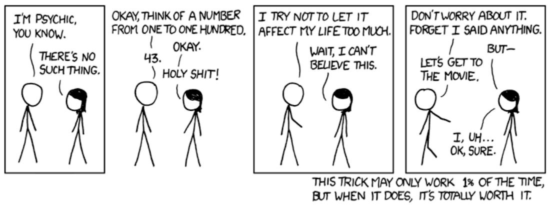

I'M PSYCHIC, YOU KNOW.
THERE'S NO SUCH THING.
OKAY, THINK OF A NUMBER FROM ONE TO ONE HUNDRED.
OKAY.
43.
HOLY SHIT!
I TRY NOT TO LET IT AFFECT MY LIFE TOO MUCH.
WAIT, I CAN'T BELIEVE THIS.
DON'T WORRY ABOUT IT. FORGET I SAID ANYTHING.
BUT—
LET'S GET TO THE MOVIE.
I, UH... OK, SURE.
THIS TRICK MAY ONLY WORK 1% OF THE TIME, BUT WHEN IT DOES, IT'S TOTALLY WORTH IT.

http://www.xkcd.com/628/

Summaries and Spelling Corection

David Kauchak
cs458
Fall 2012

adapted from:
http://www.stanford.edu/class/cs276/handouts/lecture3-tolerantretrieval.ppt
http://www.stanford.edu/class/cs276/handouts/lecture8-evaluation.ppt

Administrative

- Assignment 2
- Assignment 1
  - Overall, pretty good
  - Hard to get right!
  - Write-up:
    - be clear and concise
    - think about the point(s) that you want to make
    - justify your answer
- hw 2 back soon…

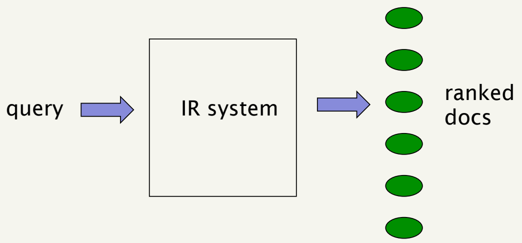
Quick recap

If we have a dictionary, with postings lists containing weights (e.g. tf-idf) explain briefly (e.g. pseudo-code) how to calculate the document similarities between a query of two words

Name two speed challenges that are faced when doing ranked retrieval vs. boolean retrieval.

One way to speed up ranked retrieval is to only perform the full ranking on a subset of the documents (inexact K). Name one method for selecting this subset of documents

So far…



what are we missing?

Today

User interface/user experience:

Once the documents are returned, how do we display them to the user?

Midleberry college
(spelling correction)

www.fordvehicles.com/cars/mustang/

en.wikipedia.org/wiki/Ford_Mustang

www.mustangseats.com/

www.mustangsurvival.com/

How is this?

2010 For Mustang | Official Site of the Ford Mustang
www.fordvehicles.com/cars/mustang/

Ford Mustang - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Ford_Mustang

Mustang Motorcycle Products, Inc.
www.mustangseats.com/

Mustang Survival Corporation
www.mustangsurvival.com/

2013 Ford Mustang | Official Site of the Ford Mustang
2013 Ford Mustang - The official homepage of the Ford Mustang | FordVehicles.com
www.fordvehicles.com/cars/mustang/

Ford Mustang - Wikipedia, the free encyclopedia
The Ford Mustang is an automobile manufactured by the Ford Motor Company. It was initially based on the second generation North American Ford Falcon, ...
en.wikipedia.org/wiki/Ford_Mustang

Mustang Motorcycle Products, Inc.
What a Difference Comfort Makes! Mustang is the world's leader in comfortable aftermarket motorcycle seats for Harley-Davidson®, Victory and Metric Cruiser ...
www.mustangseats.com/

Mustang Survival Corporation
Design, development, and manufacture of marine and aerospace safety and survival wear. Includes detailed product catalog, sizing charts, FAQs, ...
www.mustangsurvival.com/

2013 Ford Mustang | Official Site of the Ford Mustang
Warriors in Pink News SYNC News & Events
www.fordvehicles.com/cars/mustang/

Ford Mustang - Wikipedia, the free encyclopedia
I told the team that I wanted the car to appeal to women,
but I wanted men to desire it, too...
en.wikipedia.org/wiki/Ford_Mustang

Mustang Motorcycle Products, Inc.
New Tank Bibs with Pouches ...
www.mustangseats.com/

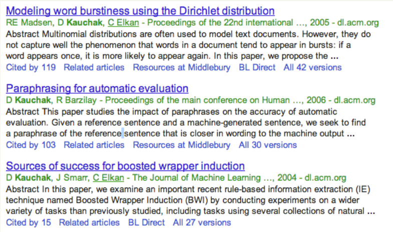
Mustang Survival Corporation
Terms of Use | Privacy Policy ...
www.mustangsurvival.com/

IR Display

In many domains, we have document metadata

web pages: titles, URLs, …

academic articles: what information do we have?



