Embeddings

Sameer Singh and Conal Sathi

BANA 290: ADVANCED DATA ANALYTICS

MACHINE LEARNING FOR TEXT

SPRING 2018

May 15, 2018

Upcoming...

Homework

- Homework 3 will be out this week
- Due in ~2 weeks: May 29, 2017
- Focused on clustering

Project

- Instructions for proposal are out
- Due tonight: May 15th
- Progress presentations in week 10 (~3 weeks)

Outline

Vector Space Models

Latent Semantic Analysis

Word Embeddings

Outline

Vector Space Models

Latent Semantic Analysis

Word Embeddings

Document Vectors

Vector Space Models!

Supervised Learning

- As features in classification
- Labels *propagate* along similar documents

Unsupervised Learning

- As distance in clustering
- Defined by similar documents

TF-IDF
Ngrams
etc.

Term-Document Matrix

Local and Global Weighting

Local Weighting

- Binary:
- Term Freq:
- Log:

Global Weighting

- Binary:
- Normal:
- IDF:

Example: Documents

c1: Human machine interface for ABC computer applications
c2: A survey of user opinion of computer system response time
c3: The EPS user interface management system
c4: System and human system engineering testing of EPS
c5: Relation of user perceived response time to error measurement

m1: The generation of random, binary, ordered trees
m2: The intersection graph of paths in trees
m3: Graph minors IV: Widths of trees and well-quasi-ordering
m4: Graph minors: A survey

From http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

Document Similarity

A survey of user opinion of computer system response time

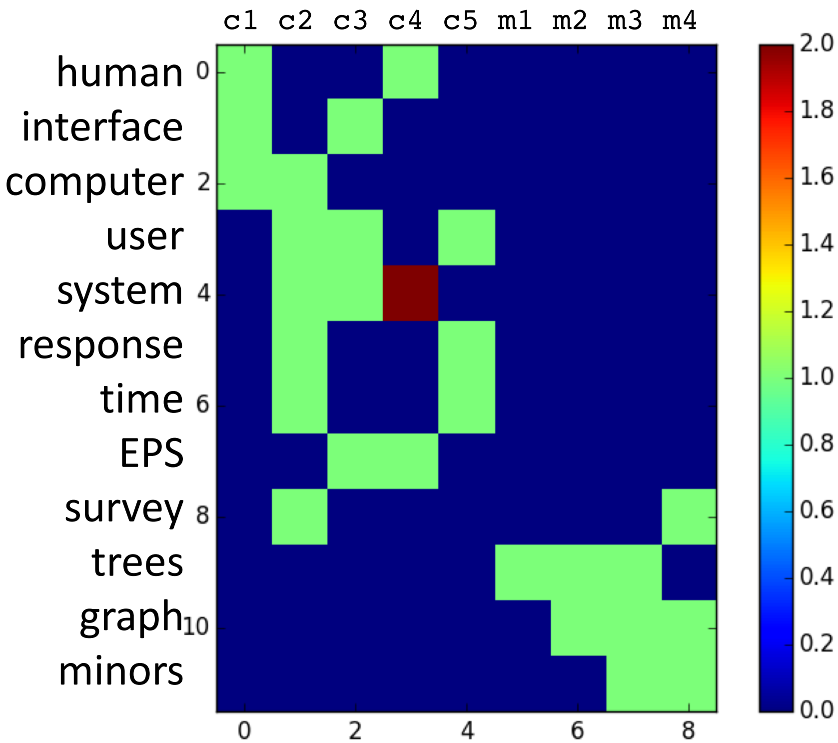
Relation of user perceived response time to error measurement

The generation of random, binary, ordered trees

Example

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Example: Word-Doc Matrix



Problem with Sparse Matrices

c2: A survey of user opinion of computer system response time

m4: Graph minors: A survey

c1: Human machine interface
for ABC computer applications

Example

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	[illegible]	[illegible]
time	0	1	0	0	1	0	0	[illegible]	[illegible]
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

r (human.user) = -.38
r (human.minors) = -.29

The Problem

Two problems that arise when using the vector space model:

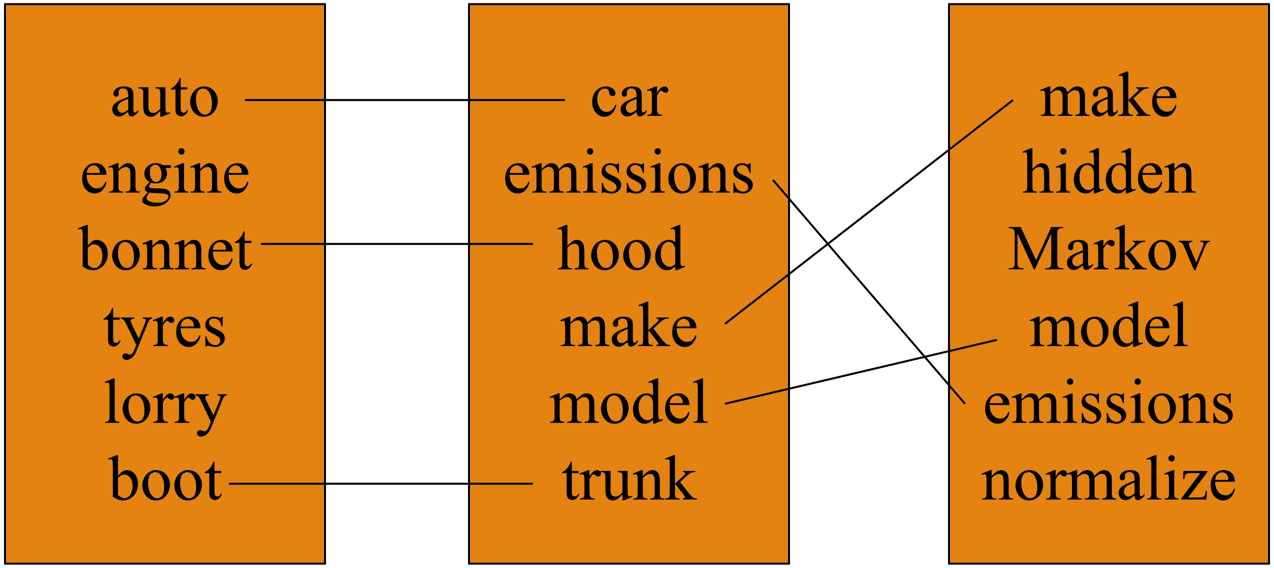
Synonymy:

