Word Embeddings (II)

CS 485, Fall 2024
Applications of Natural Language Processing

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

- What do we have
  - Dense vector model of word meanings
  - For many words, learned from a large corpus
  - Learned from principle of distributional similarity

x, y

How do we compare vectors?

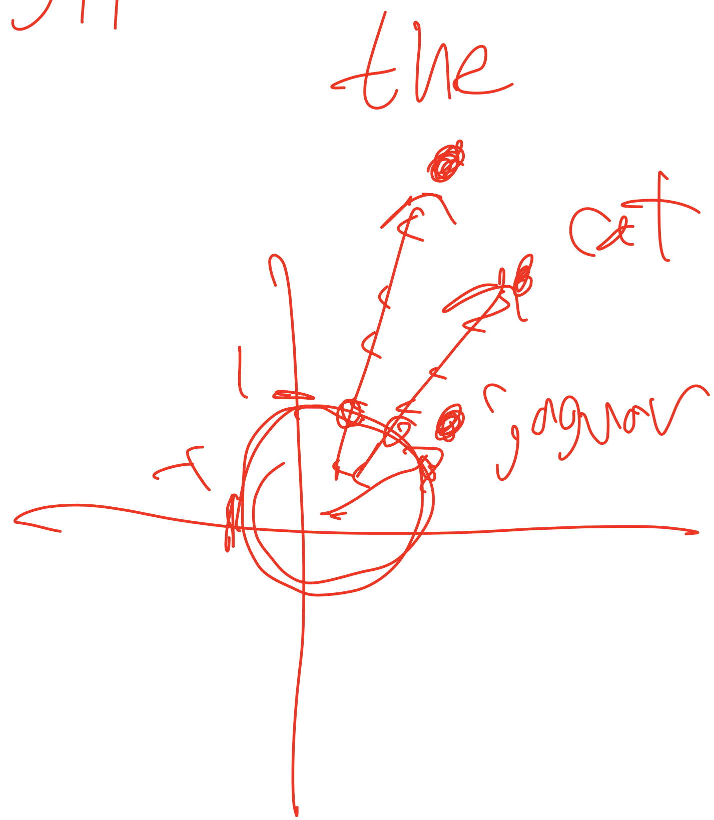
sim(x, y)

- Similarity measurements
  - Larger values → similar vectors → similar words
  - Smaller values → dissimilar vectors → dissimilar words

- Distance / dissimilarity measurements
  - *Note: distance metric requires triangle inequality*
  - Larger values → dissimilar vectors → dissimilar words
  - Smaller values → similar vectors → similar words

Euclidean Distance

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2} = \|x-y\|$$



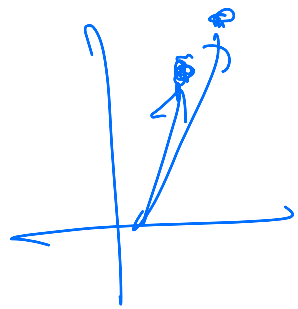
Issue: Vector length depends on frequency. More frequent words will have longer vectors.

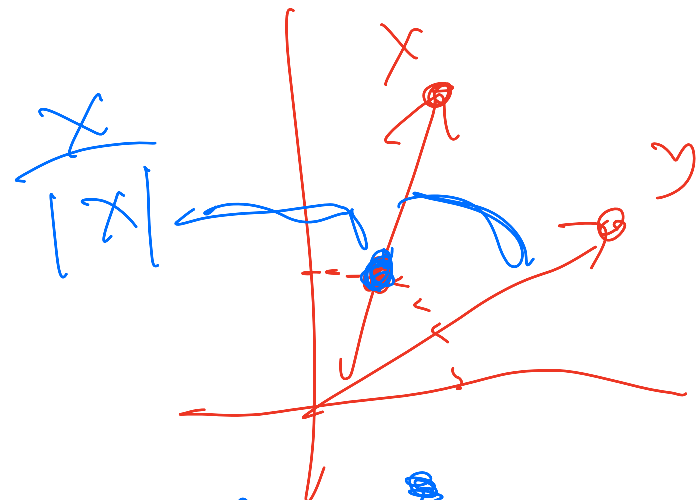
Cosine Similarity

$$s(x,y) = \frac{x \cdot y}{|x||y|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \left(\frac{x}{|x|}\right) \cdot \left(\frac{y}{|y|}\right)$$

