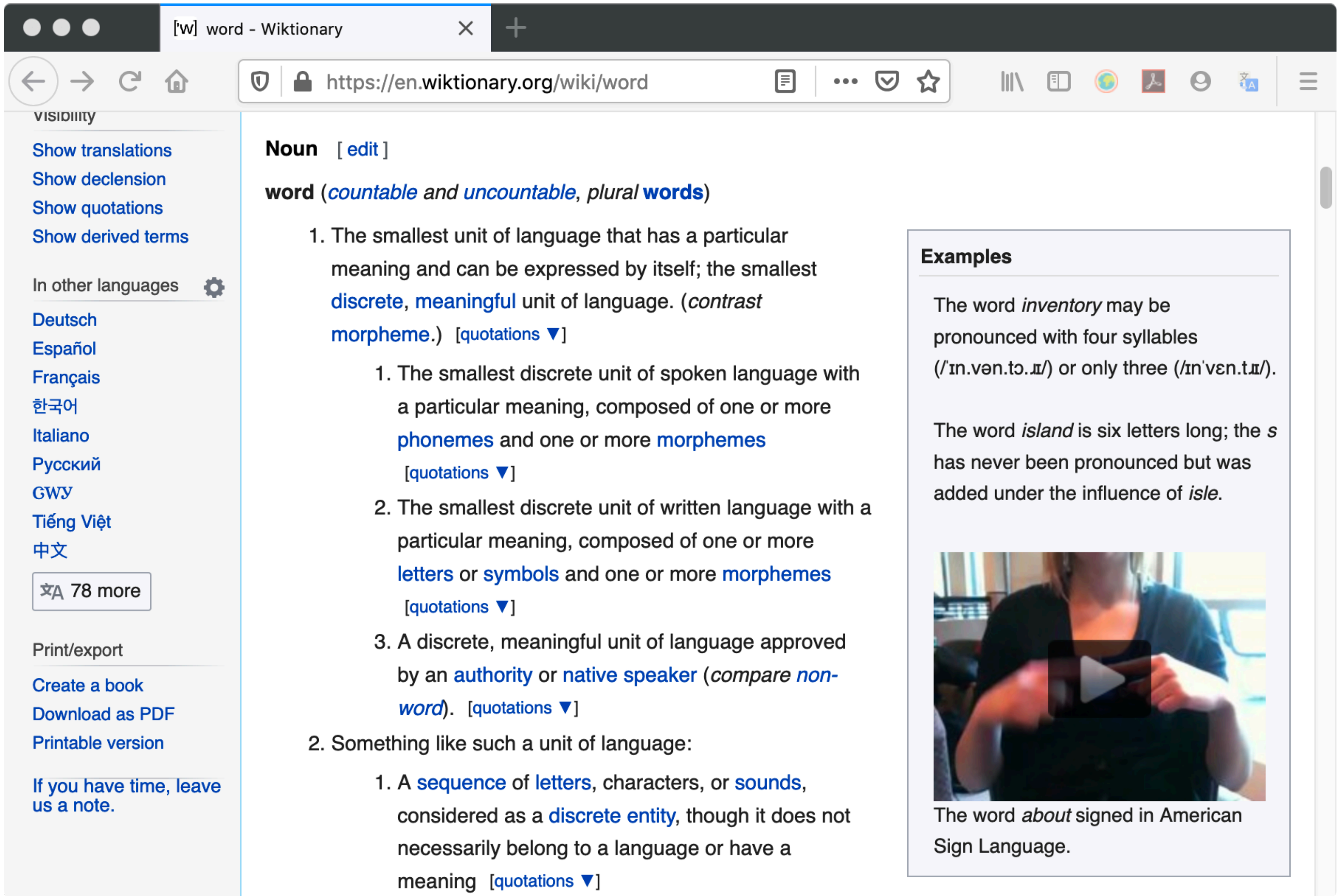
Lecture 37: word2vec and word similarity

Mark Hasegawa-Johnson

Outline

- What is a word? Lemmas, wordforms, and word sense
- Synonymy, similarity, and relatedness
- Word2vec
- Visualizations
- Bias

What is a word?



[w] word - Wiktionary

https://en.wiktionary.org/wiki/word

Visibility

Show translations
Show declension
Show quotations
Show derived terms

In other languages

Deutsch
Español
Français
한국어
Italiano
Русский
ᏣᎳᎩ
Tiếng Việt
中文

78 more

Print/export

Create a book
Download as PDF
Printable version

If you have time, leave us a note.

Noun [edit]

word (*countable and uncountable*, *plural* **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast morpheme*.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

