**Abstract**

CRF development for clinical trial database build should be driven by the study protocol. Currently, a manual review of the protocol takes place to identify the data collection needs. Tasks include reviewing study assessments and procedures to translate them into the standard library equivalents. A well-defined set of data standards, protocol templates, and review guidelines greatly expedites the process. However, this may be time consuming and lack consistency.

This poster explores the use of text mining techniques applied to a protocol in regards to clinical data collection. The aim being to highlight which standard data collection forms are applicable and where possible, what datapoints are to be collected.

**Mission One:** Identify Standard Assessments in the Protocol

- Analyze words that occur in the same section of text
- Pre-process: Protocol text divided into rows then grouped
- Filter terms of interest and highlight most common co-occurring words
- Show correlation among words (occurrence frequency relative to how often they appear separately – phi coefficient / Pearson correlation)
- Output helpful for exploration. For example, we could find the words most associated with words of interest like “dose”, “drug”, “study” etc.
- Visualize target terms or most associated words as a network to show connections

**Mission Two:** Identify Version of Form Required

- Analyze words that occur in the same section of text
- Pre-process: Protocol text divided into rows then grouped
- Filter terms of interest and highlight most common co-occurring words
- Show correlation among words (occurrence frequency relative to how often they appear separately – phi coefficient / Pearson correlation)
- Output helpful for exploration. For example, we could find the words most associated with words of interest like “dose”, “drug”, “study” etc.
- Visualize target terms or most associated words as a network to show connections

**Mission Three:** Fields within a Form

- Identify individual data points to be captured within an assessment.
- Example: Lab Chemistry – which parameters are to be collected?

Working within form standards this can be achieved using term matching. Use of a common dictionary is beneficial to increase the accuracy of terms identified using a join.

---

**Journey to the Centre of the Protocol**

Explorations in text mining for study eCRF development

**Term Frequency**

- Important meaningful words occur frequently
- Pre-process: Text separated into single words/terms
- Count and visualize as a bar chart

**Term Relationship**

- Highlight words that tend to follow each other to indicate meaning
- Pre-process: Text divided into pairs of adjacent words (bigrams)
- Visualize common protocol bigrams as a network

**Term Correlation**

- Analyze words that occur in the same section of text
- Pre-process: Protocol text divided into rows then grouped
- Filter terms of interest and highlight most common co-occurring words
- Show correlation among words (occurrence frequency relative to how often they appear separately – phi coefficient / Pearson correlation)
- Output helpful for exploration. For example, we could find the words most associated with words of interest like “dose”, “drug”, “study” etc.
- Visualize target terms or most associated words as a network to show connections

**Term Frequency**

- Important meaningful words occur frequently
- Pre-process: Text separated into single words/terms
- Count and visualize as a bar chart

**Term Relationship**

- Highlight words that tend to follow each other to indicate meaning
- Pre-process: Text divided into pairs of adjacent words (bigrams)
- Visualize common protocol bigrams as a network

**Term Correlation**

- Analyze words that occur in the same section of text
- Pre-process: Protocol text divided into rows then grouped
- Filter terms of interest and highlight most common co-occurring words
- Show correlation among words (occurrence frequency relative to how often they appear separately – phi coefficient / Pearson correlation)
- Output helpful for exploration. For example, we could find the words most associated with words of interest like “dose”, “drug”, “study” etc.
- Visualize target terms or most associated words as a network to show connections