Natural Language Processing
Lexical semantics

Sofia Serrano
sofias6@cs.washington.edu

Credit to Yulia Tsvetkov and Noah Smith for slides
Announcements

- A2 goes out on gitlab sometime today
- Reminder: the last day/time to submit A1 for credit (by tagging on GitLab) is **today at 11:59pm**, if you use the max number of late days allowed for this assignment
- Quiz 3 will be released on Canvas on Wednesday (2/1)
  - Available Wednesday 2:20pm through Thursday 2:20pm
  - 5 questions, 10 minutes
  - Will cover material from Wednesday, Friday, and today (so, language modeling and the first part of lexical semantics)
- We have a google calendar for the course now (embedded on course website)
- Midterm course eval form (online) going out sometime in the next couple of days—please let us know how we’re doing!
Language models: Conclusion (for now)
Which of the material from Wednesday and Friday only applied to *simple* language models?

Just the n-gram material! *(Since the Markovian assumption is, well, false in a language context)*

For current state-of-the-art neural language models, all of the following still apply:

- Task definition of language modeling
- Evaluation via perplexity
- Vocabulary creation considerations (e.g., `<UNK>ing`, switching to character-level modeling, or using byte-pair encoding)

Language modeling will make reappearances later in the course...
Lexical semantics
Problems with discrete representations

- Too coarse
  - expert $\leftrightarrow$ skillful
- Sparse
  - wicked, badass, ninja
- Hard to compute word relationships
  
  \[
  \text{expert} \quad [0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]
  \]
  \[
  \text{skillful} \quad [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]
  \]
- dimensionality: PTB: 50K, Google1T 13M
Lexical semantics: what do words mean?

- N-gram or text classification methods we've seen so far
  - Words are just strings (or indices $w_i$ in a vocabulary list)
  - That's not very satisfactory!
How have people thought about breaking down the meaning of a word?
Desiderata

What should a theory of word meaning do for us?

Let's look at some desiderata from **lexical semantics**, the linguistic study of word meaning.
Lexical semantics

- How should we represent the meaning of the word?
  - Words, lemmas, senses, definitions

http://www.oed.com/
A sense or "concept" is the meaning component of a word. Lemmas can be polysemous (have multiple senses).
Relation: synonymity

- Synonyms have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / H₂O
The Linguistic Principle of Contrast

Difference in form → difference in meaning

- Note that there are probably 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.
    - Water / H2O in a surfing guide?
    - my big sister ≠ my large sister
Relation: antonymy

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:
  - rise/fall, up/down
Relation: similarity

Words with similar meanings.

- Not synonyms, but sharing some element of meaning
  - car, bicycle
  - cow, horse
Ask humans how similar two words are

<table>
<thead>
<tr>
<th>word1</th>
<th>word2</th>
<th>similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanish</td>
<td>disappear</td>
<td>9.8</td>
</tr>
<tr>
<td>behave</td>
<td>obey</td>
<td>7.3</td>
</tr>
<tr>
<td>belief</td>
<td>impression</td>
<td>5.95</td>
</tr>
<tr>
<td>muscle</td>
<td>bone</td>
<td>3.65</td>
</tr>
<tr>
<td>modest</td>
<td>flexible</td>
<td>0.98</td>
</tr>
<tr>
<td>hole</td>
<td>agreement</td>
<td>0.3</td>
</tr>
</tbody>
</table>

SimLex-999 dataset (Hill et al., 2015)
Relation: word relatedness

Also called "word association"

- Words be related in any way, perhaps via a semantic frame or field
  - car, bicycle: similar
  - car, gasoline: related, not similar
Semantic field

Words that

- cover a particular semantic domain
- bear structured relations with each other

hospitals
    surgeon, scalpel, nurse, anaesthetic, hospital

restaurants
    waiter, menu, plate, food, menu, chef

houses
    door, roof, kitchen, family, bed
Relation: superordinate/ subordinate

- One sense is a subordinate (hyponym) of another if the first sense is more specific, denoting a subclass of the other
  - car is a subordinate of vehicle
  - mango is a subordinate of fruit