Examples

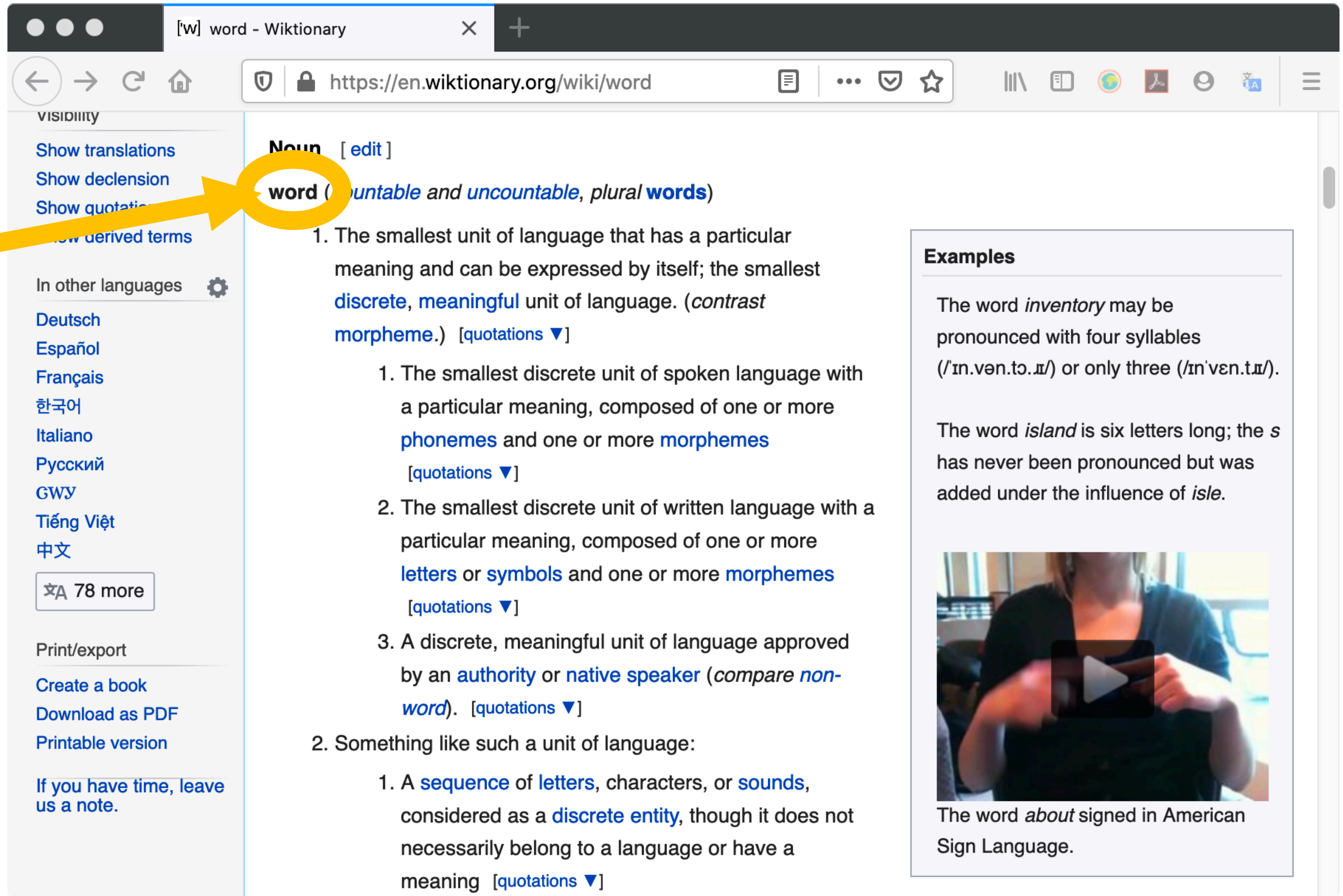
The word *inventory* may be pronounced with four syllables (/ˈɪn.vən.tɔ.ɹi/) or only three (/ɪnˈvɛn.tɹi/).

The word *island* is six letters long; the *s* has never been pronounced but was added under the influence of *isle*.

The word *about* signed in American Sign Language.

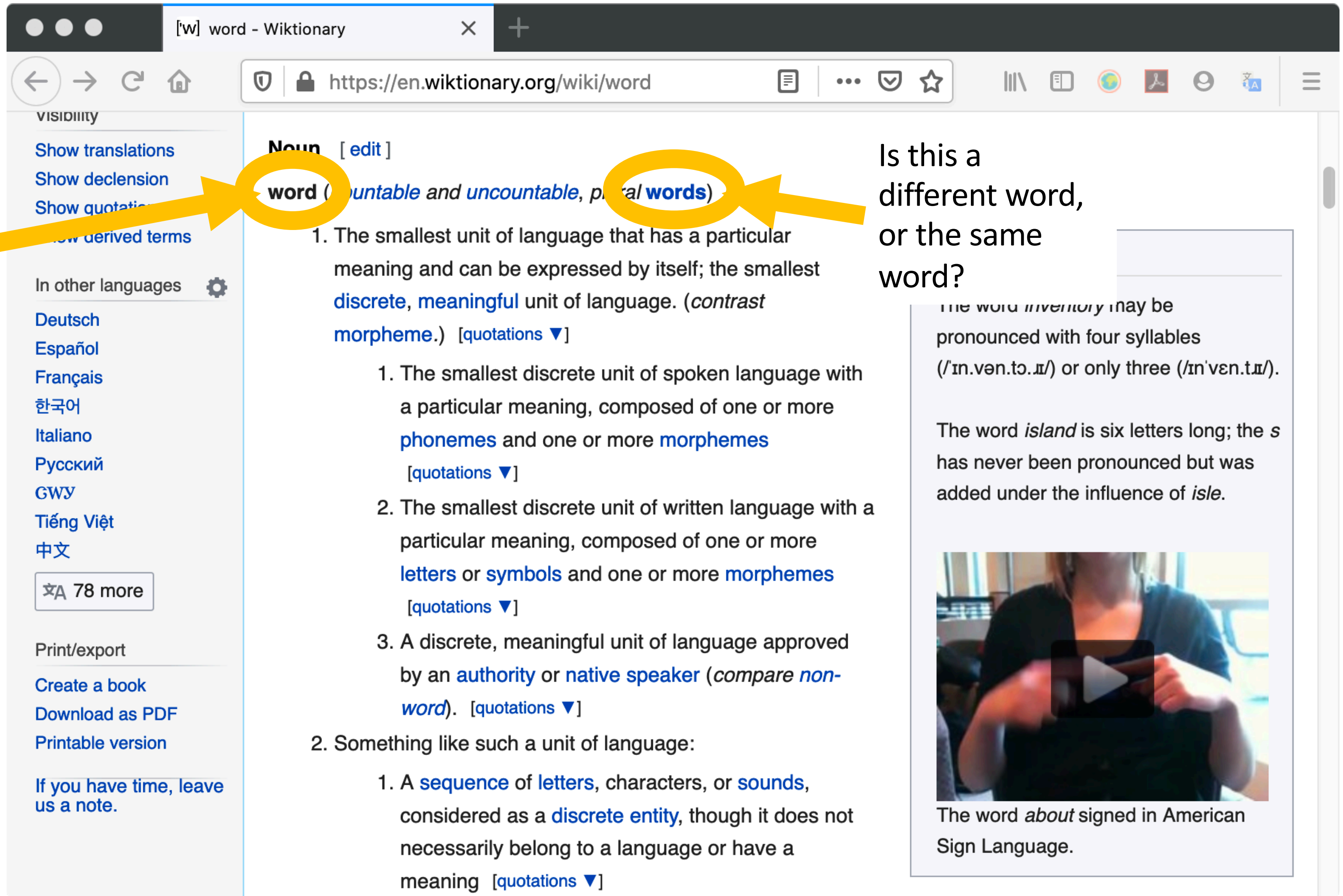
What is a word?

