Construct questions to explore unexamined data space

User poses question; assessment reveals question decomposition and synonyms

Ranks data statistics (term co-occurrence, adjacency, and frequency)

Suggests analytic question alternatives based on:
1. The original analytic question, synonyms for each question term
2. Data statistics (term co-occurrence, adjacency, and frequency)

Result: Question alternative permutations (coherence of alternatives varies)

Academic Collaboration (DO6)

Dr. Christopher Healey, NCSU Comp. Sci./Institute for Advanced Analytics (IAA)

- Sentence analysis of news articles
- Automatic generation of narrative threads spanning news articles
- Visualization and user experience
- Relating sentiment and narrative threads to original analytic question
- Stretch goal: Influence relevance-ranking of headlines/articles within prototype

Natural Language Search Prototype

- User poses question; assessment reveals question decomposition and synonyms
- Search results are rank-ordered by decomposition comparison/relevance criteria
- News articles are available through a TimeGlider widget integration

Data: 18 months of RSS headlines from Reuters (energy sector and environment) and www.eia.gov

Ingest: Subscribe to RSS feeds and parse; index each headline by stemmed nouns identified by Stanford NLP API Part of Speech (PoS) tagger using default news article model, e.g.,

```
<research, U.S. researchers see auto fuel standards driving technology>
```

For each analytic question posed, identify nouns in question (same approach as indexing) for initial retrieval; decompose the question into subject-verb-object using Stanford NLP API PoS and Dependency Parser (DepP) and custom SVO model code

Relevance-ranking of headlines

For each headline returned in initial noun retrieval,

- Compare SVO decomposition to the decomposition of the question
- Subject/object match rewarded more than verb match
- Include weighting for headline age and presence of signal words (modal) such as will/shoud vs. might/could

Results in an overall ‘relevance score’ for each headline

WordNet as a Synonym Source

WordNet is a lexical database developed at Princeton University

- Widely used within the computational linguistics community
- Provides synonyms and other relationships between English words
- (Optional) Broadens analytic questions for greater overlap with news headlines
- Key to Question Alternatives workflow
- Use synonyms, hyponyms, & hypernyms for leading senses only (estimated frequency)

Discovery Question Construction

Construct questions to explore unexamined data space

1. Anchors on underrepresented data themes (e.g., rare terms)
2. Maps rare term → headline containing the term
3. Converts (declarative) headline into a question
4. Question Alternatives → questions exploring ‘fringe’ themes

(DO3) + (DO6) Collaborative Question Construction

Man-made air pollution reduces central America rainfall

- Man-made air pollution reduces central America rainfall
- China says 90 percent of centers failed to meet air standards in 2014
- Protecting forests must become the norm in supply chains

**Discovery Questions Constructed**

- Must protecting forests become the norm in supply chains?
- Might China say 90 percent of cities failed to meet air standards in 2014?
- Might man-made air pollution reduce central America rainfall?

If discovery question is uninteresting, press Explore for a new question

- At least one headline related to each discovery question exists