- Conversely superordinate (hypernym)
  - vehicle is a superordinate of car
  - fruit is a superordinate of mango
Taxonomy

- **Superordinate**: furniture
- **Basic**: chair, lamp, table
- **Subordinate**: office chair, piano chair, rocking chair, torchiere, desk lamp, end table, coffee table
Lexical semantics

- How should we represent the meaning of the word?
  - Dictionary definition
  - Lemma and wordforms
  - Senses
  - Relationships between words or senses
  - Taxonomic relationships
  - Word similarity, word relatedness
  - Connotation and sentiment
  - Semantic frames and roles
Lexical semantics

- How should we represent the meaning of the word?
  - Dictionary definition
  - Lemma and wordforms
  - Senses
  - Relationships between words or senses
  - Taxonomic relationships
  - Word similarity, word relatedness
  - Connotation and sentiment
  - Semantic frames and roles
    - John hit Bill
    - Bill was hit by John
Lexical semantics

- How should we represent the meaning of the word?
  - Dictionary definition
  - Lemma and wordforms
  - Senses
  - Relationships between words or senses
  - Taxonomic relationships
  - Word similarity, word relatedness
  - Semantic frames and roles
  - Connotation and sentiment
    - *valence*: the pleasantness of the stimulus
    - *arousal*: the intensity of emotion
    - *dominance*: the degree of control exerted by the stimulus

<table>
<thead>
<tr>
<th>Word</th>
<th>Valence</th>
<th>Arousal</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>courageous</td>
<td>8.05</td>
<td>5.5</td>
<td>7.38</td>
</tr>
<tr>
<td>music</td>
<td>7.67</td>
<td>5.57</td>
<td>6.5</td>
</tr>
<tr>
<td>heartbeat</td>
<td>2.45</td>
<td>5.65</td>
<td>3.58</td>
</tr>
<tr>
<td>cub</td>
<td>6.71</td>
<td>3.95</td>
<td>4.24</td>
</tr>
<tr>
<td>life</td>
<td>6.68</td>
<td>5.59</td>
<td>5.89</td>
</tr>
</tbody>
</table>
Electronic Dictionaries

WordNet

https://wordnet.princeton.edu/
Electronic Dictionaries

WordNet

```python
from nltk.corpus import wordnet as wn
panda = wn.synset('panda.n.01')
hyper = lambda s: s.hypernyms()
list(panda.closure(hyper))
```

- Synset('procyonid.n.01')
- Synset('carnivore.n.01')
- Synset('placental.n.01')
- Synset('mammal.n.01')
- Synset('vertebrate.n.01')
- Synset('chordate.n.01')
- Synset('animal.n.01')
- Synset('organism.n.01')
- Synset('living_thing.n.01')
- Synset('whole.n.02')
- Synset('object.n.01')
- Synset('physical_entity.n.01')
- Synset('entity.n.01')
So what do we do?
Let's take a look at this list again

- How should we represent the meaning of the word?
  - Dictionary definition
  - Lemma and wordforms
  - Senses
  - Relationships between words or senses
  - Taxonomic relationships
  - Word similarity, word relatedness
  - Connotation and sentiment
  - Semantic frames and roles

Note: a lot of these are related to the contexts in which a word appears!
Distributional hypothesis

“The meaning of a word is its use in the language”  
[Wittgenstein PI 43]

“You shall know a word by the company it keeps”  
[Firth 1957]

If A and B have almost identical environments we say that they are synonyms.  
[Harris 1954]
Example

What does ongchoi mean?
Example

- Suppose you see these sentences:
  - Ongchoi is delicious *sautéed with garlic*.
  - Ongchoi is superb *over rice*
  - Ongchoi *leaves* with salty sauces

