WORD SIMILARITY

David Kauchak
CS159 Fall 2020

1

Admin

Assignment 4

Quiz #2 Thursday
- 1 hour (shouldn't need that long)
- Will post link on piazza
- Will be available 12:15-1:15pm on class zoom
- Open book and notes
- Class starts at 1:15pm

Assignment 5 out soon

2

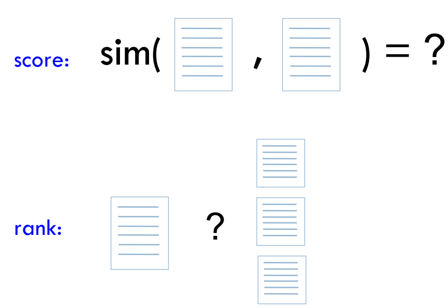
Quiz #2

Topics
- Linguistics 101
- Parsing
  - Grammars, CFGs, PCFGs
  - Top-down vs. bottom-up
  - CKY algorithm
  - Grammar learning
  - Evaluation
  - Improved models
- Text similarity
  - Will also be covered on Quiz #3, though

3

Text Similarity

A common question in NLP is how similar are texts



4

Bag of words representation

For now, let's ignore word order:

Obama said banana repeatedly last week on tv, "banana, banana, banana"

(4, 1, 1, 0, 0, 1, 0, 0, …)

banana obama said california across tv wrong capital

"Bag of words representation": multi-dimensional vector, one dimension per word in our vocabulary

Frequency of word occurrence

5

Vector based word

A

a1: When	1
a2: the	2
a3: defendant	1
a4: and	1
a5: courthouse	0
…	

B

b1: When	1
b2: the	2
b3: defendant	1
b4: and	0
b5: courthouse	1
…	

Multi-dimensional vectors, one dimension per word in our vocabulary

6

Normalized distance measures

Cosine

$$sim_{\cos}(A,B) = A \cdot B = \sum_{i=1}^{n} a'_i b'_i = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

L2

$$dist_{L2}(A,B) = \sqrt{\sum_{i=1}^{n} (a'_i - b'_i)^2}$$

L1

$$dist_{L1}(A,B) = \sum_{i=1}^{n} |a'_i - b'_i|$$

a' and b' are length normalized versions of the vectors

7

Our problems

Which of these have we addressed?

- word order
- length
- synonym
- spelling mistakes
- word importance
- word frequency

8

Word overlap problems

Treats all words the same

A: When the defendant and his lawyer walked into the court, some of the victim supporters turned their backs to him.

B: When the defendant walked into the courthouse with his attorney, the crowd truned their backs on him.

Ideas?

9

Word importance

Include a weight for each word/feature

A

a1: When	1	w1
a2: the	2	w2
a3: defendant	1	w3
a4: and	1	w4
a5: courthouse	0	w5
…		…

B

b1: When	1	w1
b2: the	2	w2
b3: defendant	1	w3
b4: and	0	w4
b5: courthouse	1	w5
…		…

10

Distance + weights

We can incorporate the weights into the distances

Think of it as either (*both work out the same*):

- preprocessing the vectors by multiplying each dimension by the weight
- incorporating it directly into the similarity measure

$$sim_{\cos}(A,B) = A \cdot B = \frac{\sum_{i=1}^{n} w_i a_i w_i b_i}{\sqrt{\sum_{i=1}^{n} (w_i a_i)^2} \sqrt{\sum_{i=1}^{n} (w_i b_i)^2}}$$

11

Idea: use corpus statistics

the

defendant

What would be a quantitative measure of word importance?

12

Document frequency

document frequency (DF) is one measure of word importance

Terms that occur in many documents are weighted less, since overlapping with these terms is very likely

- In the extreme case, take a word like the that occurs in almost EVERY document

Terms that occur in only a few documents are weighted more

13

Document vs. overall frequency

The overall frequency of a word is the number of occurrences in a dataset, counting multiple occurrences

Example:

Word	Overall frequency	Document frequency
insurance	10440	3997
try	10422	8760

Which word is a more informative (and should get a higher weight)?

