CS6370: Natural Language Processing
Project Proposal

Shashank M Patil CH18B022

Shania Mitra CH18B067

Limitations of the Current Model:

With the current model we observe that there are some tokenization and spelling errors, the correction of which is essential for the model to work satisfactorily. Examples include:

- All punctuations such as fullstops are considered tokens by the tokenizer
- Some queries and documents have erroneous words like "-dash". For example, Document **8**.
- The tokenizer fails to split across hyphens. For example, 'real-gas', 'shock-induced', 'boundary-layer' remain as it is. This causes a problem especially since keywords are matched directly. Some documents and queries contain 'boundary layer', while others contain 'boundary-layer'. This poses a problem since documents containing 'boundary-layer' are not retrieved for queries containing 'boundary layer' and vice versa, which may lead to relevant documents not being retrieved.
- Tokenizer fails to clear out noisy punctuations. For example, "/slip flow/", "/boat-tail/". In case of "/slip flow/", it is split as "/slip", "flow/" and no document with slip or flow is retrieved since, by design, the model considers "slip" and "/slip" as orthogonal dimensions.
- Spelling errors exist in the documents and queries necessitating a spell check pre-processing step. For example, in Document **74**, "turbulen coundary" is an incorrect representation of "turbulent boundary".

It can be seen that, primarily, keyword matching takes place while trying to retrieve relevant documents. Thus, very often the model does not cater to exactly what the query asks for, but rather serves documents matching in certain aspects only.

Example-1: Query 119

What is the effect of initial axisymmetric deviations from circularity on the non linear (large-deflection) load-deflection response of cylinders under hydrostatic pressure

The top-3 retrieved documents have the following IDs and titles:

- **887:** "buckling due to thermal stress of cylindrical shells subjected to axial temperature distributions ."
- **769:** "local circumferential buckling of thin circular cylindrical shells ."
- **1146:** "thermal buckling of cylinders ."

In the query the key ideas being communicated are:

- Load deflection response of cylinder
- Deviation in shape (axisymmetrically)
- Hydrostatic pressure

It must be noted that all the retrieved documents provide the user with a load deflection response, i.e., buckling for the correct object shape - cylinder. However they do not get the deviation variables and type of stress correctly. For example, in Document **887**, the deviation variable is temperature which

varies axially while the stress involved in this case is thermal stress while what is queried is hydrostatic pressure. Thus, while these documents match partially in terms of keywords, they fail to match at a concept level making the retrieved documents irrelevant.
The true documents in this case are:

- **897:** "some results on buckling and postbuckling of cylindrical shells ."
- **926:** "post buckling behaviour of circular cylinderical shells under hydrostatic pressure ."

Document **897** answers the query precisely however is still not retrieved. This may be due to the following reasons:

- While the query uses the terms axisymmetric deviations from circularity, the document uses the terms “axisymmetric deformation” and “noncircular”. Even though these words have very similar meaning, these are orthogonal dimensions in the model, by design. Thus, a measure of word relatedness could help mitigate such inaccuracies.

In Document **926**, we observe a spelling error in the spelling of cylindrical, which may have led to the words “cylinder” (present in the query) and “cylinderical”, (which remains so even after lemmatization) to be orthogonal. Since they are treated as completely different words it results in this relevant document not being retrieved. Further, in the body of the document, we notice that there is a line in the document that answers the query exactly, i.e., “calculations show that postbuckling equilibrium configurations exist for loads greater than as well as loads slightly less than the critical load calculated from small-deflection theory…….for loads corresponding to radial displacements of the order of the shell thickness, it is found that the number of circumferential waves remain essentially constant with increasing deflection and equal to the number of waves developed at buckling”
Even in its presence, the document is not retrieved. The reasons for this may be:

- As highlighted above, the tokenizer does not split across hyphens. Therefore the model is unable to make a match between “large-deflection” present in the query and “deflection” present in the document.
- Further, in the document instead of using the term “large deflection” a more explicit representation, i.e., “radial displacements of the order of the shell thickness” is used, which the model is unable to capture since it does not deal with concepts and definitions and is limited to keywords only.

Example 2: Query 116
What is the magnitude and distribution of lift over the cone and the cylindrical portion of a cone-cylinder configuration

The top-3 retrieved documents have the following IDs and titles:

- **1304:** "newtonian flow over a surface ."
- **122:** "a simplified approximate method for the calculation of the pressure around conical bodies of arbitrary shape in supersonic and hypersonic flow ."
- **196:** "pressure distributions . axially symmetric bodies in oblique flow ."

