Word Meaning

- Considers the meaning(s) of a word in addition to its written form
- Word Sense: a discrete representation of an aspect of the meaning of a word
Lexeme

- An entry in a lexicon consisting of a pair: a form with a single meaning representation
  - band (music group)
  - band (material)
  - band (wavelength)

(http://www.weiku.com/products/12426189/Polyester_Elastic_band_for_garment_underwear_shoe_bags.html)
(http://clipart.me/band-material-with-the-enthusiasm-of-the-audience-silhouette-19222)
Lemma

- The grammatical form that is used to represent a lexeme
  - Berlin
  - band
Homonymy

• Words which have similar form but different meanings
  – Homographs:
    • Berlin (Germany's capital); Berlin (music band)
    • band (music group); band (material); band (wavelength)
Homophones

- Words which have similar pronunciation but different writing and meaning
  - write
  - right
Semantics Relations

- Lexical relations among words (senses)
  - Hyponymy (is-a relation) \{parent: hypernym, child: hyponym\}
    - dog & animal

(http://animalia-life.com/dogs.html)
(http://pic-zoom.com/pictures-animals.html)
Semantics Relations

- Lexical relations among words (senses)
  - Meronymy (part-of relation)
    - arm & body

(http://www.oxfordlearnersdictionaries.com/definition/american_english/arm_1)
Semantics Relations

- Lexical relations among words (senses)
  - Synonymy
    - fall & autumn

(http://pinitgallery.com/photo/f/fall-background/7/)
Semantics Relations

- Lexical relations among words (senses)
  - Antonymy
    - tall & short

(http://pixgood.com/tall-short.html)
WordNet

- A hierarchical database of lexical relations
- Three Separate sub-databases
  - Nouns
  - Verbs
  - Adjectives and Adverbs
- Each word is annotated with a set of senses
- Available online or for download
  - http://wordnetweb.princeton.edu/perl/webwn
Word sense

- Synset (synonym set)

Noun

- S: (n) set, circle, band, lot (an unofficial association of people or groups) "the smart set goes there"; "they were an angry lot"
- S: (n) band (instrumentalists not including string players)
- S: (n) band, banding, stria, striation (a stripe or stripes of contrasting color) "chromosomes exhibit characteristic bands"; "the black and yellow banding of bees and wasps"
- S: (n) band, banding, stripe (an adornment consisting of a strip of a contrasting color or material)
- S: (n) dance band, band, dance orchestra (a group of musicians playing popular music for dancing)
- S: (n) band (a range of frequencies between two limits)
- S: (n) band (a thin flat strip of flexible material that is worn around the body or one of the limbs (especially to decorate the body))
- S: (n) isthmus, band (a cord-like tissue connecting two larger parts of an anatomical structure)
- S: (n) ring, band (jewelry consisting of a circlet of precious metal (often set with jewels) worn on the finger) "she had rings on every finger"; "he noted that she wore a wedding band"
- S: (n) band (a driving belt in machinery)
- S: (n) band (a thin flat strip or loop of flexible material that goes around or over something else, typically to hold it together or as a decoration)
- S: (n) band, ring (a strip of material attached to the leg of a bird to identify it (as in studies of bird migration))
- S: (n) band (a restraint put around something to hold it together)

Verb

- S: (v) band (bind or tie together, as with a band)
- S: (v) ring, band (attach a ring to the foot of, in order to identify) "ring birds"; "band the geese to observe their migratory patterns"
Word Relations (Hypernym)

- **S:** (n) **ring, band** (jewelry consisting of a circlet of precious metal (often set with jewels) worn on the finger) "she had rings on every finger"; "he noted that she wore a wedding band"
  - *direct hyponym / full hyponym*
    - **S:** (n) **engagement ring** (a ring given and worn as a sign of betrothal)
    - **S:** (n) **mourning ring** (a ring worn as a memorial to a dead person)
    - **S:** (n) **ringlet** (a small ring)
    - **S:** (n) **signet ring, seal ring** (a ring bearing a signet)
    - **S:** (n) **wedding ring, wedding band** (a ring (usually plain gold) given to the bride (and sometimes one is also given to the groom) at the wedding)
  - *direct hypernym / inherited hypernym / sister term*
    - **S:** (n) **jewelry, jewellery** (an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems))
Word sense disambiguation (WSD)