Modeling word burstiness using the Dirichlet distribution
RE Madsen, D **Kauchak**, C Elkan - Proceedings of the 22nd international …, 2005 - dl.acm.org
Abstract Multinomial distributions are often used to model text documents. However, they do not capture well the phenomenon that words in a document tend to appear in bursts: if a word appears once, it is more likely to appear again. In this paper, we propose the ...
Cited by 119 Related articles Resources at Middlebury BL Direct All 42 versions

Paraphrasing for automatic evaluation
D **Kauchak**, R Barzilay - Proceedings of the main conference on Human …, 2006 - dl.acm.org
Abstract This paper studies the impact of paraphrases on the accuracy of automatic evaluation. Given a reference sentence and a machine-generated sentence, we seek to find a paraphrase of the reference sentence that is closer in wording to the machine output ...
Cited by 103 Related articles Resources at Middlebury All 30 versions

Sources of success for boosted wrapper induction
D **Kauchak**, J Smarr, C Elkan - The Journal of Machine Learning …, 2004 - dl.acm.org
Abstract In this paper, we examine an important recent rule-based information extraction (IE) technique named Boosted Wrapper Induction (BWI) by conducting experiments on a wider variety of tasks than previously studied, including tasks using several collections of natural ...
Cited by 15 Related articles BL Direct All 27 versions

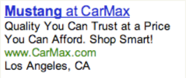
Other information

Other times, we may not have explicit meta-data, but may still want to provide additional data

- Web pages don't provide "snippets"/summaries

Even when pages do provide metadata, we may want to ignore this. Why?

The search engine may have different goals/motives than the webmasters, e.g. ads



Mustang at CarMax
Quality You Can Trust at a Price You Can Afford. Shop Smart!
www.CarMax.com
Los Angeles, CA

keyword tag

Summaries

We can generate these ourselves!

Most common (and successful) approach is to extract segments from the documents (called *extractive* in contrast with *abstractive*)

How might we identify good segments?

- Text early on in a document
- First/last sentence in a document, paragraph
- Text formatting (e.g. <h1>)
- Document frequency
- Distribution in document
- Grammatical correctness
- User query!

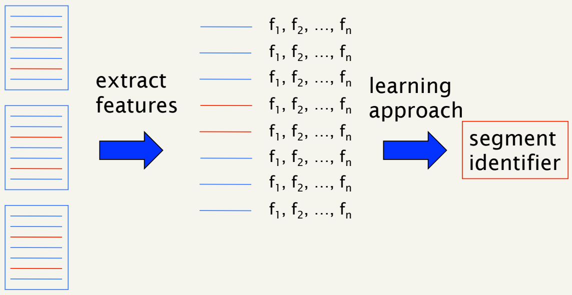
Summaries

Simplest heuristic: the first X words of the document

More sophisticated: extract from each document a set of “key” sentences

- Use heuristics to score each sentence
- Learning approach based on training data
- Summary is made up of top-scoring sentences

Segment identification



extract features

$f_1, f_2, \ldots, f_n$

$f_1, f_2, \ldots, f_n$

$f_1, f_2, \ldots, f_n$

$f_1, f_2, \ldots, f_n$

$f_1, f_2, \ldots, f_n$

$f_1, f_2, \ldots, f_n$

$f_1, f_2, \ldots, f_n$

$f_1, f_2, \ldots, f_n$

learning approach

segment identifier

hand-label “good” segments/sentences

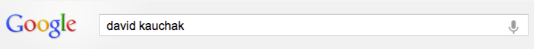
Summaries

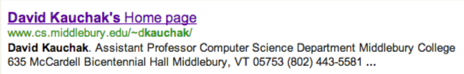
A **static summary** of a document is always the same, regardless of the query that hit the doc

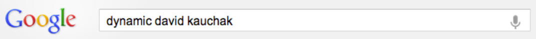
A **dynamic summary** is a *query-dependent* attempt to explain why the document was retrieved for the query at hand

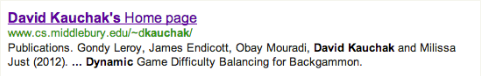
Which do most search engines use?

Summaries


david kauchak


David Kauchak's Home page
www.cs.middlebury.edu/~d**kauchak**/
David Kauchak. Assistant Professor Computer Science Department Middlebury College 635 McCardell Bicentennial Hall Middlebury, VT 05753 (802) 443-5581 **...**


dynamic david kauchak


David Kauchak's Home page
www.cs.middlebury.edu/~d**kauchak**/
Publications. Gondy Leroy, James Endicott, Obay Mouradi, **David Kauchak** and Milissa Just (2012). **...** **Dynamic** Game Difficulty Balancing for Backgammon.

Dynamic summaries

Present one or more "windows" within the document that contain several of the query terms

- "KWIC" snippets: Keyword in Context presentation

Generated in conjunction with scoring

- If query found as a phrase, all or some occurrences of the phrase in the doc
- If not, document windows that contain multiple query terms

The summary gives the entire content of the window – all terms, not only the query terms

Dynamic vs. Static

What are the benefits and challenges of each approach?

Static

- Create the summaries during indexing
- Don't need to store the documents

Dynamic

- Better user experience
- Makes the summarization process easier
- Must generate summaries on the fly and so must store documents and retrieve documents for every query!

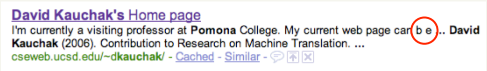
Generating dynamic summaries

If we *cache the documents* at index time, can find windows in it, cueing from hits found in the positional index

- E.g., positional index says "the query is a phrase in position 4378" so we go to this position in the cached document and stream out the content

Most often, cache only a fixed-size prefix of the doc

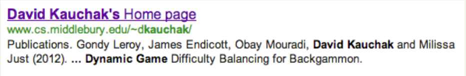
Note: Cached copy can be outdated!


