Vector-based Semantics

Credit: much of this lecture is based on Chris Potts' excellent videos for Stanford's CS 224u class

Agenda

Review of vector-based representations and similarity metrics

Evaluating the goodness of our vector representations using psycholinguistics data

What are word vectors good for (and is it really "semantics")?

- Find the most similar words
- Solve word relation puzzles like analogies
- Compose their meanings (?)

What semantics problems are difficult to do with word vectors?

- negation / antonyms / exclusion
- inference
- composition of meaning for sentences.

user - document / movie

word - document matrix

novels

	d1	d2	d3	d4	d5	...	dN
abandon	.	.	.	.	.		
abdicate							
abhore							
academic							
⋮							
zygodactyl							
zymurgy							

Compare columns

What is a document

- Shakespeare's plays ✓
- Wikipedia articles ✓
- IMDB reviews for a movie ✓
- speeches by politicians
- real estate listings

Compare the similarity of "documents"

- author attribution / plagarism detection / document de-duplication
- clustering into categories
- recommendation systems
- compare documents by "tone"
- "doc" = query

Comparing rows

- what words are similar to each other

Zellig Harris (1954)

distributional statements can cover ALL of the material of a language without requiring support from other types of information

Turney and Pantel (2010) "from freq to meaning"

if units of text have similar vectors in a text frequency matrix then they tend to have similar meanings

word x word matrix / context

	abandon	abdicate	abscond	...	zymurgy
abandon					
abdicate					
abscond					
⋮					
zymurgy					

$|V| \times |V|$

take 5k most freq.

Define a co-occurrence matrix

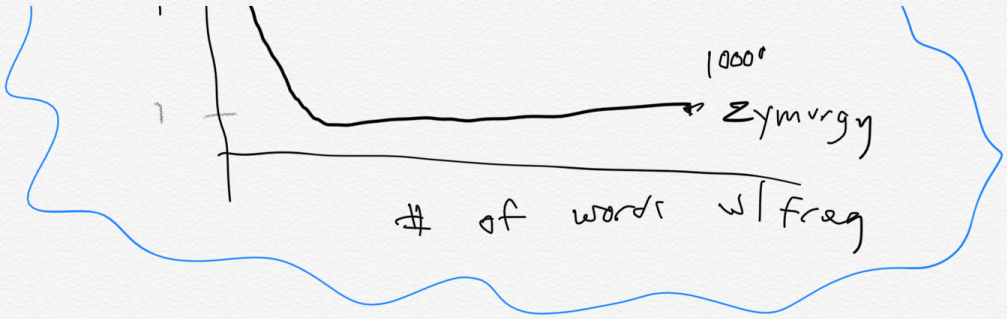
context window

w_{-2} w_{-1} target word w_{+1} w_{+2}

word2vec

Aside:

Zipfian distribution



co-occur in a document
" " " sentence
" " " window of words

more complex contexts dependency patterns

The pictures are beautiful.

subj-of-verb
adj-mod
obj-of

The pictures of the old man holding his first grandchild are beautiful

Long distance dependencies

Lin and Pantel (2001) DIRT
duty and responsibility
modified by similar sets of adjectives

additional - amod
administrative - amod
congressional - amod
constitutional - amod
⋮

objects of similar sets of verbs:

abandon - obj-of
abdicate - obj-of
assign - obj-of
assume "
breach "

par pattern co-occurrence matrix

X works with Y, X cuts Y

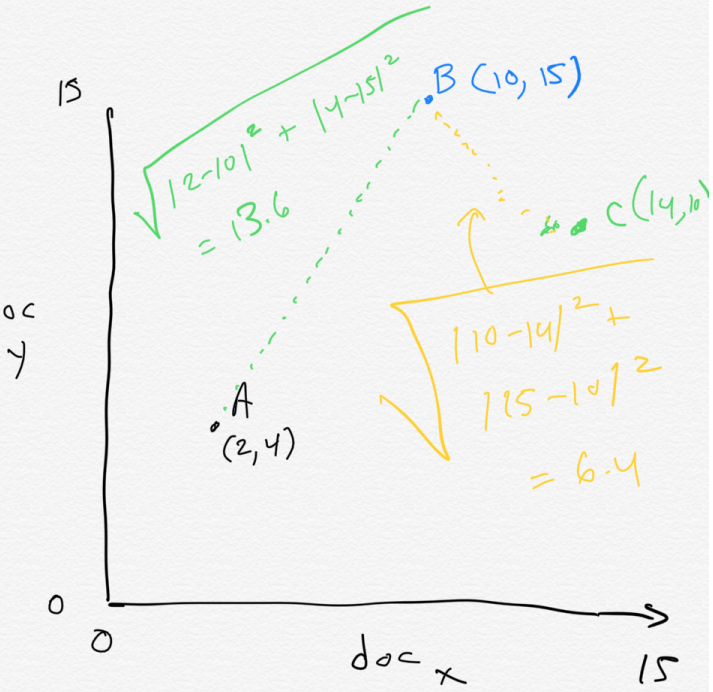
<mason, stone>
carpenter, wood
scissors, paper

