1. **Data Cleansing/Pre-processing**

After review and analysis of different approaches to data cleansing and text processing. We developed our hypothesis based on the approach used in ‘CrowdFlower winning solutions’ [Reference 1]. Here is a detailed description of our data pre-processing approach:

1. Misspelled words: As the search queries has spelling mistakes including typographical errors, we used the approach followed by ‘Stebk’ [Reference 2] for removing the spelling checks. Steubk created a simple function that uses Google to remove typos. We also did some manual spelling corrections. We also used ‘difflib.SequenceMatcher()’ [Reference 3] to find the similarities between misspelled word and correction made.
2. Stopwords: In order to create the feature vector, we have removed the words with no meaning by using the simple python library function ‘stopwords’.
3. Special Characters: The special characters were also removed by simple function.
4. Character cases (Upper case/Lower case): All upper cases were made lower.
5. Stemming:
6. Tagging Words: We used ‘NLTK.pos\_tagger()’ [Reference 4] function to separate nouns, verbs, adjectives and adverbs. This function is also used to find the most relevant words in query and title.
7. Term Frequency-Inverse Document Frequency (TF-IDF): In this project, not only words but the number of letter in the words are also important. The sum of weight in words was was calculated using the feature weights X number of letters in word. The important point to note is that we used stemmed words in TFIDF.
8. Dimensionality Reduction ---???
9. EXPLORE: Jaccard Similarity Coefficients <http://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_similarity_score.html>
10. EXPLORE: Damerau-Levenshtein <https://pypi.python.org/pypi/pyxDamerauLevenshtein>
11. **Models**
12. Xgboost:
13. Gradient Boosting Machine (GBM):
14. Random Forest:
15. Support Vector Regression:
16. ExtraTrees:
17. **Score from Model used**
18. **Conclusion**
19. **Reference**
20. <http://blog.kaggle.com/2015/07/22/crowdflower-winners-interview-3rd-place-team-quartet/>
21. <https://www.kaggle.com/steubk/home-depot-product-search-relevance/fixing-typos/comments>
22. <http://stackoverflow.com/questions/4802137/how-to-use-sequencematcher-to-find-similarity-between-two-strings>
23. <http://www.nltk.org/book/ch05.html>