IRTM Assignment 5

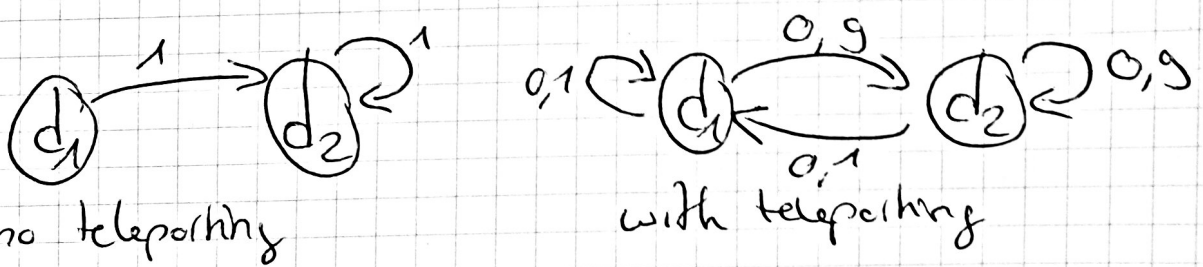
Ex 1 (Feature Selection)

The mutual Information can be defined as:

$$I(U;C) = \sum_{u \in U} \sum_{c \in C} P(u,c) \log_2 \frac{P(u,c)}{P(u)P(c)}$$ with U and C being discrete random variables

Based on this formula we can conclude that a term scores higher for class c_k if it occurs multiple documents of this class than other terms in the documents for this class c. The score also depends if a term occurs in documents which are classified differently. For our example this means the terms „coffe" and „tee" are good indicators for class c_1 and „aspirin" and „paracetamol" is a good indicator for class c_2. The mutual information score for the term „drink" for ~~docum~~ class c_1 is even higher than for ~~co~~ „coffe" and „tee". Same goes for the term „take" for class c_2 and the other terms. However there is no information that the term „I" can provide for this classification.

Ex 5 (Page Rank)



no teleporting | with teleporting

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} 0,1 & 0,9 \\ 0,1 & 0,9 \end{pmatrix}$$

	d_1	d_2	
t_0	0,5	0,5	initial page ~~rank~~ is of equal probabilistic distribution
t_1	0,1	0,9	
t_2	**0,1**	**0,9**	no change, therefore this is the page rank value.