Word Meaning and Similarity

Word Senses and Word Relations

(Reading: J+M Ch19&Ch20)
Reminder: lemma and wordform

- A **lemma** or **citation form**
  - The canonical form of words

- A **wordform**
  - The “inflected” word as it appears in text

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>banks</td>
<td>bank</td>
</tr>
<tr>
<td>sung</td>
<td>sing</td>
</tr>
<tr>
<td>sleeping</td>
<td>sleep</td>
</tr>
</tbody>
</table>
Lemmas have senses

• One lemma “bank” can have many meanings:
  • ...a bank can hold the investments in an account...
  Sense 1:
  • “...as agriculture growth on the east bank the river will shrink even more”

  Sense 2:

• Sense (or word sense)
  • A discrete representation of an aspect of a word’s meaning.

• The lemma bank here has two senses
Homonymy

Homonyms: words that share a form but have unrelated, distinct meanings:

- $\text{bank}_1$: financial institution, $\text{bank}_2$: sloping land
- $\text{bat}_1$: club for hitting a ball, $\text{bat}_2$: flying mammal

1. Homographs (bank/bank, bat/bat)

2. Homophones:
   1. Write and right
   2. Piece and peace
Homonymy causes problems for NLP applications

- Information retrieval
  - “bat care”
- Machine Translation
  - bat: murciélago (animal) or bate (for baseball)
- Text-to-Speech
  - bass (musical instrument) vs. bass (fish)
Polysemy

1. The **bank** was constructed in 1875 out of local red brick.
2. I withdrew the money from the **bank**
Are those the same sense?
   - Sense 2: “A financial institution”
   - Sense 1: “The building belonging to a financial institution”

A **polysemous** word has **related** meanings
   - Most non-rare words have multiple meanings
Lots of types of polysemy are systematic
- School, university, hospital
  - All can mean the institution or the building.

A systematic relationship:
- Building ↔ Organization

Other such kinds of systematic polysemy:
- Author (Jane Austen wrote Emma)
  ↔ Works of Author (I love Jane Austen)
- Tree (Plums have beautiful blossoms)
  ↔ Fruit (I ate a preserved plum)
How do we know when a word has more than one sense?

- The “zeugma” test: Two senses of *serve*?
  - Which flights *serve* breakfast?
  - Does Lufthansa *serve* Philadelphia?
  - *?Does Lufthansa serve breakfast and San Jose?*

- **Since this conjunction sounds weird,**
  - we say that these are **two different senses of “serve”**
Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / H\(_2\)O

- Two lexemes are synonyms
  - if they can be substituted for each other in all situations
  - If so they have the same *propositional meaning*
**Synonyms**

- But there are few (or no) examples of perfect synonymy.
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

- **Example:**
  - Water/H$_2$O
  - Big/large
  - Brave/courageous
Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How *big* is that plane?
  - Would I be flying on a *large* or small plane?
- How about here:
  - Miss Nelson became a kind of *big* sister to Benjamin.
  - Miss Nelson became a kind of *large* sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense
Antonyms

• Senses that are opposites with respect to one feature of meaning
• Otherwise, they are very similar!
  dark/light  short/long  fast/slow  rise/fall
  hot/cold    up/down    in/out

• More formally: antonyms can
  • define a binary opposition
    or be at opposite ends of a scale
    • long/short, fast/slow

  • Be reversives (reverse in direction):
    • rise/fall, up/down
Hyponymy and Hypernymy

• One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  • *car* is a hyponym of *vehicle*
  • *mango* is a hyponym of *fruit*

• Conversely **hypernym/superordinate** ("hyper is super")
  • *vehicle* is a hypernym of *car*
  • *fruit* is a hypernym of *mango*

<table>
<thead>
<tr>
<th>Superordinate/hyper</th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subordinate/hyponym</td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>
Hyponymy more formally