The concept surrounding the query is lift force on a cone-cylinder configuration. As before, the retrieved documents get the shape of the object correctly, i.e, cylinder. They, however, focus on the pressure distribution and flow over these objects as compared to the lift force and hence, do not match with the ground truth. This example reinforces the fact that this model performs only keyword matches and often tends to miss out on the overarching concept.

Thus, on the whole, taking into account word relatedness and formulating the concepts involved in both query and document could help avoid such errors.

Hypotheses:

Following are the conclusions drawn from the limitations listed above:

- Incomplete and improper tokenization, along with spelling errors and abbreviations in the queries and documents lead to a reduction in the performance of the Vector Space Model
- The vector space model does not take word-relatedness and overarching concepts into account, performing only keyword matches which results in poor performance
- All documents belong to the domain of aerodynamics, thus, phrases such as “boundary layer”, “incompressible flow”, “thermal stress”, etc. must be extracted and given weightage, since these form real concepts in theory.
- Since the documents vary greatly in length, a substitute for TF-IDF must be found such that the ‘term frequency’ expression is normalized. Further, it must be noted that documents with twice the number of terms as another document do not necessarily have to be twice as relevant. Thus, a measure to check the linearity of growth of relevance with term frequency must be put in place.

Realising the Hypotheses and Evaluation Measures:

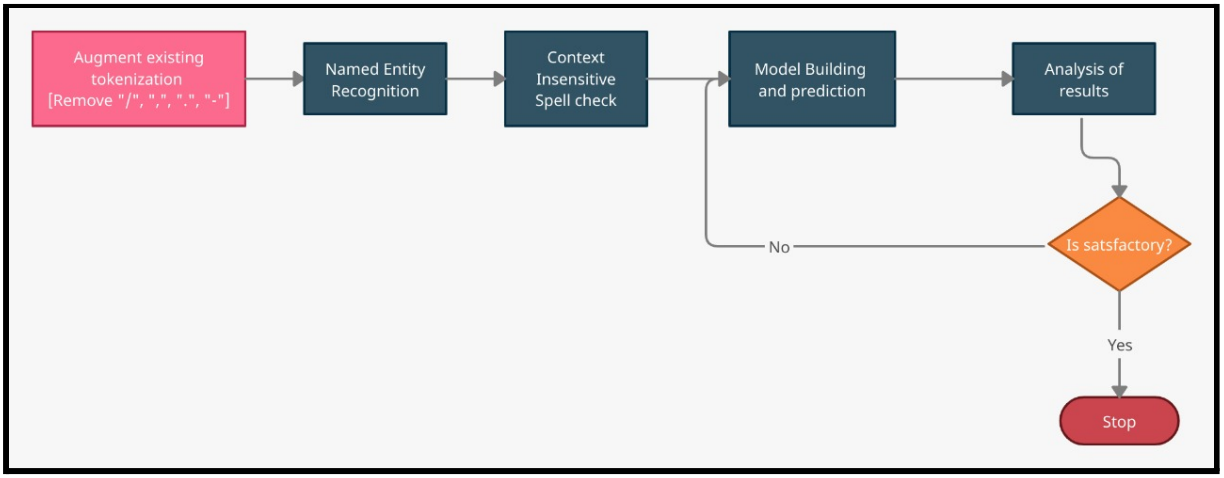

Figure 1: *Flowchart highlighting the approach to be followed*

Named Entity Recognition and Spell Check

We aim to use Named Entity Recognition, so that names (authors and scientists like Wassermann, Prandtl and terms like Reynolds, Newtonian, etc.) are not wrongly corrected as misspellings.
It must be noted that most of the spelling errors are non-dictionary words such as:

- stabilityproblems
- cylinderical
- veiocity
- aerodynamieist
- gasdynamic

Hence, we intend to proceed with a context insensitive spell check either using Shannon’s Bayesian approach or edit distance.

Phrase Extraction

PPMI (Positive Pointwise Mutual Information) may be used to extract phrases such as “boundary layer”, “shock induced”, “thermal stress”, “hydrostatic pressure”, “incompressible fluid”, “heat conduction”, “Reynolds number”, etc. Countings of occurrences and Co-occurrences of words in a text corpus can be used to approximate the probabilities appearing in the PPMI calculation.