Is this a word?



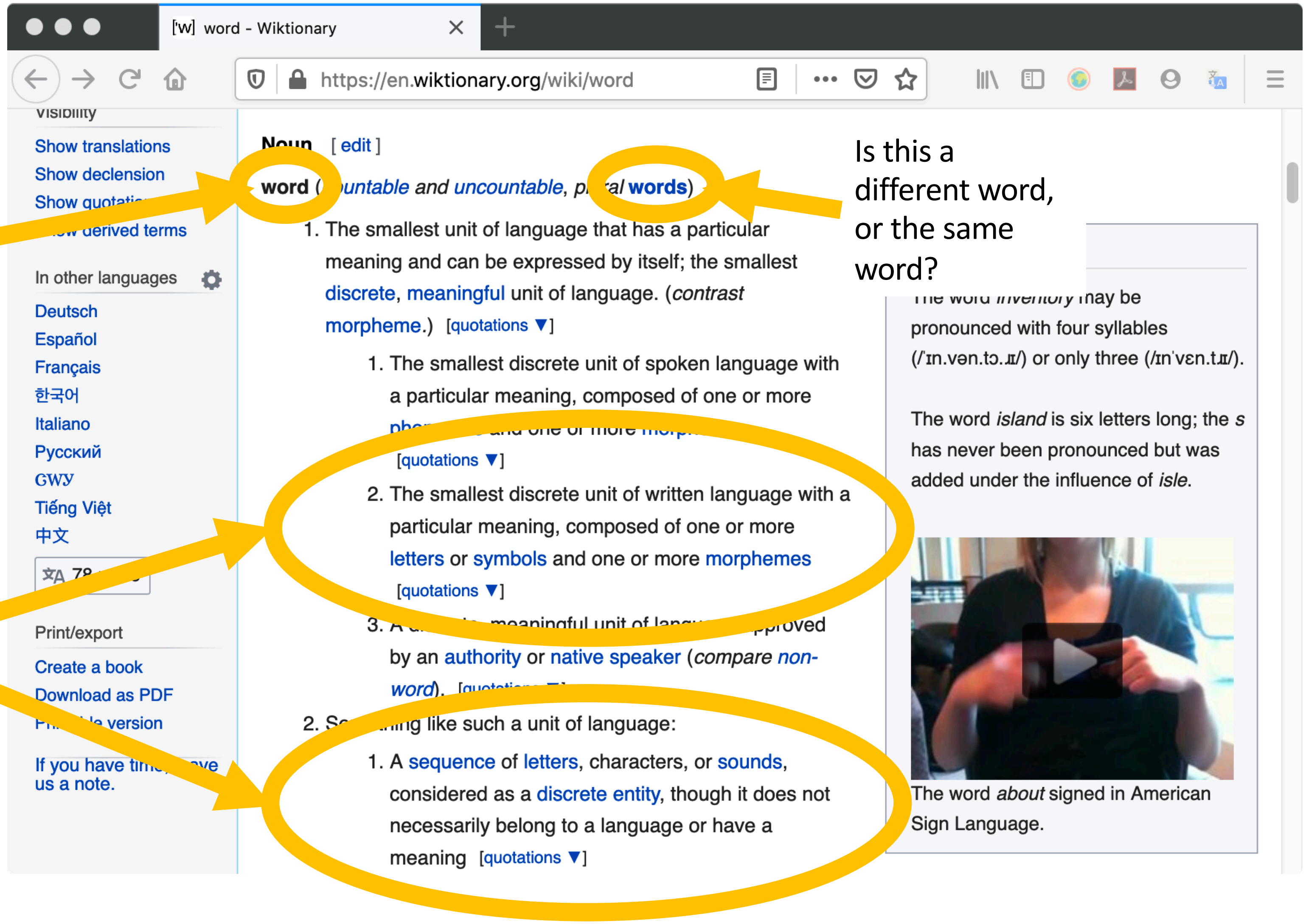
What is a word?

