# Distributional Lexical Semantics

Katharina Kann — CSCI/LING5832

# Minimally supervised word sense disambiguation

# Minimally-supervised WSD

- Famous bootstrapping algorithm: invented by Yarowsky (1995)
- Accuracy > 96%
- Uses two powerful heuristics:
    - **One sense per collocation**: nearby words provide clues to the sense of the target word, conditional on distance, order, syntactic relationship.

# Minimally-supervised WSD

- Famous bootstrapping algorithm: invented by Yarowsky (1995)
- Accuracy > 96%
- Uses two powerful heuristics:
  - **One sense per collocation**: nearby words provide clues to the sense of the target word, conditional on distance, order, syntactic relationship.
  - **One sense per discourse**: the sense of a target words is consistent within a given document.

# Goal: disambiguate *plant* in a corpus

**Step 1**: In a large corpus, identify all examples of the given polysemous word, storing their contexts as lines in an initially untagged training set.

| Sense | Training Examples (Keyword in Context) |
|---|---|
| ? | ... company said the *plant* is still operating |
| ? | Although thousands of *plant* and animal species |
| ? | ... zonal distribution of *plant* life . ... |
| ? | ... to strain microscopic *plant* life from the ... |
| ? | vinyl chloride monomer *plant* , which is ... |
| ? | and Golgi apparatus of *plant* and animal cells |
| ? | ... computer disk drive *plant* located in ... |
| ? | ... divide life into *plant* and animal kingdom |
| ? | ... close-up studies of *plant* life and natural |
| ? | ... Nissan car and truck *plant* in Japan is ... |
| ? | ... keep a manufacturing *plant* profitable without |
| ? | ... molecules found in *plant* and animal tissue |
| ? | ... union responses to *plant* closures . ... |
| ? | ... animal rather than *plant* tissues can be |
| ? | ... many dangers to *plant* and animal life |
| ? | company manufacturing *plant* is in Orlando ... |
| ? | ... growth of aquatic *plant* life in water ... |
| ? | automated manufacturing *plant* in Fremont , |
| ? | ... Animal and *plant* life are delicately |
| ? | discovered at a St. Louis *plant* manufacturing |
| ? | computer manufacturing *plant* and adjacent ... |
| ? | ... the proliferation of *plant* and animal life |
| ? | ... ... |

# Goal: disambiguate *plant* in a corpus

**Step 2**: For each possible sense of the word, identify a relatively small number of training examples representative of that sense, for example by hand tagging a subset of the training sentences.

A ~ *life*

B ~ *manufacturing*

| Sense | Training Examples (Keyword in Context) |
|---|---|
| A | used to strain microscopic ***plant*** **life** from the ... |
| A | ... zonal distribution of ***plant*** **life** . ... |
| A | close-up studies of ***plant*** **life** and natural ... |
| A | too rapid growth of aquatic ***plant*** **life** in water ... |
| A | ... the proliferation of ***plant*** and animal **life** ... |
| A | establishment phase of the ***plant*** virus **life** cycle ... |
| A | ... that divide **life** into ***plant*** and animal kingdom |
| A | ... many dangers to ***plant*** and animal **life** ... |
| A | mammals . Animal and ***plant*** **life** are delicately |
| A | beds too salty to support ***plant*** **life** . River ... |
| A | heavy seas, damage , and ***plant*** **life** growing on ... |
| A | ... ... |
| ? | ... vinyl chloride monomer ***plant*** , which is ... |
| ? | ... molecules found in ***plant*** and animal tissue |
| ? | ... Nissan car and truck ***plant*** in Japan is ... |
| ? | ... and Golgi apparatus of ***plant*** and animal cells ... |
| ? | ... union responses to ***plant*** closures . ... |
| ? | ... ... |
| ? | ... ... |
| ? | ... cell types found in the ***plant*** kingdom are ... |
| ? | ... company said the ***plant*** is still operating ... |
| ? | ... Although thousands of ***plant*** and animal species |
| ? | ... animal rather than ***plant*** tissues can be ... |
| ? | ... computer disk drive ***plant*** located in ... |
| B | ... ... |
| B | automated **manufacturing** ***plant*** in Fremont ... |
| B | ... vast **manufacturing** ***plant*** and distribution ... |
| B | chemical **manufacturing** ***plant*** , producing viscose |
| B | ... keep a **manufacturing** ***plant*** profitable without |
| B | computer **manufacturing** ***plant*** and adjacent ... |
| B | discovered at a St. Louis ***plant*** **manufacturing** |
| B | ... copper **manufacturing** ***plant*** found that they |
| B | copper wire **manufacturing** ***plant*** , for example ... |
| B | 's cement **manufacturing** ***plant*** in Alpena ... |
| B | polystyrene **manufacturing** ***plant*** at its Dow ... |
| B | company **manufacturing** ***plant*** is in Orlando ... |

# Goal: disambiguate *plant* in a corpus

**Step 3a**: Train a supervised classification algorithm on the SENSE-A/SENSE-B seed sets.

# Goal: disambiguate *plant* in a corpus

**Step 3a**: Train a supervised classification algorithm on the SENSE-A/SENSE-B seed sets.

**Step 3b**: Apply the resulting classifier to the entire sample set. Take samples that are tagged as SENSE-A or SENSE-B with probability above a certain threshold, and add those examples to the growing seed sets.

# Goal: disambiguate *plant* in a corpus

**Step 3a**: Train a supervised classification algorithm on the SENSE-A/SENSE-B seed sets.

**Step 3b**: Apply the resulting classifier to the entire sample set. Take samples that are tagged as SENSE-A or SENSE-B with probability above a certain threshold, and add those examples to the growing seed sets.

**Step 3c**: Optionally, the one-sense-per-discourse constraint can be used both to filter and augment this addition.

# Goal: disambiguate *plant* in a corpus

**Step 3a**: Train a supervised classification algorithm on the SENSE-A/SENSE-B seed sets.

**Step 3b**: Apply the resulting classifier to the entire sample set. Take samples that are tagged as SENSE-A or SENSE-B with probability above a certain threshold, and add those examples to the growing seed sets.

**Step 3c**: Optionally, the one-sense-per-discourse constraint can be used both to filter and augment this addition.

**Step 3d**: Repeat Step 3 iteratively

**Goal: disambiguate** ***plant*** **in a corpus**

**Step 4:** Stop. When the training parameters are held constant, the algorithm will converge on a stable set of remaining instances.

**Goal: disambiguate** ***plant*** **in a corpus**

**Step 5:** The resulting dataset can now be used to train a classifier and to annotate new data with sense tags and probabilities.