- **Entailment:**
  - A sense A is a hyponym of sense B if *being an A* entails *being a B*
- **Hyponymy is usually transitive**
  - (A hypo B and B hypo C entails A hypo C)
- **Another name: the IS-A hierarchy**
  - A IS-A B  (or A ISA B)
  - B subsumes A
Hyponyms and Instances

- WordNet has both classes and instances.
- An instance is an individual, a proper noun that is a unique entity.
  - San Francisco is an instance of city.
- But city is a class.
  - city is a hyponym of municipality...location...
Word Meaning and Similarity
WordNet and other Online Thesauri
Applications of Thesauri and Ontologies

- Information Extraction
- Information Retrieval
- Question Answering
- Bioinformatics and Medical Informatics
- Machine Translation
WordNet 3.0

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Some other languages available or under development
    - (Arabic, Finnish, German, Portuguese...)

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117,798</td>
</tr>
<tr>
<td>Verb</td>
<td>11,529</td>
</tr>
<tr>
<td>Adjective</td>
<td>22,479</td>
</tr>
<tr>
<td>Adverb</td>
<td>4,481</td>
</tr>
</tbody>
</table>
Senses of “bass” in Wordnet

Noun

- **S: (n) bass** (the lowest part of the musical range)
- **S: (n) bass, bass part** (the lowest part in polyphonic music)
- **S: (n) bass, basso** (an adult male singer with the lowest voice)
- **S: (n) sea bass, bass** (the lean flesh of a saltwater fish of the family Serranidae)
- **S: (n) freshwater bass, bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- **S: (n) bass, bass voice, basso** (the lowest adult male singing voice)
- **S: (n) bass** (the member with the lowest range of a family of musical instruments)
- **S: (n) bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- **S: (adj) bass, deep** (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"
How is “sense” defined in WordNet?

- **The synset (synonym set),** the set of near-synonyms, instantiates a sense or concept, with a gloss.
- Example: *chump* as a noun with the gloss:
  “a person who is gullible and easy to take advantage of”
- This sense of “chump” is shared by 9 words:
  \(chump^1, fool^2, gull^1, mark^9, patsy^1, fall\ guy^1,\)
  \(sucker^1, soft\ touch^1, mug^2\)
- Each of these senses have this same gloss
  - (Not every sense; sense 2 of gull is the aquatic bird)
WordNet Hypernym Hierarchy for “bass”

- **S**: (n) **bass, basso** (an adult male singer with the lowest voice)
  - **direct hypernym** / **inherited hypernym** / **sister term**
    - **S**: (n) **singer, vocalist, vocalizer, vocaliser** (a person who sings)
    - **S**: (n) **musician, instrumentalist, player** (someone who plays a musical instrument (as a profession))
      - **S**: (n) **performer, performing artist** (an entertainer who performs a dramatic or musical work for an audience)
        - **S**: (n) **entertainer** (a person who tries to please or amuse)
      - **S**: (n) **person, individual, someone, somebody, mortal, soul** (a human being) "there was too much for one person to do"
    - **S**: (n) **organism, being** (a living thing that has (or can develop) the ability to act or function independently)
      - **S**: (n) **living thing, animate thing** (a living (or once living) entity)
        - **S**: (n) **whole, unit** (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a unit"
      - **S**: (n) **object, physical object** (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
        - **S**: (n) **physical entity** (an entity that has physical existence)
        - **S**: (n) **entity** (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
# WordNet Noun Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Also called</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>Superordinate</td>
<td>From concepts to superordinates</td>
<td>breakfast$^1$ → meal$^1$</td>
</tr>
<tr>
<td>Hyponym</td>
<td>Subordinate</td>
<td>From concepts to subtypes</td>
<td>meal$^1$ → lunch$^1$</td>
</tr>
<tr>
<td>Member Meronym</td>
<td>Has-Member</td>
<td>From groups to their members</td>
<td>faculty$^2$ → professor$^1$</td>
</tr>
<tr>
<td>Has-Instance</td>
<td></td>
<td>From concepts to instances of the concept</td>
<td>composer$^1$ → Bach$^1$</td>
</tr>
<tr>
<td>Instance</td>
<td></td>
<td>From instances to their concepts</td>
<td>Austen$^1$ → author$^1$</td>
</tr>
<tr>
<td>Member Holonym</td>
<td>Member-Of</td>
<td>From members to their groups</td>
<td>copilot$^1$ → crew$^1$</td>
</tr>
<tr>
<td>Part Meronym</td>
<td>Has-Part</td>
<td>From wholes to parts</td>
<td>table$^2$ → leg$^3$</td>
</tr>
<tr>
<td>Part Holonym</td>
<td>Part-Of</td>
<td>From parts to wholes</td>
<td>course$^7$ → meal$^1$</td>
</tr>
<tr>
<td>Antonym</td>
<td></td>
<td>Opposites</td>
<td>leader$^1$ → follower$^1$</td>
</tr>
</tbody>
</table>
WordNet 3.0