Alternative to TF-IDF:

BM25 (Best Match 25) improves upon **TF-IDF** by casting relevance as a probability problem. The term frequency in BM25 dampens the impact of term frequency even further than traditional TF-IDF. The impact of term frequency is always increasing, but asymptotically approaches a value. Relevance always increases with increasing term frequency. In TF-IDF, the dependency is linear (A document with twice the number of words is considered twice as relevant) and constantly increases and never reaches a saturation point. However, in BM25, term frequency tends to saturate at higher occurrences, quickly hitting diminishing returns. The intuition is that more the number of terms in the document that do not match our input query - the lower the document's score should be. Thus, if a 300 page long document mentions the query term once, it is less likely to have as much to do with the query as compared to a short document which mentions the query once.

$$\text{BM25}(D,Q) = \sum_{i=1}^{n} IDF(q_i, D) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i) + k_1 \cdot (1 - b + b \cdot |D|/d_{avg}))}$$

where:

- $f(q_i, D)$ is the number of times term q_i occurs in document D.
- $|D|$ is the number of words in document D.
- d_{avg} is the average number of words per document.
- b and k_1 are hyperparameters for BM25.

$$\text{IDF}(q_i, D) = \log \frac{N - N(q_i) + 0.5}{N(q_i) + 0.5}$$

where

- $N(q_i)$ is the number of documents in the corpus that contain term q_i.

Concept Extraction

LSI (Latent Semantic Indexing) takes into account the semantic similarities between words by nature of its design because a query with terms that do not appear in a document may still end up close to a document in hyperspace. As a result, a query can return documents with terms that are semantically similar to query terms. This may not work if the document contains words that have multiple definitions (polysemy).

Evaluation

For evaluation, we intend to use the metrics **MAP (Mean Average Precision)** and **nDCG (Normalized Discounted Cumulative Gain)** as we have access to the ground truth documents relevance score. We may then compare it with our original IR system to realize the improvements made in the current IR system. In order to ensure that the proposed hypothesis is actually overcoming the limitations we had proposed, we would specifically check the retrieval performance of the system of the queries for which our original IR system didn't perform effectively. Further, clusters from the concept matrix in LSI can be extracted and evaluated. Documents from each cluster can be inspected manually or otherwise, to determine the quality of clusters and thus the effectiveness of the approach (by checking if documents that are categorised under the same concept are truly similar to each other). Efficiency can also be accounted for by measuring the time taken for Information Retrieval and comparing with the original system.

Additional Considerations:

- Assuming that a particular author performs research in a specific subdomain, when documents from an author are identified as relevant by the model, other papers written by that author could also be returned
- Similarly, bibliographies that point to certain documents in the dataset itself or documents that share the same bibliography could be returned as relevant documents
- **Skip-grams** predict surrounding words given the current word. It can hypothetically be used for query expansion because it allows search engines to recognize terms that are likely to appear to each other's contexts. We can use this in conjunction with LSI to find the relevance of the documents to a particular query.
- Another approach could be to use **ESA (Explicit Semantic Analysis)** with the research papers and articles on aerodynamics as the background knowledge

References:

1. https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=3360&context=theses
2. http://www.kmwllc.com/index.php/2020/03/20/understanding-tf-idf-and-bm25/
3. http://berlin.csie.ntnu.edu.tw/Courses/Information%20Retrieval%20and%20Extraction/2009F_Lectures/IR2009F-Lecture02-Modeling-I.pdf
4. https://aspoerri.comminfo.rutgers.edu/InfoCrystal/Ch_2.html
5. https://en.wikipedia.org/wiki/Information_retrieval
6. https://en.wikipedia.org/wiki/Topic-based_vector_space_model
7. https://www.ccs.neu.edu/home/vip/teach/IRcourse/html/schedulen_mod.html
8. https://opensourceconnections.com/blog/2015/10/16/bm25-the-next-generation-of-lucene-relevation/
9. https://abishek21.medium.com/building-your-favourite-tv-series-search-engine-information-retrieval-using-bm25-ranking-8e8c54bcdb38
10. Improvement of Vector Space Information Retrieval Model based on Supervised Learning - Xiaoying Tai
11. Relevance Analysis For Document Retrieval - Eric LaBouve
12. Comparison and Evaluation of Different Methods for the Feature Extraction from Educational Contents - Jose Aguilar
13. P. Mills. Singular value decomposition (svd) tutorial: Applications, examples, exercises, Oct 2017
14. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space.CoRR , abs/1301.3781, 2013.