Only depends on vector angle

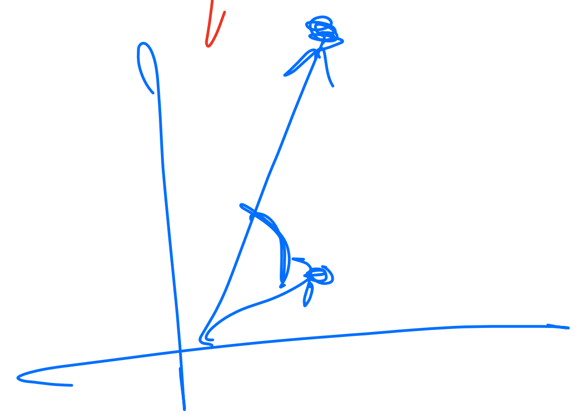
Range: $[-1, 1]$





$$s(x,x) = \frac{x \cdot x}{|x|\,|x|} = \frac{\sum_i x_i^2}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i x_i^2}}$$

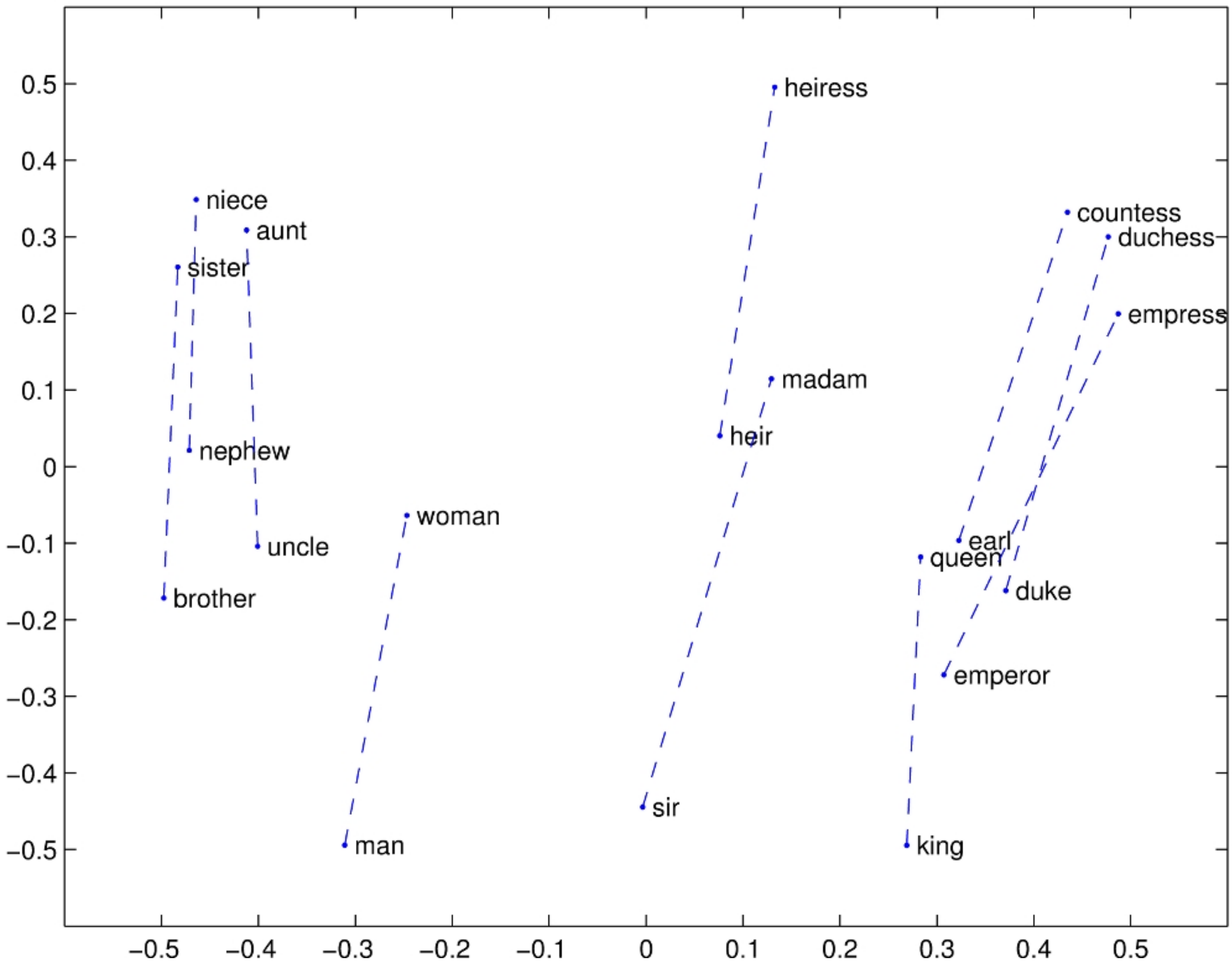
$$= \frac{\sum_i x_i^2}{\sum_i x_i^2} = 1$$



