POS-tagging and Hidden Markov Models unit:

- The POS-tagging problem (Monday)
- HMM definition and problem statements (Wednesday)
- Solution to Problem 1 (forward algorithm) and POS application (Wednesday and Friday)
- Solution to Problem 2 (Viterbi algorithm) and POS application (Friday and next week Monday)
- Solution to Problem 3 (EM/Baum-Welch algorithm) and linguistic application (next week Monday and Wednesday)

Universal		Penn Treebank		
ADJ	Adjective	JJ	Adjective	*yellow*
		JJR	Comparative adjective	*bigger*
		JJS	Superlative adjective	*wildest*
ADP	Adposition	IN	Preposition	*of, in , by*
		RP	Particle	*up, off*
ADV	Adverb	RB	Adverb	*quickly*
		RBR	Comparative adverb	*faster*
		RBS	Superlative adverb	*fastest*
		WRB	Wh-adverb	*how, where*
CONJ	Conjunction	CC	Coordinating conjunction	*and, but, or*

	Universal		**Penn Treebank**	
DET	Determiner, article	DT	Determiner	*a, the*
		PDT	Predeterminer	*all, both*
		PRP$	Posessive pronoun	*your, one's*
		WDT	Wh-determiner	*which, that*
		WP$	Wh-possessive	*whose*
NOUN	Noun	NN	Singular or mass noun	*llama*
		NNP	Proper noun, singular	*IBM*
		NNPS	Noun, plural	*llamas*
NUM	Numeral	CD	Cardinal number	*one, two*
PRT	Particle	POS	Possessive ending	*'s*
		TO	"to" [Infinitive marker]	*to*
PRON	Pronoun	EX	Existential "there"	*there*
		PRP	Personal pronoun	*I, you, he*
		WP	Wh-pronoun	*what, who*

	Universal		**Penn Treebank**	
VERB	Verb	MD	Modal *can, should*	
		VB	Verb base	*eat*
		VBD	Verb past tense	*ate*
		VBG	Verb gerund	*eating*
		VBN	Verb past participle	*eaten*
		VBP	Verb non-3sp	*eat*
		VBZ	Verb 3sp	*eats*
.	Puntuation mark	(none)		
X	Other	FW	Foreign word	*mea culpa*
		LS	List item marker	*1, 2, One*
		SYM	Symbol	*+, %, &*
		UH	Interjection	*ah, oops*

PRON	VERB	PRT	VERB	ADP	DET	ADJ	NOUN
I	rose	to	saw	off	the	still	rose

PRON	PRON	VERB	ADV	VERB	ADP	DET	NOUN
that	I	saw	still	grew	by	the	still.

Suppose we want to determine the average annual temperature at a particular location on earth over a series of years.

To simplify the problem, we consider only two annual temparatures, "hot" and "cold." Suppose that evidence indicates that the probability of a hot year followed by another hot year is 0.7 and the probability that a cold year is followed by another cold year is 0.6.

Also suppose that research indicates a correlation between the size of tree growth rings and temparature. For simplicity, we consider only three different tree ring sizes: small, medium, and large. Finally suppose hot years are more likely to result in large tree rings, cold years in small.

	H	C	S	M	L
H	0.7	0.3	0.1	0.4	0.5
C	0.4	0.6	0.7	0.2	0.1

Mark Stamp, "A Revealing Introduction to Hidden Markov Models". Abridged.

Let Q be a set of N states types. Use $i, j, ii, jj \in [0, N)$ to index into Q.
Let V be a set of M symbols types. Use $k \in$ to index into V.

Let $\bar{S}$ be a sequence of T state tokens and $\bar{\mathcal{O}}$ be a sequence of T observation tokens.
Use $t \in [0, T)$ to index into $\bar{\mathcal{O}}$ and $\bar{S}$

Thus $\bar{\mathcal{O}} = \langle \mathcal{O}_0, \mathcal{O}_1, \ldots \mathcal{O}_{T-1} \rangle$ is a sequence of observation tokens, e.g., $\mathcal{O}_t = v_k$,
and $\bar{S} = \langle S_0, S_1, \ldots S_{T-1} \rangle$ is a sequence of state tokens, e.g., $S_t = q_j$.

A **hidden Markov model** is a triple $\lambda = (A, B, \pi)$ where

- A is an $N \times N$ matrix of state transition probabilities: $a_{ij} = P(S_{t+1} = q_j \mid S_t = q_i)$
- B is an $N \times M$ matrix of emission (or observation) probabilities:
  $b_j(k) = P(\mathcal{O}_t = v_k \mid S_t = q_j)$
- $\boldsymbol{\pi}$ is the initial state distribution. $\pi_i = P(S_0 = q_i)$

Four HMM problems:

Problem 0. Given $\bar{\mathcal{O}}$ together with $\bar{S}$, compute $\lambda = (A, B, \pi)$ most likely to have produced those sequences.
[Solution: MLE, possibly with smoothing.]