• Where it is:
  • [http://wordnetweb.princeton.edu/perl/webwn](http://wordnetweb.princeton.edu/perl/webwn)

• Libraries
  • Python: WordNet from NLTK
    • [http://www.nltk.org/Home](http://www.nltk.org/Home)
  • Java:
    • JWNL, extJWNL on sourceforge
Word Meaning and Similarity

Word Similarity: Thesaurus Methods
Word Similarity

- **Synonymy**: a binary relation
  - Two words are either synonymous or not
- **Similarity** (or distance): a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between senses
  - The word “bank” is not similar to the word “slope”
  - Bank$^1$ is similar to fund$^3$
  - Bank$^2$ is similar to slope$^5$
- But we’ll compute similarity over both words and senses
Why word similarity

- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading
- Plagiarism detection
- Document clustering
Word similarity and word relatedness

- We often distinguish **word similarity** from **word relatedness**
  - **Similar words**: near-synonyms
  - **Related words**: can be related any way
    - car, bicycle: *similar*
    - car, gasoline: *related*, not similar
Two classes of similarity algorithms

- Thesaurus-based algorithms
  - Are words “nearby” in hypernym hierarchy?
  - Do words have similar glosses (definitions)?
- Distributional algorithms
  - Do words have similar distributional contexts?
Path based similarity

- Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  - =have a short path between them
  - concepts have path 1 to themselves
Refinements to path-based similarity

- \( \text{pathlen}(c_1, c_2) = 1 + \) number of edges in the shortest path in the hypernym graph between sense nodes \( c_1 \) and \( c_2 \)

- \( \text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)} \)

- ranges from 0 to 1 (identity)

- \( \text{wordsim}(w_1, w_2) = \max_{c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)} \text{sim}(c_1, c_2) \)
Example: path-based similarity

\[ \text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)} \]

\[
\begin{align*}
\text{simpath}(\text{nickel}, \text{coin}) &= \frac{1}{2} = .5 \\
\text{simpath}(\text{fund}, \text{budget}) &= \frac{1}{2} = .5 \\
\text{simpath}(\text{nickel}, \text{currency}) &= \frac{1}{4} = .25 \\
\text{simpath}(\text{nickel}, \text{money}) &= \frac{1}{6} = .17 \\
\text{simpath}(\text{coinage}, \text{Richter scale}) &= \frac{1}{6} = .17
\end{align*}
\]
Problem with basic path-based similarity

• Assumes each link represents a uniform distance
  • But \textit{nickel} to \textit{money} seems to us to be closer than \textit{nickel} to \textit{standard}
  • Nodes high in the hierarchy are very abstract

• We instead want a metric that
  • Represents the cost of each edge independently
  • Words connected only through abstract nodes
    • are less similar
Let’s define $P(c)$ as:

