Small TP - Text mining

You Zuo & Jiangnan Huang

January 25, 2021

1 Goal

a. To determine the most frequently discussed geographic location

b. Establish the social or family relationships (friend, sister...) between the characters mentioned.

2 Method

2.1 To achieve goal a

1. Data prepossessing: removing all the punctuation like comma, quotation mark and period etc.
2. Implementation of word segmentation
3. Implementation of named entity recognition
4. Filter the results of NER with only types of location and organization (organizations could sometimes refer to locations)
5. Count the number of occurrence of locations (including co-reference resolution)
6. Sort in descending order and pick the first one.

2.2 To achieve goal b

1. Data prepossessing: removing all the punctuation like comma, quotation mark and period, etc.
2. Implementation of word segmentation.

3. Implementation of named entity recognition.(these 3 steps are the same as above.)

4. Filter the results of NER with only types of relations (friend, sister,stc) and there characters corresponded.(This is the most important and hard step)

5. Establish a database(like RDF) with the relation pairs and at the same time eliminate the duplication.