- Figure out the meaning of a word in a certain context
Motivation: Information retrieval & question answering

- Berlin is the capital of Germany.
- Berlin may also refer to:

**Individuals**
- Berlin (surname)
- Berlin Ndebe-Nlome (born 1987), Cameroonian football player
- Berlin, former stage name for professional wrestler Alex Wright

**Places**
- Canada
  - Berlin, former name of Kitchener, Ontario
    - Berlin to Kitchener name change
- United States
  - Berlin, California, the former name of Genevra, California
  - Berlin, Connecticut
    - Berlin (Amtrak station), rail station in Berlin, Connecticut
  - Berlin, Georgia
  - Berlin, Illinois
  - Berlin, Indiana, extinct town
  - Berlin, Kentucky
  - Berlin, Maryland
Motivation: Machine translation

I get money from the bank.

The bank of the river was very nice.

Ich Geld von der Bank.

Die Ufer des Flusses war sehr schön.

(http://translate.google.de)
Motivation: Speech synthesis

- Eggs have a high protein content.
- She was content to step down after four years as chief executive.
Word Sense Disambiguation

- Input
  - A word
  - The context of the word
  - Set of potential senses for the word

- Output
  - The best sense of the word for this context

The bank of the river was nice.

bank

Ufer

Bank
Approaches

- Thesaurus-based
- Supervised learning
- Semi-supervised learning
Thesaurus-based

- Extract sense definitions from existing sources
  - Dictionaries
  - Thesauri
  - Wikipedia

<table>
<thead>
<tr>
<th>Science and technology</th>
<th>[edit]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAND (application), a private space for groups</td>
<td></td>
</tr>
<tr>
<td>Band (mathematics), an idempotent semigroup</td>
<td></td>
</tr>
<tr>
<td>Band (radio), a range of frequencies or wavelengths used in radio transmission and radar</td>
<td></td>
</tr>
<tr>
<td>Band cell, a type of white blood cell</td>
<td></td>
</tr>
<tr>
<td>Gastric band, a weight-control measure</td>
<td></td>
</tr>
<tr>
<td>Bird banding, placing numbered bands of metal on birds' legs for identification</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organizations</th>
<th>[edit]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band (channel), nickname of Brazilian broadcast television network Rede Bandeirantes</td>
<td></td>
</tr>
<tr>
<td>Bands (Italian Army irregulars), military units once in the service of the Italian Regio Esercito</td>
<td></td>
</tr>
<tr>
<td>The Band (professional wrestling), the Total Nonstop Wrestling name for the professional wrestling stable New World Order</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Music</th>
<th>[edit]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band (music), a group of people who perform instrumental or vocal music</td>
<td></td>
</tr>
<tr>
<td>Concert band, an ensemble of woodwind, brass, and percussion instruments</td>
<td></td>
</tr>
<tr>
<td>School band, a group of student musicians who rehearse and perform instrumental music together</td>
<td></td>
</tr>
<tr>
<td>Marching band, a group of instrumental musicians who generally perform outdoors incorporating some type of marching</td>
<td></td>
</tr>
<tr>
<td>Jazz band, a musical ensemble that plays jazz music</td>
<td></td>
</tr>
<tr>
<td>The Band, a Canadian-American rock and roll group</td>
<td></td>
</tr>
<tr>
<td><em>The Band</em> (album), its eponymous album released in 1969</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clothing, jewelry, and accessories</th>
<th>[edit]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bands (neckwear), two pieces of cloth fitted around the neck as part of formal clothing for clergy, academics, and lawyers</td>
<td></td>
</tr>
<tr>
<td>Bandolier or bandoleer, an ammunition belt</td>
<td></td>
</tr>
<tr>
<td>Wedding band, a metal ring indicating the wearer is married</td>
<td></td>
</tr>
<tr>
<td>Belt (clothing), a flexible band or strap, typically made of leather or heavy cloth, and worn around the waist</td>
<td></td>
</tr>
<tr>
<td>Strap, an elongated flap or ribbon, usually of fabric or leather</td>
<td></td>
</tr>
</tbody>
</table>
The Lesk Algorithm