- The probability that a randomly selected word in a corpus is an instance of concept $c$
- Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
  - for a given concept, each observed noun is either
    - a member of that concept with probability $P(c)$
    - not a member of that concept with probability $1-P(c)$
- All words are members of the root node (Entity)
  - $P(\text{root})=1$
- The lower a node in hierarchy, the lower its probability
Information content similarity

Train by counting in a corpus

- Each instance of hill counts toward frequency of natural elevation, geological formation, entity, etc.
- Let words(c) be the set of all words that are children of node c + c itself
  - words(“geo-formation”) = {hill, ridge, grotto, coast, cave, shore, natural elevation, geo-formation}
  - words(“natural elevation”) = {hill, ridge, natural elevation}

\[
P(c) = \frac{\sum \text{count}(w)}{N} \quad \text{for } w \in \text{words}(c)
\]
Information content similarity

- WordNet hierarchy augmented with probabilities $P(c)$


```
entity 0.395

  inanimate-object 0.167

  natural-object 0.0163

    geological-formation 0.00176

    0.000113 natural-elevation shore 0.0000836

    0.0000189 hill coast 0.0000216
```
Information content: definitions

• Information content:
  \[ IC(c) = -\log P(c) \]

• Most informative subsumer (Lowest common subsumer)
  \[ LCS(c_1, c_2) = \]
  The most informative (lowest) node in the hierarchy subsuming both \( c_1 \) and \( c_2 \)
Using information content for similarity: the Resnik method


• The similarity between two words is related to their common information

• The more two words have in common, the more similar they are

• Resnik: measure common information as:
  • The information content of the lowest common subsumer (LCS) of the two nodes

  \[ \text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2)) \]
Dekang Lin method

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. ICML

- Intuition: Similarity between A and B is not just what they have in common
- The more **differences** between A and B, the less similar they are:
  - Commonality: the more A and B have in common, the more similar they are
  - Difference: the more differences between A and B, the less similar
- Commonality: IC(common(A,B))
- Difference: IC(description(A,B)) - IC(common(A,B))
Dekang Lin similarity theorem

- The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are:

\[
sim_{Lin}(A, B) \propto \frac{IC(\text{common}(A, B))}{IC(\text{description}(A, B))}
\]

- Lin (altering Resnik) defines IC(\text{common}(A,B)) as 2 x information of the LCS:

\[
sim_{Lin}(c_1, c_2) = \frac{2 \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)}
\]
Lin similarity function

\[ sim_{Lin}(A, B) = \frac{2 \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \]

\[ sim_{Lin}(\text{hill}, \text{coast}) = \frac{2 \log P(\text{geological-formation})}{\log P(\text{hill}) + \log P(\text{coast})} \]

\[ = \frac{2 \ln 0.00176}{\ln 0.0000189 + \ln 0.0000216} \]

\[ = .59 \]
The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at *glosses*
- Two concepts are similar if their glosses contain similar words
  - *Drawing paper*: *paper* that is *specially prepared* for use in drafting
  - *Decal*: the art of transferring designs from *specially prepared* paper to a wood or glass or metal surface
- For each *n*-word phrase that’s in both glosses
  - Add a score of $n^2$
  - *Paper* and *specially prepared* for $1 + 2^2 = 5$
- Compute overlap also for other relations
  - glosses of hypernyms and hyponyms
Summary: thesaurus-based similarity

\[
\text{sim}_{\text{path}}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)}
\]

\[
\text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2))
\]

\[
\text{sim}_{\text{lin}}(c_1, c_2) = \frac{2\log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)}
\]

\[
\text{sim}_{\text{jiangconrath}}(c_1, c_2) = \frac{1}{\log P(c_1) + \log P(c_2) - 2\log P(\text{LCS}(c_1, c_2))}
\]

\[
\text{sim}_{\text{eLesk}}(c_1, c_2) = \sum_{r, q \in \text{RELS}} \text{overlap(gloss}(r(c_1)), \text{gloss}(q(c_2))))
\]
Libraries for computing thesaurus-based similarity

