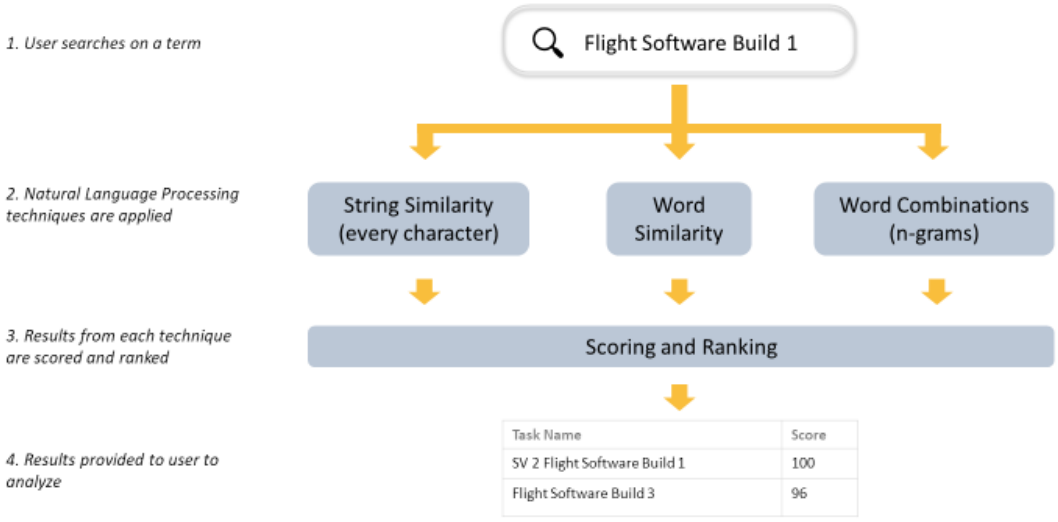
Ranking Matches for Task Search



Some of the NLP-based methods used behind the scenes include these standard preprocessing techniques:

- Tokenization – breaking strings/sentences into words
- Lemmatization – identifying the root of a word
- Parts of Speech – using a Hidden Markov Model to identify the parts of speech for a word given preceding words
- Synonyms – identifying synonyms for a given word and looking for those synonyms in a potential matching task
- Abbreviation/Acronym identification – identifying if a word is an abbreviation for another word or stands for a set of words.
- Spell Checking – identifying and correcting common spelling errors

Some of the techniques used for matching include:

- Edit Distance[10]
- Percent of common words/synonyms
- Percent of common N-grams[11]

Importantly, the techniques used consider word order, usage of common phrases, and other characteristics of a task/metadata description. Again, since these are likely to have been written by two different people, the language used to describe each will require evaluating many different characteristics (features in ML language) of the strings to identify the best potential match.

FUTURE DEVELOPMENT

The IMS Database is currently a prototype, but has already shown value in enabling schedulers to have an additional, important input into their analysis. It also serves as an invaluable source of data for future research and analysis of schedule data. Areas of current research include:

- Use of Neural Networks (Deep Learning) to facilitate improved matching and scalability.
- Final definitions of the data and metadata to be extracted from and stored for each schedule snapshot.