- many ways to refer to the same object, e.g. car and automobile
- leads to poor recall

Polysemy:

- most words have more than one distinct meaning, e.g. model, python, chip
- leads to poor precision

The Problem



Synonymy

Will have small cosine

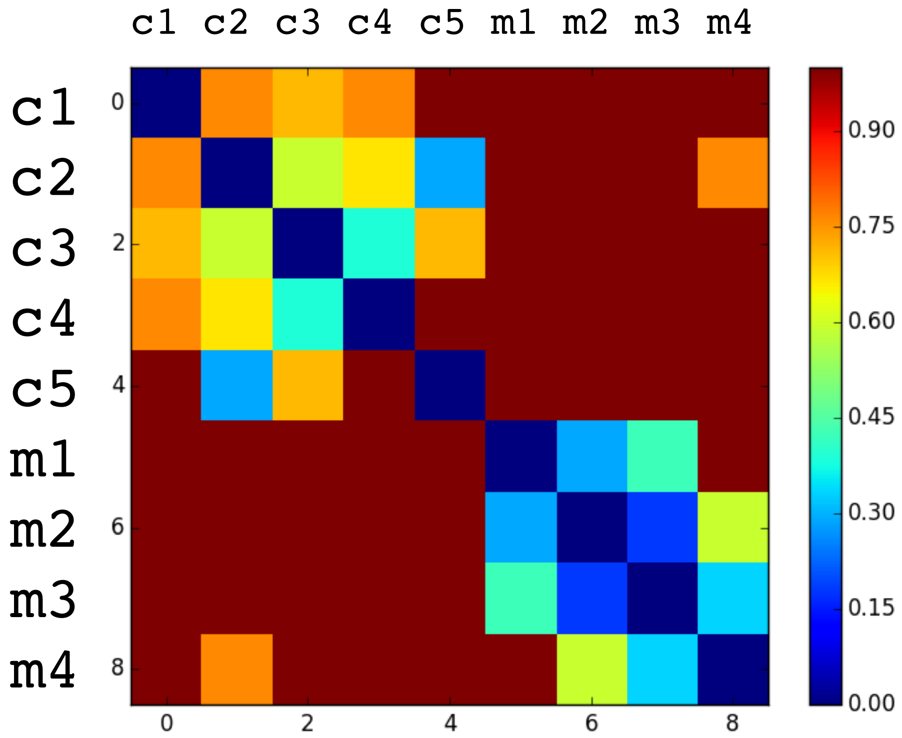
but are related

Polysemy

Will have large cosine

but not truly related

Example: Distance Matrix



Going from Sparse to Dense

Outline

Vector Space Models

Latent Semantic Analysis

Word Embeddings

Latent Semantic Analysis (LSA)

Term-Document Matrix

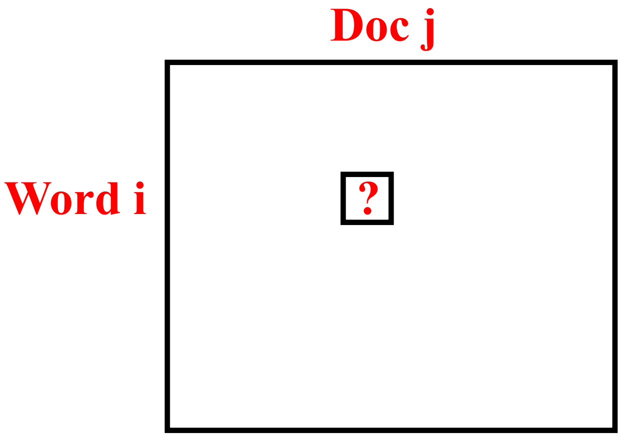
- "Word i appears" = row i
- "in document j" = column j

Huge matrix (mostly zeros)

- Treat zeros as if word is not relevant?