Is this a word?



Is this a different word, or the same word?

What is a word?



Is this a word?

Is this a different word, or the same word?

Are these the same word, or different words?

Lemma

A lemma is what humans usually think of as a “word.” It is defined to be the form of the word that appears in a dictionary.

- Other wordforms that can be easily predicted from the lemma need not be listed.

word (*countable and uncountable*, *plural* **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast* morpheme.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

Wordform

A wordform is a unique sequence of characters.

- Wordforms are much easier for computers to find than lemmas, therefore most automatic processing deals with wordforms.
- …however, we lose something. "dog" and "dogs" become completely unrelated – as unrelated as "dog" and "exaggerate."

word (*countable and uncountable*, *plural* **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast* morpheme.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

Word sense

Often, a word has different meanings that are completely unrelated. We think of them as different words, that just happen to be spelled and pronounced the same way.

We say that these are different "senses" of the same word.

The Bank of England. By Diliff - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=40912212

The Bank of the Thames. By Diliff - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=3639626

Wordform, lemma, and word sense

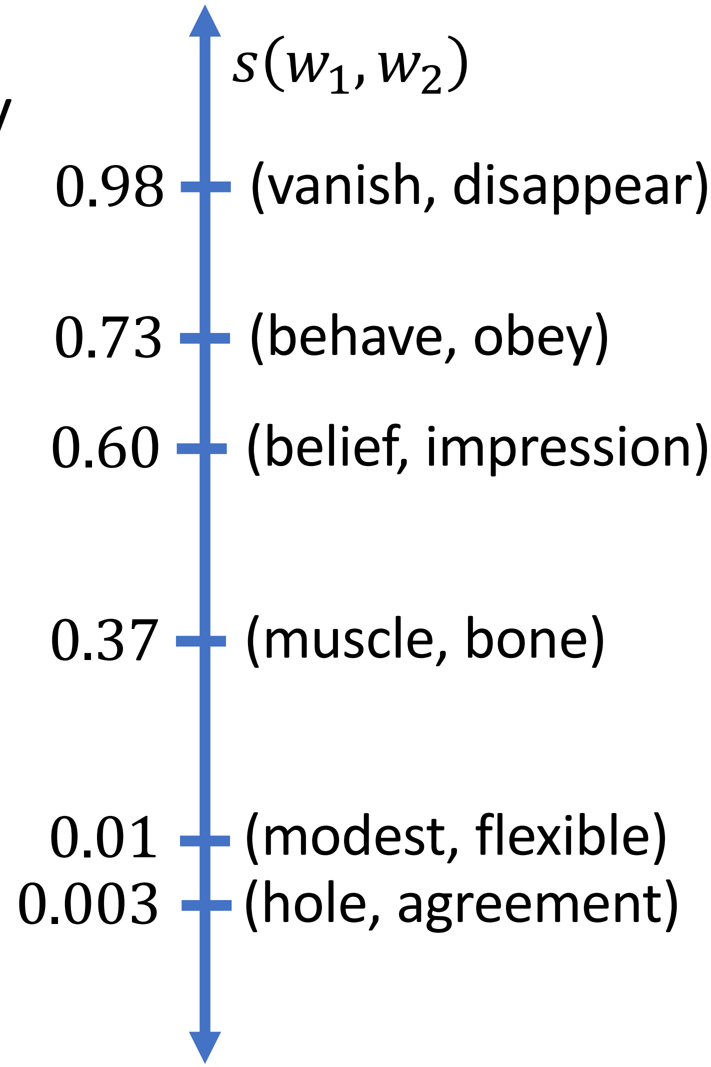
- wordform
  - easy for a computer to work with: just look for space-bounded sequences of characters
- lemma
  - This is what humans think of as a word. A cluster of wordforms whose spellings, pronunciations, and meanings can all be derived from one another by applying simple rules.
- word sense
  - A meaning so distinct from the other meanings of the word that it’s hard to consider them the same word.