14

Document frequency

Word	Collection frequency	Document frequency
insurance	10440	3997
try	10422	8760

Document frequency is often related to word importance, but we want an actual weight. Problems?

$$sim_{\cos}(A,B) = A \cdot B = \frac{\sum_{i=1}^{n} w_i a_i w_i b_i}{\sqrt{\sum_{i=1}^{n} (w_i a_i)^2} \sqrt{\sum_{i=1}^{n} (w_i b_i)^2}}$$

15

From document frequency to weight

Word	Collection frequency	Document frequency
insurance	10440	3997
try	10422	8760

weight and document frequency are **inversely** related

- higher document frequency should have lower weight and vice versa

document frequency is unbounded

document frequency will change depending on the size of the data set (i.e. the number of documents)

16

Inverse document frequency

$$\text{idf}_w = \log \frac{N}{\text{df}_w}$$

N: # of documents in dataset

df_w: document frequency of w

IDF is inversely correlated with DF
- higher DF results in lower IDF

N incorporates a dataset dependent normalizer

log dampens the overall weight

17

IDF example, suppose N=1 million

term	df_t	idf_t
calpurnia	1	
animal	100	
sunday	1,000	
fly	10,000	
under	100,000	
the	1,000,000	

What are the IDFs assuming log base 10?

18

IDF example, suppose N=1 million

term	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1,000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

There is one idf value/weight for each word

19

IDF example, suppose N=1 million

term	df_t	idf_t
calpurnia	1	
animal	100	
sunday	1,000	
fly	10,000	
under	100,000	
the	1,000,000	

What if we didn't use the log to dampen the weighting?

20

IDF example, suppose N=1 million

term	df_t	idf_t
calpurnia	1	1,000,000
animal	100	10,000
sunday	1,000	1,000
fly	10,000	100
under	100,000	10
the	1,000,000	1

What if we didn't use the log to dampen the weighting?

21

TF-IDF

One of the most common weighting schemes

TF = term frequency

IDF = inverse document frequency

$$a'_i = \underbrace{a_i}_{\text{TF}} \times \underbrace{\log N / df_i}_{\text{IDF (word importance weight)}}$$

We can then use this with any of our similarity measures!

22

Stoplists: extreme weighting

Some words like 'a' and 'the' will occur in almost every document

- IDF will be 0 for any word that occurs in all documents
- For words that occur in almost all of the documents, they will be nearly 0

A ***stoplist*** is a list of words that should **not** be considered (in this case, similarity calculations)

- Sometimes this is the *n* most frequent words
- Often, it's a list of a few hundred words manually created

23

Stoplist

I	all-over	around	beneath	due	go
a	almost	as	beside	durin	goddamn
aboard	along	aside	besides	during	goody
about	alongside	astride	between	each	gosh
above	altho	at	bewteen	eh	half
across	although	atop	beyond	either	have
after	amid	avec	bi	en	he
afterwards	amidst	away	both	every	hell
against	among	back	but	ever	her
agin	amongst	be	by	everyone	herself
ago	an	because	ca.	everything	hey
agreed-upon	and	before	de	except	him
ah	another	beforehand	des	far	himself
alas	any	behind	despite	fer	his
albeit	anyone	behynde	do	for	ho
all	anything	below	down	from	how

If most of these end up with low weights anyway, why use a stoplist?

24

Stoplists

Two main benefits

- More fine grained control: some words may not be frequent, but may not have any content value (alas, teh, gosh)
- Often does contain many frequent words, which can drastically reduce our storage and computation

Any downsides to using a stoplist?

- For some applications, some stop words may be important

25

Our problems

Which of these have we addressed?

- word order
- length
- synonym
- spelling mistakes
- word importance
- word frequency

A model of word similarity!

26

Word overlap problems

A: When the defendant and his *lawyer* walked into the *court*, some of the victim supporters *turned* their backs to him.

B: When the defendant walked into the *courthouse* with his *attorney*, the crowd *truned* their backs on him.