- Select the sense whose definition shares the most words with the word’s context

```plaintext
function SIMPLIFIED LESK(word, sentence) returns best sense of word
    best-sense ← most frequent sense for word
    max-overlap ← 0
    context ← set of words in sentence
    for each sense in senses of word do
        signature ← set of words in the gloss and examples of sense
        overlap ← COMPUTEOVERLAP (signature, context)
        if overlap > max-overlap then
            max-overlap ← overlap
            best-sense ← sense
    end return (best-sense)
```

(http://en.wikipedia.org/wiki/Lesk_algorithm)
The Lesk Algorithm

- Simple to implement
- No training data needed, „only“ a lexicon
- Relatively bad results
Supervised Learning

- Training data:
  - A corpus in which each occurrence of the ambiguous word "w" is annotated with its correct sense
    - SemCor: 234,000 sense-tagged from Brown corpus
    - SENSEVAL-1: 34 target words
    - SENSEVAL-2: 73 target words
    - SENSEVAL-3: 57 target words (2081 sense-tagged)
<s snum=2>
<wf cmd=tag pos=NNP>Mr. Hawksley</wf>
<wf cmd=done pos=VB lemma=say wnsn=1 lexsn=2:32:00::>said</wf>
<wf cmd=tag pos=NN>yesterday</wf>
<wf cmd=ignore pos=PRP>he</wf>
<wf cmd=ignore pos=MD>would</wf>
<wf cmd=done pos=VB ot=metaphor>be</wf>
<wf cmd=tag pos=JJ>willing</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=done pos=VB lemma=go wnsn=1 lexsn=2:38:00::>go</wf>
<wf cmd=ignore pos=IN>before</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=tag pos=NN>city_council</wf>
<punct>``</punct>
<wf cmd=ignore pos=CC>or</wf>
<wf cmd=tag pos=NN>anyone</wf>
<wf cmd=tag pos=RB>else</wf>
<wf cmd=tag pos=RB>locally</wf>
<punct>'</punct>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=done pos=VB lemma=outline wnsn=1 lexsn=2:32:00::>outline</wf>
<wf cmd=ignore pos=PRP$>his</wf>
<wf cmd=tag pos=NN>proposal</wf>
<wf cmd=ignore pos=IN>at</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=tag pos=RBS>earliest</wf>
<wf cmd=tag pos=JJ>possible</wf>
<wf cmd=tag pos=NN>time</wf>
<punct>.</punct>
</s>
Feature Selection

• Use the words in the context with a specific window size
  
  – Collocation
    
    • Consider all words in a window (as well as their POS) and their position:

\[
\{W_{n-3}, P_{n-3}, W_{n-2}, P_{n-2}, W_{n-1}, P_{n-1}, W_{n+1}, P_{n+1}, W_{n+2}, P_{n+2}, W_{n+3}, P_{n+3}\}
\]
Collocation: example

- band:
  "There would be equal access to all currencies financial instruments and financial services dash and no major constitutional change. As realignments become more rare and exchange rates waver in narrower bands the system could evolve into one of fixed exchange rates."

- Window size: +/- 3
- Context: waver in narrower bands the system could
- \{W_{n-3}, P_{n-3}, W_{n-2}, P_{n-2}, W_{n-1}, P_{n-1}, W_{n+1}, P_{n+1}, W_{n+2}, P_{n+2}, W_{n+3}, P_{n+3}\}
- \{waver, NN, in , IN , narrower, JJ, the, DT, system, NN , could, MD\}
Feature Selection

- Use the words in the context with a specific window size
  - Bag-of-word
    - Consider the frequent words regardless their position
    - Derive a set of k most frequent words in the window from the training corpus
    - Represent each word in the data as a k-dimension vector
    - Find the frequency of the selected words in the context of the current observation

