Textual Analysis Steps

1. Pre-processing:
   1. Convert .pdf to .jpg
2. Machine Readable Text
   1. Tools: Google Cloud Vision API / Google Tesseract
      1. Look into [doc2text](https://github.com/jlsutherland/doc2text) package: Recommended by Elliot Ass
   2. Other Steps:
      1. Fix OCR errors:
         1. [Regular expression](https://docs.python.org/2/howto/regex.html#regex-howto) in python
      2. Remove extraneous content: conditional random fields
3. Split Into Sections
   1. Overview
   2. Union Offer
   3. City Offer
   4. Award
4. Prepare for analysis
   1. Segment text into sentences (spaCY)
   2. Encode sentences
5. Locate protections/provisions within arbitration awards
   1. For each protection identified by Rushin want to know
      1. Whether the union proposed a change
      2. Whether the city agreed to the change
      3. Whether the change was awarded
   2. Potential Machine Learning Techniques to identify presence of provisions:
      1. Word/N-gram spotting
      2. Text Categorization
         1. Manual training
      3. Topic Modeling
   3. A good place to start may be to use Genism Topic Modeling to identify the types of changes proposed by unions/cities and what changes were awarded and how these vary over time