David Kauchak's Home page
I'm currently a visiting professor at **Pomona** College. My current web page can b e ... **David Kauchak** (2006). Contribution to Research on Machine Translation. **...**
cseweb.ucsd.edu/~d**kauchak**/ - Cached - Similar -

Dynamic summaries

Producing good dynamic summaries is a tricky optimization problem

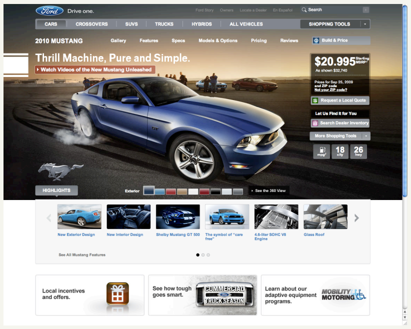
- The real estate for the summary is normally small and fixed
- Want short item, so show as many KWIC matches as possible, and perhaps other things like title



David Kauchak's Home page
www.cs.middlebury.edu/~**dkauchak**/
Publications. Gondy Leroy, James Endicott, Obay Mouradi, **David Kauchak** and Milissa Just (2012). ... **Dynamic Game** Difficulty Balancing for Backgammon.

Users really like snippets, even if they complicate IR system design

Challenge…



Drive one.
CARS
CROSSOVERS
SUVS
TRUCKS
HYBRIDS
ALL VEHICLES
SHOPPING TOOLS
2010 MUSTANG
Thrill Machine, Pure and Simple.
Watch Videos of the New Mustang Unleashed
$20,995
Request a Local Quote
Search Dealer Inventory
HIGHLIGHTS
Local incentives and offers.
See how tough goes smart.
Learn about our adaptive equipment programs.

