Discussion

Pre-workshop

- Revise “language models”.
- What is “POS tagging”? What kinds of information can be used to POS tag a sentence?

Workshop

1. What are the assumptions that go into a “Hidden Markov Model”? How can an HMM be use to “POS tag” text? For the purposes of POS tagging:
   (a) How can the initial state probabilities $\pi$ be estimated?
   (b) How can the transition probabilities $A$ be estimated?
   (c) How can the emission probabilities $O$ be estimated?

2. Consider using the following Hidden Markov Model to tag the sentence silver wheels turn:

   $\pi_{[JJ, NN, VBP]} = [0.3, 0.4, 0.3]$

<table>
<thead>
<tr>
<th></th>
<th>JJ</th>
<th>NN</th>
<th>VBP</th>
<th>O</th>
<th>silver</th>
<th>wheels</th>
<th>turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
<td>JJ</td>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>NN</td>
<td>0.1</td>
<td>0.4</td>
<td>0.5</td>
<td>NN</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>VBP</td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
<td>VBP</td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

   (a) Visualise the HMM as a graph.
   (b) “Unroll” the graph for this sentence.
   (c) Find the most likely tag sequence.

Post-workshop

- Use the HMM above to tag the following sentences:
  - turn silver wheels
  - turn wheels silver
  - wheels turn silver
Programming

Pre-workshop

- There are numerous pre-implemented HMM models, for example, in scikit-learn (http://scikit-learn.org/stable/modules/hmm.html, although it’s scheduled to be removed). Download an HMM and familiarise yourself with setting up an HMM tagger.

Workshop

1. Estimate the HMM parameters for the very small HMM above, by counting frequencies in a tagged corpus, like Brown (nltk.corpus.brown.tagged_words).
2. Use the parameter values you estimated from the corpus to tag the four sentences above.

Post-workshop

- Try building a non-trivial HMM to tag some texts. Compare its output to a Unigram or Bigram tagger. What kinds of words are getting better tags. Why?