\{0, 0, 0, 0, 0, 1, 0, 0, 1, ..., \}
Bag-of-words: example

• band:
  „There would be equal access to all currencies financial instruments and financial services dash and no major constitutional change. As realignments become more rare and exchange rates waver in narrower bands the system could evolve into one of fixed exchange rates.“

• Window size: +/- 3
• Context: waver in narrower bands the system could
• k frequent words for „band“:
  - \{circle, dance, group, jewelery, music, narrow, ring, rubber, wave\}
  - \{0, 0, 0, 0, 0, 1, 0, 0, 1\}
Naïve Bayes Classification

• Choose the best sense $\hat{s}$ out of all possible senses $s_i$ for a feature vector $\vec{f}$ of the word $w$

$$\hat{s} = \arg\max_{s_i} P (s_i | \vec{f})$$

$$\hat{s} = \arg\max_{s_i} \frac{P (\vec{f} | s_i) P (s_i)}{P (\vec{f})}$$

$P (\vec{f})$ has no effect

$$\hat{s} = \arg\max_{s_i} P (\vec{f} | s_i) P (s_i)$$
Naïve Bayes Classification

\[ \hat{s} = \arg\max_{s_i} P(\vec{f} | s_i) P(\ s_i) \]

Likelihood probability

Prior probability

\[ \hat{s} = \arg\max_{s_i} P(s_i) \prod_{j=1}^{m} P( f_j | s_i) \]

\[ P(s_i) = \frac{\#(s_i)}{\#(w)} \]

\#(s_i): number of times the sense \( s_i \) is used for the word \( w \) in the training data

\#(w): the total number of samples for the word \( w \)
Naïve Bayes Classification

\[ \hat{s} = \arg\max_{s_i} P(\vec{f} | s_i) P(s_i) \]

Prior probability

Likelihood probability

\[ \hat{s} = \arg\max_{s_i} P(s_i) \prod_{j=1}^{m} P(f_j | s_i) \]

\[ P(f_j | s_i) = \frac{\#(f_j, s_i)}{\# s_i} \]

\#(f_j, s_i): the number of times the feature f_j occurred for the sense s_i of word w

\#(s_i): the total number of samples of w with the sense s_i in the training data
Semi-supervised Learning

- A small amount of labeled data
- A large amount of unlabeled data

Solution:
- Find the similarity between the labeled and unlabeled data
- Predict the labels of the unlabeled data
Semi-supervised Learning

- For each sense of „band“:
  - Select the most important word which frequently co-occurs with the target word only for this particular sense
    - „play“ (music)
    - „elastic“ (rubber)
    - „spectrum“ (range)
Semi-supervised Learning

- For each sense of „band“:
  - Find the sentences from unlabeled data which contain the target word and the selected word

For example the Jamaican reggae musician Bob Marley and his band The Wailers were known to play the concerts ....

A rubber band, also known as a binder, elastic band, lackey band, laggy band, "gum band", or elastic, is a short length of rubber and latex, elastic in nature and formed ...

The band spectrum is the combination of many different spectral lines

Semi-supervised Learning

- For each sense,
  - Label the sentence with the corresponding sense
  - Add the new labeled sentences to the training data
Word similarity

- Task
  - Find the similarity between two words in a wide range of relations (e.g., relatedness)
  - Different of synonymy
  - Being defined with a score (degree of similarity)
Word similarity

bank $\leftrightarrow$ fund

0.8

car $\leftrightarrow$ bicycle

0.5

car $\leftrightarrow$ gasoline

0.2
Motivation: Information retrieval & Question Answering
Motivation: Document categorization

<table>
<thead>
<tr>
<th>Category</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggested for you</td>
<td></td>
</tr>
<tr>
<td>World</td>
<td></td>
</tr>
<tr>
<td>U.S.</td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td></td>
</tr>
<tr>
<td>Entertainment</td>
<td></td>
</tr>
<tr>
<td>Sports</td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td></td>
</tr>
</tbody>
</table>
Motivation: Machine translation, summarization, text generation

