Advanced Features

1. Token Features
   - **cwc_min**: This is the ratio of the number of common words to the length of the smaller question
   - **cwc_max**: This is the ratio of the number of common words to the length of the larger question
   - **csc_min**: This is the ratio of the number of common stop words to the smaller stop word count among the two questions
   - **csc_max**: This is the ratio of the number of common stop words to the larger stop word count among the two questions
   - **ctc_min**: This is the ratio of the number of common tokens to the smaller token count among the two questions
   - **ctc_max**: This is the ratio of the number of common tokens to the larger token count among the two questions
   - **last_word_eq**: 1 if the last word in the two questions is same, 0 otherwise
   - **first_word_eq**: 1 if the first word in the two questions is same, 0 otherwise
2. Length Based Features
   - **mean_len**: Mean of the length of the two questions (number of words)
   - **abs_len_diff**: Absolute difference between the length of the two questions (number of words)
   - **longest_substr_ratio**: Ratio of the length of the longest substring among the two questions to the length of the smaller question
3. Fuzzy Features
   - **fuzz_ratio**: fuzz_ratio score from fuzzywuzzy
   - **fuzz_partial_ratio**: fuzz_partial_ratio from fuzzywuzzy
   - **token_sort_ratio**: token_sort_ratio from fuzzywuzzy
   - **token_set_ratio**: token_set_ratio from fuzzywuzzy

FuzzyWuzzy: Fuzzy String Matching in Python - ChairNerd

https://stackoverflow.com/a/19794953 - Search

https://en.wikipedia.org/wiki/Wikipedia%3aList_of_English_contractions