PCA/SVD on this matrix provides a new representation

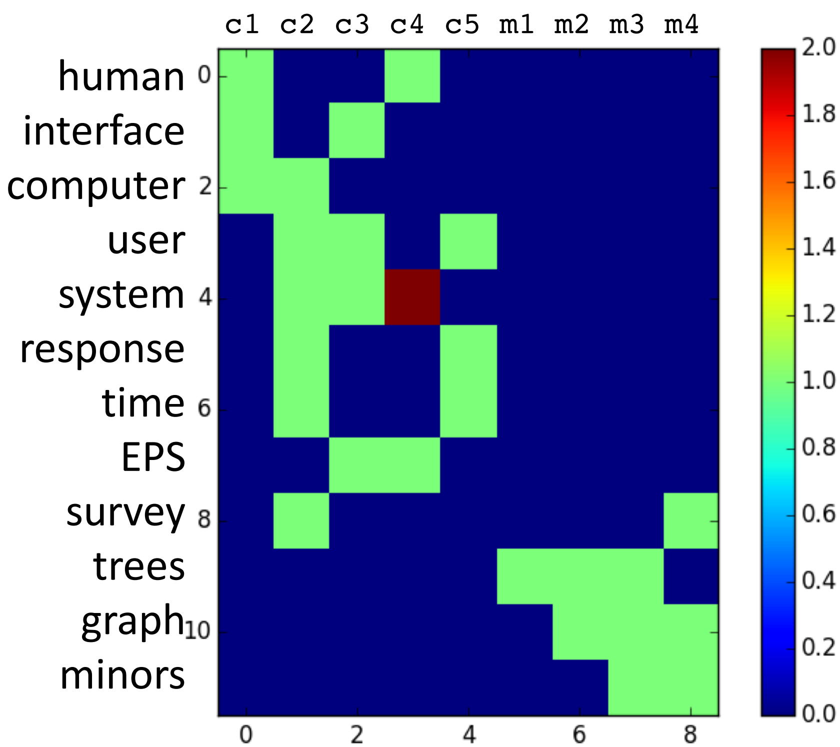
- Latent = "hidden"
  - Consider which other words "could have appeared"
- Semantic = "topics"
  - Fuzzy search ("concept" instead of "word" matching)



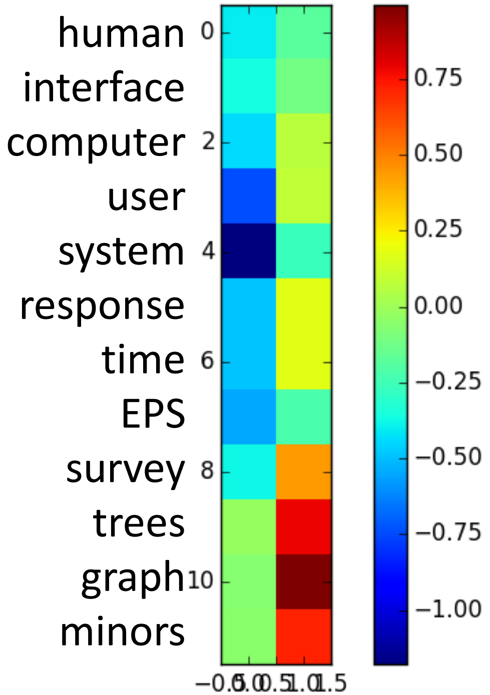
Singular Value Decomp (SVD)

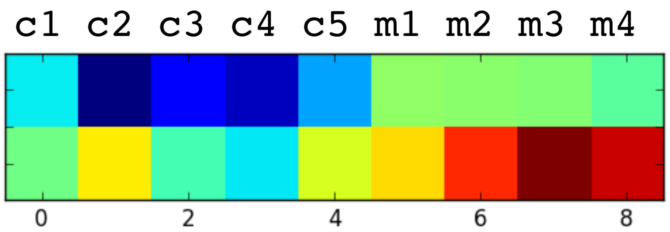
Latent Semantic Analysis (LSA)

