DISTRIBUTIONAL WORD SIMILARITY

David Kauchak
CS159 Fall 2014

Admin

Assignment 5 out

Word similarity

How similar are two words?

score: $\text{sim}(w_1, w_2) = ?$

rank: w ?

w_1
w_2
w_3

list: w_1 and w_2 are synonyms

Word similarity

Four categories of approaches (maybe more)

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

Word similarity

Four general categories

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

Dictionary-based similarity

Word	Dictionary blurb
aardvark	a large, nocturnal, burrowing mammal, *Orycteropus afer, ofcentral and southern Africa, feeding on ants and termites andhaving a long, extensile tongue, strong claws, and long ears.*
beagle	One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat.
dog	Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.

Dictionary-based similarity

Utilize our text similarity measures

sim(*dog*, *beagle*) =

sim(One of a breed of small hounds having long ears, short legs, and a usually black, tan, and white coat. ,

Any carnivore of the family Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.)

Dictionary-based similarity

–noun

1. a domesticated canid, *Canis familiaris,* bred in many varieties.
2. any carnivore of the dogfamily Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.
3. the male of such an animal.
4. any of various animals resembling a dog.
5. a despicable man or youth.
6. *Informal* . a fellow in general: *a lucky dog.*
7. *dogs, Slang* . feet.
8. *Slang* .
   a. something worthless or of extremely poor quality: *That used car you bought is a dog.*
   b. an utter failure; flop: *Critics say his new play is a dog.*
9. *Slang* . an ugly, boring, or crude person.
10. *Slang* . hot dog.
11. (*initial capital letter*) *Astronomy* . either of two constellations, Canis Major or Canis Minor.
12. *Machinery* .
    a. any of various mechanical devices, as for gripping or holding something.
    b. a projection on a moving part for moving steadily or for tripping another part with which it engages.
13. Also called **gripper**, **nipper.** *Metalworking* . a device on a drawbench for drawing the work through the die.
14. a cramp binding together two timbers.
15. an iron bar driven into a stone or timber to provide a means of lifting it.
16. an andiron; firedog.
17. *Meteorology* . a sundog or fogdog.
18. a word formerly used in communications to represent the letter D.

What about words that have multiple senses/parts of speech?

Dictionary-based similarity

–noun
1. a domesticated canid, *Canis familiaris*, bred in many varieties.
2. any carnivore of the dogfamily Canidae, having prominent canine teeth and, in the wild state, a long and slender muzzle, a deep-chested muscular body, a bushy tail, and large, erect ears. Compare canid.
3. the male of such an animal.
4. any of various animals resembling a dog.
5. a despicable man or youth.
6. *Informal* . a fellow in general: *a lucky dog.*
7. dogs, *Slang* . feet.
8. *Slang* .
   a. something worthless or of extremely poor quality: *That used car you bought is a dog.*
   b. an utter failure; flop: *Critics say his new play is a dog.*
9. *Slang* . an ugly, boring, or crude person.
10. *Slang* . hot dog.
11. (*initial capital letter*) *Astronomy* . either of two constellations, Canis Major or Canis Minor.
12. *Machinery* .
    a. any of various mechanical devices, as for gripping or holding something.
    b. a projection on a moving part for moving steadily or for tripping another part with which it engages.
13. Also called **gripper**, **nipper**, *Metalworking* . a device on a drawbench for drawing the work through the die.
14. a cramp binding together two timbers.
15. an iron bar driven into a stone or timber to provide a means of lifting it.
16. an andiron; firedog.
17. *Meteorology* . a sundog or fogdog.
18. a word formerly used in communications to represent the letter D.

1. part of speech tagging
2. word sense disambiguation
3. most frequent sense
4. average similarity between all senses
5. max similarity between all senses
6. sum of similarity between all senses

Dictionary + WordNet

WordNet also includes a "gloss" similar to a dictionary definition

Other variants include the overlap of the word senses as well as those word senses that are related (e.g. hypernym, hyponym, etc.)

- incorporates some of the path information as well
- Banerjee and Pedersen, 2003

Word similarity

Four general categories

- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

Corpus-based approaches

Word

ANY blurb with the word

aardvark

beagle

Ideas?

dog

Corpus-based

The ***Beagle*** is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg

Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern ***Beagle*** can be traced in Ancient Greece[2] back to around the 5th century BC.

From medieval times, ***beagle*** was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard ***Beagle*** type was beginning to develop: the distinction between the North Country Beagle and Southern

Corpus-based: feature extraction

The ***Beagle*** is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg

We'd like to utilize or vector-based approach

How could we we create a vector from these occurrences?