- And you've also seen these:
  - ...spinach *sautéed with garlic over rice*
  - Chard stems and *leaves* are delicious
  - Collard greens and other *salty* leafy greens
Ongchoi: Ipomoea aquatica "Water Spinach"

Ongchoi is a leafy green like spinach, chard, or collard greens.
Model of meaning focusing on similarity

- Each word = a vector
  - not just “word” or word45.
  - similar words are “nearby in space”
  - We build this space automatically by seeing which words are nearby in text
We define meaning of a word as a vector

- Called an "embedding" because it's embedded into a space
- The standard way to represent meaning in NLP

Every modern NLP algorithm uses embeddings as the representation of word meaning
Intuition: why vectors?

Consider sentiment analysis:

- **With words**, a feature is a word identity
  - Feature 5: 'The previous word was "terrible"'
  - requires *exact same word* to be in training and test

- **With embeddings:**
  - Feature is a word vector
  - 'The previous word was vector [35,22,17...]
  - Now in the test set we might see a similar vector [34,21,14]
  - Whether we can pull this new vector together depends on the method we use
  - Gives us a chance to generalize to similar but unseen-in-training words!
There are many kinds of embeddings

- **Count-based**
  - Words are represented by a simple function of the counts of nearby words

- **Class-based**
  - Representation is created through hierarchical clustering, Brown clusters

- **Distributed prediction-based (type) embeddings**
  - Representation is created by training a classifier to distinguish nearby and far-away words: word2vec, fasttext

- **Distributed contextual (token) embeddings from language models**
  - ELMo, BERT
We'll discuss different kinds of embeddings

- **Sparse vectors**
  - Like TF-IDF: Information Retrieval workhorse!
  - Common baseline models
  - Words are represented by (a simple function of) the counts of nearby words

- **Dense vectors**
  - Dimensionality reduction
    - Latent Semantic Analysis (LSA)
  - Word2vec
    - Representation is created by training a classifier to predict whether a word is likely to appear nearby
    - Later we'll discuss extensions called contextual embeddings
Sparse vectors
Term-document matrix

<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>0</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>clown</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Context = appearing in the same document.
## Term-document matrix

<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>0</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>clown</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Each document is represented by a vector of words
Vectors are the basis of information retrieval

<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>0</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>soldier</td>
<td>2</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>clown</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- Vectors are similar for the two comedies
- Different than the history
- Comedies have more fools and wit and fewer battles.
Visualizing Document Vectors

Henry V [4, 13]  
Julius Caesar [1, 7]  
As You Like It [36, 1]  
Twelfth Night [58, 0]
**Words can be vectors too**

<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>0</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>good</td>
<td>114</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>clown</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- battle is "the kind of word that occurs in Julius Caesar and Henry V"
- fool is "the kind of word that occurs in comedies, especially Twelfth Night"
More common: word-word matrix ("term-context matrix")

<table>
<thead>
<tr>
<th></th>
<th>knife</th>
<th>dog</th>
<th>sword</th>
<th>love</th>
<th>like</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>dog</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>sword</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>love</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>like</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

- Two words are “similar” in meaning if their context vectors are similar
  - Similarity == relatedness
Term-context matrix

Two words are similar in meaning if their context vectors are similar.

- Football is traditionally followed by cherry pie, a traditional dessert.
- Often mixed, such as strawberry rhubarb pie. Apple pie.
- Computer peripherals and personal assistants. These devices usually:
  - Digital information available on the internet

<table>
<thead>
<tr>
<th></th>
<th>aardvark</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>cherry</td>
<td>0</td>
<td>...</td>
<td>2</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>strawberry</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>...</td>
<td>1670</td>
<td>1683</td>
<td>85</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>...</td>
<td>3325</td>
<td>3982</td>
<td>378</td>
</tr>
</tbody>
</table>
## Count-based representations

<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>0</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>good</td>
<td>114</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>wit</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- **Counts: term-frequency**
  - remove stop words
  - use $\log_{10}(tf)$
  - normalize by document length
But raw frequency is a bad representation