Example: Term-Doc Matrix

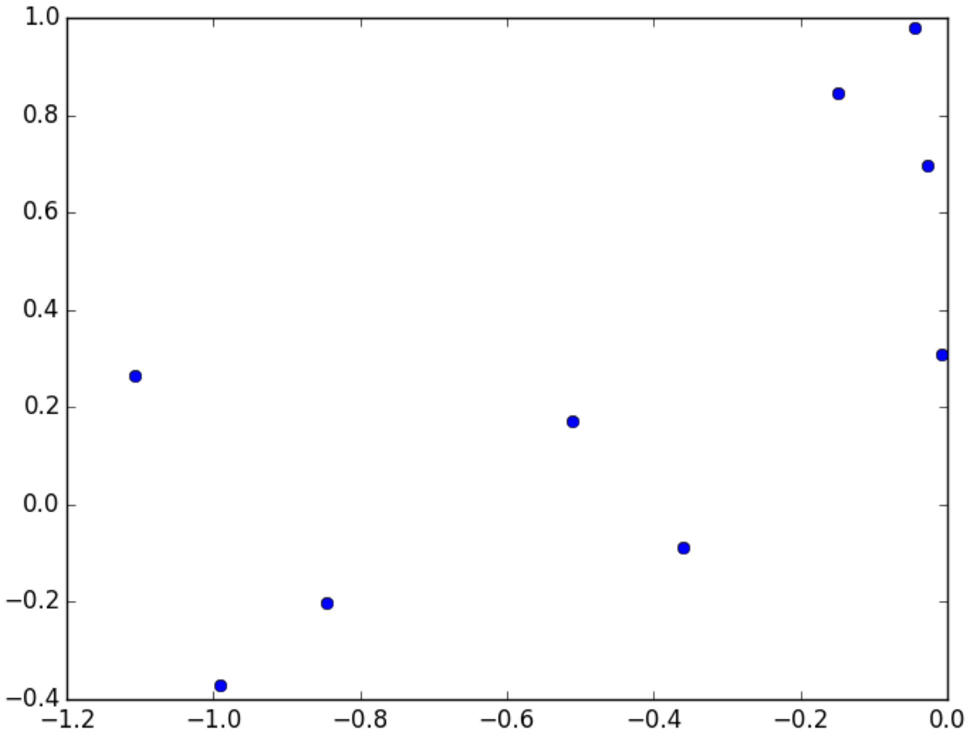


Example: Decomposition

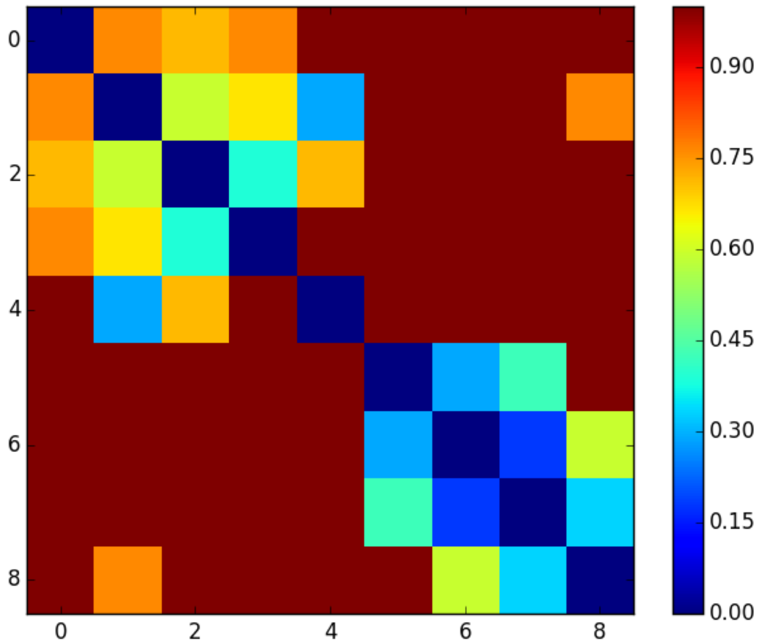


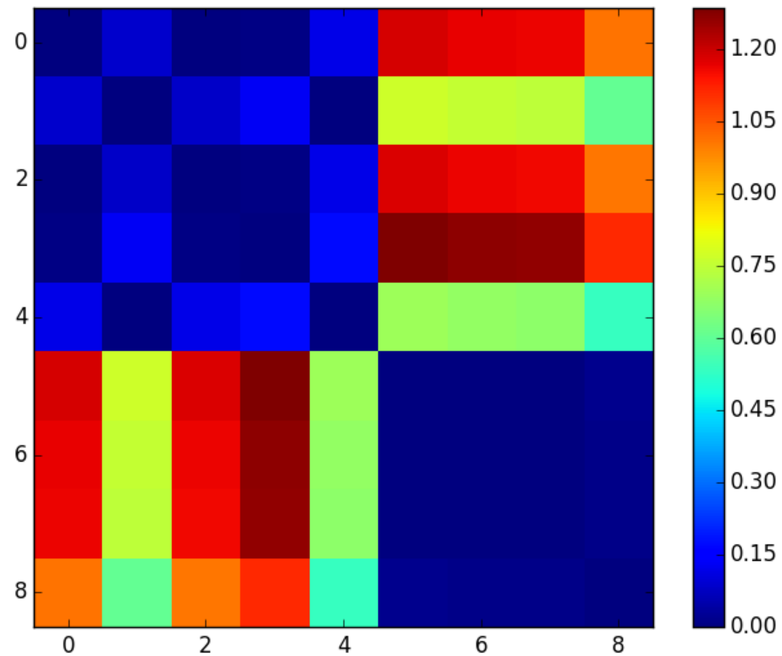


New Document Vectors

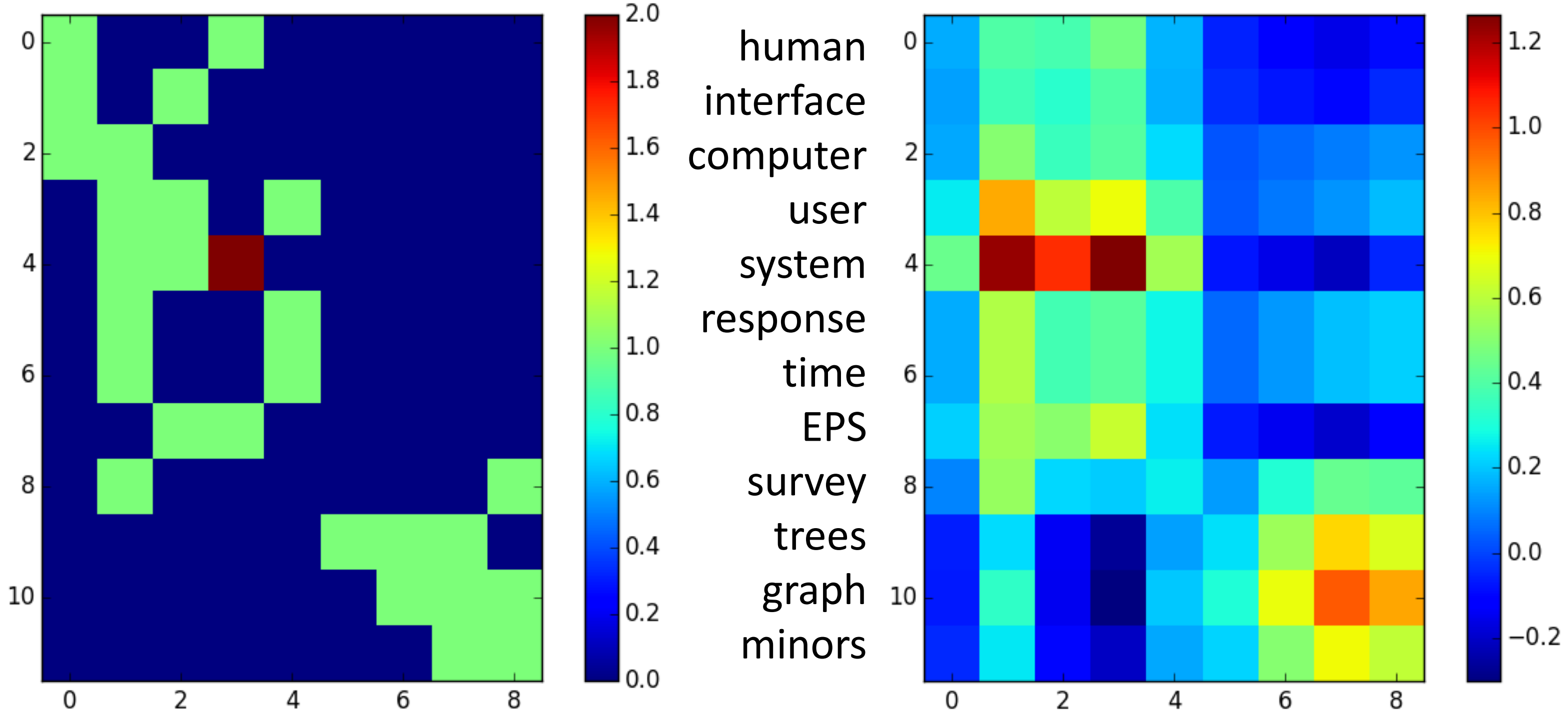


Example: Distance Matrix





Example: Reconstruction



In-Class Activity 1

Outline

Vector Space Models

Latent Semantic Analysis

Word Embeddings

Let’s look at words

A bottle of tezguino is on the table.
Everybody likes tezguino.
Tezguino makes you drunk.
We make tezguino out of corn.

What does tezguino mean?
Loud, motor oil, tortillas, choices, wine

You shall know a word by the company keeps.
(Firth, 1957)

Term-Context Matrix

C1: A bottle of _______ is on the table.
C2: Everybody likes _______.
C3: ______ makes you drunk.
C4: We make ______ out of corn.

	C1	C2	C3	C4
tezguino				
loud				
motor oil				
tortillas				
choices				
wine				

What is a “Context”?

Can be anything you want!

- Entire contents of the sentence
- One word before and after
- Words in the same sentence
- Document it appears in
- Many other variations…

A bottle of tezguino is on the table.
Tezguino makes you drunk.
…
I had a fancy bottle of wine and got drunk last night!
The terrible wine is on the table.

What is a “Context”?

Can be anything you want!