- Substitute of one word for other in some contexts
  - „The bank is on the left bank of the river“
  - „The financial institution is on the left bank of the river“
Motivation: Word clustering
Approaches

• Thesaurus-based
  - Based on their distance in a thesaurus
  - Based on their definition in a thesaurus (gloss)

• Distributional
  - Based on the similarity between their contexts
Thesaurus-based Methods

- Two concepts (sense) are similar if they are “nearby” (short path in the hypernym hierarchy)
Path-base Similarity

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

- \( \text{sim}_{\text{path}}(c_1, c_2) = - \log \text{pathlen}(c_1, c_2) \)

- \( \text{wordsim}(w_1, w_2) = \max_{c_1 \in \text{senses}(w_1), \ c_2 \in \text{senses}(w_2)} \text{sim}(c_1, c_2) \)

when we have no knowledge about the exact sense
(which is the case when processing general text)
Path-base Similarity

- Shortcoming
  - Assumes that each link represents a uniform distance
  - "nickel" to "money" seems closer than "nickel" to "standard"
Path-base Similarity

- Use a metric which represents the cost of each edge independently.
  ⇒ Words connected only through abstract nodes are less similar.
Information Content Similarity

- Assign a probability $P(c)$ to each node of thesaurus
  - $P(c)$ is the probability that a randomly selected word in a corpus is an instance of concept $c$
    - $P(\text{root}) = 1$, since all words are subsumed by the root concept
  - The probability is trained by counting the words in a corpus
  - The lower a concept in the hierarchy, the lower its probability

\[ P(c) = \frac{\sum_{w \in \text{words}(c)} \# w}{N} \]

- $\text{words}(c)$ is the set of words subsumed by concept $c$
- $N$ is the total number of words in the corpus that are available in thesaurus
Information Content Similarity

words(coin) = {nickel, dime}
words(coinage) = {nickel, dime, coin}
words(money) = {budget, fund}
words(medium of exchange) = {nickel, dime, coin, coinage, currency, budget, fund, money}
Information Content Similarity

- Augment each concept in the hierarchy with a probability $P(c)$
Information Content Similarity

- Information Content (self-information):
  \[ IC(c) = - \log P(c) \]
Information Content Similarity

- Lowest common subsumer:
  \[ \text{LCS}(c_1, c_2) = \text{the lowest node that subsumes } c_1 \text{ and } c_2 \]
Information Content Similarity

- Resnik similarity
  - Measure the common amount of information by the information content of the lowest common subsumer of the two concepts

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

\[ \text{sim}_{\text{resnik}}(\text{dime}, \text{nickel}) = - \log P(\text{coin}) \]
Information Content Similarity

- Lin similarity
  - Measure the difference between two concepts in addition to their commonality

\[
\text{similarity}_{\text{LIN}}(c_1, c_2) = \frac{2 \log P(LCS(c_1, c_2))}{\log P(c_1) \log P(c_2)}
\]

\[
\text{similarity}_{\text{LIN}}(\text{dime}, \text{nickel}) = \frac{2 \log P(\text{coin})}{\log P(\text{dime}) \log P(\text{nickel})}
\]
Information Content Similarity

- Jiang-Conrath similarity

\[
similarity_{JC}(c_1, c_2) = \frac{1}{\log P(c_1) + \log P(c_2) - 2 \log P(LCS(c_1, c_2))}
\]
Extended Lesk

- Look at word definitions in thesaurus (gloss)
- Measure the similarity based on the number of common words in their definition
- Add a score of $n^2$ for each $n$-word phrase that occurs in both glosses

$$\text{similarity}_{e\text{Lesk}} = \sum_{r,q \in \text{RELS}} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))$$
Extended Lesk

- Compute overlap for other relations as well (gloss of hypernyms and hyponyms)

\[
\text{similarity}(A,B) = \text{overlap}(\text{gloss}(A),\text{gloss}(B)) \\
+ \text{overlap}(\text{gloss}(\text{hypo}(A)),\text{gloss}(\text{hypo}(B))) \\
+ \text{overlap}(\text{gloss}(A),\text{gloss}(\text{hypo}(B))) \\
+ \text{overlap}(\text{gloss}(\text{hypo}(A)),\text{gloss}(B))
\]
Extended Lesk (example)

