Questions in preparation of lecture 4

Joris van Vugt, s4279859

Question 1

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

An advantage of rule based systems is that they are transparent to humans. We can come up with rules based on our own knowledge and rules are easily interpreted by others.
A disadvantage of rule based systems is that they require a lot of effort. A lot of rules need to be created which may require expert knowledge and rules are often text-specific (i.e. they don't generalize to other texts).

Question 2

Named entity recognition often depends on gazetteers. What is a gazetteer?

A gazetteer is a list of names of people, organizations, places, etc.

Question 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.*

Token	BIO-encoding
Orange	B-TIT
Is	I-TIT
the	I-TIT
New	I-TIT
Black	I-TIT
is	O
an	O
American	O
comedy-drama	O

web	O
television	O
series	O
created	O
by	O
Jenji	B-PER
Kohan	I-PER
.	O

Question 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?*

This pattern can occur in a regular sentence without the ORG-LOC relation. For example, when the comma is used to seperate parts of a sentence.