# How to evaluate?

# How to evaluate?

Intrinsic evaluation:

- Sense accuracy
- On a held-out part of the corpus

# How to evaluate?

Intrinsic evaluation:

- Sense accuracy
- On a held-out part of the corpus

Extrinsic evaluation:

- Use WSD in downstream tasks and evaluate final performance

# How to evaluate?


Intrinsic evaluation:

- Sense accuracy
- On a held-out part of the corpus

Extrinsic evaluation:

- Use WSD in downstream tasks and evaluate final performance

# Hypernym detection

# Task Definition

- Given: a corpus
- Task: Find pairs of terms that are in a hyponym-hypernym relationship

- Examples:
  - Dog is a **<?>** of animal
  - Food is a **<?>** of pasta

# Hearst Patterns (Hearst, 1992)

**Idea**: Pairs of words that are in hyponym-hypernym relationships tend to occur in certain lexico-syntactic patterns.

# Hearst Patterns (Hearst, 1992)

**Idea**: Pairs of words that are in hyponym-hypernym relationships tend to occur in certain lexico-syntactic patterns.

The bow lute, such as the Bambara ndang,
is plucked and has an individual
curved neck for each string.

# Hearst Patterns (Hearst, 1992)

**Idea**: Pairs of words that are in hyponym-hypernym relationships tend to occur in certain lexico-syntactic patterns.

The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string.

(1a) $NP_0$ *such as* $\{NP_1, NP_2 \ldots, (and \mid or)\}$ $NP_n$

are such that they imply

(1b) *for all* $NP_i$, $1 \leq i \leq n$, *hyponym*($NP_i$, $NP_0$)

Thus from sentence (S1) we conclude

*hyponym*("Bambara ndang", "bow lute").

# Hearst Patterns (Hearst, 1992)

(2) *such NP as {NP ,}* {(or | and)} NP*

... works by such authors as Herrick, Goldsmith, and Shakespeare.

==> *hyponym("author", "Herrick"),*
*hyponym("author", "Goldsmith"),*
*hyponym("author", "Shakespeare")*

# Hearst Patterns (Hearst, 1992)

(3) *NP* {, *NP*}* {,} *or other NP*

Bruises, wounds, broken bones or other injuries ...

$\Longrightarrow$ *hyponym*("*bruise*", "*injury*"),
*hyponym*("*wound*", "*injury*"),
*hyponym*("*broken bone*", "*injury*")

# Hearst Patterns (Hearst, 1992)

(4) *NP* {, *NP*}* {,} *and other NP*

... temples, treasuries, and other important civic buildings.

$\Longrightarrow$ *hyponym*("*temple*", "*civic building*"), *hyponym*("*treasury*", "*civic building*")

# Hearst Patterns (Hearst, 1992)

(5) *NP* {,} *including* {*NP* ,}* {*or* | *and*} *NP*

`All common-law countries, including Canada and England ...`

$\Longrightarrow$ *hyponym*(*"Canada"*, *"common-law country"*), *hyponym*(*"England"*, *"common-law country"*)

# Hearst Patterns (Hearst, 1992)

(6) *NP* {,} *especially* {*NP* ,}* {*or* | *and*} *NP*

... most European countries, especially France, England, and Spain.

$\Longrightarrow$ *hyponym*(*"France"*, *"European country"*),
*hyponym*(*"England"*, *"European country"*),
*hyponym*(*"Spain"*, *"European country"*)

# Overview: Hearst Patterns

- NP such as {NP}* {and|or} NP
- such NP as {NP ,}* {or|and} NP
- NP {, NP}* {,} or other NP
- NP {, NP}* {,} and other NP
- NP {,} including {NP, }* {or|and} NP
- NP {,} especially {NP ,}* {or|and} NP

# Distributional Lexical Semantics

# The Big Question

What kind of thing is the meaning of a word?

# The Big Question

~~What kind of thing is the meaning of a word?~~

What can you do with a word if you know its meaning?

# Distributional Lexical Semantics

**Semantics**:

The study of how meaning is expressed in language.

**Lexical**:

Involving words (rather than sentences, etc.).

**Distributional**:

Involving *distributions*.

**Distributional (Lexical) Semantics**:

Learning about the meanings of words by how those words are distributed in texts.

# Distributional Lexical Semantics

**John Rupert Firth, (1957, 'A synopsis of linguistic theory')**

"You shall know a word by the company it keeps."

# Distributional Lexical Semantics

## John Rupert Firth, (1957, 'A synopsis of linguistic theory')

"You shall know a word by the company it keeps."

## Zellig Harris (1954, 'Distributional structure')

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

# Distributional Lexical Semantics

### John Rupert Firth, (1957, 'A synopsis of linguistic theory')

"You shall know a word by the company it keeps."

### Zellig Harris (1954, 'Distributional structure')

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

### Turney and Pantel (2010, 'From frequency to meaning')

"If units of text have similar vectors in a text frequency matrix, then they tend to have similar meanings."

# Example

“<word> are performed around the world.”

# Example

“<word> are performed around the world.”

“Several shorter <word> on Broadway and in the West End have been presented in one act in recent decades.”

# Example

“<word> are performed around the world.”

“Several shorter <word> on Broadway and in the West End have been presented in one act in recent decades.”

“Some of the most famous <word> through the decades that followed include West Side Story (1957), [...] and Hamilton (2015).”

# Example

“<word> are performed around the world.”

“Several shorter <word> on Broadway and in the West End have been presented in one act in recent decades.”

“Some of the most famous <word> through the decades that followed include West Side Story (1957), [...] and Hamilton (2015).”

What is <word>?

# Another Question

How do we represent words inside an NLP system?

Some options:

- String of letters
- Indices
- One-hot vectors
    - No information about words

# Another Question

How do we represent words inside an NLP system?

Some options:

- ...or maybe some other type of vectors
  - Maybe with more information?

How should our word vectors/word embeddings/vector representations of words be?

- Words with similar properties should have similar embeddings
- Words with dissimilar properties should have dissimilar embeddings

# Where is this useful?

# Question Answering

"**fast**" is similar to "**rapid**"

"**tall**" is similar to "**height**"

Question answering:

*Q: "How **tall** is Mt. Everest?"*

*Candidate A: "The official **height** of Mount Everest is 29029 feet"*

# Plagiarism Detection

**MAINFRAMES**

Mainframes are primarily referred to large computers with rapid, advanced processing capabilities that can execute and perform tasks equivalent to many Personal Computers (PCs) machines networked together. It is characterized with high quantity Random Access Memory (RAM), very large secondary storage devices, and high-speed processors to cater for the needs of the computers under its service.

