2/14/24

Subject, Action, Attribute Extraction Algorithm Utilizing spaCy

1) We split the document into a list of sentences.
2) For each sentence, we iterate over tokens and do:
    a) If the token's dependency is tagged as "nsubj," we store token in variable "subject" and do:
        i) If the token has parents, we iterate through each ancestor and do:
            (a) If the token's dependency is tagged "ROOT" or "AUX" and is not already in array "actions," we **add this token to array "actions"**.
            (b) we then iterate though this token's children and do:
                (i) If the token's dependency is tagged "conj" and the token is not in actions, this **token is added to array "actions"**
                (ii) If the token's dependency is tagged "dobj" the token is added to array "attributes." We then iterate through this tokens children and do:
                    a. If the token's dependency is tagged "appos" it is **added to array "attributes."**
                        i. We then check this same token for children with dependencies tagged "conj." If this is true, we call conjTrail() to recursively traverse the rest of the tree looking for more dependencies tagged with "conj" and adding them to the list until no more tokens are found with the "conj" dependency. **this list is returned and concatenated to "attributes"**
                (iii) We then check the token for the dependency "prep" with a child that has the dependency "pobj" using prepToPobj(). If found this **child token is added to array "attributes".**
                (iv) If the token's dependency is tagged "acomp" we iterate through this tokens children and do:
                    1. We then check the token for the dependency tagged "prep" with a child tagged with "pobj" using prepToPobj(). If found this **child token is added to array "attributes".**
                (v) If the token's dependency is tagged "xcomp," we iterate through the token's children searching for a child with a dependency tagged "dobj" and **add it to the array "attributes."**
                    1. We then check the token for the dependency tagged "prep" with a child tagged with "pobj" using prepToPobj(). If found this **child token is added to array "attributes".**
    b) We continue to iterate through the sentence looking for another subject, if another is found, step a is repeated, if not, the subject, attributes, and actions are output, and we move onto the next sentence.