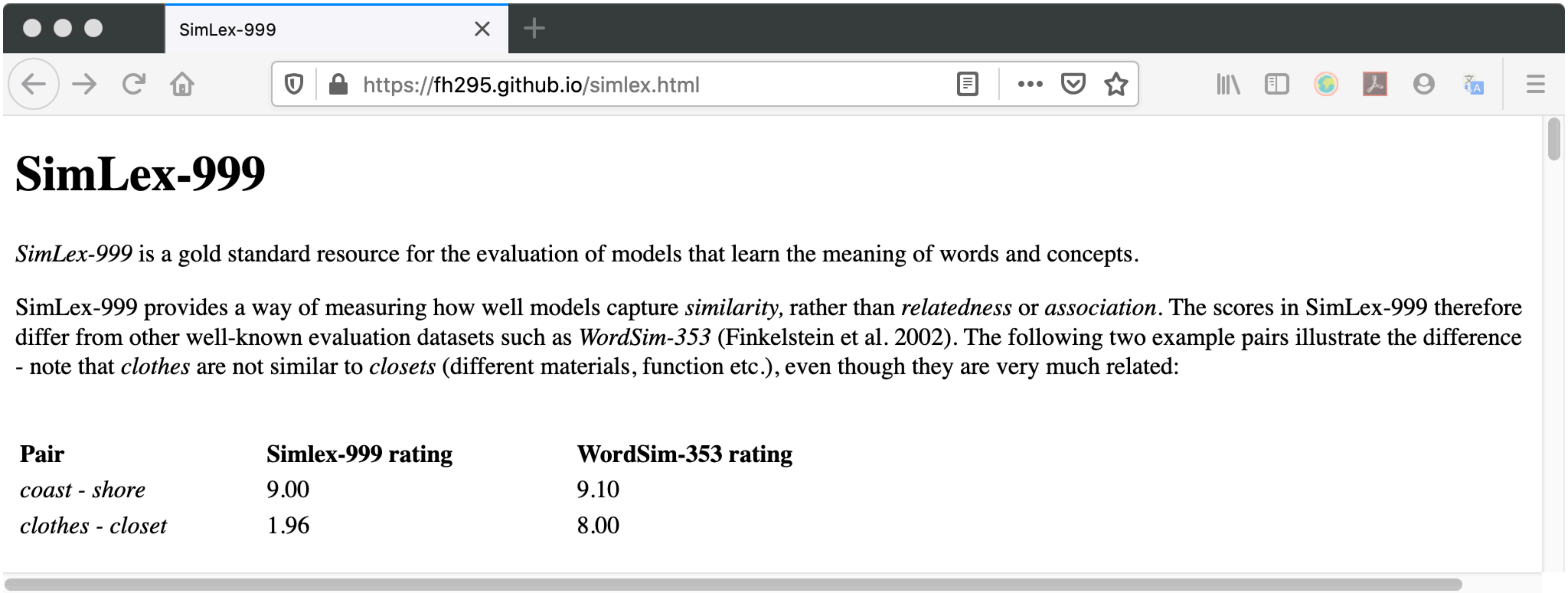
Outline

- What is a word? Lemmas, wordforms, and word sense
- Synonymy, similarity, and relatedness
- Word2vec
- Visualizations
- Bias

Synonymy and similarity

- Words are "synonyms" if they have exactly the same meaning.
- No words ever have ***exactly*** the same meaning, so no two words are ever exactly synonyms.
- We prefer to talk about word similarity, $0 \le s(w_1, w_2) \le 1$
  - $s(w_1, w_2) = 1$: w_1 and w_2 are perfect synonyms. Never happens in practice, but sometimes close.
  - $s(w_1, w_2) = 0$: w_1 and w_2 are completely different.





SimLex-999

https://fh295.github.io/simlex.html

SimLex-999

SimLex-999 is a gold standard resource for the evaluation of models that learn the meaning of words and concepts.

SimLex-999 provides a way of measuring how well models capture *similarity,* rather than *relatedness* or *association*. The scores in SimLex-999 therefore differ from other well-known evaluation datasets such as *WordSim-353* (Finkelstein et al. 2002). The following two example pairs illustrate the difference - note that *clothes* are not similar to *closets* (different materials, function etc.), even though they are very much related:

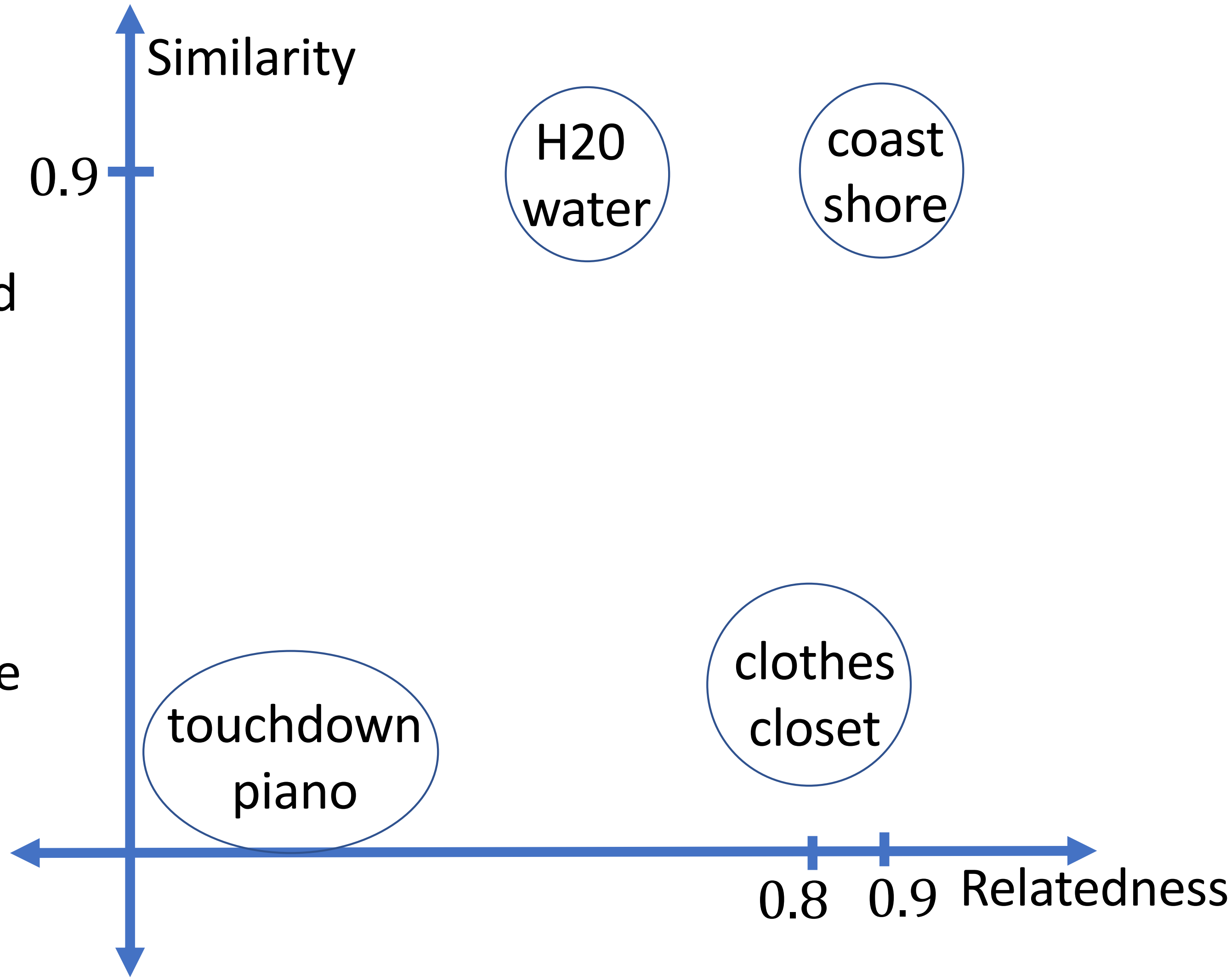
Pair	Simlex-999 rating	WordSim-353 rating
coast - shore	9.00	9.10
clothes - closet	1.96	8.00