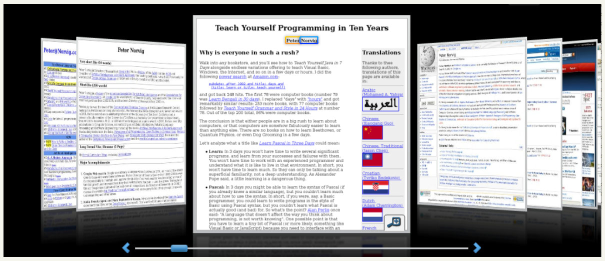
Challenge…

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/
xhtml1/DTD/xhtml1-strict.dtd"><html xmlns="http://www.w3.org/1999/xhtml"
xml:lang="en" lang="en"><head><script type="text/javascript">var __params =
{};__params.site = "bs"; // Used in DHTML Form library to identify brandsites
pages__params.model = "Mustang2010";__params.modelName =
"Mustang";__params.year = "2010";__params.make = "Ford";__params.segment =
"cars";__params.baseURL = "http://www.fordvehicles.com";__params.canonicalURL = "/
cars/mustang/";__params.anchorPage = "page";__params.domain="fordvehicles.com";</
script><script type="text/javascript" src="http://www.fordvehicles.com/ngtemplates/
ngassets/com/forddirect/ng/log4javascript.js?gtmo=ngbs"></script><script type="text/
javascript">log4javascript.setEnabled(false);var log = log ||
log4javascript.getDefaultLogger();if ( log4javascript.isEnabled() ) {log.info("Log
initialized");}</script><script language="javascript" type="text/
javascript">document.domain = "fordvehicles.com";</script><script type="text/
javascript">var akamaiQueryStringFound = false;var isCookieEnabled = false;/*Checking
For QueryString Parameters Being Present*/if (__params && __params.gtmo &&
__params.gtmo === "ngbs") {akamaiQueryStringFound = true;}/*Checking For Cookies
Being Enabled*/var cookieenabled = false;document.cookie = "testcookie=val";if
(document.cookie.indexOf("testcookie=") === -1) {isCookieEnabled = false;} else
{isCookieEnabled = true;}/*Redirection Check and Redirecting if required*/// Commenting
out the redirection logic for v0.27/*if ((!akamaiQueryStringFound) && (!isCookieEnabled))
{window.location.replace("http://www2.fordvehicles.com");}
```

Alternative results presentations?

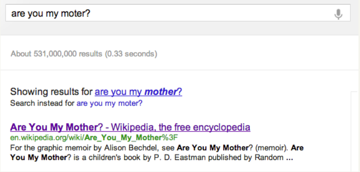
An active area of HCI research

An alternative: http://www.searchme.com/ copies the idea of Apple's Cover Flow for search results



Teach Yourself Programming in Ten Years
Peter Norvig
Why is everyone in such a rush?
Translations

Spelling correction

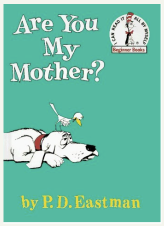


are you my moter?

About 531,000,000 results (0.33 seconds)

Showing results for are you my ***mother***?
Search instead for are you my moter?

Are You My Mother? - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Are_You_My_Mother**%3F
For the graphic memoir by Alison Bechdel, see **Are You My Mother**? (memoir). **Are You My Mother**? is a children's book by P. D. Eastman published by Random **...**