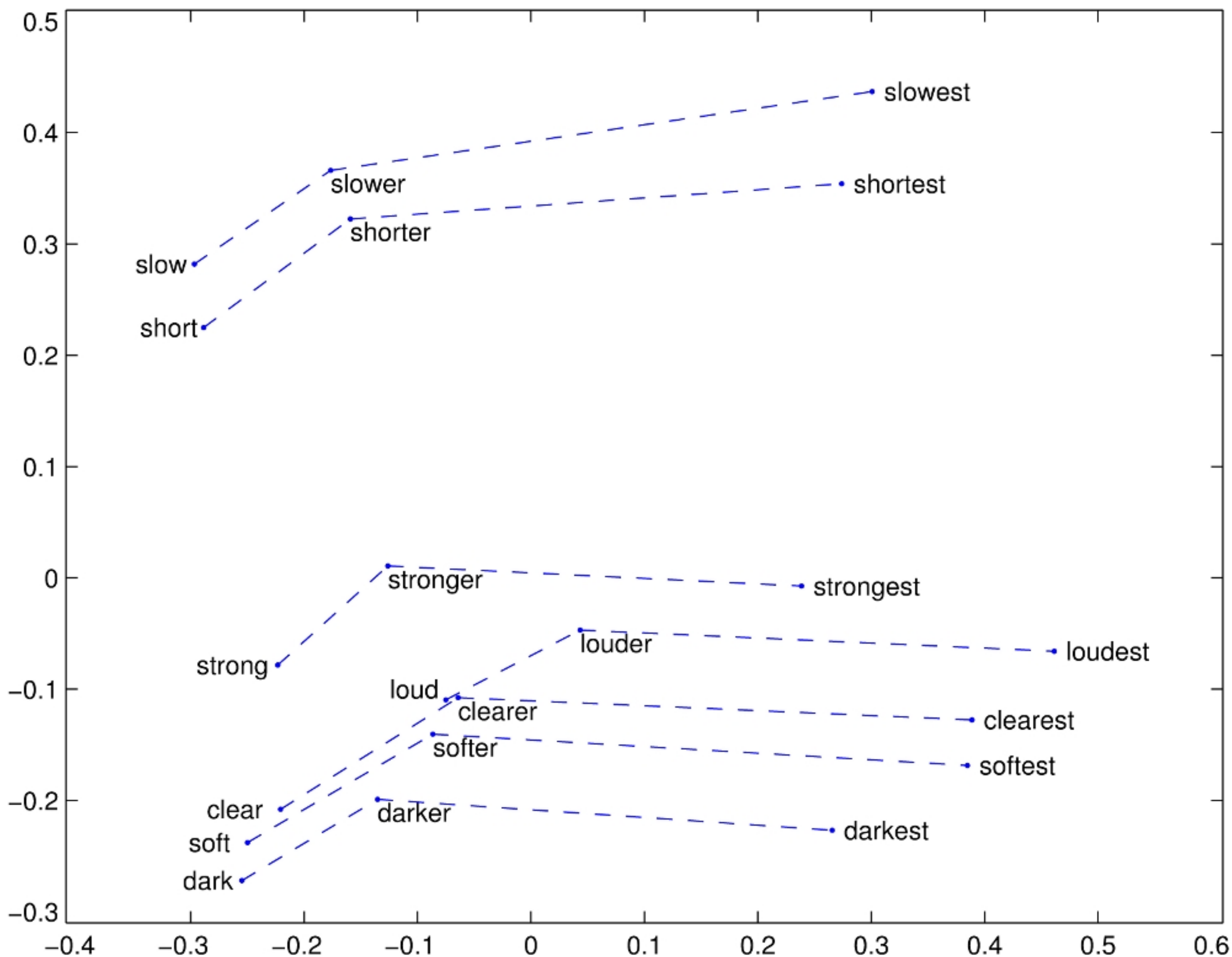
What does it learn?

