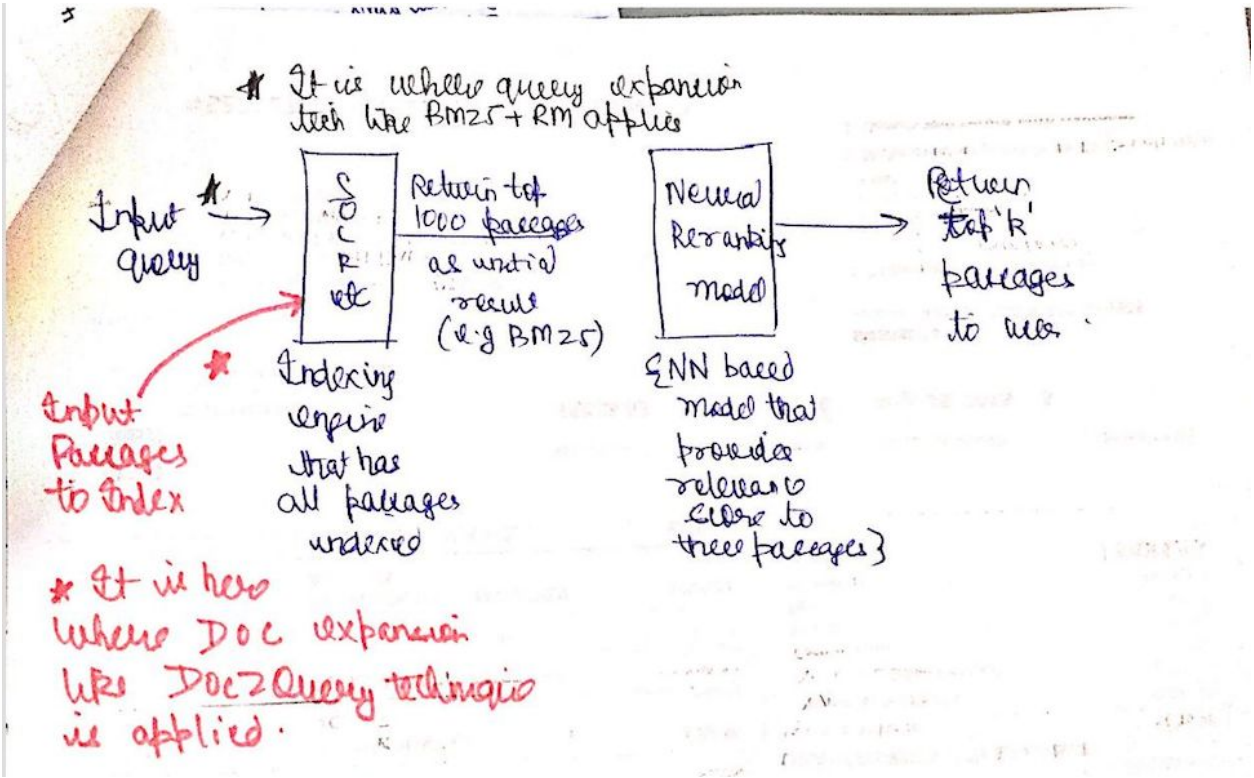
Name: Document Expansion by Query Prediction

Link: https://arxiv.org/pdf/1904.08375.pdf

GitHub: https://github.com/nyu-dl/dl4ir-doc2query

A. Introduction.

- Idea is to enhance passage/ document representation BEFORE indexing them in SOLR, LUCENE etc.
- There are 2 ways to handle VOCAB MISMATCH (problem wherein the user enters query in terms that are different from those used in relevant documents):
  - Query expansion e.g. BM25+RM relevance feedback
  - Doc/ Passage expansion: e.g. Doc2Query **[FOCUS AREA]**



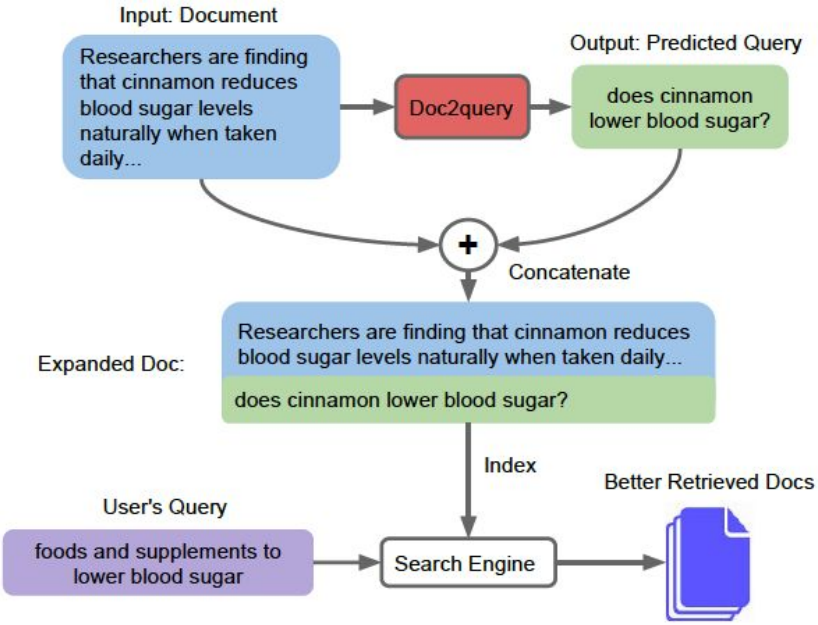
- Authors claim following contributions:
  - 1st ever successful application of Doc expansion with NN
  - Portray that Doc expansion is better than Query expansion
  - No need for an expensive re-ranking NN method application.

B. Description:

a. Uses **Transformer seq2seq** model. Idea is to generate questions that can be answered by the passage/ document.

b. **Data**: MS-Marco and TREC-CAR datasets as **(target query, relevant document) pairs**

c. **Input**: Target query and document segmented using BPE
d. **Output**: seq2seq model which can generate queries that can be answered from the document/ passage
e. Once a model is trained, predict top k (10) queries that can be asked from this passage. Append these queries to document/ passage. Then INDEX it



C. **My Assessment/ Analysis**
   a. CODE on Github
   b. Claim that the first successful method to use Doc expansion.
   c. Document/ Passage text contains much richer signal as compared to query. Hence higher performance as compared to Query expansion kind of makes sense.
   d. A good point is that it is done before indexing and results are good to be used even without using sophisticated NN based re-ranker models.

D. **LIMITATIONS/ CONFUSIONS for me**
   a. Not very convinced on how they measure that output query by transformer model is of good quality. They have specified that they measure BLEU scores, but I am still not convinced.
   b. Not certain that it will apply to the insurance domain as I need to have extensive annotated data so as to train the seq2seq model.
   c. Not very clear whether they append all the 10 predicted queries one after the other or what?