- Algorithms that try to estimate the similarity of two wordforms can be tested on databases such as SimLex-999.
- Humans rated the similarity of each word pair on a 10-point scale.

Similarity vs. Relatedness

Similar: words can be used interchangeably in most contexts

Related: there is some connection between the two words, such that they tend to appear in the same documents.



Similarity: The Internet is the database

Similarity = words can be used interchangeably in most contexts

How do we measure that in practice?

Answer: extract examples of word w_1, +/- N words (N=2 or 3):

…hot, although iced **<u>coffee</u>** is a popular…

…indicate that moderate **<u>coffee</u>** consumption is benign…

…and of w_2:

…consumed as iced **<u>tea</u>**. Sweet tea is…

…national average of **<u>tea</u>** consumption in Ireland…

The words "iced" and "consumption" appear in both contexts, so we can conclude that $s(\text{coffea}, \text{tea}) > 0$. No other words are shared, so we can conclude $s(\text{coffee}, \text{tea}) < 1$.

Outline

- What is a word? Lemmas, wordforms, and word sense
- Synonymy, similarity, and relatedness
- Word2vec
- Visualizations
- Bias

word2vec

- <u>**word2vec**</u>, or skip-gram, is an algorithm for training real-valued vectors to represent each word.
- If word w_1 is represented by vector $\vec{v}_1 = [v_{11}, \dots, v_{1D}]$, we say that $\vec{v}_1$ is the D-dimensional <u>**embedding**</u> of word w_1.
- The general area of <u>**vector semantics**</u> (represent the meaning of a word as a vector) goes back to the 1950s, in the field of information retrieval (more about that in the next lecture).
- word2vec is an algorithm for learning those vectors using a one-layer neural network, in such a way that similar words are close together in the vector space.

cosine similarity

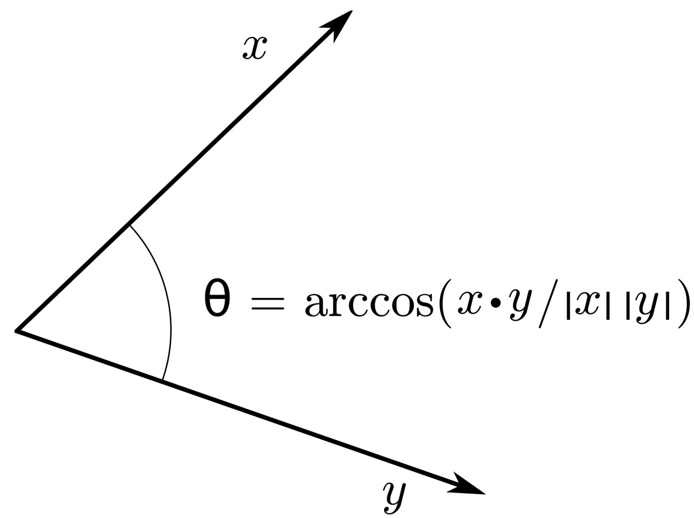
If words w_1 and w_2 are similar, w_1 is represented by vector $\vec{v}_1$, and w_2 by vector $\vec{v}_2$, then the angle between the two vectors should be small.

Angle between two vectors can be measured by their dot product:

$$\cos\theta = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1||\vec{v}_2|}$$

where

$$\vec{v}_1 \cdot \vec{v}_2 = \sum_{d=1}^{D} v_{1d} v_{2d} \,, \qquad |\vec{v}_1| = \sqrt{\sum_{d=1}^{D} v_{1d}^2}$$



Word2vec: context probability

The key innovation of word2vec is the idea of representing similarity as the probability that words w_1 and w_2 could occur in the same context, and of estimating the probability using a sigmoid.

Consider the "…hot although iced **<u>coffee</u>** is a popular…".

Define the target word to be w =coffee.

Define the context words c_{-3} =hot, c_{-2} =although, …, c_3 =popular.