- The co-occurrence matrices we have seen represent each cell by word frequencies
- Frequency is clearly useful; if sugar appears in a lot of the same context as cake, that's useful information
- But overly frequent words like the, it, or they are not very informative about the context
- How can we balance these two conflicting constraints?
Two common solutions for word weighting

tf-idf: tf-idf value for word $t$ in document $d$:

$$w_{t,d} = tf_{t,d} \times idf_t$$

Words like “the” or “it” have very low idf

PMI: Pointwise mutual information

$$PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

See if words like “good” appear more often with “great” than we would expect by chance
What to do with words that are evenly distributed across many documents?

\[ tf_{t,d} = \log_{10}(\text{count}(t,d) + 1) \]

\[ \text{idf}_i = \log \left( \frac{N}{\text{df}_i} \right) \]

Total # of docs in collection

# of docs that have word i

Words like "the" or "good" have very low idf

\[ w_{t,d} = tf_{t,d} \times \text{idf}_t \]
In word-context matrix
Do words \( w \) and \( c \) co-occur more than if they were independent?

\[
\text{PMI}(w, c) = \log_2 \frac{P(w, c)}{P(w)P(c)}
\]

\[
\text{PPMI}(w, c) = \max(\log_2 \frac{P(w, c)}{P(w)P(c)}, 0)
\]

PMI is biased toward infrequent events
- Very rare words have very high PMI values
- Give rare words slightly higher probabilities \( \alpha = 0.75 \)

\[
\text{PPMI}_\alpha(w, c) = \max(\log_2 \frac{P(w, c)}{P(w)P_\alpha(c)}, 0)
\]

