Eric March Scope

Why are you getting the results you’re getting?

you already have the tokenizing down.

keep really good notes on what we have done

might be useful to look at segmenting the text –

python can segment the text into sentences and do part of speech tagging.

Frame – kinds of sentence templates.

training sets – corpus of ones that we know – the quoted ones. and then test it against unquoted ones.

1. segmenting the text – into sentences. part of speech tagging.
2. we’ll want to pull out a corpus of quoted quotes and a corpus of unquoted quotes. that includes the context around them.
3. training set vs test sets.
4. couple of machine learning techniques that we might be able to use. a rolling window. train it on the ones with quotes and then apply it to the untrained ones

NLTK – natural language toolkit. open it up and look at the code. has a part of speech tagger.

use that to make the training corpus.

try and frontload as much of this as possible to the beginning of the year.