Use a naïve Bayes model of the context probability:

$$p(c_{-3},\dots,c_3|w) = \prod_{\substack{i\neq 0\\ i=-3}}^{3} p(c_i|w)$$

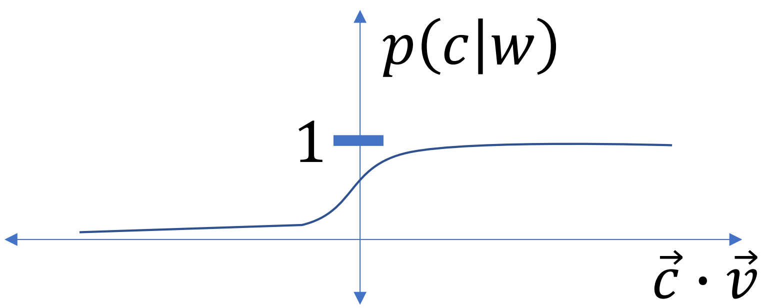
word2vec: context probability

Now suppose we want to embed w =coffee with a vector $\vec{v}$.

…and we want to embed c_{-3} =hot with a vector $\vec{c}$.

Define the probability that "hot" occurs within +/-N words of "coffee" to be just a sigmoid:

$$p(c|w) = \frac{1}{1 + e^{-\vec{c}\cdot\vec{v}}}$$



word2vec: training

We train the neural network by listing, as positive examples, the words that occur in the context of "w =coffee," e.g.,

$$\mathcal{D}_+(w) = \{\text{hot}, \text{although}, \text{iced}, \text{moderate}, \text{the}, \text{hot}, \text{consumption}, \ldots\}$$

Create a negative database by selecting words at random from the vocabulary, each word in proportion to its frequency in the whole dataset:

$$\mathcal{D}_-(w) = \{\text{aardvark}, \text{dog}, \text{gazebo}, \text{the}, \text{precipitates}, \ldots\}$$

word2vec: training

The coefficients $\vec{v}_i = [v_{i1}, \ldots, v_{iD}]$ for each vector are then learned in order to maximize the log probability of the dataset:

$$\mathcal{L} = \ln p(\mathrm{Data}) = \sum_{w \in \mathcal{V}} \ln p(\mathcal{D}_+(w)|w) + \sum_{w \in \mathcal{V}} \ln p(\mathcal{D}_-(w)|w)$$

$$= \sum_{w \in \mathcal{V}} \sum_{c \in \mathcal{D}_+(w)} \ln p(c|w) + \sum_{w \in \mathcal{V}} \sum_{c \in \mathcal{D}_-(w)} \ln\big(1 - p(c|w)\big)$$

$$= \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_+(w)} \ln \frac{1}{1 + e^{-\vec{c}\cdot\vec{v}}} + \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_-(w)} \ln\left(1 - \frac{1}{1 + e^{-\vec{c}\cdot\vec{v}}}\right)$$

$$\mathcal{L} = \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_+(w)} \ln \frac{1}{1 + e^{-\vec{c}\cdot\vec{v}}} + \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_-(w)} \ln \frac{1}{1 + e^{\vec{c}\cdot\vec{v}}}$$

word2vec: training

The coefficients $\vec{v}_i = [v_{i1}, \ldots, v_{iD}]$ for each vector are then learned in order to maximize the log probability of the dataset:

$$v_{id} \leftarrow v_{id} + \eta \frac{d\mathcal{L}}{dv_{id}}$$

$$= v_{id} + \eta \frac{d}{dv_{id}} \left(\sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_+(w)} \ln \frac{1}{1 + e^{-\vec{c} \cdot \vec{v}}} + \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_-(w)} \ln \frac{1}{1 + e^{\vec{c} \cdot \vec{v}}} \right)$$

There's one more issue to consider here: if the word coffee occurs as a center word (w=coffee) or a context word (c=coffee), should those vectors ($\vec{v}$ and $\vec{c}$, respectively) be the same vector, or different vectors? The results are slightly different; which one is better depends on the application for which you're training word2vec.