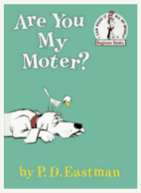


Spell correction

How might we utilize spelling correction?

Two common uses:

- Correcting user queries to retrieve “right” answers
- Correcting documents being indexed



Document correction

Especially needed for OCR’ed documents

- Correction algorithms are tuned for this
- Can use domain-specific knowledge
  - E.g., OCR can confuse O and D more often than it would confuse O and I (adjacent on the keyboard)

Web pages and even printed material have typos

Often we don’t change the documents but aim to fix the query-document mapping

Query misspellings

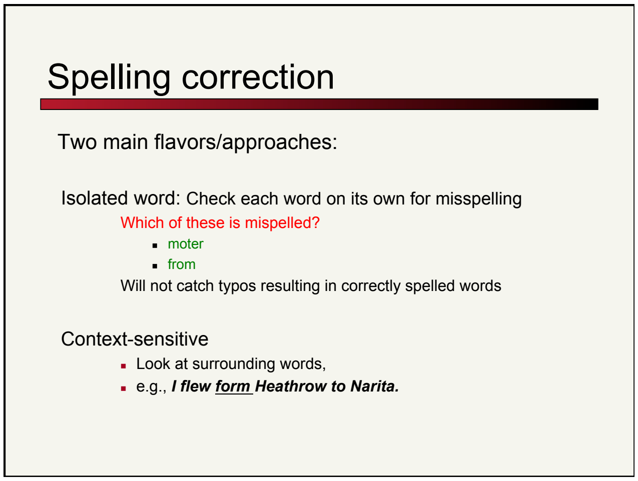
Our principal focus here

- e.g., the query ***Alanis Morisett***

What should/can we do?

- Retrieve documents indexed by the correct spelling
- Return several suggested alternative queries with the correct spelling
  - *Did you mean … ?*
- Return results for the incorrect spelling
- Some combination

Advantages/disadvantages?



Spelling correction

Two main flavors/approaches:

Isolated word: Check each word on its own for misspelling

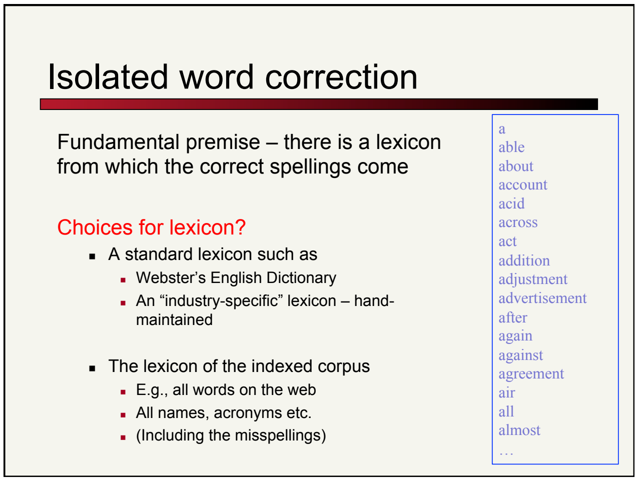
Which of these is mispelled?

- moter
- from

Will not catch typos resulting in correctly spelled words

Context-sensitive

- Look at surrounding words,
- e.g., ***I flew <u>form</u> Heathrow to Narita.***



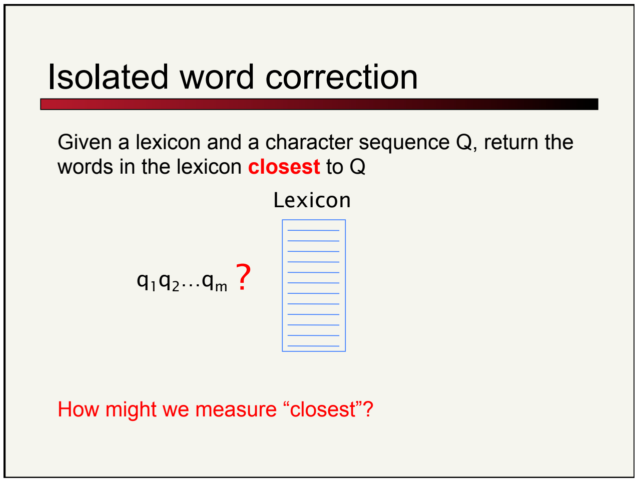
Isolated word correction

Fundamental premise – there is a lexicon from which the correct spellings come

Choices for lexicon?

- A standard lexicon such as
  - Webster's English Dictionary
  - An "industry-specific" lexicon – hand-maintained
- The lexicon of the indexed corpus
  - E.g., all words on the web
  - All names, acronyms etc.
  - (Including the misspellings)

a
able
about
account
acid
across
act
addition
adjustment
advertisement
after
again
against
agreement
air
all
almost
…



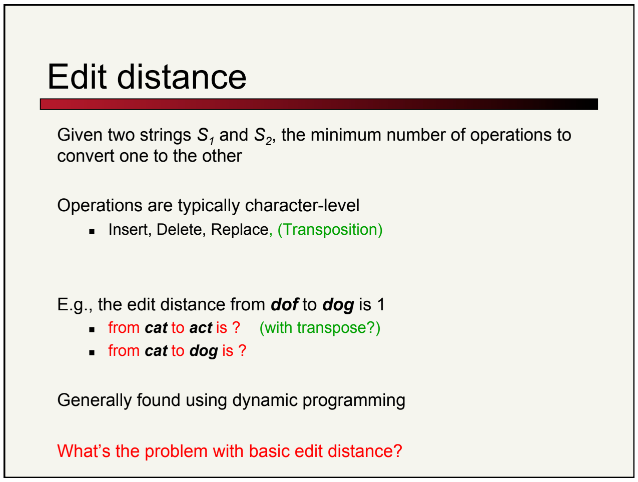
Isolated word correction

Given a lexicon and a character sequence Q, return the words in the lexicon **closest** to Q

Lexicon

$q_1 q_2 \ldots q_m$?

How might we measure "closest"?



Edit distance

Given two strings S_1 and S_2, the minimum number of operations to convert one to the other

Operations are typically character-level

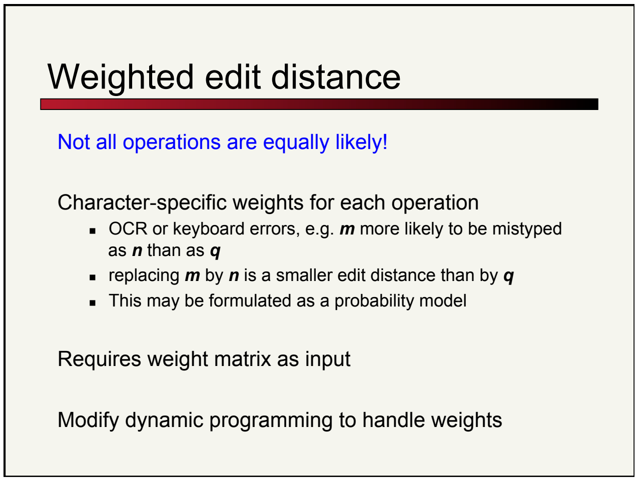
- Insert, Delete, Replace, (Transposition)

E.g., the edit distance from ***dof*** to ***dog*** is 1

- from ***cat*** to ***act*** is ? (with transpose?)
