Methodology – The following section explains the process along with the code component responsible for the process. Also, please be noted that the steps form a pipeline in the explained order:

1. **Stop_words creation:**

    - Creates stop words from English language and HTML tokens since initial exploration of data revealed contamination with HTML tags

2. **filter_data:** Filters the data from other language sentences considering only English

3. **Clean_data:** Cleans the raw data with the following methods

    - Create tokens after removing stop_words, applying lower case, filtering string through removing numbers and special characters, and checking for token_size

4. **get_domain_stop_words:** Extends the stop_words list through fetching corpus based stop words which are filtered based on the word_count of each token in the corpus

5. **remove_domain_specific_stop_words:** Filters the tokens based on the new domain based stop words

6. **create_corpus:** This is transformation which is specific for modeling where it returns the following:

    - id2word: Maps each unique token to a number
    - corpus: transforms each sentence in the format below:

      [(numerical key of each token, number of times the token appears in the sentence)]

7. **find_optimal_topic:** Finds the optimal number of topics for the model based on the corpus and calculating the COHERENCE score for the mentioned range of topics

8. **model:** creates a LDA – Latent Dirichlet Allocation model based on the optimal number of topic from the previous step

   Intuition of the LDA algorithm: LDA is an unsupervised generative probabilistic model where the distribution is modeled as p(X|y). In LDA the documents are represented as a mixture of hidden topics and each topic is represented as a distribution of words hence the model is used to find the hidden topic or topics in each sentence/document etc.

9. **write_topics**: Adds a new column "topic to the data frame, where the topics are modeled in the previous step using LDA.

10. **write_op_file**: Creates a JSON file and writes the op in the JSON with the given format.

Reasons for choosing LDA:

- LDA is more dynamic in allowing the words and topics to be modeled on the corpus when compared to methods like Non-negative matrix factorization (NMF) where the probability vectors are fixed.
- Easy improvements like guided LDA, Named entity recognition based etc.

Future Work:

- Improve the model with LDA2vec, where the ideas of LDA and word to vector modeling is combined where the model creates a context vector which is the sum of word vector and document vector.