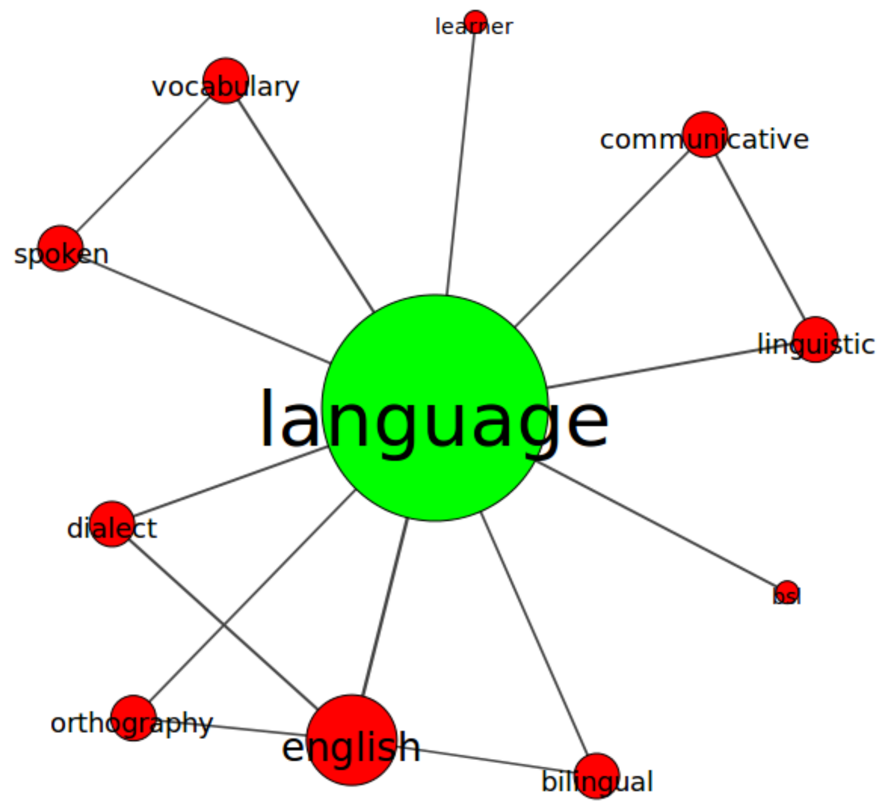
Outline

- What is a word? Lemmas, wordforms, and word sense
- Synonymy, similarity, and relatedness
- Word2vec
- Visualizations
- Bias

Visualizations: Similarity

Andrei Kutuzov and Erik Velidal (2016) visualized the degree to which models like word2vec capture similarity by learning graphs:

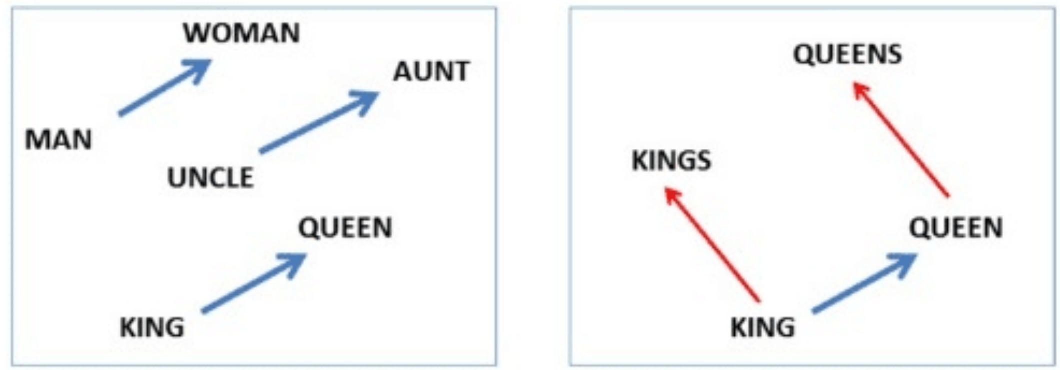
- The graph connects each word is connected to its K nearest neighbors, i.e., the K other vectors for which $|\vec{v}_1 - \vec{v}_2|$ is smallest.
- Each of the edges is then evaluated to see if the resulting words are similar or not.



Andrei Kutuzov and Erik Velidal (2016) https://www.mn.uio.no/ifi/studier/masteroppgaver/ltg/graph-of-embeddings.html

Visualizations: Relatedness

vec("man") – vec("king") + vec("woman") = vec("queen")



Christian S. Perone, "Voynich Manuscript: word vectors and t-SNE visualization of some patterns," in *Terra Incognita*, 16/01/2016, http://blog.christianperone.com/2016/01/voynich-manuscript-word-vectors-and-t-sne-visualization-of-some-patterns/.

Mikolov (2013) showed that word2vec captures similarity relationships among words. For example, the difference between the vectors for "woman" and "man" is roughly the same as the difference between the vectors for "queen" and "king." Perone (2016) showed that this effect works differently depending on the training corpus: in his blog post, he looks at word relatedness in the 15th century Voynich manuscript.

Outline

- What is a word? Lemmas, wordforms, and word sense
- Synonymy, similarity, and relatedness
- Word2vec
- Visualizations
- Bias

Learning biased analogies from data

- It's useful that algorithms like word2vec learn appropriate analogies, like "Paris → France as Tokyo → Japan" and "kings → king as queens → queen."
- Unfortunately, it also learns other analogies that were implied in the training corpus, but that are invalid analogies.
- The paper that first demonstrated that problem was named after one of the worst such discovered analogies:

"Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," Bolukbasi et al., 2016

Biased analogies

Bolukbasi et al. defined a “male-female” continuum by subtracting vec(female)-vec(male), vec(woman)-vec(man), and so on, then averaging these difference vectors.

They then took all of the words whose dictionary definitions included gender-specific language (man, woman), and considered those to be the gender-specific words (words for which a gender difference is appropriate).

All other words were considered gender-neutral (any difference on the male-female dimension is inappropriate).

The result is a second dimension: the appropriateness of a gender bias.

The Male-Female vs. Neutral-Specific Space

Here's the resulting 2D space, from Bolukbasi et al., 2016:

Outline

- What is a word? Lemmas, wordforms, and word sense
- Synonymy, similarity, and relatedness
- Word2vec: maximize

$$\mathcal{L} = \sum_{\vec{v}\in\mathcal{V}} \sum_{\vec{c}\in\mathcal{D}_+(w)} \ln \frac{1}{1+e^{-\vec{c}\cdot\vec{v}}} + \sum_{w\in\mathcal{V}} \sum_{\vec{c}\in\mathcal{D}_-(w)} \ln \frac{1}{1+e^{\vec{c}\cdot\vec{v}}}$$

- Visualizations
  - Similarity: K-nearest neighbor graph structure
  - Relatedness: analogies are shown as directions in the vector space!
- Bias
  - Bias can be reduced by learning a direction that should not depend on the female-male axis, and then squashing the female-male axis to zero for words that should be gender-neutral.