\[
P_\alpha(c) = \frac{\text{count}(c)^\alpha}{\sum_c \text{count}(c)^\alpha}
\]
<table>
<thead>
<tr>
<th># name</th>
<th>formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Joint probability</td>
<td>$p(x</td>
</tr>
<tr>
<td>2. Conditional probability</td>
<td>$p(y</td>
</tr>
<tr>
<td>3. Reverse cond. probability</td>
<td>$p(x</td>
</tr>
<tr>
<td>4. Pointwise mutual inf. (MI)</td>
<td>$\log \frac{p(x</td>
</tr>
<tr>
<td>5. Mutual dependency (MD)</td>
<td>$\log \frac{p(x,y)}{p(x)p(y)}$</td>
</tr>
<tr>
<td>6. Log frequency biased MD</td>
<td>$\log x + \log y$</td>
</tr>
<tr>
<td>7. Normalized expectation</td>
<td>$\frac{\log x + \log y}{x + y}$</td>
</tr>
<tr>
<td>8. Mutual information</td>
<td>$\sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$</td>
</tr>
<tr>
<td>9. Salience</td>
<td>$\log p(x</td>
</tr>
<tr>
<td>10. Pearson's $\chi^2$ test</td>
<td>$\sum (f(x,y) - \hat{f}(x,y))^2 / \hat{f}(x,y)$</td>
</tr>
<tr>
<td>11. Fisher's exact test</td>
<td>$N! / (n!m!) \prod (x+y)!$</td>
</tr>
<tr>
<td>12. t test</td>
<td>$\frac{\hat{f}(x,y) - \hat{f}(x,y)}{\sqrt{\hat{f}(x,y) \hat{f}(x,y)}}$</td>
</tr>
<tr>
<td>13. z score</td>
<td>$\sqrt{\hat{f}(x,y) \hat{f}(x,y)}$</td>
</tr>
<tr>
<td>14. Poisson significance</td>
<td>$\frac{\log N}{N}$</td>
</tr>
<tr>
<td>15. Log likelihood ratio</td>
<td>$-2 \sum f(x,y) \log \frac{\hat{f}(x,y)}{\hat{f}(x,y)}$</td>
</tr>
<tr>
<td>16. Squared log likelihood ratio</td>
<td>$-2 \sum f(x,y)^2 \log \frac{\hat{f}(x,y)^2}{\hat{f}(x,y)^2}$</td>
</tr>
<tr>
<td>17. Russel-Rao</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>18. Sokal-Michener</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>19. Rogers-Tanimoto</td>
<td>$\frac{(2a-c)(a+d-c)(b+d-c)}{a+b+c+d}$</td>
</tr>
<tr>
<td>20. Hamann</td>
<td>$\frac{a}{a+b+c}$</td>
</tr>
<tr>
<td>21. Third Sokal-Sneath</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>22. Jaccard</td>
<td>$\frac{a}{a+b+c}$</td>
</tr>
<tr>
<td>23. First Kulczynsky</td>
<td>$\frac{a}{a+b}$</td>
</tr>
<tr>
<td>24. Second Sokal-Sneath</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>25. Second Kulczynski</td>
<td>$\frac{a}{a+b}$</td>
</tr>
<tr>
<td>26. Fourth Sokal-Sneath</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>27. Odds ratio</td>
<td>$\frac{a+b}{a+c}$</td>
</tr>
<tr>
<td>28. Yule's $\omega$</td>
<td>$\frac{a}{a+b+c}$</td>
</tr>
<tr>
<td>29. Yule's $Q$</td>
<td>$\frac{a}{a+b+c}$</td>
</tr>
<tr>
<td>30. Driver-Kroebner</td>
<td>$\sqrt{\frac{a}{a+b+c}}$</td>
</tr>
<tr>
<td>31. Fifth Sokal-Sneath</td>
<td>$(\sqrt{\frac{a+b}{a+c+d} \cdot \frac{a+b}{b+c+d}})$</td>
</tr>
<tr>
<td>32. Pearson</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>33. Baroni-Urni</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>34. Braun-Blanquet</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>35. Simpson</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>36. Michael</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>37. Mountford</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>38. Fager</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>39. Unigram subtuple</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>40. U cost</td>
<td>$\log(1 + \frac{a+b+c-d}{a+b+c})$</td>
</tr>
<tr>
<td>41. S cost</td>
<td>$\log(1 + \frac{a+b+c-d}{a+b+c})$</td>
</tr>
<tr>
<td>42. R cost</td>
<td>$\log(1 + \frac{a+b+c-d}{a+b+c})$</td>
</tr>
<tr>
<td>43. T combined cost</td>
<td>$\sqrt{U \times S \times R}$</td>
</tr>
<tr>
<td>44. Phi</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>45. Kappa</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>46. J measure</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>47. Gini index</td>
<td>$\frac{a+b+c-d}{a+b+c}$</td>
</tr>
<tr>
<td>48. Confidence</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>49. Laplace</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>50. Conviction</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>51. Platersky-Shapiro</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>52. Certainty factor</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>53. Added value (AV)</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>54. Collective strength</td>
<td>$\max(p(x</td>
</tr>
<tr>
<td>55. Klossgen</td>
<td>$\sqrt{p(x</td>
</tr>
</tbody>
</table>
Dense vectors (part 1)
Dimensionality Reduction

  - High dimensionality of word--document matrix
    - Sparsity
    - The order of rows and columns doesn’t matter
- Goal:
  - Good similarity measure for words or documents
  - Dense representation
- Sparse vs Dense vectors
  - Short vectors may be easier to use as features in machine learning (less weights to tune)
  - Dense vectors may generalize better than storing explicit counts
    - They may do better at capturing synonymy
    - In practice, they work better
Singular Value Decomposition (SVD)

- Solution idea:
  - Find a projection into a low-dimensional space (~300 dim)
  - That gives us a best separation between features

\[ \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \]

- \( \mathbf{A} \): \( m \times n \) matrix of documents
- \( \mathbf{U} \): \( m \times r \) matrix, orthogonal
- \( \Sigma \): \( r \times r \) diagonal matrix, sorted
- \( \mathbf{V}^T \): \( r \times n \) matrix, orthogonal
Truncated SVD