Consisting of advanced components, mainframes have the capability of running multiple large applications required by many and most enterprises and organizations. This is one of its advantages. Mainframes are also suitable to cater for those applications (programs) or files that are of very high

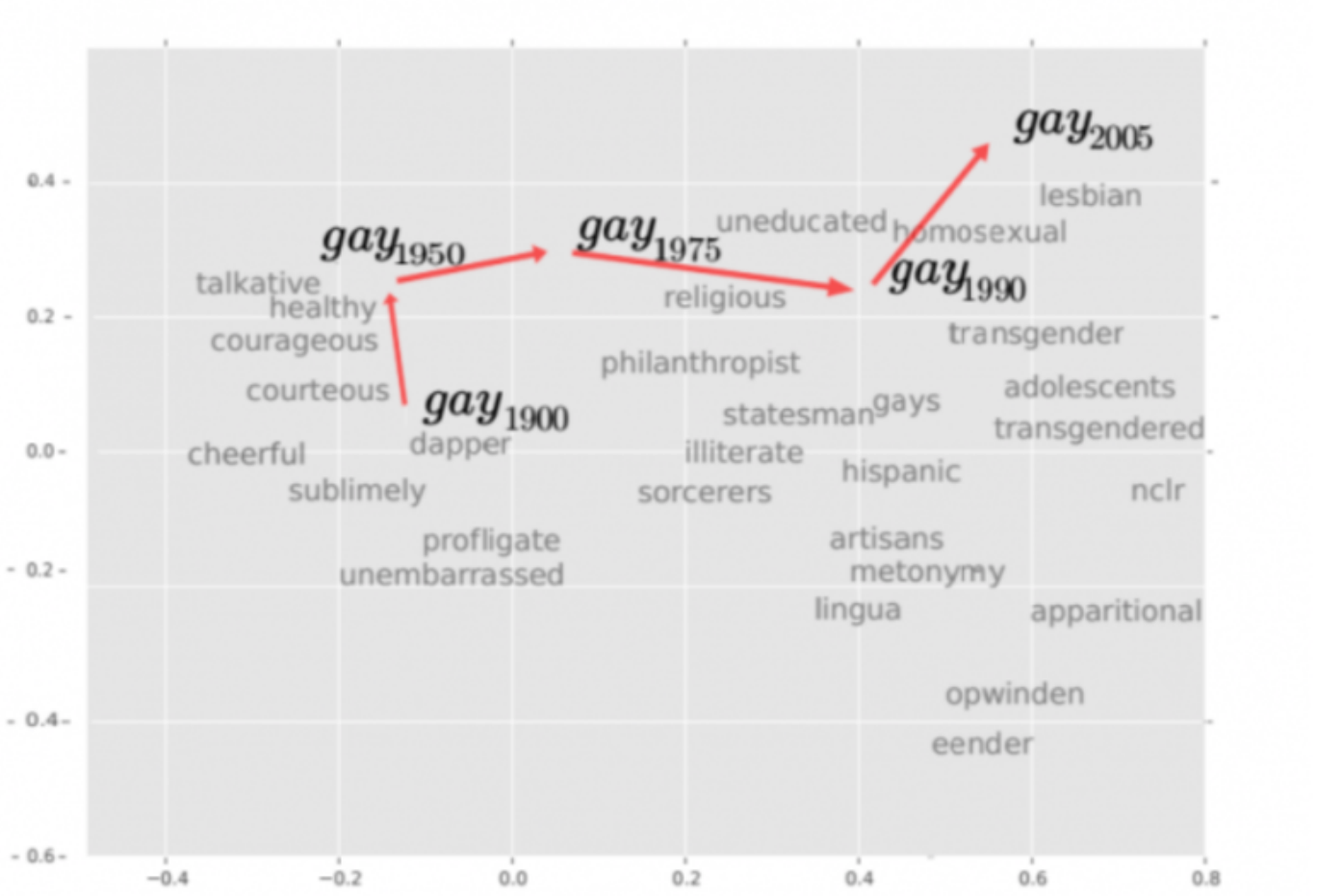
**MAINFRAMES**

Mainframes usually are referred those computers with fast, advanced processing capabilities that could perform by itself tasks that may require a lot of Personal Computers (PC) Machines. Usually mainframes would have lots of RAMs, very large secondary storage devices, and very fast processors to cater for the needs of those computers under its service.

Due to the advanced components mainframes have, these computers have the capability of running multiple large applications required by most enterprises, which is one of its advantage. Mainframes are also suitable to cater for those applications or files that are of very large demand

# What else can we do with these?

Kulkarni, Al-Rfou, Perozzi, & Skiena (2015)



# The basics: Distributional semantics

**Main idea:**

Combine the distributional hypothesis with the idea of vectors to represent words!

# Input: A Word–Document Matrix

| | d1 | d2 | d3 | d4 | d5 | d6 | d7 | d8 | d9 | d10 |
|---|---|---|---|---|---|---|---|---|---|---|
| against | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 2 | 3 | 0 |
| age | 0 | 0 | 0 | 1 | 0 | 3 | 1 | 0 | 4 | 0 |
| agent | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ages | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| ago | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 | 0 |
| agree | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ahead | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ain’t | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| air | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| aka | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

# Or...: A Word–Word Co-Occurrence Matrix

| | against | age | agent | ages | ago | agree | ahead | ain't | air | aka | al |
|---|---|---|---|---|---|---|---|---|---|---|---|
| against | 2003 | 90 | 39 | 20 | 88 | 57 | 33 | 15 | 58 | 22 | 24 |
| age | 90 | 1492 | 14 | 39 | 71 | 38 | 12 | 4 | 18 | 4 | 39 |
| agent | 39 | 14 | 507 | 2 | 21 | 5 | 10 | 3 | 9 | 8 | 25 |
| ages | 20 | 39 | 2 | 290 | 32 | 5 | 4 | 3 | 6 | 1 | 6 |
| ago | 88 | 71 | 21 | 32 | 1164 | 37 | 25 | 11 | 34 | 11 | 38 |
| agree | 57 | 38 | 5 | 5 | 37 | 627 | 12 | 2 | 16 | 19 | 14 |
| ahead | 33 | 12 | 10 | 4 | 25 | 12 | 429 | 4 | 12 | 10 | 7 |
| ain't | 15 | 4 | 3 | 3 | 11 | 2 | 4 | 166 | 0 | 3 | 3 |
| air | 58 | 18 | 9 | 6 | 34 | 16 | 12 | 0 | 746 | 5 | 11 |
| aka | 22 | 4 | 8 | 1 | 11 | 19 | 10 | 3 | 5 | 261 | 9 |
| al | 24 | 39 | 25 | 6 | 38 | 14 | 7 | 3 | 11 | 9 | 861 |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | **Betty** | **has** | **brown** | **hair** | **and** | **John** | **black** |
|---|---|---|---|---|---|---|---|
| **Betty** | | | | | | | |
| **has** | | | | | | | |
| **brown** | | | | | | | |
| **hair** | | | | | | | |
| **and** | | | | | | | |
| **John** | | | | | | | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | | | | | | | |
| **brown** | | | | | | | |
| **hair** | | | | | | | |
| **and** | | | | | | | |
| **John** | | | | | | | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 1 | 1 | | | | |
| **brown** | | | | | | | |
| **hair** | | | | | | | |
| **and** | | | | | | | |
| **John** | | | | | | | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 1 | 1 | | | | |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | | | | | |
| **and** | | | | | | | |
| **John** | | | | | | | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 1 | 1 | | | | |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | 1 | 1 | 1 | | |
| **and** | | | | | | | |
| **John** | | | | | | | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 1 | 1 | | | | |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | 1 | 1 | 1 | | |
| **and** | | | | 1 | 1 | 1 | |
| **John** | | | | | | | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 1 | 1 | | | | |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | 1 | 1 | 1 | | |
| **and** | | | | 1 | 1 | 1 | |
| **John** | | 1 | | | 1 | 1 | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 2 | 1 | | | 1 | 1 |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | 1 | 1 | 1 | | |
| **and** | | | | 1 | 1 | 1 | |
| **John** | | 1 | | | 1 | 1 | |
| **black** | | | | | | | |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 2 | 1 | | | 1 | 1 |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | 1 | 1 | 1 | | |
| **and** | | | | 1 | 1 | 1 | |
| **John** | | 1 | | | 1 | 1 | |
| **black** | | 1 | | 1 | | | 1 |