- **NLTK**

- **WordNet::Similarity**
  - Web-based interface:
    - [http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi](http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi)
Evaluating similarity

- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings

- Extrinsic (task-based, end-to-end) Evaluation:
  - Spelling error detection
  - WSD
  - Essay grading
  - Taking TOEFL multiple-choice vocabulary tests

*Levied* is closest in meaning to:
  imposed, believed, requested, correlated
Word Meaning and Similarity

Word Similarity: Distributional Similarity (I)
Problems with thesaurus-based meaning

• We don’t have a thesaurus for every language
• Even if we do, they have problems with recall
  • Many words are missing
  • Most (if not all) phrases are missing
  • Some connections between senses are missing
  • Thesauri work less well for verbs, adjectives
    • Adjectives and verbs have less structured hyponymy relations
Distributional models of meaning

- Offer much higher recall than hand-built thesauri
  - Although they tend to have lower precision
- Zellig Harris (1954): “oculist and eye-doctor ... occur in almost the same environments.... If A and B have almost identical environments we say that they are synonyms.
- Firth (1957): “You shall know a word by the company it keeps!”
Intuition of distributional word similarity

• Example:

A dish of *Esquites* is on the table.
Everybody likes *Esquites*.
*Esquites* is very delicious.
We make *Esquites* out of corn.

• From context words humans can guess *Esquites* means
  • A corn based food

• Intuition for algorithm:
  • Two words are similar if they have similar word contexts.
Term-document matrix

- Each cell: count of term $t$ in a document $d$: $tf_{t,d}$
  - Each document is a count vector in $\mathbb{N}^v$: a column below

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### Term-document matrix

- Each cell: count of term $t$ in a document $d$: $\text{tf}_{t,d}$
  - Each document is a count vector in $\mathbb{N}^v$: a column below

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Term-document matrix

- Two documents are similar if their vectors are similar

<table>
<thead>
<tr>
<th>Term</th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Term-document matrix

- Two documents are similar if their vectors are similar

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The words in a term-document matrix

- Each word is a count vector in $\mathbb{N}^D$: a row below

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The words in a term-document matrix

- Each word is a count vector in $\mathbb{N}^D$: a row below

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The words in a term-document matrix

- Two **words** are similar if their vectors are similar

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The words in a term-document matrix

- Two **words** are similar if their vectors are similar

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>fool</td>
<td>37</td>
<td>58</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>clown</td>
<td>6</td>
<td>117</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The Term-Context matrix

• Instead of using entire documents, use smaller contexts
  • Paragraph
  • Window of 10 words
• A word is now defined by a vector over counts of context words
Sample contexts: 20 words (Brown corpus)

• equal amount of sugar, a sliced lemon, a tablespoonful of apricot preserve or jam, a pinch each of clove and nutmeg,

• on board for their enjoyment. Cautiously she sampled her first pineapple and another fruit whose taste she likened to that of

• of a recursive type well suited to programming on the digital computer. In finding the optimal R-stage policy from that of

• substantially affect commerce, for the purpose of gathering data and information necessary for the study authorized in the first section of this
Term-context matrix for word similarity

- Two **words** are similar in meaning if their context vectors are similar

| Word     | aardvark | computer | data | pinch | result | sugar | ...
|----------|----------|----------|------|-------|--------|-------|-------
| apricot  | 0        | 0        | 0    | 1     | 0      | 1     |       |
| pineapple| 0        | 0        | 0    | 1     | 0      | 1     |       |
| digital  | 0        | 2        | 1    | 0     | 1      | 0     |       |
| information | 0    | 1        | 6    | 0     | 4      | 0     |       |
Term-context matrix for word similarity

- Two **words** are similar in meaning if their context vectors are similar

|        | aardvark | computer | data | pinch | result | sugar | ...
|--------|----------|----------|------|-------|--------|-------|---
| apricot| 0        | 0        | 0    | 1     | 0      | 1     |   |
| pineapple| 0    | 0        | 0    | 1     | 0      | 1     |   |
| digital | 0        | 2        | 1    | 0     | 1      | 0     |   |
| information| 0 | 1        | 6    | 0     | 4      | 0     |   |
Should we use raw counts?

