NLP Methodology Research:

Solution/Algorithm Options:

* N-gram (bi-gram) Contextual Analysis
* Noun phrase extraction
* Synonym and library extraction
* LDA Topic Modeling
* Pre-trained classification algorithm

1. **Bigram Contextual Analysis (2) (Option A with Pre trained, B rule based)**
   1. Summary: A bigram is a 2 word window size. This pulls a little bit of context surrounding a word. You can dial in the results better by using stopwords to remove fluff and can decide whether to just pull nouns and their surrounding words. The words around the nouns pull in further context. You could also do nouns and adjectives and their surrounding words.
   2. Pros: Conceptually quite easy, leaves great room for partnering with other solutions, good base case option
   3. Cons: Noisy data, longer processing time,
2. **Noun phrase extraction (4)**
   1. Summary: With a noun phrase, it’s similar to NER (named entity recognition) where it will pull more than just a noun itself but all the words around a phrase that make it a noun. This gives you better content around the full noun meaning that just a small piece of it. Once the phrases are pulled, they would need to then be categorized of their respective topics. This would need to be not on a pre-set number of categories or we will miss our subtopics. This goes for both of these solutions. This could be done with clustering, classification, or a modern tool like Hugging Face.
   2. Pros: Simple, more directed than full bigram, removes more of the fluff
   3. Cons: Chance to lose some content, still some noisy data, still longer processing time
3. **Synonym and Library Extraction (1)**
   1. Summary: Creating a library of main words around our pre-selected life events and then using synonym extraction to pull the library words and words that are near them. Synonym detection can be done in a few ways: word embeddings, NLTK Wordnet, BERT, and so forth
   2. Pros: Rule based and manual, More control over what is pulled and no need to post train or classify for the topic area
   3. Cons: Rule based and manual, not as innovative, would take longer to generate the libraries, not as scalable
4. **LDA Topic Modeling (5)**
   1. Summary: Most popular methodology of topic modeling. It looks at the words within the document and sets a probability that it belongs to one of the (pre-set number) topics that the set of documents are funneled into. The number of topics can be variable to the “best” with the set of topics. All documents are assigned into a topic one of which is undefined.
   2. Pros: LDA is quite good at doing topic modeling based on similar word usage, also well established with many good tools and code already created to solve the problem.
   3. Cons: Would require all books to fall into a topic even if they do not cover a life event, Creates major topics around what’s thematic in the group of documents, sub-topics could be overpowered by the major topics in the summary of the book or what is talked about more
5. **Pre-trained classification algorithm (3)**
   1. Summary: Use an algorithm that was trained by another group around similar topics such as ours. This goes all the way from CBOW (continuous bag of words) or Skip-gram options. Includes tokenization and preprocessing of the words. Can use ELMO, or some of the other tools listed prior. Would take research into the one that works best for our context and that can do subtopic not just overall topic
   2. Pros: Option 1 & 2 in most cases need this option to better contain their results. Have good accuracy. Widely used
   3. Cons: Lots of research and testing needed to verify the classification algorithm. May require manual data labeling to evaluate the success of the algorithm and be sure that it is accurately working. Could end up halfway into this option and have to pivot (this con could be a con for others as well)

References:

* <https://www.datasciencecentral.com/nlp-picks-bestsellers-a-lesson-in-using-nlp-for-hidden-feature-ex/>
* <https://www.lexalytics.com/resources/Lexalytics-WP-Context.pdf>
* <https://monkeylearn.com/topic-analysis/>