Assignment 2 Text mining - Tanja Crijns s4204999

1.

Early information extraction systems were rule-based systems. Give one advantage and one disadvantage of rule-based methods. Advantage rule based information extraction:

- Easy to interpret
- Easy to add domain knowledge

Disadvantage rule based information extraction:

- It is heuristic based
- A lot of manual work
- Often not generalizable

2.

Named entity recognition often depends on gazetteers. What is a gazetteer?
A gazetteer is a list of entities like; countries, names, animals etc.

3.

Show what the fragment "Orange Is the New Black is an American comedy-drama web television series created by Jenji Kohan." looks like with BIO-encoding. Use "TIT" as entity type for titles and "PER" for person names. Don't label other entities.

B-TIT - Orange
I-TIT - Is
I-TIT - the
I-TIT - New
I-TIT - Black
O - is
O - an
O - American
O - comedy-drama
O - web
O - television
O - series
O - created

O - by
B-PER - Jenji
I-PER - Kohan
O - .

4.

Section 3.3.1: "An important step in bootstrapping methods is to evaluate the quality of extraction patterns so as not to include many noisy patterns during the extraction process. For example, from the seed entity pair ⟨Google,Mountain View⟩ we may also find “Google, Mountain View” in the corpus. However, the pattern "ORG, LOC" is not a reliable one and thus should not be used." Why is "ORG, LOC" not a reliable pattern?

"ORG, LOC" is too general, a lot of other "X,Y" combinations could fit this pattern without X an Y necessarily being a organization and a location. The earlier mentioned "LOC-based ORG" is more fitted for the specific combination of ORG and LOC.