- Demo: GLOVE embedding similarities
  - fasttext, glove, and word2vec are most-often used pretrained word embeddings

https://nlp.stanford.edu/projects/glove/



https://nlp.stanford.edu/projects/glove/

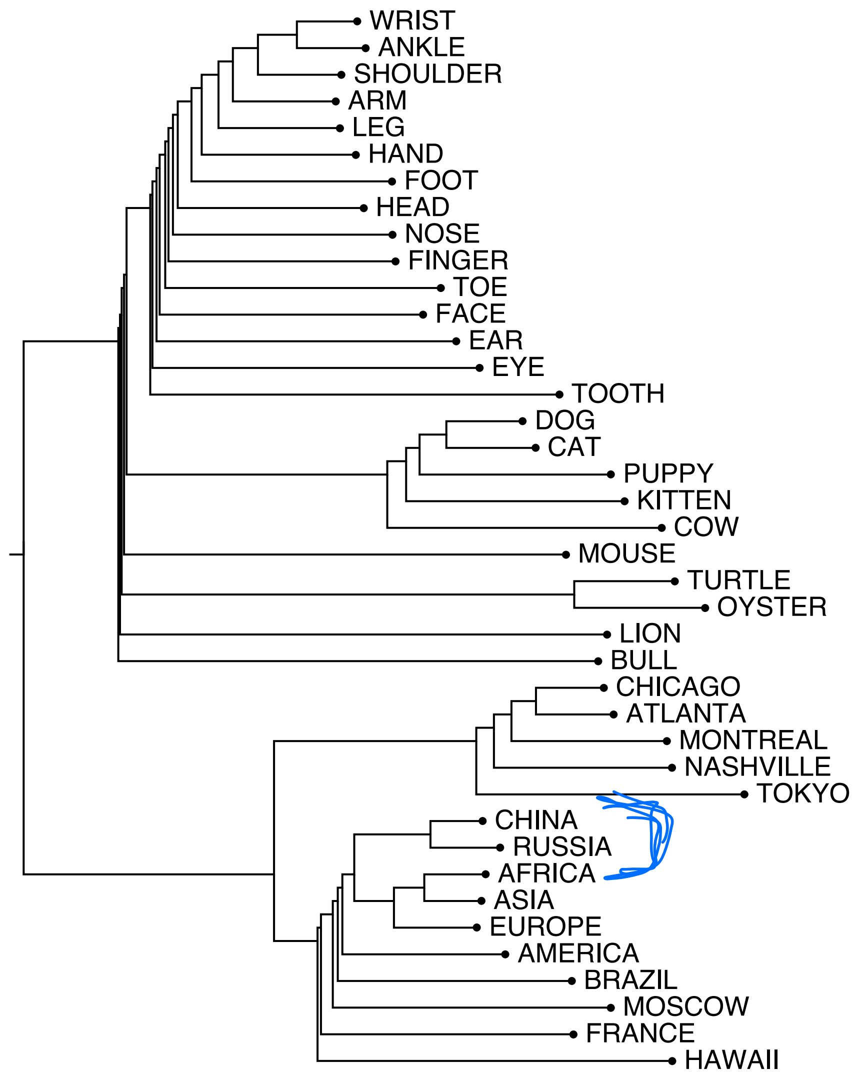


embeddings may have larger-scale semantic structure?