We can approximate the full matrix by only considering the leftmost $k$ terms in the diagonal matrix (the $k$ largest singular values).

$$\begin{align*}
A_{m \times n} \approx U_{m \times k} \Sigma_{k \times k} V_{k \times n}^T \\
k \ll m, n
\end{align*}$$
Latent Semantic Analysis

<table>
<thead>
<tr>
<th>#0</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
</tr>
</thead>
<tbody>
<tr>
<td>we</td>
<td>music</td>
<td>company</td>
<td>how</td>
<td>program</td>
<td>10</td>
</tr>
<tr>
<td>said</td>
<td>film</td>
<td>mr</td>
<td>what</td>
<td>project</td>
<td>30</td>
</tr>
<tr>
<td>have</td>
<td>theater</td>
<td>its</td>
<td>about</td>
<td>russian</td>
<td>11</td>
</tr>
<tr>
<td>they</td>
<td>mr</td>
<td>inc</td>
<td>their</td>
<td>space</td>
<td>12</td>
</tr>
<tr>
<td>not</td>
<td>this</td>
<td>stock</td>
<td>or</td>
<td>russia</td>
<td>15</td>
</tr>
<tr>
<td>but</td>
<td>who</td>
<td>companies</td>
<td>this</td>
<td>center</td>
<td>13</td>
</tr>
<tr>
<td>be</td>
<td>movie</td>
<td>sales</td>
<td>are</td>
<td>programs</td>
<td>14</td>
</tr>
<tr>
<td>do</td>
<td>which</td>
<td>shares</td>
<td>history</td>
<td>clark</td>
<td>20</td>
</tr>
<tr>
<td>he</td>
<td>show</td>
<td>said</td>
<td>be</td>
<td>aircraft</td>
<td>sept</td>
</tr>
<tr>
<td>this</td>
<td>about</td>
<td>business</td>
<td>social</td>
<td>ballet</td>
<td>16</td>
</tr>
<tr>
<td>there</td>
<td>dance</td>
<td>share</td>
<td>these</td>
<td>its</td>
<td>25</td>
</tr>
<tr>
<td>you</td>
<td>its</td>
<td>chief</td>
<td>other</td>
<td>projects</td>
<td>17</td>
</tr>
<tr>
<td>are</td>
<td>disney</td>
<td>executive</td>
<td>research</td>
<td>orchestra</td>
<td>18</td>
</tr>
<tr>
<td>what</td>
<td>play</td>
<td>president</td>
<td>writes</td>
<td>development</td>
<td>19</td>
</tr>
<tr>
<td>if</td>
<td>production</td>
<td>group</td>
<td>language</td>
<td>work</td>
<td>21</td>
</tr>
</tbody>
</table>

*(Deerwester et al., 1990)*
How do we tell whether a set of word embeddings is any good?
The dot product between two vectors is a scalar:

\[
\text{dot product}(v, w) = v \cdot w = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N
\]

- The dot product tends to be high when the two vectors have large values in the same dimensions
- Dot product can thus be a useful similarity metric between vectors
Problem with raw dot-product

- Dot product favors long vectors
  - Dot product is higher if a vector is longer (has higher values in many dimension)
  - Vector length:
    \[
    |v| = \sqrt{\sum_{i=1}^{N} v_i^2}
    \]

- Frequent words (of, the, you) have long vectors (since they occur many times with other words).
  - So dot product overly favors frequent words
Alternative: cosine for computing word similarity

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

Based on the definition of the dot product between two vectors \( \mathbf{a} \) and \( \mathbf{b} \)

\[
\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta
\]

\[
\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| |\mathbf{b}|} = \cos \theta
\]
Cosine as a similarity metric

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

- But since raw frequency values are non-negative, the cosine for term-term matrix vectors ranges from 0–1
Cosine examples

