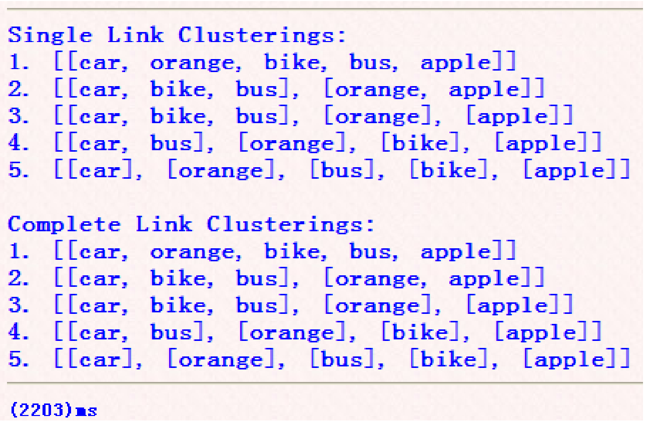
Clustering Methods

Exercises 4/7, 27.3.2017

1. At first step, agglomerative clustering merges two objects that are most similar. In the example below, car and bus are. Later in the process it needs to calculate also similarities of clusters. How to calculate similarity between [car, bus], [bike] and [orange, apple]?



```
Single Link Clusterings:
1. [[car, orange, bike, bus, apple]]
2. [[car, bike, bus], [orange, apple]]
3. [[car, bike, bus], [orange], [apple]]
4. [[car, bus], [orange], [bike], [apple]]
5. [[car], [orange], [bus], [bike], [apple]]

Complete Link Clusterings:
1. [[car, orange, bike, bus, apple]]
2. [[car, bike, bus], [orange, apple]]
3. [[car, bike, bus], [orange], [apple]]
4. [[car, bus], [orange], [bike], [apple]]
5. [[car], [orange], [bus], [bike], [apple]]

(2203)ms
```

2. Calculate Monge-Elkan similarity between strings “orange bicycle” and “apple bus”.

3. What class of similarity measures is more suitable for web page clustering? Give examples.

4. Answer the following questions about inverse term frequency (idf):

   a. What is the idf of a term that occurs in every web page?
   b. Why is the idf of a term always finite?
   c. Can the tf-idf weight of a term in a web page exceed 1?