# Or...: A Word–Word Co-Occurrence Matrix

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 2 | 1 | | | 1 | 1 |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | 1 | 2 | 1 | | 1 |
| **and** | | | | 1 | 1 | 1 | |
| **John** | | 1 | | | 1 | 1 | |
| **black** | | 1 | | 1 | | | 1 |

# The Data

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| <s> | <s> | <unk> | communications | pittsburgh | acquired | <unk> | & | co. |
| investment | managemen | inc. | a | pittsburgh | firm | that | runs | a |
| <s> | mr. | allen | 's | pittsburgh | firm | advanced | investment | management |
| look | stupid | <unk> | former | pittsburgh | <unk> | second | <unk> | <unk> |
| through | the | university | of | pittsburgh | law | school | <s> | <s> |
| with | the | university | of | pittsburgh | <s> | <s> | <s> | <s> |
| <unk> | he | heads | the | pittsburgh | branch | of | the | committee |
| at | the | university | of | pittsburgh | earn | up | to | $ |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| for | society | corp. | a | cleveland | bank | said | demand | for |
| as | washington | <unk> | r.i. | cleveland | <unk> | n.c. | minneapolis | and |
| <s> | <s> | <unk> | a | cleveland | merchant | bank | owns | about |
| new | stadiums | ranging | from | cleveland | to | san | antonio | and |
| <s> | the | philadelphia | and | cleveland | districts | for | example | reported |
| mcdonald | & | co. | in | cleveland | said | <unk> | 's | unanticipated |
| <unk> | tumor | at | the | cleveland | clinic | in | N | <s> |
| at | mcdonald | & | co. | cleveland | <s> | <s> | <s> | <s> |

# Co-occurrence

How do we define “co-occurrence”? (What is the *context*?)

- Documents
- Sentences
- Tweets
- Windows
  - Size 1
  - Size 5
  - …
- …

**Tip:** Smaller contexts lead to more focus on syntax, larger contexts lead to more focus on semantics!

# Measuring similarity

# Measuring Similarity

Betty has brown hair and John has black hair.

| | Betty | has | brown | hair | and | John | black |
|---|---|---|---|---|---|---|---|
| **Betty** | 1 | 1 | | | | | |
| **has** | 1 | 2 | 1 | | | 1 | 1 |
| **brown** | | 1 | 1 | 1 | | | |
| **hair** | | | 1 | 2 | 1 | | 1 |
| **and** | | | | 1 | 1 | 1 | |
| **John** | | 1 | | | 1 | 1 | |
| **black** | | 1 | | 1 | | | 1 |

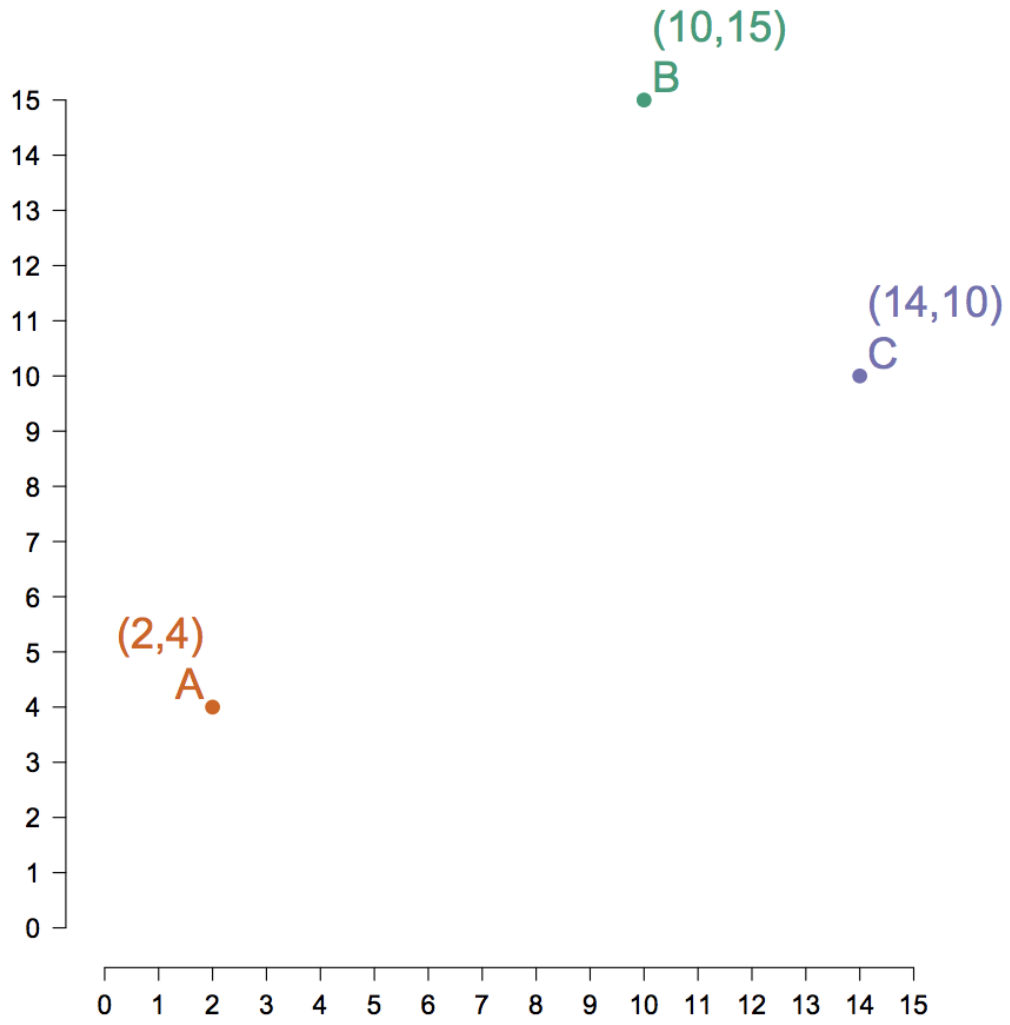
How similar are: *brown* and *black*? *brown* and *and*?