- Entire contents of the sentence
  - Unlikely to occur again!
- One word before and after
- Words in the same sentence
- Document ID it appears in
- Many other variations…

A bottle of tezguino is on the table.
Tezguino makes you drunk.
…
I had a fancy bottle of wine and got drunk last night!
The terrible wine is on the table.

	C1	C2	C3	C4
tezguino				
wine				

What is a “Context”?

Can be anything you want!

- Entire contents of the sentence
- One word before and after
  - Or n-words
- Words in the same sentence
- Document it appears in
- Many other variations…

A bottle of tezguino is on the table.
Tezguino makes you drunk.
…
I had a fancy bottle of wine and got drunk last night!
The terrible wine is on the table.

	bottle-of	is-on	makes-you	and-got	the-terrible
tezguino					
wine					

What is a “Context”?

Can be anything you want!

- Entire contents of the sentence
- One word before and after
- Words in the same sentence
  - Filter: nouns and verbs?
  - Bag of words in a window
- Document it appears in
- Many other variations…

A bottle of tezguino is on the table.
Tezguino makes you drunk.
…
I had a fancy bottle of wine and got drunk last night!
The terrible wine is on the table.

	bottle	table	you	drunk	fancy	night	terrible
tezguino							
wine							

What is a “Context”?

Can be anything you want!