- from ***cat*** to ***dog*** is ?

Generally found using dynamic programming

What's the problem with basic edit distance?



Weighted edit distance

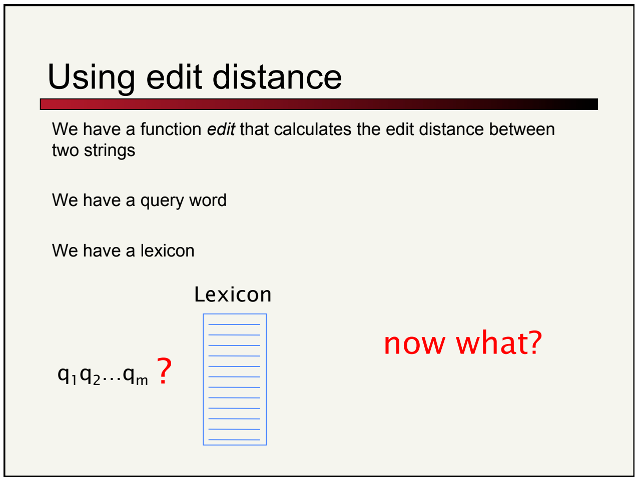
Not all operations are equally likely!

Character-specific weights for each operation

- OCR or keyboard errors, e.g. ***m*** more likely to be mistyped as ***n*** than as ***q***
- replacing ***m*** by ***n*** is a smaller edit distance than by ***q***
- This may be formulated as a probability model

Requires weight matrix as input

Modify dynamic programming to handle weights



Using edit distance

We have a function *edit* that calculates the edit distance between two strings

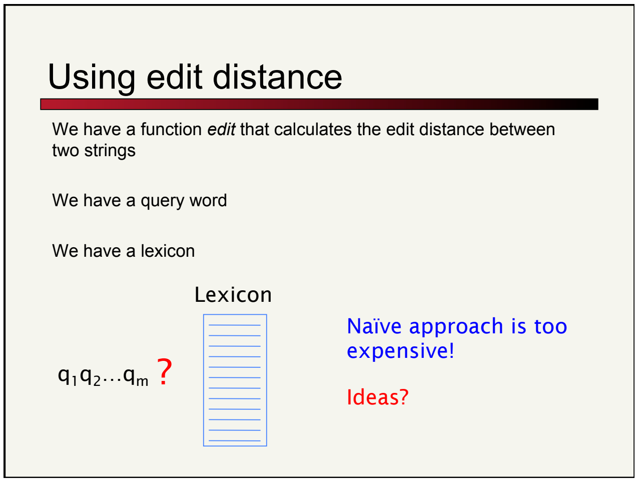
We have a query word

We have a lexicon

Lexicon

$q_1q_2\ldots q_m$?

now what?



Using edit distance

We have a function *edit* that calculates the edit distance between two strings

We have a query word

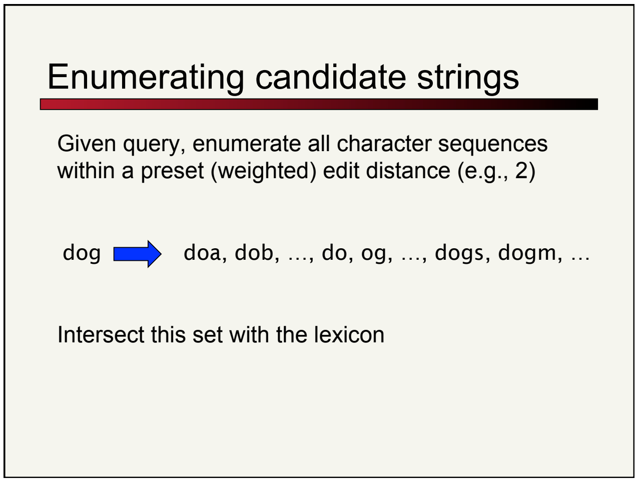
We have a lexicon

Lexicon

$q_1q_2\ldots q_m$?

Naïve approach is too expensive!

Ideas?

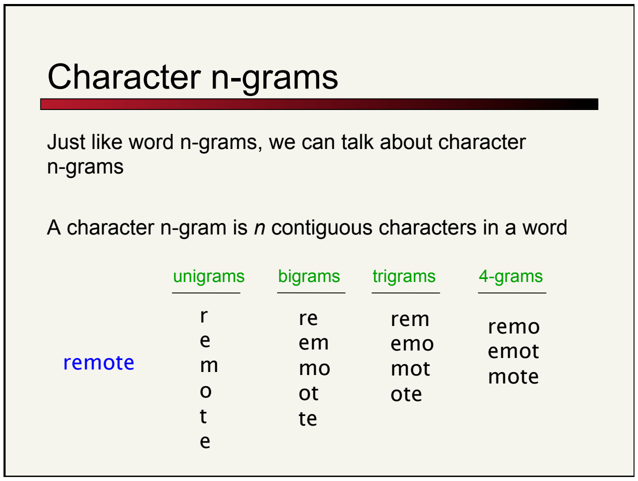


Enumerating candidate strings

Given query, enumerate all character sequences within a preset (weighted) edit distance (e.g., 2)

dog → doa, dob, …, do, og, …, dogs, dogm, …

Intersect this set with the lexicon

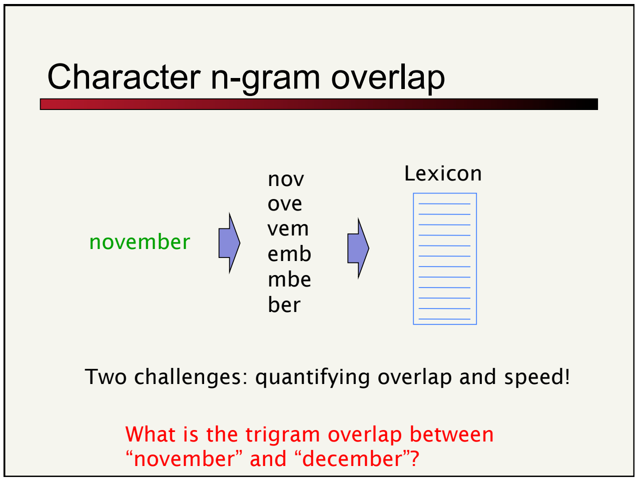


Character n-grams

Just like word n-grams, we can talk about character n-grams

A character n-gram is *n* contiguous characters in a word

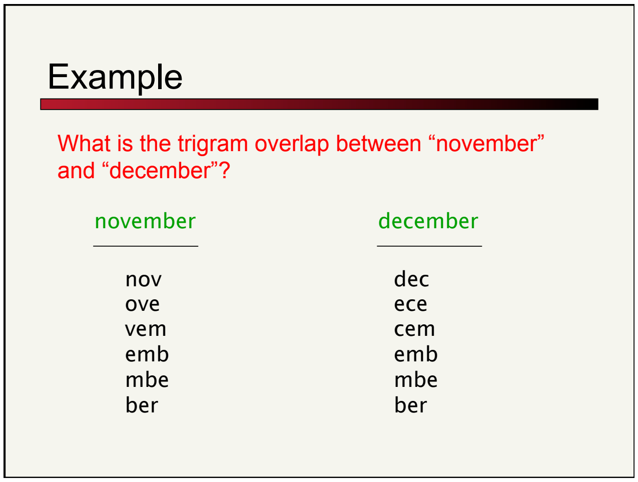
	unigrams	bigrams	trigrams	4-grams
remote	r	re	rem	remo
	e	em	emo	emot
	m	mo	mot	mote
	o	ot	ote	
	t	te		
	e			



Character n-gram overlap

Two challenges: quantifying overlap and speed!

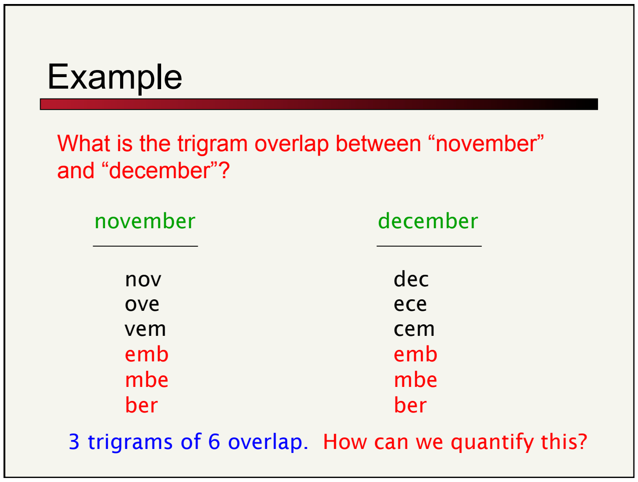
What is the trigram overlap between "november" and "december"?



Example

What is the trigram overlap between "november" and "december"?

november	december
nov	dec
ove	ece
vem	cem
emb	emb
mbe	mbe
ber	ber

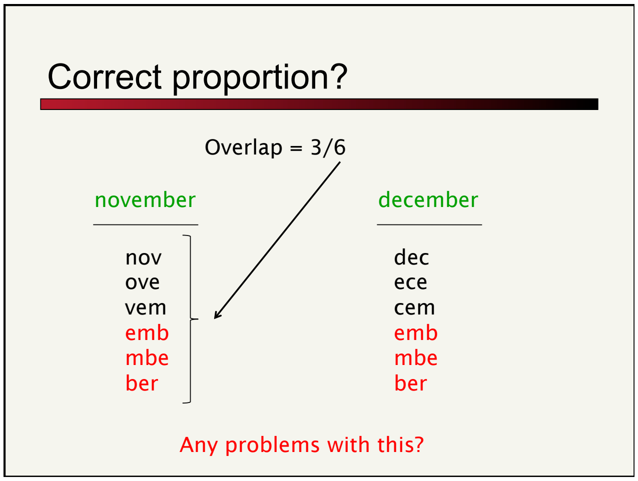


Example

What is the trigram overlap between "november" and "december"?

november	december
nov	dec
ove	ece
vem	cem
emb	emb
mbe	mbe
ber	ber

3 trigrams of 6 overlap. How can we quantify this?

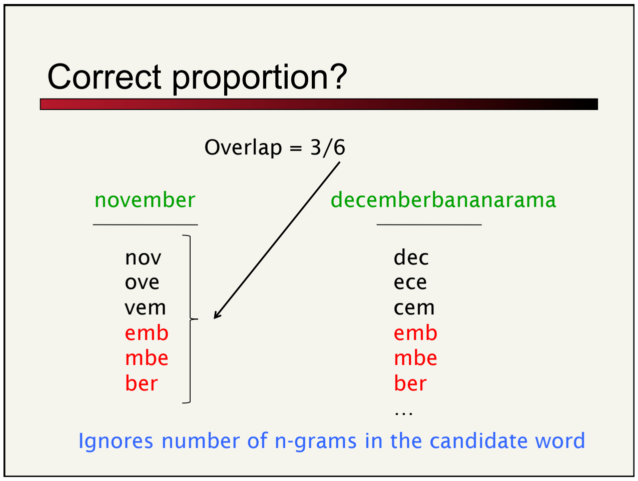


Correct proportion?

Overlap = 3/6

november	december
nov	dec
ove	ece
vem	cem
emb	emb
mbe	mbe
ber	ber

Any problems with this?

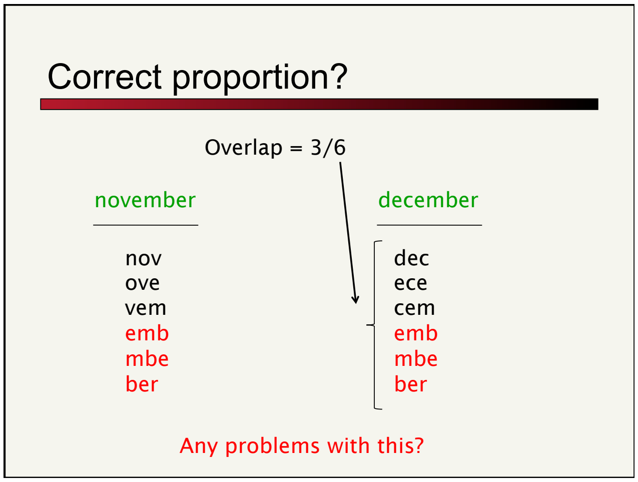


Correct proportion?

Overlap = 3/6

november	decemberbananarama
nov	dec
ove	ece