# Measuring Similarity

| | $d_x$ | $d_y$ |
|---|---|---|
| *A* | 2 | 4 |
| *B* | 10 | 15 |
| *C* | 14 | 10 |

# Measuring Similarity

| | $d_x$ | $d_y$ |
|---|---|---|
| *A* | 2 | 4 |
| *B* | 10 | 15 |
| *C* | 14 | 10 |



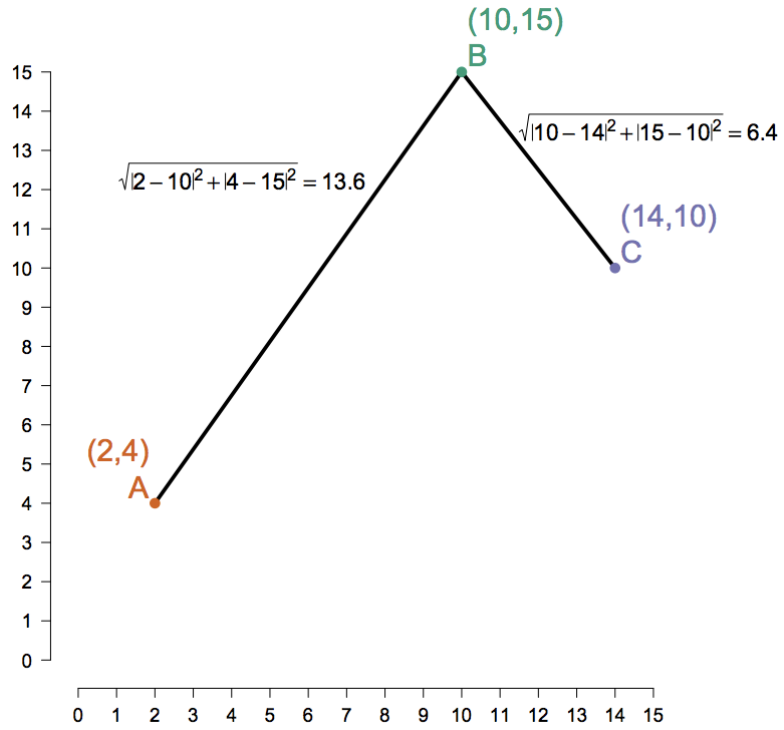
# Euclidean Distance

## Definition

Between vectors $u$ and $v$ of dimension $n$:

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

| | $d_x$ | $d_y$ |
|---|---|---|
| $A$ | 2 | 4 |
| $B$ | 10 | 15 |
| $C$ | 14 | 10 |



# Length (L2) Normalization

**Definition**

Given a vector $u$ of dimension $n$, the normalization of $u$ is a vector $\hat{u}$ of dimension $n$ obtained by dividing each element of $u$ by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.

| | $d_x$ | $d_y$ |
|---|---|---|
| $A$ | 2 | 4 |
| $B$ | 10 | 15 |
| $C$ | 14 | 10 |

L2 norm the rows

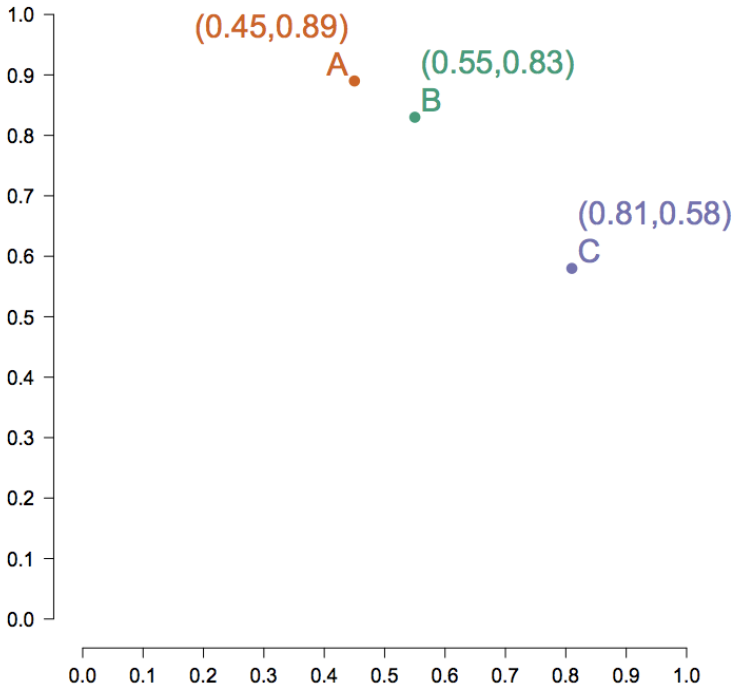
$\Rightarrow$

| | $d_x$ | $d_y$ |
|---|---|---|
| $A$ | 0.45 | 0.89 |
| $B$ | 0.55 | 0.83 |
| $C$ | 0.81 | 0.58 |

# Length (L2) Normalization

## Definition

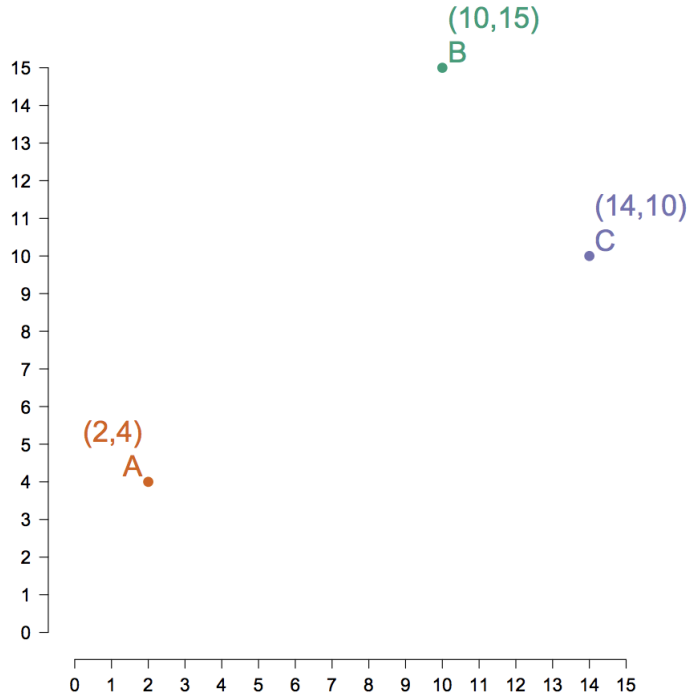
Given a vector $u$ of dimension $n$, the normalization of $u$ is a vector $\hat{u}$ of dimension $n$ obtained by dividing each element of $u$ by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.