- Entire contents of the sentence
- One word before and after
- Words in the same sentence
- Document it appears in
    - Term-document matrix!
    - Latent Semantic Analysis
- Many other variations…

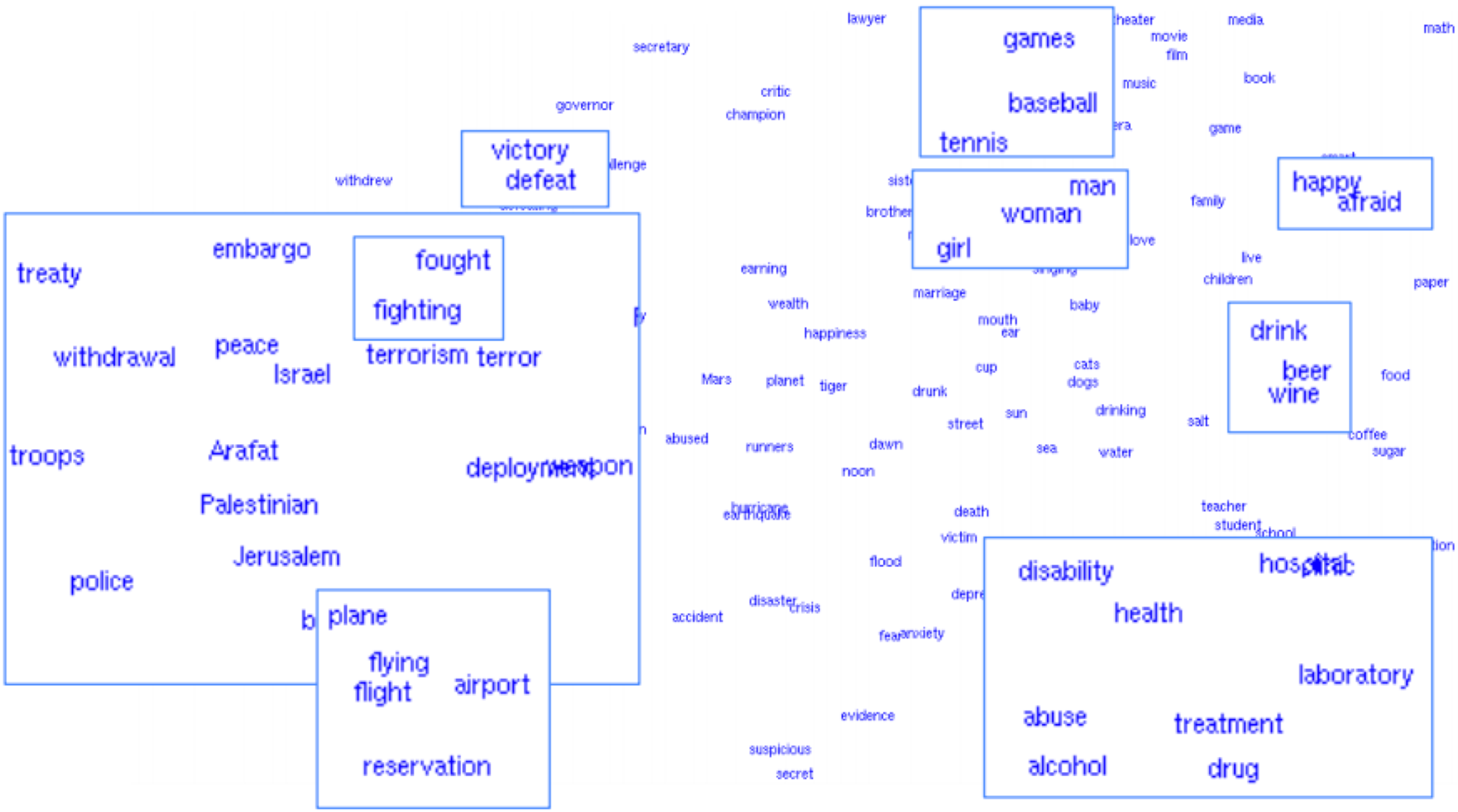
A bottle of tezguino is on the table.
Tezguino makes you drunk.
…
I had a fancy bottle of wine and got drunk last night!
The terrible wine is on the table.

	D1	D2	D3	D4
tezguino				
table				
bottle				
drunk				
wine				

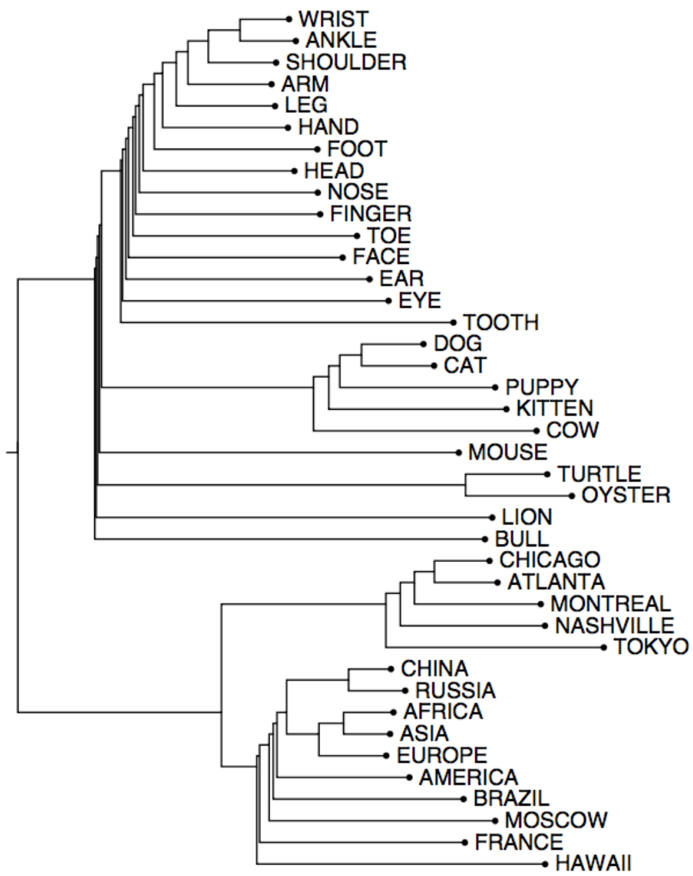
What are word embeddings?

