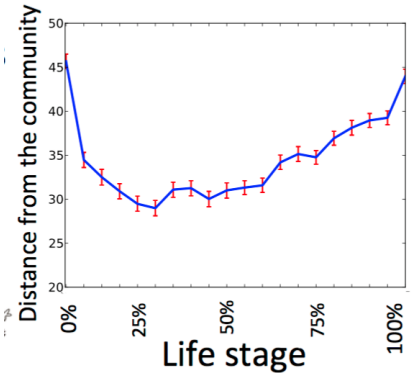
1 (By way of semi-contrast) What we've seen before: lists of Fightin' Words

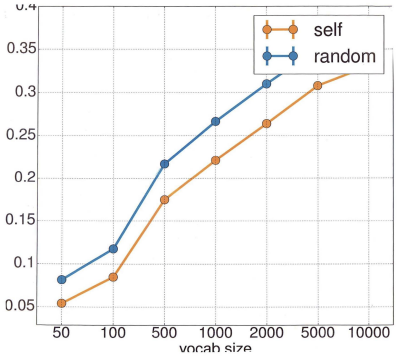
Earlier lecture: Fightin' words introduced as an exploratory method for getting lists of words that seem to distinguish two language samples (e.g., "babies" vs. "women" in 106th U.S. Senate Republican vs. Democrat speeches on Abortion). I think should you apply this method in your course project and future research (unless inapplicable, but usually it's applicable).

2 Motivation for our new unit on language models

The Fightin' words analysis depended on having estimate $p(v)$ for words v. Where did those ps come from?

Also, in many settings, we want to have at hand a way to compute *a single number* that tells us the degree to which one language "source" differs from another.





Left: figure 6(a) in No country for old members: User lifecycle and linguistic change in online communities, Danescu-Niculescu-Mizil et al., WWW 2013. For each month, a bigram language model is constructed for the user's postings; and a bigram language model is constructed for the posts for a randomly selected 500 users active that month (using only two posts per user, so the size of the samples is the same across months). Plotted is the *cross-entropy* between the two distributions

Right: Figure by Chenhao Tan. Is a user more influenced by their "own language" or their "current environment"? A language model is constructed from a user u posting in r/food (the "current environment"). We compare the *Jensen-Shannon divergence* (y-axis, down=more similar) of this language model against (1) "self", a language model built on that u's postings but in in r/movie, and against (2) "random", a language model built from some other randomly drawn user $\hat{u}$ posting in r/food. (The x-axis is for different choices of vocabulary size. Note that it is not a linear scale.)

Further example (inspired by course projects): instead of looking for whether the topics, according to a topic model, for two parts/samples of language are different, use an "easier" language model class and compare language models for the two parts/samples.

There are several ways to measure the difference between two (estimated) distributions.

3 Language models

A *language models (LM)* is a means for assigning probabilities over strings (word sequences).

Examples of classes of language models: n-gram models; topic models; Hidden Markov Models; probabilistic context-free grammars; GPT-3 and cousins.

I tend to think of language models as the (imaginary) source underlying a given language sample, and learning a language model as attempting to learn parameters (characteristics) of that source.

4 How do you build a language model?

- Use a toolkit to finetune an existing pretrained language model to your data
- Use a toolkit to learn one from just your data (plus maybe some priors you know)
- Build one of your own on your data (plus maybe some priors you know) [example: sorta like Fightin' Words with uninformative prior]
- Interpolate between a language model built on your data plus a language model built on "generic data" [example: sorta like Fightin' Words with informative dirichlet prior on other Senate speeches]

5 General education: Estimating a multinomial's categorical distribution.

Learning goal: appreciation for first-principles derivation

Assume a fixed non-empty finite vocabulary $V = \{v_i\}$.
The multinomial has the following parameters:

- L, the number of draws (the sample length)
- $\overrightarrow{\phi} \in \Re^{|V|}$, where $\sum_i \phi_i = 1$ and for all i, $\phi_i \geq 0$. This *categorical* distribution on just V (not V^*) specifies probabilities on the sides of the "die" whose sides are labeled with the vocabulary items v_i.

We are given a sample $S = w_1 \ldots w_L$, $w_k \in V$, and collect the counts S_i for each word v_i.

Maximum-likelihood estimate: find $\overrightarrow{\phi}$ that maximizes

$$\text{(some constant with factorials?)} \times \prod_i \phi_i^{S_i} \qquad \ldots? \tag{1}$$

Maximum a posteriori estimate: assuming Dirichlet prior's parameter vector $\overrightarrow{\alpha}$ is fixed, find the $\overrightarrow{\phi}$ that maximizes

$$\text{Prob}(\overrightarrow{\phi} \text{drawn according to} \overrightarrow{\alpha}) \cdot \text{(some constant with factorials?)} \cdot \prod_i \phi_i^{S_i} \qquad \ldots? \tag{2}$$