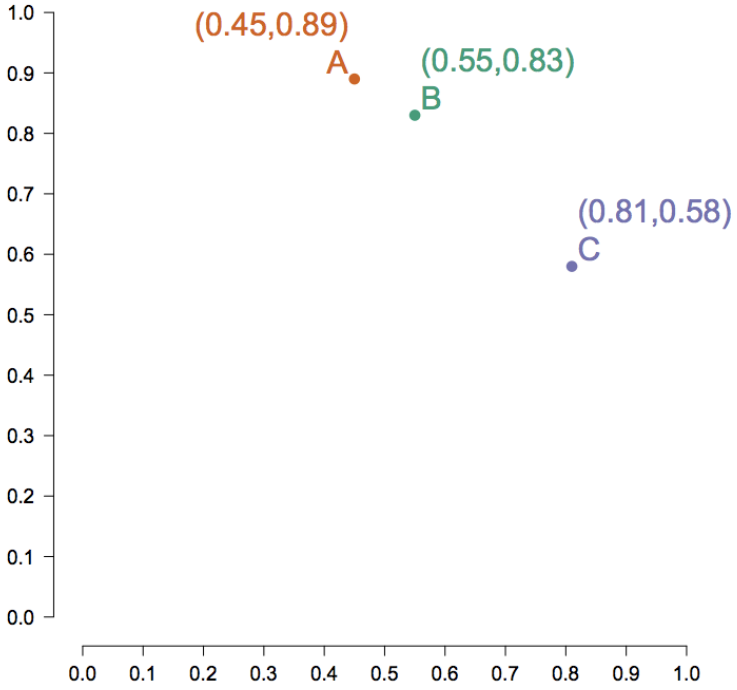


# Length (L2) Normalization

## Definition

Given a vector $u$ of dimension $n$, the normalization of $u$ is a vector $\hat{u}$ of dimension $n$ obtained by dividing each element of $u$ by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.



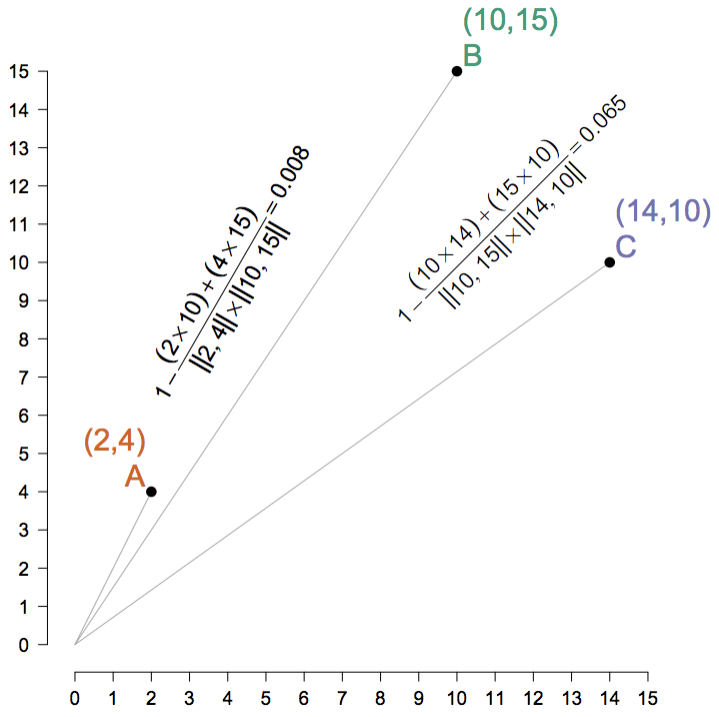



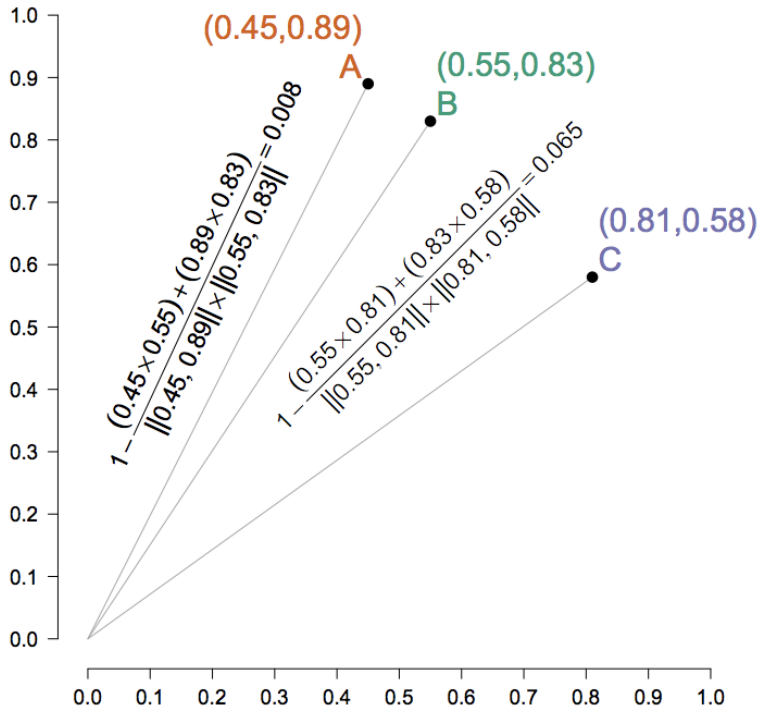
# Cosine Distance

## Definition (Cosine distance)

Between vectors $u$ and $v$ of dimension $n$:

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



# Picking a Distance Metric

| | $d_x$ | $d_y$ |
|---|---|---|
| $A$ | 2 | 4 |
| $B$ | 10 | 15 |
| $C$ | 14 | 10 |

$$\|A\| = 4.47$$
$$\|B\| = 18.03$$
$$\|C\| = 17.20$$

| | $A$ and $B$ closer than $B$ and $C$? |
|---|---|
| Euclidean distance | No |
| Cosine distance | Yes |

# In-Class Exercise

***a cat chases a black bird a white bird and a yellow bird***

# In-Class Exercise

***a cat chases a black bird a white bird and a yellow bird***

1. Construct the co-occurrence matrix for window size 1.
2. Compute the Euclidean distance between the *black* and *white,* and *black* and *cat*.
3. Compute the cosine distance between the *black* and *white,* and *black* and *cat*.

