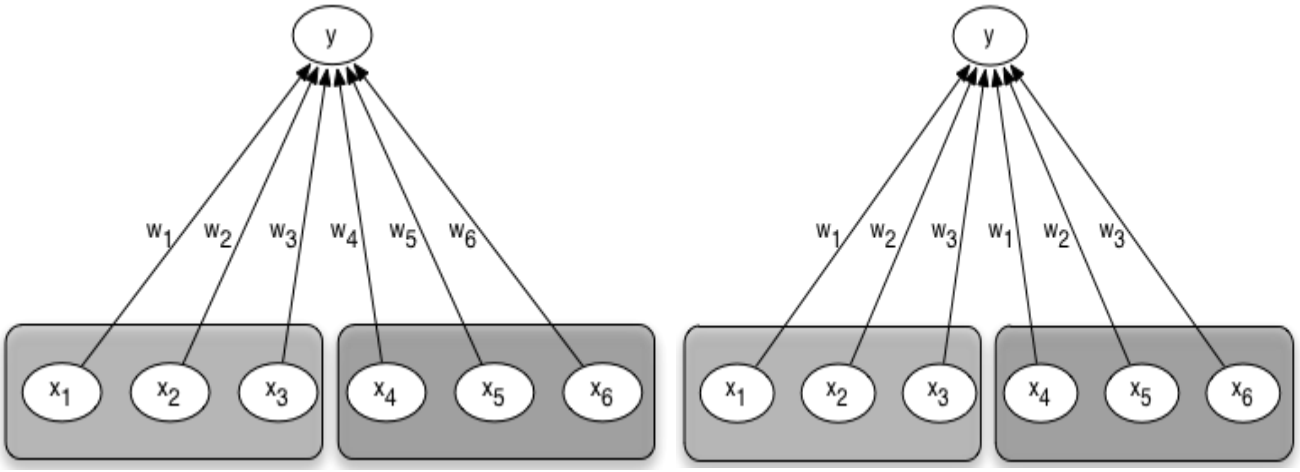
Natural language processing/Obdelava naravnega jezika, 2nd cycle, written exam, 25 January 2017

Each student may have up to 10 sheets of notes in A4 format, other literature is not allowed. Use of any electronic device is considered cheating. All five questions count equally. Duration: 90 minutes.

Oral exam for students who wish to improve their grade and have achieved at least 50% of points in written exam, will take place on Monday, 30 January 2017 at 12:00 in room of Prof Robnik Šikonja. This will also be an opportunity to look into the written exam results.

1. Build a trigram language model for a given document corpus. Suggest at least two solutions how to solve problem of word trigrams that do not appear in corpus. How can you assure that the sum of all possible returned trigram probability scores equals 1.

2. Assume we have a dictionary composed of three words ("a", "b" in "c") and we want to use neural network to predict the next word based on previous two words. The words are encoded as vectors with 3 dimensions, where one component is 1, and the others are 0. For example, word "c" is represented as (0, 0, 1). Neural network has therefore 6-dimensional input (two three-dimensional vectors). For two previous words "a" and "c" the input is (1, 0, 0, 0, 0, 1). Left-hand side of figure below shows our example where each hidden neuron has 6 input weights w_1, w_2, w_3, w_4, w_5 and w_6, where w_i is a weight from i-th input. For a larger context the input will be longer. In case of k previous words and a vocabulary with V words there will be kV input weights for each hidden unit. If we have m hidden units the network will have to learn mkV weights which increases chances of overfitting. A way to reduce the number of parameters is to tie certain weights together, so that they share a parameter. One possibility is to tie the weights coming from input units that correspond to the same word but at different context positions. In our example that would mean that $w_1 = w_4$, $w_2 = w_5$ in $w_3 = w_6$, as shown on right-hand side of figure below.
   Are there any significant problems with this approach, i.e., do we lose any significant information? If yes, what do we loose and why? If not, why not and what do we gain by weight tying?



3. List and explain language resources, which are needed to build POS tagger for morphologically rich language such as Slovene. Justify the listed resources.

4. Write a pseudocode for detector of words with similar meaning based on WordNet. As a criterion, use an appropriate distance over all meanings of a word. Graphically illustrate the workings of your algorithm.

5. Describe the big five personality dimensions and explain how we can analyse them automatically.