Back to SVD?

Example Word Projection

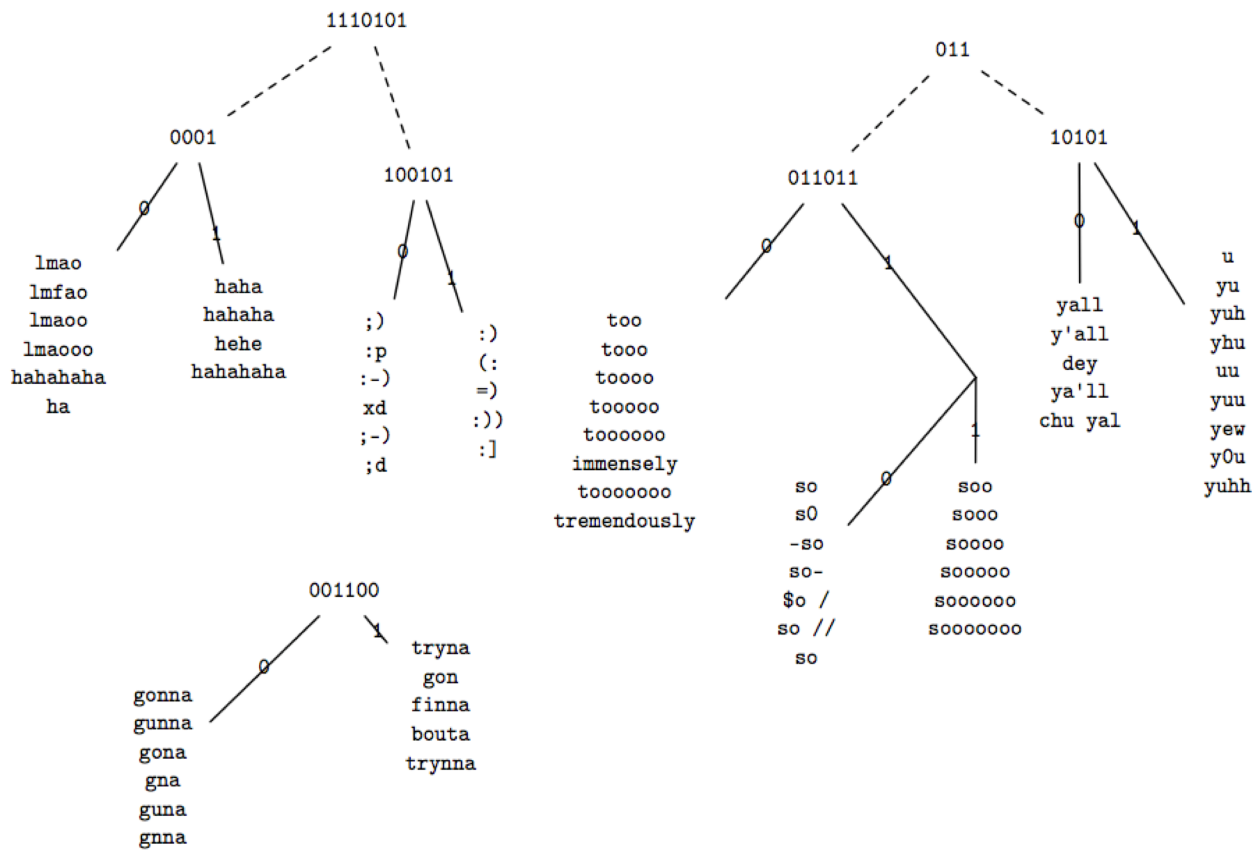


Clustering?



WRIST
ANKLE
SHOULDER
ARM
LEG
HAND
FOOT
HEAD
NOSE
FINGER
TOE
FACE
EAR
EYE
TOOTH
DOG
CAT
PUPPY
KITTEN
COW
MOUSE
TURTLE
OYSTER
LION
BULL
CHICAGO
ATLANTA
MONTREAL
NASHVILLE
TOKYO
CHINA
RUSSIA
AFRICA
ASIA
EUROPE
AMERICA
BRAZIL
MOSCOW
FRANCE
HAWAII

Clustering for Twitter



http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

Problem with SVD (& Clustering)

Computational Complexity

- SVD: $O(mn^2)$
- Clustering: $O(knm)$ per iteration, or $O(n^3)$
- But, n can be 100,000!

"One shot"

- Difficult to add new documents or words
- Cannot work with streaming data

Predict surrounding words

A bottle of tezguino is on the table.

u

v

Similar Meaning = Close

Target Word	BoW5	BoW2
	nightwing	superman
	aquaman	superboy
batman	catwoman	aquaman
	superman	catwoman
	manhunter	batgirl
	dumbledore	evernight
	hallows	sunnydale
hogwarts	half-blood	garderobe
	malfoy	blandings
	snape	collinwood
	nondeterministic	non-deterministic
	non-deterministic	finite-state
turing	computability	nondeterministic
	deterministic	buchi
	finite-state	primality

Target Word	BoW5	BoW2
	gainesville	fla
	fla	alabama
florida	jacksonville	gainesville
	tampa	tallahassee
	lauderdale	texas
	aspect-oriented	aspect-oriented
	smalltalk	event-driven
object-oriented	event-driven	objective-c
	prolog	dataflow
	domain-specific	4gl
	singing	singing
	dance	dance
dancing	dances	dances
	dancers	breakdancing
	tap-dancing	clowning

Similar Meaning = Close

The
Sicilian
gelato
was
extremely
rich.