• For the term-document matrix
  • We use \textit{tf-idf} instead of raw term counts

• For the term-context matrix
  • \textbf{Positive Pointwise Mutual Information (PPMI) is common}
Pointwise Mutual Information

- **Pointwise mutual information:**
  - Do events $x$ and $y$ co-occur more than if they were independent?

  \[
  \text{PMI}(X, Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}
  \]

- **PMI between two words:** (Church & Hanks 1989)
  - Do words $x$ and $y$ co-occur more than if they were independent?

  \[
  \text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}
  \]

- **Positive PMI between two words** (Niwa & Nitta 1994)
  - Replace all PMI values less than 0 with zero
Computing PPMI on a term-context matrix

- Matrix $F$ with $W$ rows (words) and $C$ columns (contexts)
- $f_{ij}$ is # of times $w_i$ occurs in context $c_j$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$
$$p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$
$$p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}$$

$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_{i*}p_{*j}}$$

$$ppmi_{ij} = \begin{cases} pmi_{ij} & \text{if } pmi_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$
\[ p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}} \]  

Count(w,context)  

\[ \begin{array}{cccccc}
\text{computer} & \text{data} & \text{pinch} & \text{result} & \text{sugar} \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 \\
2 & 1 & 0 & 1 & 0 \\
1 & 6 & 0 & 4 & 0 \\
\end{array} \]

\[ p(w=\text{information},c=\text{data}) = \frac{6}{19} = .32 \]
\[ p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}} \]

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ p(w=\text{information}, c=\text{data}) = \frac{6}{19} = .32 \]

\[ p(w=\text{information}) = \frac{11}{19} = .58 \]
\[ p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}} \]

where:
- \( W \) is the number of words in the dictionary,
- \( C \) is the number of contexts.

**Count(\( w \), context)**

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ p(w=\text{information},c=\text{data}) = \frac{6}{19} = .32 \]
\[ p(w=\text{information}) = \frac{11}{19} = .58 \]
\[ p(c=\text{data}) = \frac{7}{19} = .37 \]
\[
p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}
\]

- apricot
- pineapple
- digital
- information

\[
p(w=\text{information}, c=\text{data}) = \frac{6}{19} = .32
\]
\[
p(w=\text{information}) = \frac{11}{19} = .58
\]
\[
p(c=\text{data}) = \frac{7}{19} = .37
\]

\[
p(w) = \frac{\sum_{i=1}^{W} f_{ij}}{N}
\]
\[
p(c_j) = \frac{\sum_{i=1}^{C} f_{ij}}{N}
\]

\[
\begin{array}{cccccc}
\text{computer} & \text{data} & \text{pinch} & \text{result} & \text{sugar} \\
apricot & 0.00 & 0.00 & 0.05 & 0.00 & 0.05 \\
pineapple & 0.00 & 0.00 & 0.05 & 0.00 & 0.05 \\
digital & 0.11 & 0.05 & 0.00 & 0.05 & 0.00 \\
information & 0.05 & 0.32 & 0.00 & 0.21 & 0.00 \\
\end{array}
\]

\[
\begin{array}{cccccc}
\text{computer} & \text{data} & \text{pinch} & \text{result} & \text{sugar} \\
p(w, \text{context}) & 0.16 & 0.37 & 0.11 & 0.26 & 0.11 \\
\end{array}
\]
\[ pmi_{ij} = \log_2 \frac{p_{ij}}{p_i p_j} \]