Vector Comparison

pos movie review

words	doc_x	doc_y
A	2	4
B	10	15
C	14	10

A = "superb" +/low freq
B = "good" +/high freq
C = "disappointing" −



Euclidean distance

vectors u, v of dimension N

$$\sqrt{\sum_{i=1}^{N} |u_i - v_i|^2}$$

Goal: related words are close
unrelated words are far apart

Vector L2 (length) normalization

normalization of u is a vector $\hat{u}$ obtained by dividing each element of u by $\|u\|$

L2 length

$$\|u\| = \sqrt{\sum_{i=1}^{N} u_i^2}$$

	doc_x	doc_y	
A	0.45	0.89	$\sqrt{2^2 + 4^2} = 4.47$

Vector L2 (length) normalization

normalization of u is a vector $\hat{u}$ obtained by dividing each element of u by $\|u\|$ L2 length

$$\|u\| = \sqrt{\sum_{i=1}^{N} u_i^2}$$

	doc x	doc y
A	0.45	0.89
B	0.55	0.83
C	0.81	0.58

$\sqrt{2^2 + 4^2} = 4.47$

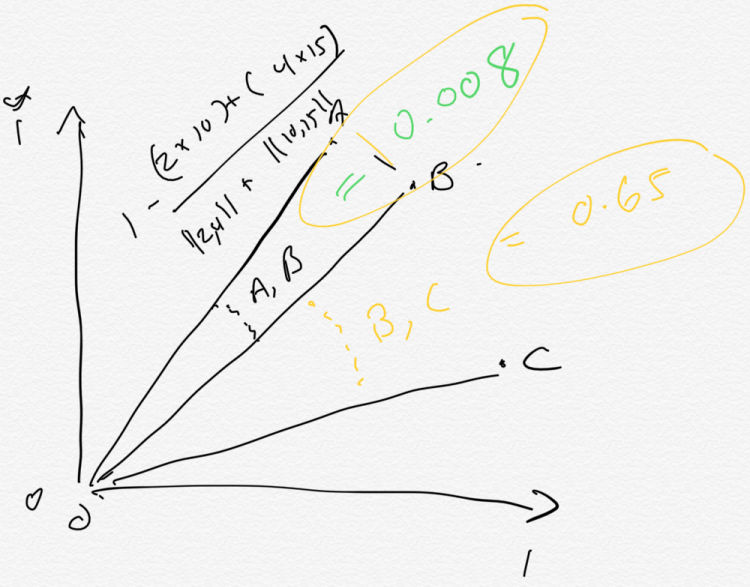
$\hat{A} = \left[2/4.47 , 4/4.47 \right]$

cosine distance

$$1 - \frac{\sum_{i=1}^{N} u_i \times v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \times \sqrt{\sum_{i=1}^{N} v_i^2}}$$

dot product

norm quantities



What information do we lose when we discard freq?

Distributional inclusion hypothesis

A is a B

dire wolf is a animal

How do I make a nice graph w/ >>2 dim?

t-SNE

Matrix type

Vector comparisons

cosine
Manhattan distance
KL-divergence
JS distance
DICE coefficient
Jaccard
..

Reweighting

length norm.
probabilities
TF-IDF
PMI
positive PMI

Dim. Reduction

LSA
PCA
LDA

GloVe
word2vec
CBOW
skip grams

How many dimensions should we use?
10? 50? 100?

mods to input

tokenization/normalization
annotation/tagger
parsing, feature selection

cluster texts by author/time/etc

How to pick the right combination?
Trial and error.

Evaluating word vectors

2 kinds of evaluation
① extrinsic evaluation = task-based
② intrinsic

Psycholinguistic experiment

cos sim			mean 10 judges
	Love	sex	6.8
	tiger	cat	1.3
	tiger	tiger	10
	fertility	egg	6.7
	stock	egg	1.8
	professor	cucumber	0.3

GOLD STANDARD

WordSim 353

Compute correlation

- Spearman's rank correlation coefficient
- Kendall's tau "

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{\# of discordant pairs})}{N(N-1)/2}$$

total pairs

human ordering

sex, love > prof., cucumber

system ordering predicts

< ⇒ discordant pair

> ⇒ concordant pair

range -1 to 1

wordvectors.org

- community based eval
- leaderboard
- visualizations t-SNE

wordsim 353 plus several other human similarity benchmark sets

Does similarity == meaning / semantics?

word vectors fail to capture real implications

dog(x) → animal(x)

dog(x) ⇒ ¬ gorilla(x)

Does similarity == meaning / semantics?

benchmark sets

word vectors fail to capture logical implications

$dog(x) \rightarrow animal(x)$

$dog(x) \Rightarrow \neg\, gorilla(x)$

word vectors for antonyms or logically exclusive things are often very similar

sim(boys , girls) high sim

cats dogs high sim

France Germany high sim

rise fall high sim

big- man green- ?

Next time:
How Word2vec can be used to solve analogies!