Problem 1. Given $\lambda = (A, B, \pi)$ and $\bar{\mathcal{O}}$, compute the probability that λ assigns to $\bar{\mathcal{O}}$.
[Solution: The forward algorithm.]

Problem 2. Given $\lambda = (A, B, \pi)$ and $\bar{\mathcal{O}}$, find $\bar{S}$ that maximizes the probability that λ assigns to $\bar{\mathcal{O}}$.
[Solution: The Viterbi algorithm.]

Problem 3. Given $\bar{\mathcal{O}}$, M (or V), and N, find $\lambda = (A, B, \pi)$ that maximizes the likelihood of $\bar{\mathcal{O}}$.
[Solution: The Baum-Welch algorithm, a version of EM.]

$$\alpha_t(i) = P(\bar{\mathcal{O}}[:t+1], S_t = q_i \mid \lambda) = \begin{cases} \pi_i \cdot b_i(\mathcal{O}_0) & \text{if } t = 0 \\ \left(\sum_{j=0}^{N-1} \alpha_{t-1}(j) \cdot a_{ji} \right) \cdot b_i(\mathcal{O}_t) & \text{otherwise} \end{cases}$$

$$\beta_t(i) = P(\bar{\mathcal{O}}[t+1:] \mid S_t = q_i) = \begin{cases} 1 & \text{if } t = T-1 \\ \sum_{j=0}^{N-1} a_{ij} \cdot b_j(\mathcal{O}_{t+1}) \cdot \beta_{t+1}(j) & \text{if } t < T-1 \end{cases}$$

$$\delta_t(i) = \max_{\bar{S}[:t+1]} P(\bar{\mathcal{O}}[:t+1], \bar{S}[:t+1] \mid S_t = q_i)$$

$$= \begin{cases} \pi_i \cdot b_i(\mathcal{O}_0) & \text{if } t = 0 \\ \\ \left(\max_{0 \le j < N} \delta_{t-1}(j) \cdot a_{ji}\right) \cdot b_i(\mathcal{O}_t) & \text{otherwise} \end{cases}$$

$$\psi_t(i) = \operatorname*{argmax}_{q_j} P(S_{t-1} = q_j, S_t = q_i \mid \bar{\mathcal{O}}[:t+1])$$

$$= \begin{cases} \texttt{None} & \text{if } t = 0 \\ \\ \operatorname*{argmax}_{0 \le j < N} \delta_{t-1}(j) \cdot a_{ji} & \text{if } t > 0 \end{cases}$$

$$
\begin{aligned}
\lg \sum_{i=0}^{n-1} x_i &= \lg(x_0 + x_i + \cdots + x_{n-1}) \\
&= \lg x_0 + \lg\left(1 + \sum_{i=1}^{n-1} \frac{x_i}{x_0}\right) \\
&= \lg x_0 + \lg\left(1 + \sum_{i=1}^{n-1} 2^{\lg x_i - \lg x_0}\right)
\end{aligned}
$$

$$
\begin{aligned}
\xi_t(i,j) &= P(S_t = q_i, S_{t+1} = q_j \mid \bar{\mathcal{O}}, \lambda) \\
&= \frac{P(S_t = q_i, S_{t+1} = q_j, \bar{\mathcal{O}} \mid \lambda)}{P(\bar{\mathcal{O}} \mid \lambda)} \\
&= \frac{\alpha_t(i) \cdot a_{ij} \cdot b_j(\mathcal{O}_{t+1}) \cdot \beta_{t+1}(j)}{\sum_{ii} \sum_{jj} \alpha_t(ii) \cdot a_{ii\ jj} \cdot b_{jj}(\mathcal{O}_{t+1}) \cdot \beta_{t+1}(jj)}
\end{aligned}
$$

$$
\begin{aligned}
\gamma_t(i) &= P(S_t = q_i \mid \bar{\mathcal{O}}, \lambda) \\
&= \sum_{j=0}^{N-1} P(S_t = q_i, S_{t+1} = q_j \mid \bar{\mathcal{O}}, \lambda) \\
&= \sum_{j=0}^{N-1} \xi_t(i,j)
\end{aligned}
$$

$$\pi_i = \gamma_0(i)$$

$$a_{ij} = \frac{\text{expected transitions from } q_i \text{ to } q_j}{\text{expected transitions from } q_i} = \frac{\sum_{t=0}^{T-2} \xi_t(i,j)}{\sum_{t=0}^{T-2} \gamma_t(i)}$$

$$b_i(k) = \frac{\text{expected times } q_i \text{ emits } v_k}{\text{expected times in } q_i} = \frac{\sum_{t=0}^{T-2} \{\gamma_t(i) \mid \mathcal{O}_t = v_k\}}{\sum_{t=0}^{T-2} \gamma_t(i)}$$

Coming up:

- Language model programming assignment (Mon, Sept 25)

- POS quiz (Thurs, Sept 21)
- Reading from J&M, Sections 8.(0–4) (Fri, Sept 22)
- HMM quiz (Tues, Sept 26)
- HMM programming assignment (Wed, Oct 4)