# Distributed vs. distributional representations

Distributed: A concept is represented as continuous activation levels in a number of elements

[Contrast: local]

Distributional: Meaning is represented by contexts of use

[Contrast: denotational]

# Many Design Choices

tokenization
annotation
tagging
parsing
feature selection
⋮
cluster texts by date/author/discourse context/...
⇓ ⇙

| Matrix type | Reweighting | Dimensionality reduction | Vector comparison |
|---|---|---|---|
| word × document | probabilities | LSA | Euclidean |
| word × word | length normalization | PLSA | Cosine |
| word × search proximity | TF-IDF | LDA | Dice |
| adj. × modified noun | PMI | PCA | Jaccard |
| word × dependency rel. | Positive PMI | IS | KL |
| ⋮ | ⋮ | ⋮ | ⋮ |

# Word2vec & co.

# word2vec, fastText and GloVe

- Algorithms to create dense word embeddings
- Fast
- Efficient to train
- Code and pre-trained embeddings are available online

# word2vec, fastText and GloVe

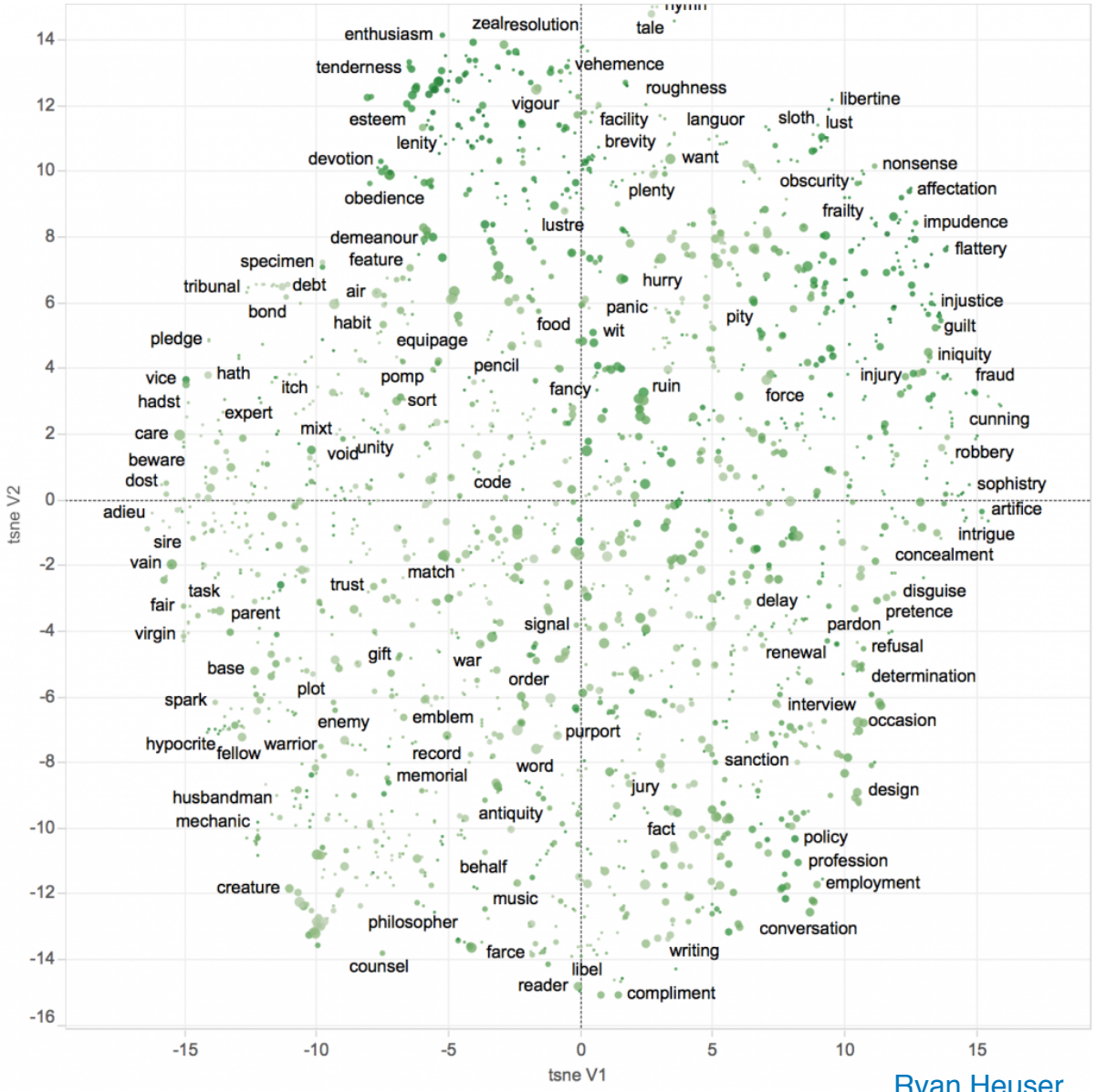
*Highlights*:

- Use minibatch stochastic gradient descent (SGD) to efficiently *approximate* a typical distributional semantics pipeline (see Levy & Goldberg 2014).
- Scale it up *massively*:
  - Largest commonly used public GloVe package is a model with 13 billion parameters trained on 840 billion words of text!

# Results: Neighbors (word2vec, Mikolov+ ‘13)

| target: | Redmond | Havel | ninjutsu | graffiti | capitulate |
|---|---|---|---|---|---|
| | Redmond Wash. | Vaclav Havel | ninja | spray paint | capitulation |
| | Redmond Washington | president Vaclav Havel | martial arts | grafitti | capitulated |
| | Microsoft | Velvet Revolution | swordsmanship | taggers | capitulating |

# The Output




Ryan Heuser

# Evaluation of embedding models

# Evaluation Methods

Intrinsic evaluation:

- Word similarity
- Word analogy

Extrinsic evaluation:

- Use embeddings in down-stream task like question answering or natural language inference