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

<table>
<thead>
<tr>
<th></th>
<th>pie</th>
<th>data</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>cherry</td>
<td>442</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>digital</td>
<td>114</td>
<td>80</td>
<td>62</td>
</tr>
<tr>
<td>information</td>
<td>36</td>
<td>58</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\cos(\text{cherry, information}) = \frac{442 \cdot 5 + 8 \cdot 3982 + 2 \cdot 3325}{\sqrt{442^2 + 8^2 + 2^2 \sqrt{5^2 + 3982^2 + 3325^2}}} = .017
\]

\[
\cos(\text{digital, information}) = \frac{5 \cdot 5 + 1683 \cdot 3982 + 1670 \cdot 3325}{\sqrt{5^2 + 1683^2 + 1670^2 \sqrt{5^2 + 3982^2 + 3325^2}}} = .996
\]
Visualizing angles

Dimension 1: ‘pie’

Dimension 2: ‘computer’

cherry

digital

information
## Evaluation

- Intrinsic
- Extrinsic
- Qualitative

<table>
<thead>
<tr>
<th>WORD</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>...</th>
<th>d50</th>
</tr>
</thead>
<tbody>
<tr>
<td>summer</td>
<td>0.12</td>
<td>0.21</td>
<td>0.07</td>
<td>0.25</td>
<td>0.33</td>
<td>...</td>
<td>0.51</td>
</tr>
<tr>
<td>spring</td>
<td>0.19</td>
<td>0.57</td>
<td>0.99</td>
<td>0.30</td>
<td>0.02</td>
<td>...</td>
<td>0.73</td>
</tr>
<tr>
<td>fall</td>
<td>0.53</td>
<td>0.77</td>
<td>0.43</td>
<td>0.20</td>
<td>0.29</td>
<td>...</td>
<td>0.85</td>
</tr>
<tr>
<td>light</td>
<td>0.00</td>
<td>0.68</td>
<td>0.84</td>
<td>0.45</td>
<td>0.11</td>
<td>...</td>
<td>0.03</td>
</tr>
<tr>
<td>clear</td>
<td>0.27</td>
<td>0.50</td>
<td>0.21</td>
<td>0.56</td>
<td>0.25</td>
<td>...</td>
<td>0.32</td>
</tr>
<tr>
<td>blizzard</td>
<td>0.15</td>
<td>0.05</td>
<td>0.64</td>
<td>0.17</td>
<td>0.99</td>
<td>...</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Extrinsic Evaluation

- Chunking
- POS tagging
- Parsing
- MT
- SRL
- Topic categorization
- Sentiment analysis
- Metaphor detection
- etc.
Intrinsic Evaluation

- **WS-353** ([Finkelstein et al. ‘02](#))
- **MEN-3k** ([Bruni et al. ‘12](#))
- **SimLex-999 dataset** ([Hill et al., 2015](#))

<table>
<thead>
<tr>
<th>word1</th>
<th>word2</th>
<th>similarity (humans)</th>
<th>similarity (embeddings)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanish</td>
<td>disappear</td>
<td>9.8</td>
<td>1.1</td>
</tr>
<tr>
<td>behave</td>
<td>obey</td>
<td>7.3</td>
<td>0.5</td>
</tr>
<tr>
<td>belief</td>
<td>impression</td>
<td>5.95</td>
<td>0.3</td>
</tr>
<tr>
<td>muscle</td>
<td>bone</td>
<td>3.65</td>
<td>1.7</td>
</tr>
<tr>
<td>modest</td>
<td>flexible</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>hole</td>
<td>agreement</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Spearman's rho (human ranks, model ranks)
Visualisation

- Visualizing Data using t-SNE (van der Maaten & Hinton '08)

(Faruqui et al., 2014)

Figure 6.5: Monolingual (top) and multilingual (bottom; marked with apostrophe) word projections of the antonyms (shown in red) and synonyms of “beautiful”.

- Visualizing Data using t-SNE (van der Maaten & Hinton '08)