vem	cem
emb	emb
mbe	mbe
ber	ber
	…

Ignores number of n-grams in the candidate word

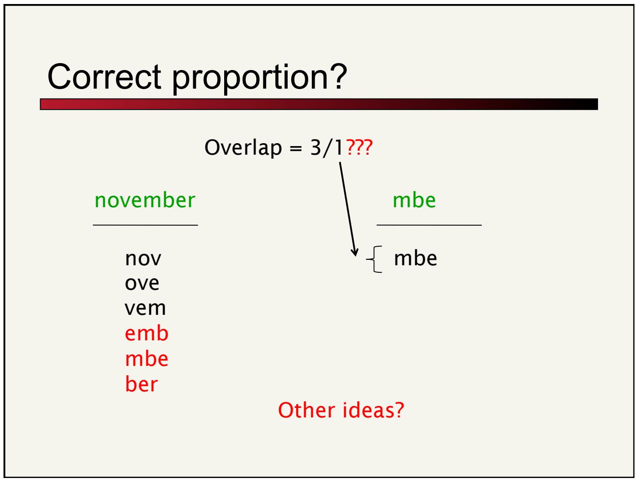


Correct proportion?

Overlap = 3/6

november	december
nov	dec
ove	ece
vem	cem
emb	emb
mbe	mbe
ber	ber

Any problems with this?



Correct proportion?

Overlap = 3/1???

november	mbe
nov	mbe
ove	
vem	
emb	
mbe	
ber	

Other ideas?

One option – Jaccard coefficient

Let *X* and *Y* be two sets; then the J.C. is

$$|X \cap Y| / |X \cup Y|$$

What does this mean?

$\lvert X \cap Y\rvert$	number of overlapping n-grams
$\lvert X \cup Y\rvert$	total n-grams between the two

Example

november	december
nov	dec
ove	ece
vem	cem
emb	emb
mbe	mbe
ber	ber

$|X \cap Y|$ 3

$|X \cup Y|$ 9

JC = 1/3

Jaccard coefficient

Equals 1 when *X* and *Y* have the same elements and zero when they are disjoint

X and *Y* don’t have to be the same size

Always assigns a number between 0 and 1

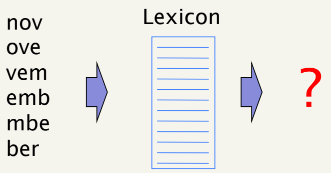
Threshold to decide if you have a match

- E.g., if J.C. > 0.8, declare a match

Efficiency

We have all the n-grams for our query word

How can we efficiently compute the words in our lexicon that have non-zero n-gram overlap with our query word?

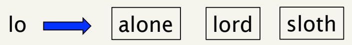


Efficiency

We have all the n-grams for our query word

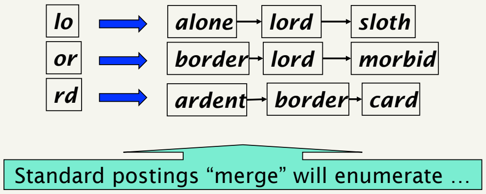
How can we efficiently compute the words in our lexicon that have non-zero n-gram overlap with our query word?

Index the words by n-grams!



Matching trigrams

Consider the query ***lord*** – we wish to identify words matching 2 of its 3 bigrams (***lo, or, rd***)



Standard postings "merge" will enumerate …

Adapt this to using Jaccard (or another) measure.

Context-sensitive spell correction

Text: ***I flew <u>from</u> Heathrow to Narita.***

Consider the phrase query
"flew <u>form</u> Heathrow"

We'd like to respond: Did you mean "***flew from Heathrow***"?

How might you do this?

Context-sensitive correction

Similar to isolated correction, but incorporate surrounding context

Retrieve dictionary terms close to each query term (e.g. isolated spelling correction)

Try all possible resulting phrases with one word "fixed" at a time

- ***flew from heathrow***
- ***fled form heathrow***
- ***flea form heathrow***

Rank alternatives based on frequency in corpus

Can we do this efficiently?

Another approach?

What do you think the search engines actually do?

Often a combined approach

Generally, context-sensitive correction