# Word Similarity

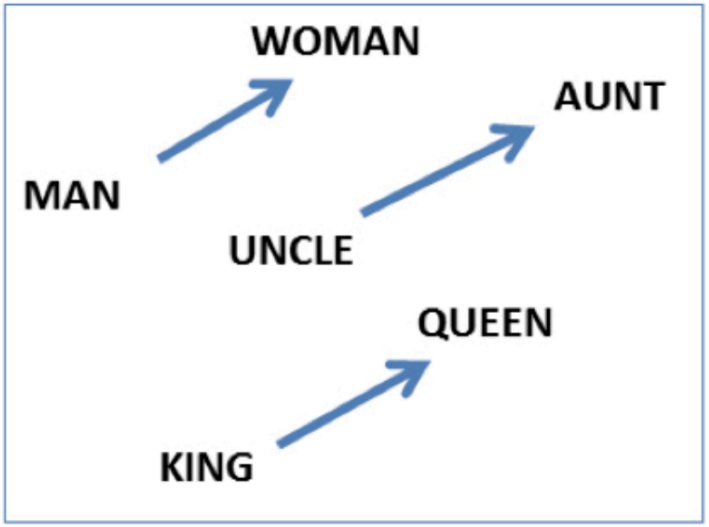
wordsim_similarity_goldstandard.txt

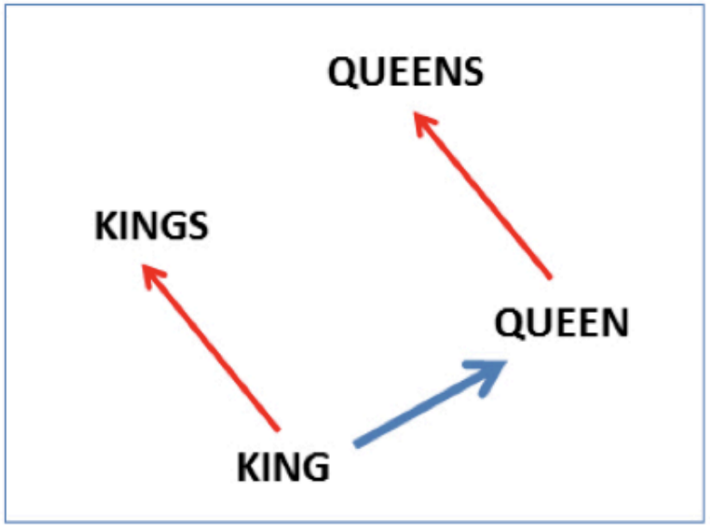
```
mile    kilometer       8.66
skin    eye     6.22
Japanese        American        6.50
century year    7.59
announcement    news    7.56
doctor  personnel       5.00
Harvard Yale    8.13
hospital        infrastructure  4.63
life    death   7.88
travel  activity        5.00
type    kind    8.97
street  place   6.44
street  avenue  8.88
street  block   6.88
cell    phone   7.81
dividend        payment 7.63
calculation     computation     8.44
profit  loss    7.63
dollar  yen     7.78
dollar  buck    9.22
phone   equipment       7.13
liquid  water   7.89
marathon        sprint  7.47
seafood food    8.34
seafood lobster 8.70
lobster food    7.81
lobster wine    5.70
championship    tournament      8.36
man     woman   8.30
```

# Word Analogy (word2vec, Mikolov+ ‘13)

vector(*‘king’*) - vector(*‘man’*) + vector(*‘woman’*) ≈ vector(‘queen’)

vector(*‘Paris’*) - vector(*‘France’*) + vector(*‘Italy’*) ≈ vector(‘Rome’)






**Tip:** If you want to try this, you should exclude all 3 original vectors from the pool of candidates!

# A Few Results

| win | Method | WordSim Similarity | WordSim Relatedness | Bruni et al. MEN | Radinsky et al. M. Turk | Luong et al. Rare Words | Hill et al. SimLex | Google Add / Mul | MSR Add / Mul |
|---|---|---|---|---|---|---|---|---|---|
| 2 | PPMI | .732 | **.699** | .744 | .654 | .457 | .382 | .552 / .677 | .306 / .535 |
| | SVD | .772 | .671 | **.777** | .647 | **.508** | .425 | .554 / .591 | .408 / .468 |
| | SGNS | **.789** | .675 | .773 | **.661** | .449 | **.433** | .676 / **.689** | .617 / **.644** |
| | GloVe | .720 | .605 | .728 | .606 | .389 | .388 | .649 / .666 | .540 / .591 |
| 5 | PPMI | .732 | **.706** | .738 | **.668** | .442 | .360 | .518 / .649 | .277 / .467 |
| | SVD | .764 | .679 | **.776** | .639 | **.499** | **.416** | .532 / .569 | .369 / .424 |
| | SGNS | **.772** | .690 | .772 | .663 | .454 | .403 | .692 / **.714** | .605 / **.645** |
| | GloVe | .745 | .617 | .746 | .631 | .416 | .389 | .700 / .712 | .541 / .599 |
| 10 | PPMI | .735 | **.701** | .741 | .663 | .235 | .336 | .532 / .605 | .249 / .353 |
| | SVD | .766 | .681 | .770 | .628 | **.312** | .419 | .526 / .562 | .356 / .406 |
| | SGNS | **.794** | .700 | **.775** | **.678** | .281 | **.422** | .694 / .710 | .520 / **.557** |
| | GloVe | .746 | .643 | .754 | .616 | .266 | .375 | .702 / **.712** | .463 / .519 |
| 10 | SGNS-LS | .766 | .681 | **.781** | **.689** | **.451** | .414 | .739 / **.758** | .690 / **.729** |
| | GloVe-LS | .678 | .624 | .752 | .639 | .361 | .371 | .732 / .750 | .628 / .685 |

Table 5: Performance of each method across different tasks using 2-fold cross-validation for hyperparameter tuning. Configurations on large-scale (LS) corpora are also presented for comparison.

Levy et al. ‘15: Improving Distributional Similarity with Lessons Learned from Word Embeddings

# Results: Extrinsic Evaluation

| Dataset | Random | GloVe |
|---|---|---|
| SST-2 | 84.2 | 88.4 |
| SST-5 | 48.6 | 53.5 |
| IMDb | 88.4 | 91.1 |
| TREC-6 | 88.9 | 94.9 |
| TREC-50 | 81.9 | 89.2 |
| SNLI | 82.3 | 87.7 |
| SQuAD | 65.4 | 76.0 |

| Model | Squad |
|---|---|
| GloVe Wiki + news | 77.7% |
| fastText Wiki + news | 78.8% |
| GloVe Crawl | 78.9% |
| fastText Crawl | **79.8%** |

McCann+ ‘17; Mikolov+ ‘17

# Problems and Challenges

# How about Antonyms?

- Antonyms appear in very similar contexts.
- Their word embeddings are similar.
- However, meaning can be very different, depending on the task at hand.

# How about Antonyms?

- Antonyms appear in very similar contexts.
- Their word embeddings are similar.
- However, meaning can be very different, depending on the task at hand.

The movie I watched yesterday was really good.

vs.

The movie I watched yesterday was really bad.

# How about Antonyms?

- Antonyms appear in very similar contexts.
- Their word embeddings are similar.
- However, meaning can be very different, depending on the task at hand.

Give me the red block! vs. Give me the green block!

# Pitfalls of Unsupervised Learning

Word similarity:

- Occupations most similar to *she*:
  - *nurse*, *librarian, nanny, stylist, dancer*
- Occupations most similar to *he*:
  - *architect, captain, philosopher, legend, hero*

Source: Bolukbasi et al. ‘16, Quantifying and Reducing Stereotypes in Word Embeddings

# Pitfalls of Unsupervised Learning

Word analogy:

- doctor - father + mother: nurse

Source: Bolukbasi et al. ‘16, Quantifying and Reducing Stereotypes in Word Embeddings

# Pitfalls of Unsupervised Learning

Additionally:

- African American names like have a higher GloVe cosine with unpleasant words.
- European American names (‘Brad’, ‘Greg’, ‘Courtney’) have a higher cosine with pleasant words.

Source: Bolukbasi et al. ‘16, Quantifying and Reducing Stereotypes in Word Embeddings

# Pitfalls of Unsupervised Learning

Impossible to avoid these issues altogether when learning from naturally occurring text.

Mitigating bias will usually require identifying explicitly, and the best method will depend on the task at hand.

Source: Bolukbasi et al. ‘16, Quantifying and Reducing Stereotypes in Word Embeddings

# Wrapping up

- Discussed today:
  - WSD/hypernym detection
  - Co-occurrence matrices*
  - Popular algorithms for word embedding creation
  - Word embedding evaluation
  - Pitfalls of unsupervised learning

- On Wednesday: Probabilistic language modeling