The
Italian
ice-cream
was
very
velvety.

https://siddhant7.github.io/Vector-Representation-of-Words/

Similar Meaning = Close

Target Word	BoW5	BoW2	Target Word	BoW5	BoW2
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	object-oriented	aspect-oriented smalltalk event-driven prolog domain-specific	aspect-oriented event-driven objective-c dataflow 4gl
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality	dancing	singing dance dances dancers tap-dancing	singing dance dances breakdancing clowning

Similar Meaning = Close

The
Sicilian
gelato
was
extremely
rich.

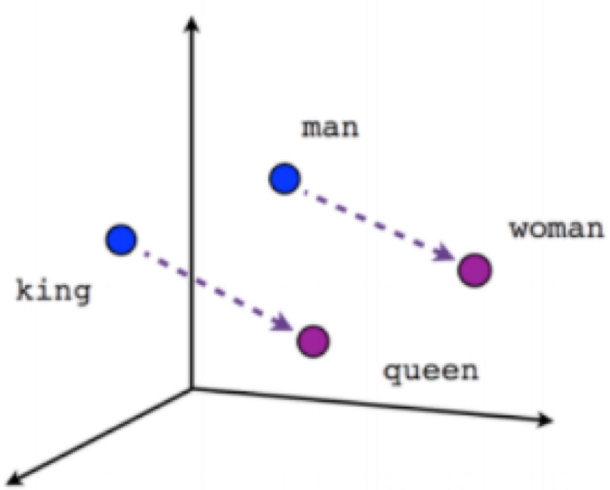
The
Italian
ice-cream
was
very
velvety.

https://siddhant7.github.io/Vector-Representation-of-Words/

Vectors “know” Gender

male : female :: King : queen

King - male + female queen

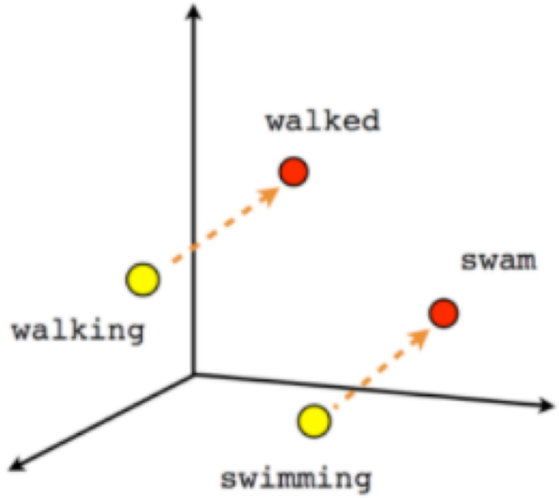


https://siddhant7.github.io/Vector-Representation-of-Words/

They “know” Tenses!

walking : walked :: swimming : swam

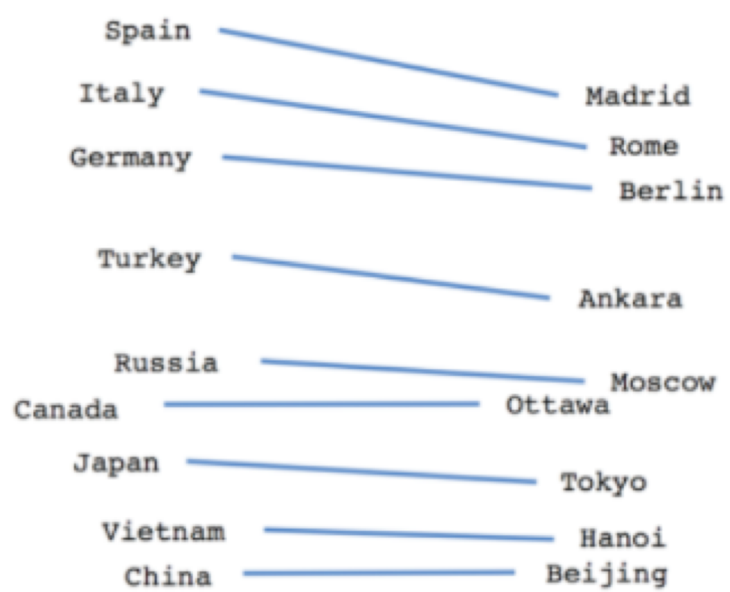
swimming – walking + walked swam



https://siddhant7.github.io/Vector-Representation-of-Words/

They “know” Facts!

Capital – Country + Spain   Madrid



https://siddhant7.github.io/Vector-Representation-of-Words/

Word embeddings

Variations

- Skip-gram: predict context from word
- CBOW: predict word from context bag of words
- Dependencies: a better description of context

Uses

- Similarity
- Grammar
- Analogies
- Odd one out

Demo:
https://rare-technologies.com/word2vec-tutorial/#bonus_app

Back to document vectors?

Average

Max-pooling

In-Class Activity 2