- 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
- common phrases: specially prepared and paper

\[
similarity_{eLesk} = 1^2 + 2^2 = 1 + 4 = 5
\]
Available Libraries

- WordNet::Similarity
  - Source:
    - http://wn-similarity.sourceforge.net/
  - Web-based interface:
    - http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi
Thesaurus-based Methods

- **Shortcomings**
  - Many words are missing in thesaurus
  - Only use hyponym info
    - Might useful for nouns, but weak for adjectives, adverbs, and verbs
  - Many languages or domains have no thesaurus

- **Alternative**
  - Distributional methods for word similarity
Distributional Methods

- Use context information to find the similarity between words
- Guess the meaning of a word based on its context
Distributional Methods

• tezgüino?
  - A bottle of tezgüino is on the table
  - Everybody likes tezgüino
  - Tezgüino makes you drunk
  - We make tezgüino out of corn

What is tezgüino?
Context Representations

• Consider a target term t

• Build a vocabulary of M words (\{w_1, w_2, w_3, ..., w_M\})

• Create a vector for t with M features (t = \{f_1, f_2, f_3, ..., f_M\})

• \(f_i\) means the number of times the word \(w_i\) occurs in the context of t
Context Representations

- tezgüino?
  - A bottle of tezgüino is on the table
  - Everybody likes tezgüino
  - Tezgüino makes you drunk
  - We make tezgüino out of corn

- $t = \text{tezgüino}$

  vocab = \{ book, bottle, city, drunk, like, water, \ldots \}

  $t = \{ 0, 1, 0, 1, 1, 0, \ldots \}$
Word vector or Word embeddings

- Frequently used in many neural networks architectures, e.g., morphology, language models
- Available tools: word2vec, GloVe

Figure 4: Bidirectional Multi-Window LSTM model

(https://www3.nd.edu/~dchiang/papers/vaswani-emnlp13.pdf)

(https://www3.nd.edu/~dchiang/papers/window-lstm-morph-segmentation.pdf)
Context Representations

- Term-term matrix
  - The number of times the context word „c“ appear close to the term „t“ within a window

<table>
<thead>
<tr>
<th>term / word</th>
<th>art</th>
<th>boil</th>
<th>data</th>
<th>function</th>
<th>large</th>
<th>sugar</th>
<th>summarize</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Context Representations

- Goal: find a good metric that based on the vectors of these four words shows
  - [apricot, pineapple] and [digital, information] to be highly similar
  - the other four pairs to be less similar

<table>
<thead>
<tr>
<th></th>
<th>art</th>
<th>boil</th>
<th>data</th>
<th>function</th>
<th>large</th>
<th>sugar</th>
<th>summarize</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Distributional similarity

- Size of the context:
  - How are the co-occurrence terms defined? (What is a neighbor?)
    - Window of k words
    - Sentence
    - Paragraph
    - Document
Distributional similarity

- Weights: How are terms weighted?
  - Binary
    - 1, if two words co-occur (no matter how often)
    - 0, otherwise

<table>
<thead>
<tr>
<th>term / word</th>
<th>art</th>
<th>boil</th>
<th>data</th>
<th>function</th>
<th>large</th>
<th>sugar</th>
<th>summarize</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Distributional similarity

- Weights: How are terms weighted?
  - Frequency
    - Number of times two words co-occur with respect to the total size of the corpus

\[
P(t, c) = \frac{\#(t, c)}{N}
\]
# Distributional similarity

<table>
<thead>
<tr>
<th></th>
<th>art</th>
<th>boil</th>
<th>data</th>
<th>function</th>
<th>large</th>
<th>sugar</th>
<th>summarize</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>art</th>
<th>boil</th>
<th>data</th>
<th>function</th>
<th>large</th>
<th>sugar</th>
<th>summarize</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.071</td>
<td>0</td>
<td>0.035</td>
</tr>
<tr>
<td>pineapple</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0.035</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>0.035</td>
<td>0.035</td>
<td>0.107</td>
<td>0.035</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>0.321</td>
<td>0.035</td>
<td>0.035</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0.071</td>
<td>0</td>
</tr>
</tbody>
</table>
Distributional similarity