- Hierarchical distributional word clusters, trained from tweets: http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html
- What distinctions is it learning?

optimistic ---- pessimistic

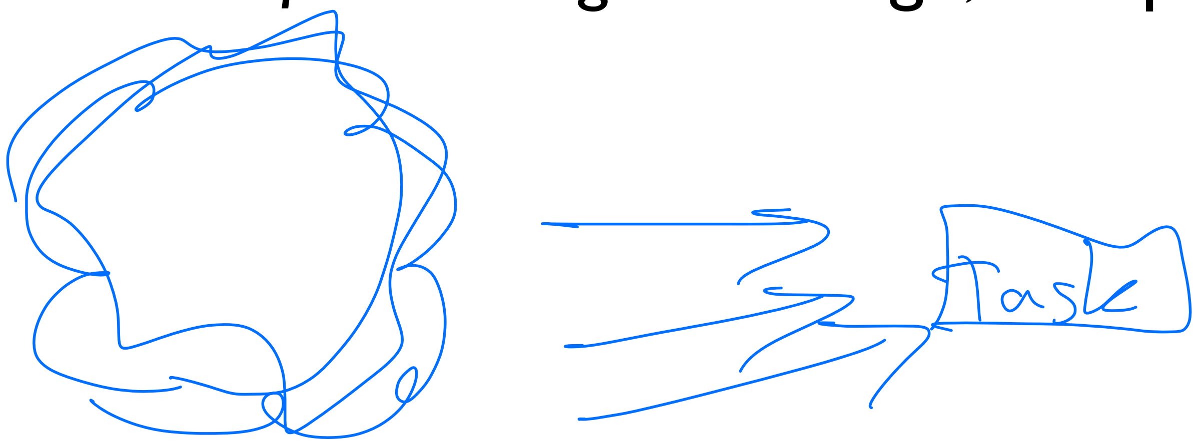
embeddings may have larger-scale semantic structure?



SLP3 ch. 6

ok so what can we do with them?

- *Transfer learning* from large, unsup. corpus

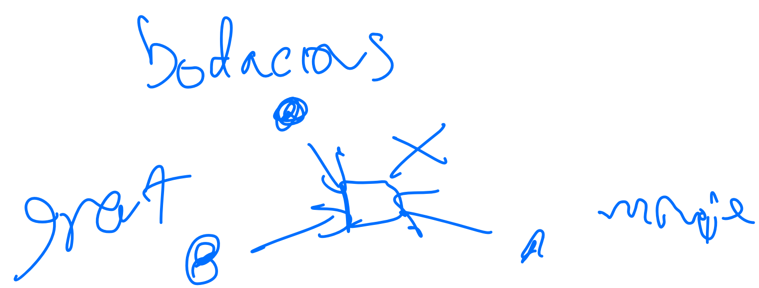


- Document embeddings
  - 1. Supervised learning: Bag-of-Embeddings logreg
    - labeled train docs->labeled new docs
  - 2. Unsupervised learning / exploratory analysis
    - docs->[analysis]
- Wordlist-based inferences
  - 3. Semi-automatic dictionary expansion
    - (words->words)
  - 4. DDR: Distrib. Dict. Representations
    - (words->docs)

(1) Sup. learning with document embedding

- Instead of bag-of-words, can we derive a latent embedding of a document/sentence?
  - "Bag of embeddings" or "averaged word embeddings" representation
  - You can use it just like a BOW logistic regression - it's just a different type of feature vector
  - Pros/cons?
- Especially for shorter texts, BoE LR typically outperforms BOW LR.