<table>
<thead>
<tr>
<th>Word</th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
<th>p(context)</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>digital</td>
<td>0.11</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.21</td>
</tr>
<tr>
<td>information</td>
<td>0.05</td>
<td>0.32</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>0.58</td>
</tr>
</tbody>
</table>
$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_i p_j}$$

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
<th>p(w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>digital</td>
<td>0.11</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.21</td>
</tr>
<tr>
<td>information</td>
<td>0.05</td>
<td>0.32</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>0.58</td>
</tr>
<tr>
<td>p(context)</td>
<td>0.16</td>
<td>0.37</td>
<td>0.11</td>
<td>0.26</td>
<td>0.11</td>
<td></td>
</tr>
</tbody>
</table>

- \( pmi(\text{information}, \text{data}) = \log_2 ( .32 / (.37 * .58) ) = .58 \)

\( (.57 \text{ using full precision}) \)
\[ pmi_{ij} = \log_2 \frac{p_{ij}}{p_i \cdot p_{*j}} \]

### Example Calculations

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
<th>p(w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>digital</td>
<td>0.11</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.21</td>
</tr>
<tr>
<td>information</td>
<td>0.05</td>
<td>0.32</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>0.58</td>
</tr>
</tbody>
</table>

\[
p(context) = 0.16 \quad 0.37 \quad 0.11 \quad 0.26 \quad 0.11
\]

- \( pmi(\text{information, data}) = \log_2 (\frac{0.32}{0.37 \cdot 0.58}) = 0.57 \)
Weighing PMI

- PMI is biased toward infrequent events
- Various weighting schemes help alleviate this
  - See Turney and Pantel (2010)
- Add-one smoothing can also help
<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>pineapple</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>digital</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>information</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>p(w,context) [add-2]</th>
<th>p(w)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>computer</td>
<td>data</td>
</tr>
<tr>
<td>apricot</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>digital</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>information</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td>p(context)</td>
<td>0.19</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>computer</td>
<td>data</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td>apricot</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>pineapple</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>digital</td>
<td>1.66</td>
<td>0.00</td>
</tr>
<tr>
<td>information</td>
<td>0.00</td>
<td>0.57</td>
</tr>
</tbody>
</table>

**PPMI(w,context) [add-2]**

<table>
<thead>
<tr>
<th></th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.00</td>
<td>0.00</td>
<td>0.56</td>
<td>0.00</td>
<td>0.56</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.00</td>
<td>0.00</td>
<td>0.56</td>
<td>0.00</td>
<td>0.56</td>
</tr>
<tr>
<td>digital</td>
<td>0.62</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>information</td>
<td>0.00</td>
<td>0.58</td>
<td>0.00</td>
<td>0.37</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Word Meaning and Similarity

Word Similarity: Distributional Similarity (II)
Using syntax to define a word’s context

• Zellig Harris (1968)
  • “The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities”

• Two words are similar if they have similar parse contexts

• **Duty** and **responsibility** (Chris Callison-Burch’s example)

<table>
<thead>
<tr>
<th>Modified by adjectives</th>
<th>additional, administrative, assumed, collective, congressional, constitutional ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects of verbs</td>
<td>assert, assign, assume, attend to, avoid, become, breach ...</td>
</tr>
</tbody>
</table>
Co-occurrence vectors based on syntactic dependencies

Dekang Lin, 1998 “Automatic Retrieval and Clustering of Similar Words”

- The contexts $C$ are different dependency relations
  - Subject-of- “absorb”
  - Prepositional-object of “inside”

- Counts for the word “cell”:

<table>
<thead>
<tr>
<th></th>
<th>subj-of, absorb</th>
<th>subj-of, adapt</th>
<th>subj-of, behave</th>
<th>pobj-of, inside</th>
<th>pobj-of, into</th>
<th>nmod-of, abnormality</th>
<th>nmod-of, anemia</th>
<th>nmod-of, architecture</th>
<th>obj-of, attack</th>
<th>obj-of, call</th>
<th>obj-of, come from</th>
<th>obj-of, decorate</th>
<th>nmod, bacteria</th>
<th>nmod, body</th>
<th>nmod, bone marrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>30</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
PMI applied to dependency relations

