

The “Purpose” and “Background” sections of each grant call were concatenated (in the cases in which only one exists, no concatenation was made) and embedded by the SPECTER model and then by the all-mpnet-base-v2 model and stored separately. The “abstract” sections of each paper were similarly embedded by both models and stored.

Each of these four databases was used to create its respective Qdrant collection.

Matching was then done by embedding the content of a given URL (which was retrieved by either scraping or by using Semantic Scholar’s API), or in the case of free text, by simply embedding the free text itself with the chosen model and finding the best matches for it in the appropriate Qdrant collection. Matches are based on the cosine distance function. The members in the collection whose embeddings have the lowest cosine distance between them and the embedding of the query are considered matches.

A simple GUI was built for the matching system. It allows for the insertion of grant call and paper URLs and free text for both. The model embeddings and number of matches to be found can be chosen as well. Relevant information for matches that were found is presented at the bottom. Buttons for moving between matches are included.

Regarding the data itself, there is a slight mismatch between many of the grants and the topics of many of the papers, as some of the grant calls pertain to very broad topics while most of the papers are about the research of specific conditions and drugs. Also, many of the grant calls pertain more to technical matters than to actual research, so not many matching papers exist for them. Nevertheless, appropriate matches were found for the big majority of grants and papers I tested for.

The overall quality of the matches was good. The top few matches were almost always very relevant to the query, though for the queries with the more esoteric topics, the system was struggling to find the requested number of relevant matches, if any at all. Between SPECTER’s embeddings and all-mpnet-base-v2’s performance was nearly identical. Similar matches were found while using both types of embeddings, though it seemed SPECTER’s embeddings were a bit more robust and consistent with their matches.

Below are some examples of successful matches made by the matching system:

* Using SPECTER, the "Notice of Special Interest (NOSI): Tackling Acquisition of Language in Kids (TALK) R01 Research Projects" grant call was matched to the "Late Language Emergence" paper, and vice versa.
* Using all-mpnet-base-v2, the "Notice of Special Interest (NOSI): Public Health Research on Cannabis" grant call was matched to the "Medical Cannabis Use among Adults in the Southeastern United States." paper, and vice versa.
* Using SPECTER, the free text "immune cell-based therapy for treatment of acute leukemia" was matched to the "State-of-Art of Cellular Therapy for Acute Leukemia" paper.
* Using SPECTER, the free text "Alzheimer's disease in Africa" was matched to the "Building Neuroscience Research Infrastructure for Alzheimer's Disease (AD) and AD-Related Dementias (ADRD) in Africa (UG3/UH3 Clinical Trial Not Allowed)" grant call.

Below are some examples of non-successful matches made by the matching system:

* Using SPECTER, the "Curve sprinting in soccer: relationship with linear sprints and vertical jump performance" paper was matched to the "Notice of Intent to Publish a Funding Opportunity Announcement for Analyses to Determine Diagnostic Cut-Points for Older Adults with Low Muscle Mass or Strength (U01)" grant call. The reason for this non-successful match probably lies in the fact that this paper is about the research of a very specific condition (curve sprinting in soccer), so as I mentioned before – there is no existing grant call that is specific enough to be a sufficient match for it.
* Using SPECTER, the "Notice of Change: Extension of PAR-22-028 "Ultra-Rare Gene-based Therapy (URGenT) Network Resource Access (X01, Clinical Trial Not Allowed)" by one Council Round" grant call was matched to the "Return of individual research results from genomic research: A systematic review of stakeholder perspectives" paper. The reason for this non-successful match probably lies in the fact that this grant call pertains almost entirely to a technical matter (date extension) than to actual research, so as I mentioned before – there is no existing paper that matches it enough.