One overlooked resource so far…

Query logs

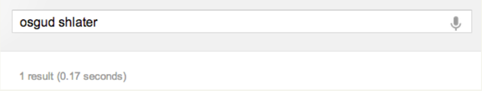
AnonID	Query	QueryTime	ItemRank	ClickURL
2524140	osgood-schlatter syndrome	2006-05-18 15:07:58	1	http://www.medic8.com
2524140	osgood-schlatter syndrome	2006-05-18 15:07:58	2	http://www.disability.vic.gov.au
2524140	osgood-schlatter syndrome	2006-05-18 15:07:58	3	http://www.emedicine.com
2524140	evergreen real estate co.	2006-05-19 09:33:08	4	http://www.homegain.com
2524140	evergreen real estate co. sc	2006-05-19 09:33:42	3	http://www.sciway.net
2524140	evergreen real estate co. sc	2006-05-19 09:33:42	3	http://www.sciway.net
2524140	evergreen real estate co. sc	2006-05-19 09:33:42	7	http://www.eraevergreen.com
2524140	westgatevacationvillas	2006-05-19 18:41:35	1	http://www.vacationrentals.com
2524140	westgatevacationvillas	2006-05-19 18:41:35	2	http://www.aberfoyleholidays.com
2524140	westgatevacationvillas	2006-05-19 18:41:35	4	http://www.funtastik.com
2524140	westgate vacation villas	2006-05-19 18:44:07	2	http://www.westgateresorts.com
2524140	hilton head vacation	2006-05-19 20:37:12	1	http://www.vacationcompany.com
2524140	hilton head vacation	2006-05-19 20:37:12	2	http://www.hiltonheadvacation.com

How might we use query logs to assist in spelling correction?

Query logs

Find similar queries
“flew form heathrow” and “flew from heathrow”

Query logs contain a temporal component!



osgud shlater

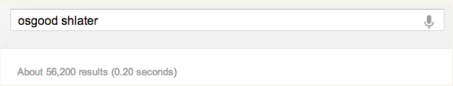
1 result (0.17 seconds)

Attempt 1: one doc retrieved, don’t click on any docs

Query logs

Find similar queries
“flew form heathrow” and “flew from heathrow”

Query logs contain a temporal component!



osgood shlater

About 56,200 results (0.20 seconds)

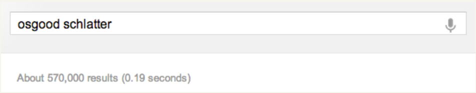
Attempt 2: may docs retrieved
click on one doc, but quickly issue another query

Query logs

Find similar queries

“flew form heathrow” and “flew from heathrow”

Query logs contain a temporal component!



osgood schlatter

About 570,000 results (0.19 seconds)

Attempt 3: even more docs retrieved
click on one doc, then no more activity

General issues in spell correction

Do we enumerate multiple alternatives for “Did you mean?”

Need to figure out which to present to the user

Use heuristics

- The alternative hitting most docs
- Query log analysis + tweaking
  - For especially popular, topical queries

Spell-correction is computationally expensive

- Avoid running routinely on every query?
- Run only on queries that matched few docs