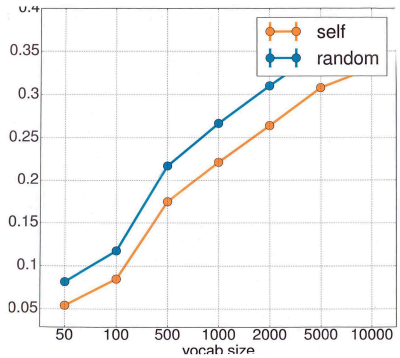
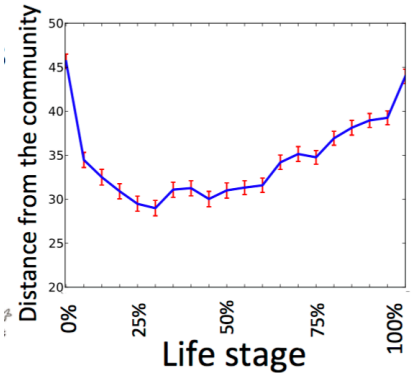
CS/IS 6742: NLP and Social Interaction, Fall 2017
Oct. 26, 2017: Reference sheet for lecture 19, "How different are two language models? (cont.)"

1 Motivating examples





Left: Figure by Chenhao Tan. Is a user more influenced by their "home language" or their "current environment"? A language model is constructed from a user u posting in r/food. We compare the *Jensen-Shannon divergence* (y-axis) of this language model against (1) *self*, a language model built on that u's postings 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.)

Right: figure 6(a) in No country for old members: User lifecycle and linguistic change in online communities, Danescu-Niculescu-Mizil et al., WWW 2013. (Title is a clickable link.)

2 Measuring the difference between two "single-word" distributions

2.1 Vocabulary issues

We will restrict attention to language models where it is sensible to consider an induced distribution, a "next-single-item" distribution, on just V.

But what is a "word?"

Two sets from a preliminary analysis by a student in class:

yahoo, 2000, re, guys, yahoo.com, kind, won, thought, %, le, street, provider, series, night, where, don, august, dorland, percent, cents, in, partnerships, first, went, down, water, officer, nov., might, stocks, wrote, another, car, probably, reserved, jane, doing, la, story, ?, care, profits, told, like, god, ∗, friend, get, one, fund, merger, fastow, two, that, pretty, back, much, news, they, good, ., u.s., ventures, game, executives, thanksgiving, dow, house, raised, executive, chief, times, almost, funding, little, a, after, says, when, billion, tell, up, could, life, thing, capital, former, ever, −, still, ceo, big, man, partners, some, fell, even, then, got, how,

...

incumbent, warrant, fate, dolores, l=ike, ups, counts, on=, distinction, iv, stations, witter, hewlett, maley, voting, 11:31, =market, idaho, isthat, unregulated, unbelievable, worship, holden, powergen, exclusively, canal, promptly, ramirez, spun, unlimited, embarrassment, mines, competitiveness, jason.leopold, desire, citi, instructed, aronowitz/hou/ect, pastoria, appointment, 11:09, noble, errol, 7,500, kathleen, reversal, metering, totake, brink, sits, monique, rep, insult, nasty, taxation, proves, followers, listening, 10:20, thestate, 9:12, minimal, icould, 12, bcf, protections, advancing, 24., studio, relevantaffiliate, whats, shoot, 12:26, deposit, 135, 11:40, unhappy, gibson, grief, eugene, hunt, totals, handles, endless, pge.com, pot, ranch, evident, feed, mich., 77, georgia, frustrated, swift, 97, scottsdale, fidelity, wreck, coles, attorney−client, 77002, rents, printing, 2:20, thesame, weeks=, thave, improved, =this, accompanied, day−to−day, since=, lag, checks, strengthened

We restrict attention to $q(\cdot)$ and $r(\cdot)$ over "vocabulary" $V = \{v_i\}$. We write q_i and r_i for $q(v_i)$ and $r(v_i)$.

2.2 Examining the behavior of various distance functions

The *surprisal*[1]:

$$-\log(r_i) = \log \frac{1}{r_i} \tag{1}$$

can be thought of as how *surprised* we should be from the perspective of using r as a model to see v_i, or r's *surprisedness* or *surprisingness* for v_i. The base of the log is customarily taken to be 2, which makes this surprisingness number interpretable as a number of bits of information.[2]

2.2.1 Cross-entropy

If we considered the "reference" distribution to be q, then the *cross-entropy*

$$H(q||r) = \sum_i q_i \log \frac{1}{r_i} \tag{2}$$

is the expected surprisedness for r with respect to reference distribution q.

If the "reference" distribution is taken to be the one induced from the empirical counts from a sample $S = w_1 w_2 \ldots$, where each $w_k \in V$ and the length of the sample is L, then this can be refactored as:

$$\hat{H}_S(r) = \frac{1}{L} \sum_{k=1}^{L} \log \frac{1}{r(w_k)} \tag{3}$$

2.2.2 KL-Divergence

$$D(q||r) = \sum_i q_i \log \frac{q_i}{r_i} \tag{4}$$

2.2.3 Jensen-Shannon divergence

See Lin, Jianhua. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37(1): 145-151. Let $\text{avg}_{q,r}$ be the average distribution between q and r.

$$JS(q, r) = \frac{1}{2} \left[D(q||\text{avg}_{q,r}) + D(r||\text{avg}_{q,r}) \right] \tag{5}$$

2.2.4 Skew divergence

See Lee, Lillian. 1999. Measures of distributional similarity. In *Proceedings of the ACL*, 25-32.

$$\text{skew}_\beta(q||r) = D(q||\beta \cdot r + (1 - \beta)q) \tag{6}$$

Values used include $\beta = .99$.

[1] According to Wikipedia, the term was coined in Tribus, 1961, *Thermostatics and Thermodynamics*

[2] Indeed, a much more common interpretation of equation 1 is as a number of bits needed to encode v_i assuming the distribution r over V.