$$X = \frac{1}{N_{tok}} \sum_{w \in doc} \text{embedding}(w)$$



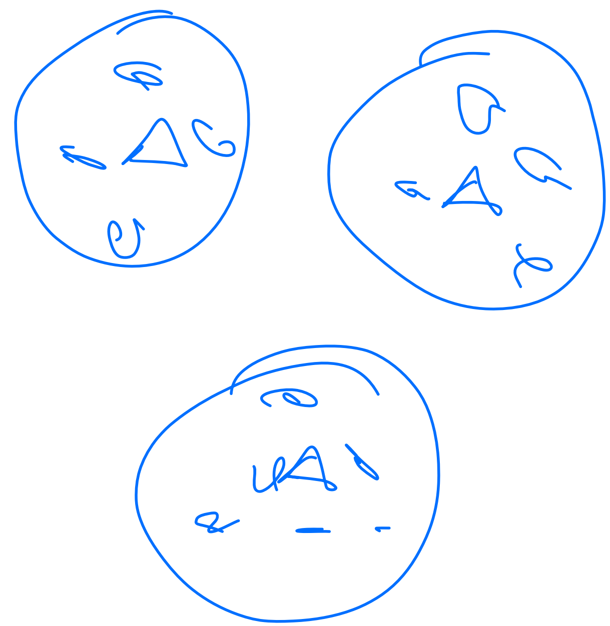
See: Arora et al. 2017

(2) Unsup. learning with document embedding

- Example: tweets about mass shootings (Demszky et al. 2019)
  1. Average word embeddings => tweet embeddings
  2. Cluster tweets (k-means)
  3. Interpret clusters' words (closest to centroid)

Topic	10 Nearest Stems
news (19%)	break, custodi, #breakingnew, #updat, confirm, fatal, multipl, updat, unconfirm, sever
investigation (9%)	suspect, arrest, alleg, apprehend, custodi, charg, accus, prosecutor, #break, ap
shooter's identity & ideology (11%)	extremist, radic, racist, ideolog, label, rhetor, wing, blm, islamist, christian
victims & location (4%)	bar, thousand, california, calif, among, los, southern, veteran, angel, via
laws & policy (14%)	sensibl, regul, requir, access, abid, #gunreformnow, legisl, argument, allow, #guncontolnow
solidarity (13%)	affect, senseless, ach, heart, heartbroken, sadden, faculti, pray, #prayer, deepest
remembrance (6%)	honor, memori, tuesday, candlelight, flown, vigil, gather, observ, honour, capitol
other (23%)	dude, yeah, eat, huh, gonna, ain, shit, ass, damn, guess

Table 1: Our eight topics (with their average proportions across events) and nearest-neighbor stem embeddings to the cluster centroids. Topic names were manually assigned based on inspecting the tweets.



Application: keyword expansion

- I have a few keywords for my task. Are there any I missed?
- Automated or semi-automated new terms from embedding neighbors

- Other non-embedding lexical resources can do this too (e.g. WordNet), but word embeddings typically cover a *lot* of diverse vocabulary

Application: doc sim to words

- Given a word list to represent a concept, can we score a document for how much it expresses that concept?
    - Count based approach?

Application: doc sim to words

- Given a word list to represent a concept, can we score a document for how much it expresses that concept?
- DDR is a very simple embedding approach:
  - Average the word lists embeddings to create a concept vector
  - Average a doc's words to create a document vector
  - Apply cosine similarity!
- Supplying a set of keywords is *low-supervision*, or low-expertise, approach compared to labeling docs
  - Though you don't get a nice logreg probability (until you label some...)

[Garten et al. 2018]

