# Lab - 2: RAG

Mihir Gohel (117284926)*
*Stony Brook University*
*CS519, Data Science fundamentals*

## I. INTRO

All the files to the Lab2 of CS519 are available at Link.

## II. PART 2 - TESTING

Here we'll take 5 queries as input and tell our model trained with docs to fetch the best matching files. For every single query we'll fetch top 5 matched files and calculate the value of precision for k=1,3,5. We also calculate the value of recall for k=5. The values of the same are given in the Table. (I).

**Discussion questions**:
**Query length**: It's better to have a longer query length than a shorter one. More the context more it can match it with indices or nearer it gets to other indices. If the query ends up being very small, the output will be quite unreliable because of less content present in the query.
**Documentation coverage**: This definitely affects the retrieval quality. There are certain documents which has got definitions and a lot more context, so whenever we search for any specific keyword there's a high probability that we'll end up retrieving that specific doc.
**Exact vs Fuzzy matching**: The embeddings that we used in this lab does handle the fuzzy matching. Well it's better if there's no spelling mistakes, but even if there's any, it's not going to change our result by much, because our embedding vectors can handle that case.
**Variation across query types**: Other than natural language query type, all the other types behaved in a similar fashion. We can retrieve relevant data file according to the query, but that retrieval is not super promising too. Out of all Natural language queries performed the worst. The retrieved data file was not relevant while performing natural language query searching.
**Role of Chat GPT**: Chat GPT was almost accurate as a judge. Out of many iterations performed there was error in only 1 to 2 files while performing judgment. It's pretty much spot-on.
**Why might we use different models for code and for documentation?** : The model for code is good at understanding the code structure, functions, and variables. On the contrary the model for docs is better at understanding the natural language better. So, using a different model for docs and code gives us the best possible embedding vectors respectively.
**Why might we choose different chunk lengths for code vs. docs?** : Code are seen in smaller blocks, such as functions. we don't need too big of a chunk size for context matching. Whereas for documents, they include paragraphs, which demands more context matching, and because of that reason it's better to have bigger chunk length.

## III. PART 4: CODE ONLY

Here we'll take 5 queries as input and tell our model trained with code to fetch the best matching files. For every single query we'll fetch top 5 matched files and calculate the value of precision for k=1,3,5. We also calculate the value of recall for k=5. The values of the same are given in the Table. (II).

## IV. PART 4: FUSION

Here we'll take top 50 for both docs and code. Then we'll calculate RRF score for all those 100 files. We'll then sort them and take the top 5 files, and then calcuate the Precision and Recall for the resultant fusion. Data for the same is given in Table. (III)

| Query | Precision@1 | Precision@3 | Precision@5 | Recall |
|---|---|---|---|---|
| What is the purpose of the scipy.signal module? | 0.0 | 0.33 | 0.20 | 1.0 |
| How to use scipy.stats.norm.interval? | 1.0 | 0.33 | 0.20 | 1.0 |
| sparse matrix multiplication | 0.0 | 0.33 | 0.20 | 1.0 |
| How can I find the roots of a polynomial equation in SciPy? | 0.0 | 0.0 | 0.0 | 0.0 |
| linear algebrra solver | 0.0 | 0.33 | 0.20 | 1.0 |

TABLE I: Docs model. Precision at k=1,3,5 and Recall at k=5 for different queries

| Query | Precision@1 | Precision@3 | Precision@5 | Recall@5 |
|---|---|---|---|---|
| How to call scipy.optimize.minimize with constraints? | 1.0 | 0.33 | 0.20 | 1.0 |
| Example usage of scipy.sparse.linalg.spsolve | 0.0 | 0.33 | 0.20 | 0.5 |
| What are the parameters for scipy.fft.fft2? | 0.0 | 0.0 | 0.20 | 1.0 |
| How to handle LinAlgError in SciPy? | 0.0 | 0.33 | 0.20 | 1.0 |
| Show me code for scipy.signal.convolve2d | 1.0 | 0.33 | 0.40 | 1.0 |

TABLE II: Precision at k=1,3,5 and Recall at k=5 for selected queries

| Query | Precision@1 | Precision@3 | Precision@5 | Recall@5 |
|---|---|---|---|---|
| Explain how to perform a t-test in SciPy and interpret the results. | 1.0 | 0.666667 | 0.4 | 0.133333 |
| How to use scipy.optimize.root to find the roots of a function? | 1.0 | 1.0 | 1.0 | 0.087719 |
| What are the different interpolation methods available in SciPy? | 1.0 | 1.0 | 0.6 | 0.090909 |

TABLE III: Precision at k=1,3,5 and Recall at k=5 for different queries