Strategies for post-processing

1 Exactly filter out:

Filter out CITY annotations when it exactly appears in filter_out list. Currently the exact filter_out list contains country names, continent names, and some names extracted by Alix.

2 Partly filter out:

Filter out CITY annotations when part of the annotated content is in the part filter out list. E.g. <NER:CITY>Fountain Street </NER:CITY> would be filtered out as “street” appears in the list. Currently the part filter out list contains words that is commonly used for location but not for city mentions.

3 CITY inside ORG:

Inside ORG annotations, find city names based on a 56000 city name list

4 <> <>, <>XX<> (e.g. <NE:CITY>New York</NE:CITY>,<NE:CITY> NY</NE:CITY>)

For State abbreviation, still in construction.

Data(list) for strategies above:

filter_out list [strategy 1]; word_filter_out list [strategy 2]; city name list [strategy 3]

Pipeline

Right now the best pipeline I can think of looks like:

3 -> 1 -> 2

Thus after finding city mentions inside ORG annotations, there might be a chance to filter them out in the strategies coming after.