- collect word counts from all documents with the word in it
- collect word counts from all sentences with the word in it
- collect all word counts from all words within ***X*** words of the word
- collect all words counts from words in specific relationship: subject-object, etc.

Word-context co-occurrence vectors

The ***Beagle*** is a breed of small to medium-sized dog. A member of the Hound Group, it is similar in appearance to the Foxhound but smaller, with shorter leg

Beagles are intelligent, and are popular as pets because of their size, even temper, and lack of inherited health problems.

Dogs of similar size and purpose to the modern ***Beagle*** can be traced in Ancient Greece[2] back to around the 5th century BC.

From medieval times, ***beagle*** was used as a generic description for the smaller hounds, though these dogs differed considerably from the modern breed.

In the 1840s, a standard ***Beagle*** type was beginning to develop: the distinction between the North Country Beagle and Southern

Word-context co-occurrence vectors

The ***Beagle*** is a breed

Beagles are intelligent, and

to the modern ***Beagle*** can be traced

From medieval times, ***beagle*** was used as

1840s, a standard ***Beagle*** type was beginning

the:	2
is:	1
a:	2
breed:	1
are:	1
intelligent:	1
and:	1
to:	1
modern:	1
…	

Often do some preprocessing like lowercasing and removing stop words

Corpus-based similarity

sim(*dog*, *beagle*) =

sim(*context_vector(dog)*, *context_vector(beagle)*)

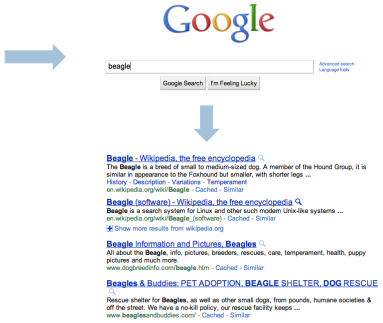
the:	5	the:	2
is:	1	is:	1
a:	4	a:	2
breeds:	2	breed:	1
are:	1	are:	1
intelligent:	5	intelligent:	1
…		and:	1
		to:	1
		modern:	1
		…	

Web-based similarity

Ideas?

Web-based similarity

beagle



Web-based similarity

Concatenate the snippets for the top ***N*** results

Concatenate the web page text for the top ***N*** results

Another feature weighting

TF- IDF weighting takes into account the general importance of a feature

For distributional similarity, we have the feature (f_i), but we also have the word itself (w) that we can use for information

sim(*context_vector(**dog**)*, *context_vector(**beagle**)*)

the:	5	the:	2
is:	1	is:	1
a:	4	a:	2
breeds:	2	breed:	1
are:	1	are:	1
intelligent:	5	intelligent:	1
…		and:	1
		to:	1
		modern:	1
		…	

Another feature weighting

Feature weighting ideas given this additional information?

sim(*context_vector(**dog**)*, *context_vector(**beagle**)*)

the:	5	the:	2
is:	1	is:	1
a:	4	a:	2
breeds:	2	breed:	1
are:	1	are:	1
intelligent:	5	intelligent:	1
…		and:	1
		to:	1
		modern:	1
		…	

Another feature weighting

count *how likely* feature f_i and word w are to occur together

- incorporates co-occurrence
- but also incorporates how often w and f_i occur in other instances

sim(*context_vector(**dog**)*, *context_vector(**beagle**)*)

Does IDF capture this?

Not really. IDF only accounts for f_i regardless of w

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

When will this be high and when will this be low?

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

if x and y are independent (i.e. one occurring doesn't impact the other occurring) then:

$$p(x,y) =$$

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

if x and y are independent (i.e. one occurring doesn't impact the other occurring) then:

$$p(x,y) = p(x)p(y)$$

What does this do to the sum?

Mutual information

A bit more probability ☺

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

if they are dependent then:

$$p(x,y) = p(x)p(y \mid x) = p(y)p(x \mid y)$$

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(y \mid x)}{p(y)}$$

Mutual information

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(y \mid x)}{p(y)}$$

What is this asking?
When is this high?

How much more likely are we to see y given x has a particular value!

Point-wise mutual information

Mutual information

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

How related are two variables (i.e. over all possible values/events)

Point-wise mutual information

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

How related are two particular events/ values

PMI weighting

Mutual information is often used for feature selection in many problem areas

PMI weighting weights co-occurrences based on their correlation (i.e. high PMI)

context_vector(beagle)

the:	2
is:	1
a:	2
breed:	1
are:	1
intelligent:	1
and:	1
to:	1
modern:	1
…	

$\log \frac{p(beagle,the)}{p(beagle)p(the)}$

$\log \frac{p(beagle,breed)}{p(beagle)p(breed)}$

How do we calculate these?