• Weights: How are terms weighted?
  – Pointwise Mutual information
    • Number of times two words co-occur, compared with what we would expect if they were independent

\[ PMI(t, c) = \log \frac{P(t, c)}{P(t)P(c)} \]
Pointwise Mutual Information

<table>
<thead>
<tr>
<th></th>
<th>art</th>
<th>boil</th>
<th>data</th>
<th>function</th>
<th>large</th>
<th>sugar</th>
<th>summarize</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.071</td>
<td>0</td>
<td>0.035</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0.035</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.107</td>
<td>0.035</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>0</td>
<td>0.321</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0.071</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
P(\text{digital, summarize}) = 0.035
P(\text{information, function}) = 0.035
\]

\[
P(\text{digital, summarize}) = P(\text{information, function})
\]

\[
\text{PMI(\text{digital, summarize}) =?}
\text{PMI(\text{information, function}) =?}
\]
### Pointwise Mutual Information

<table>
<thead>
<tr>
<th></th>
<th>art</th>
<th>boil</th>
<th>data</th>
<th>function</th>
<th>large</th>
<th>sugar</th>
<th>summarize</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.071</td>
<td>0</td>
<td>0.035</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0.035</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>0</td>
<td>0.035</td>
<td>0.107</td>
<td>0.035</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>0</td>
<td>0.321</td>
<td>0.035</td>
<td>0.035</td>
<td>0</td>
<td>0.071</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
P(\text{digital, summarize}) = 0.035
\]
\[
P(\text{information, function}) = 0.035
\]
\[
P(\text{digital}) = 0.212
\]
\[
P(\text{summarize}) = 0.106
\]
\[
P(\text{function}) = 0.142
\]
\[
P(\text{information}) = 0.462
\]

\[
\text{PMI}(\text{digital, summarize}) = \frac{P(\text{digital, summarize})}{P(\text{digital}) \cdot P(\text{summarize})} = \frac{0.035}{0.212 \cdot 0.106} = 1.557
\]

\[
\text{PMI}(\text{information, function}) = \frac{P(\text{information, function})}{P(\text{information}) \cdot P(\text{function})} = \frac{0.035}{0.462 \cdot 0.142} = 0.533
\]

\[
\text{PMI}(\text{digital, summarize}) > \text{PMI}(\text{information, function})
\]
Distributional similarity

• Weights: How are terms weighted?
  - t-test statistic
    • How much more frequent the association is than by chance?

\[
t - \text{test}(t, c) = \frac{P(t, c) - P(t)P(c)}{\sqrt{P(t)P(c)}}
\]
Distributional similarity

- Vector similarity: Which vector distance metric should be used?
  - Cosine
    \[
    \text{similarity}_{cosine}(\vec{v}, \vec{w}) = \frac{\sum_i v_i \times w_i}{\sqrt{\sum_i v_i^2} \sqrt{\sum_i w_i^2}}
    \]
  - Jaccard, Tanimoto, min/max
    \[
    \text{similarity}_{jaccard}(\vec{v}, \vec{w}) = \frac{\sum_i \min(v_i, w_i)}{\sum_i \max(v_i, w_i)}
    \]
  - Dice
    \[
    \text{similarity}_{dice}(\vec{v}, \vec{w}) = \frac{2 \cdot \sum_i \min(v_i, w_i)}{\sum_i (v_i + w_i)}
    \]
Summary

- Semantics
  - Senses, relations
- Word disambiguation
  - Thesaurus-based, (semi-) supervised learning
- Word similarity
  - Thesaurus-based
  - Distributional
    - Features, weighting schemes and similarity algorithms
Further Reading

• Speech and Language Processing (3rd edition draft)
  - [https://web.stanford.edu/~jurafsky/slp3/](https://web.stanford.edu/~jurafsky/slp3/)
  - Chapters 15 and 17