LIWC "posemo" list

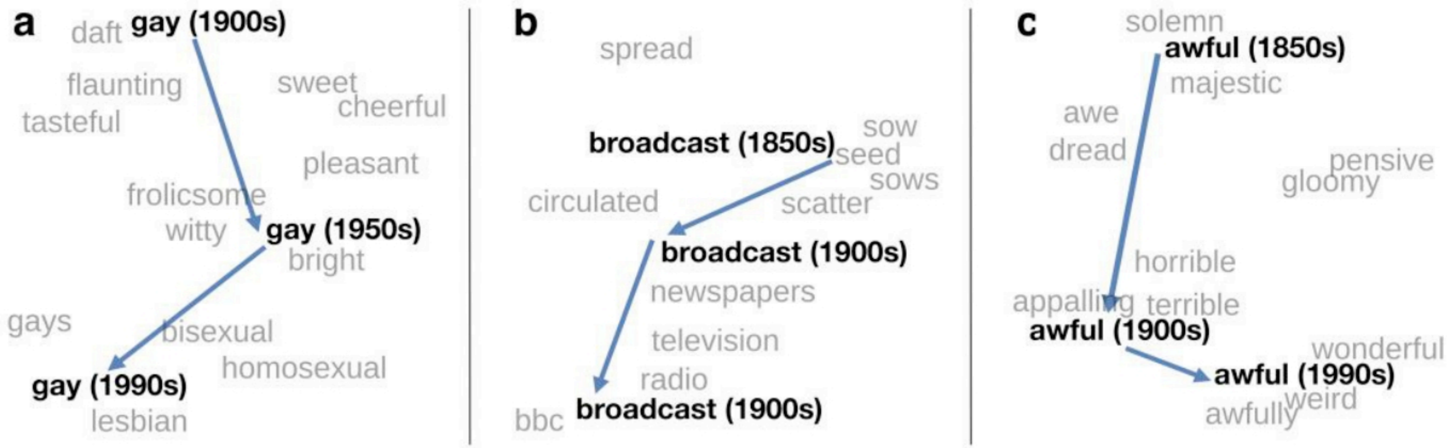
accept, accepta*, accepted, accepting, accepts, active*, admir*, ador*, advantag*, adventur*, affection*, agree, agreeab*, agreed, agreeing, agreement*, agrees, alright*, amaz*, amor*, amus*, aok, appreciat*, assur*, attachment*, attract*, award*, awesome, beaut*, beloved, benefic*, benefit, benefits, benefitt*, benevolen*, benign*, best, better, bless*, bold*, bonus*, brave*, bright*, brillian*, calm*, care, cared, carefree, careful*, cares, caring, casual, casually, certain*, challeng*, champ*, charit*, charm*, cheer*, cherish*, chuckl*, clever*, comed*, comfort*, commitment*, compassion*, compliment*, confidence, confident, confidently, considerate, contented*, contentment, convinc*, cool, courag*, create*, creati*, credit*, cute*, cutie*, daring, darlin*, dear*, definite, definitely, delectabl*, delicate*, delicious*, deligh*, determina*, determined, devot*, digni*, divin*, dynam*, eager*, ease*, easie*, easily, easiness, easing, easy*, ecsta*, efficien*, elegan*, encourag*, energ*, engag*, enjoy*, entertain*, enthus*, excel*, excit*, fab, fabulous*, faith*, fantastic*, favor*, favour*, fearless*, festiv*, fiesta*, fine, flatter*, flawless*, flexib*, flirt*, fond, fondly, fondness, forgave, forgiv*, free, freeb*, freed*, freeing, freely, freeness, freer, frees*, friend*, fun, funn*, genero*, gentle, gentler, gentlest, gently, giggl*, giver*, giving, glad, gladly, glamor*, glamour*, glori*, glory, good, goodness, gorgeous*, grace, graced, graceful*, graces, graci*, grand, grande*, gratef*, grati*, great, grin, grinn*, grins, ha, haha*, handsom*, happi*, happy, harmless*, harmon*, heartfelt, heartwarm*, heaven*, heh*, helper*, helpful*, helping, helps, hero*, hilarious, hoho*, honest*, honor*, honour*, hope, hoped, hopeful, hopefully, hopefulness, hopes, hoping, hug, hugg*, hugs, humor*, humour*, hurra*, ideal*, importan*, impress*, improve*, improving, incentive*, innocen*, inspir*, intell*, interest*, invigor*, joke*, joking, joll*, joy*, keen*, kidding, kind, kindly, kindn*, kiss*, laidback, laugh*, libert*, like, likeab*, liked, likes, liking, livel*, lmao, lol, love, loved, lovely, lover*, loves, loving*, loyal*, luck, lucked, lucki*, lucks, lucky, madly, magnific*, merit*, merr*, neat*, nice*, nurtur*, ok, okay, okays, oks, openminded*, openness, opport*, optimal*, optimi*, original, outgoing, painl*, palatabl*, paradise, partie*, party*, passion*, peace*, perfect*, play, played, playful*, playing, plays, pleasant*, please*, pleasing, pleasur*, popular*, positiv*, prais*, precious*, prettie*, pretty, pride, privileg*, prize*, profit*, promis*, proud*, radian*, readiness, ready, reassur*, relax*, relief, reliev*, resolv*, respect, revigor*, reward*, rich*, rofl, romanc*, romantic*, safe*, satisf*, save, scrumptious*, secur*, sentimental*, share, shared, shares, sharing, silli*, silly, sincer*, smart*, smil*, sociab*, soulmate*, special, splend*, strength*, strong*, succeed*, success*, sunnier, sunniest, sunny, sunshin*, super, superior*, support, supported, supporter*, supporting, supportive*, supports, suprem*, sure*, surpris*, sweet, sweetheart*, sweetie*, sweetly, sweetness*, sweets, talent*, tehe, tender*, terrific*, thank, thanked, thankf*, thanks, thoughtful*, thrill*, toleran*, tranquil*, treasur*, treat, triumph*, true, trueness, truer, truest, truly, trust*, truth*, useful*, valuabl*, value, valued, values, valuing, vigor*, vigour*, virtue*, virtuo*, vital*, warm*, wealth*, welcom*, well, win, winn*, wins, wisdom, wise*, won, wonderf*, worship*, worthwhile, wow*, yay, yays