27

Word similarity

How similar are two words?

score: $sim(w_1, w_2) = ?$

rank: w ? w_1 w_2 w_3

applications?

list: w_1 and w_2 are synonyms

28

Word similarity applications

General text similarity

Thesaurus generation

Automatic evaluation

Text-to-text
- paraphrasing
- summarization
- machine translation

information retrieval (search)

29

Word similarity

How similar are two words?

score: $sim(w_1, w_2) = ?$

rank: w ? w_1 w_2 w_3

ideas? useful resources?

list: w_1 and w_2 are synonyms

30

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

31

Character-based similarity

sim(*turned*, *truned*) = ?

How might we do this using only the words (i.e. no outside resources?

32

Edit distance (Levenshtein distance)

The edit distance between w_1 and w_2 is the minimum number of operations to transform w_1 into w_2

Operations:
- insertion
- deletion
- substitution

EDIT(turned, truned) = ?
EDIT(computer, commuter) = ?
EDIT(banana, apple) = ?
EDIT(wombat, worcester) = ?

33

Edit distance

EDIT(turned, truned) = 2
- delete u
- insert u

EDIT(computer, commuter) = 1
- replace p with m

EDIT(banana, apple) = 5
- delete b
- replace n with p
- replace a with p
- replace n with l
- replace a with e

EDIT(wombat, worcester) = 6

34

Better edit distance

Are all operations equally likely?
- No

Improvement: give different weights to different operations
- replacing a for e is more likely than z for y

Ideas for weightings?
- Learn from actual data (known typos, known similar words)
- Intuitions: phonetics
- Intuitions: keyboard configuration

35

Vector character-based word similarity

sim(*turned*, *truned*) = ?

Any way to leverage our vector-based similarity approaches from last time?

36

Vector character-based word similarity

sim(*turned*, *truned*) = ?

Generate a feature vector based on the characters
(or could also use the set based measures at the character level)

a:	0	a:	0
b:	0	b:	0
c:	0	c:	0
d:	1	d:	1
e:	1	e:	1
f:	0	f:	0
g:	0	g:	0
…		…	

problems?

37

Vector character-based word similarity

sim(*restful*, *fluster*) = ?

Character level loses a lot of information

a:	0	a:	0
b:	0	b:	0
c:	0	c:	0
d:	1	d:	1
e:	1	e:	1
f:	0	f:	0
g:	0	g:	0
…		…	

ideas?

38

Vector character-based word similarity

sim(*restful*, *fluster*) = ?

Use character bigrams or even trigrams

aa:	0	aa:	0
ab:	0	ab:	0
ac:	0	ac:	0
…		…	
es:	1	er:	1
…		…	
fu:	1	fl:	1
…		…	
re:	1	lu:	1
…		…	

39

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

40

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

41

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.

42

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.)

43

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?

44

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

45

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

46

Word similarity

Four general categories

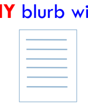
- Character-based
  - turned vs. truned
  - cognates (night, nacht, nicht, natt, nat, noc, noch)
- Semantic web-based (e.g. WordNet)
- Dictionary-based
- Distributional similarity-based
  - similar words occur in similar contexts

47

Corpus-based approaches

Word — **ANY** blurb with the word

aardvark



beagle

Ideas?

dog

48

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

49

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 our 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.

50

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

51

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

52

Corpus-based similarity

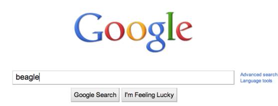
sim(*dog*, *beagle*) =

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
		…	

53

Web-based similarity

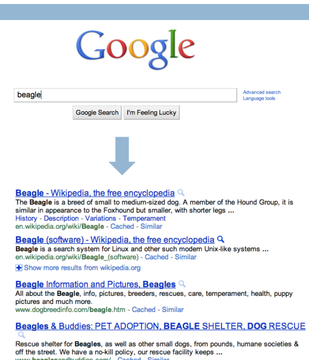


Ideas?

54

Web-based similarity

beagle



55

Web-based similarity

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

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

56

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
		…	

57

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
		…	

58

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

59

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?

60

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:

$$p(x,y) =$$

61

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?

62

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:

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

$$\Downarrow$$

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

63

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!

64

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

65

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?

66