Hindle, Don. 1990. Noun Classification from Predicate-Argument Structure. ACL

<table>
<thead>
<tr>
<th>Object of “drink”</th>
<th>Count</th>
<th>PMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>tea</td>
<td>2</td>
<td>11.8</td>
</tr>
<tr>
<td>liquid</td>
<td>2</td>
<td>10.5</td>
</tr>
<tr>
<td>coffee</td>
<td>2</td>
<td>9.3</td>
</tr>
<tr>
<td>anything</td>
<td>3</td>
<td>5.2</td>
</tr>
<tr>
<td>it</td>
<td>3</td>
<td>1.3</td>
</tr>
</tbody>
</table>

- “Drink it” more common than “drink coffee”
- But “coffee” is a better “drinkable” thing than “it”
Cosine for computing similarity

\[
\cos(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{||\vec{v}|| \cdot ||\vec{w}||} = \frac{\vec{v}}{||\vec{v}||} \cdot \frac{\vec{w}}{||\vec{w}||} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}
\]

\(v_i\) is the PPMI value for word \(v\) in context \(i\)

\(w_i\) is the PPMI value for word \(w\) in context \(i\).

\(\text{Cos}(\vec{v}, \vec{w})\) is the cosine similarity of \(\vec{v}\) and \(\vec{w}\)
Cosine as a similarity metric

-1: vectors point in opposite directions
+1: vectors point in same directions
0: vectors are orthogonal

Raw frequency or PPMI are non-negative, so cosine range 0-1
Which pair of words is more similar?

\[
\cos(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{||\vec{v}|| ||\vec{w}||} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}
\]

\[
\cos(\text{apricot, information}) = \frac{1+0+0}{\sqrt{1+0+0} \sqrt{1+36+1}} = \frac{1}{\sqrt{38}} = .16
\]

<table>
<thead>
<tr>
<th></th>
<th>large</th>
<th>data</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>information</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
\[ \cos(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \|\vec{w}\|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}} \]

Which pair of words is more similar?

\[ \cos(\text{apricot}, \text{information}) = \frac{1+0+0}{\sqrt{1+0+0} \sqrt{1+36+1}} = \frac{1}{\sqrt{38}} = .16 \]

\[ \cos(\text{digital}, \text{information}) = \frac{0+6+2}{\sqrt{0+1+4} \sqrt{1+36+1}} = \frac{8}{\sqrt{38 \sqrt{5}}} = .58 \]
\[ \cos(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \|\vec{w}\|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}} \]

Which pair of words is more similar?

\[
\begin{align*}
\text{cosine}(\text{apricot, information}) &= \frac{1 + 0 + 0}{\sqrt{1 + 0 + 0} \sqrt{1 + 36 + 1}} = \frac{1}{\sqrt{38}} = .16 \\
\text{cosine}(\text{digital, information}) &= \frac{0 + 6 + 2}{\sqrt{0 + 1 + 4} \sqrt{1 + 36 + 1}} = \frac{8}{\sqrt{38} \sqrt{5}} = .58 \\
\text{cosine}(\text{apricot, digital}) &= \frac{0 + 0 + 0}{\sqrt{1 + 0 + 0} \sqrt{0 + 1 + 4}} = 0 
\end{align*}
\]
Evaluating similarity
(the same as for thesaurus-based)

• Intrinsic Evaluation:
  • Correlation between algorithm and human word similarity ratings

• Extrinsic (task-based, end-to-end) Evaluation:
  • Spelling error detection, WSD, essay grading
  • Taking TOEFL multiple-choice vocabulary tests

Levied is closest in meaning to which of these: imposed, believed, requested, correlated