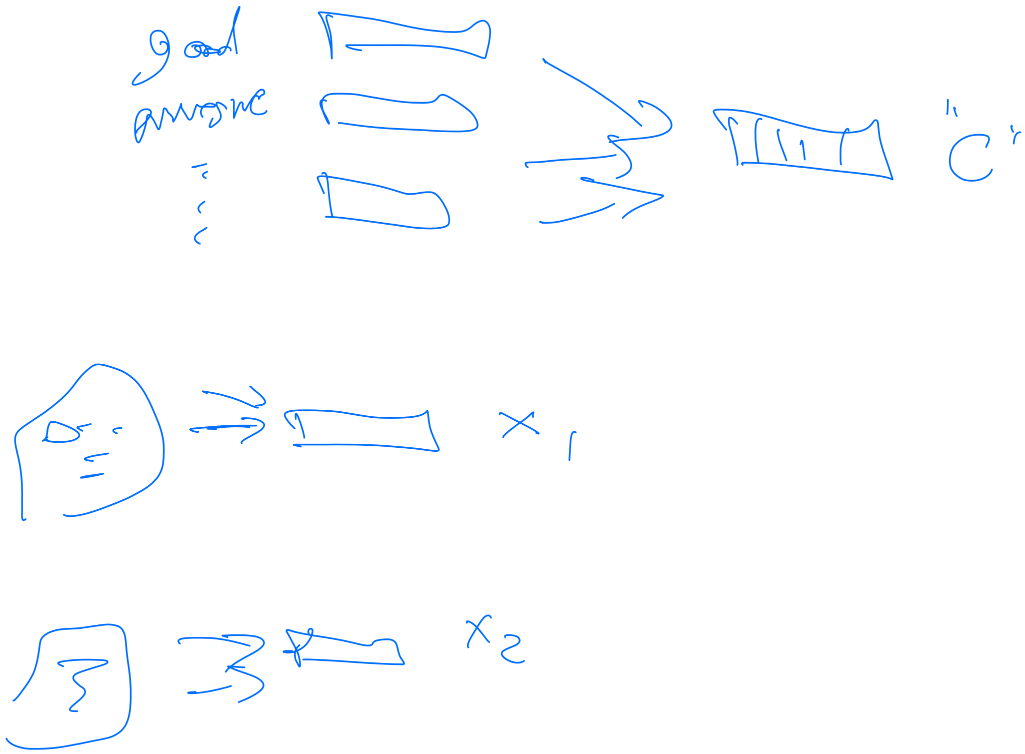
Fig. 4 Nearest neighbors of the LIWC positive emotions dictionary

Pretraining corpus is key

- Language models—this week, word embeddings learned via LMs—enable *transfer learning* from the pretraining corpus, to whatever your desired end-task is
- Ideally: train on domain-specific corpus.
  Usually: use Wikipedia + random web pages
  (is this good??)
- The content of the pretraining corpus is very important!!
  - The best word embedding releases document and explore the implications of how they chose their pretraining corpus.

Word use over time
[Hamilton et al. 2016]




good
awesome
C
x1
x2