

THIS TRICK MAY ONLY WORK 1% OF THE TIME,
BUT WHEN IT DOES, IT'S TOTALLY WORTH IT.

Results Summaries

Spelling Correction

David Kauchak

cs160

Fall 2009

adapted from:

<http://www.stanford.edu/class/cs276/handouts/lecture3-tolerantretrieval.ppt>

<http://www.stanford.edu/class/cs276/handouts/lecture8-evaluation.ppt>

Administrative

- Course feedback

Google

Search

[Advanced Search](#)

www.fordvehicles.com/cars/mustang/

en.wikipedia.org/wiki/Ford_Mustang

www.mustangseats.com/

www.mustangsurvival.com/

[Advanced Search](#)

www.fordvehicles.com/cars/mustang/

en.wikipedia.org/wiki/Ford_Mustang

www.mustangseats.com/

www.mustangsurvival.com/

[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/

[Advanced Search](#)

[2010 For Mustang | Official Site of the Ford Mustang](#)

2010 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/

[Advanced Search](#)

[2010 For 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 metadata about the documents we're retrieving
- For web pages, we have titles and URLs
- For other collections, we may have other types of information
- For academic articles, what information do we have?

[PDF] ► [Modeling word burstiness using the Dirichlet distribution](#)

RE Madsen, D **Kauchak**, C Elkan - MACHINE LEARNING-INTERNATIONAL WORKSHOP ..., 2005 - cseweb.ucsd.edu

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 ...

[Cited by 53](#) - [Related articles](#) - [View as HTML](#) - [BL Direct](#) - [All 22 versions](#)

[Paraphrasing for automatic evaluation](#) - ► [upenn.edu](#) [PDF]

D **Kauchak**, R Barzilay - Proceedings of the main conference on Human Language ..., 2006 - portal.acm.org

This paper studies the impact of para- phrases on the accuracy of automatic eval- uation. Given a reference sentence and a machine-generated sentence, we seek to find a paraphrase of the reference sen- tence that is closer in ...

[Cited by 41](#) - [Related articles](#) - [Resources @ My Library](#) - [All 16 versions](#)

[PDF] ► [Sources of success for information extraction methods](#) - [Full-Text @ My Library](#)

D **Kauchak**, J Smarr, C Elkan - Journal of Machine Learning Research, 2004 - www-cse.ucsd.edu

Page 1. Sources of Success for Information Extraction Methods **David Kauchak** Joseph Smarr Charles Elkan Dept. of Computer Science Symbolic Systems Program Dept. of Computer Science UC San Diego Stanford University UC San Diego ...

[Cited by 9](#) - [Related articles](#) - [View as HTML](#) - [All 10 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

[09 Ford Mustang Prices](#)

Compare Prices on a New **Mustang**
Get a Low Price Today!

FordMustang.Motortree.com

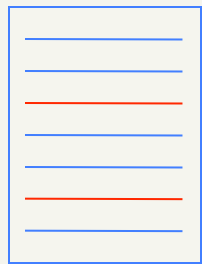
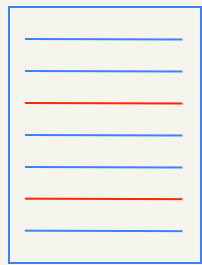
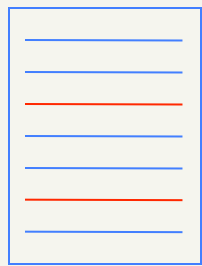
Summaries

- We can generate these ourselves!
- Lots of summarization techniques
- Most common (and successful) approach is to extract segments from the documents
- 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!

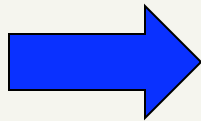
Summaries

- In typical systems, the summary is a subset of the document
- Simplest heuristic: the first 50 (or so – this can be varied) words of the document
- More sophisticated: extract from each document a set of “key” sentences
 - Simple NLP heuristics to score each sentence
 - Learning approach based on training data
 - Summary is made up of top-scoring sentences

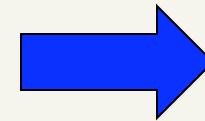
Segment identification



extract
features



learning
approach



segment
identifier

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

[David Kauchak's Home page](#)

Rasmus E. Madsen, **David Kauchak** and Charles Elkan (2005). Modeling Word Burstiness Using the Dirichlet Distribution. In Proceedings of the Twenty-Second ...

cseweb.ucsd.edu/~dkauchak/ - [Cached](#) - [Similar](#) -   

[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/~dkauchak/ - [Cached](#) - [Similar](#) -   

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 itself 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
 - Unfortunately, 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 be .. David Kauchak (2006). Contribution to Research on Machine Translation. ...

cseweb.ucsd.edu/~dkauchak/ - [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](#)

Rasmus E. Madsen, **David Kauchak** and Charles Elkan (2005). Modeling Word Burstiness Using the Dirichlet Distribution. In Proceedings of the Twenty-Second ...

cseweb.ucsd.edu/~dkauchak/ - [Cached](#) - [Similar](#) -   

- But users really like snippets, even if they complicate IR system design

Challenge...

Drive one. Ford Story Owners Locate a Dealer En Español Search

CARS CROSSOVERS SUVS TRUCKS HYBRIDS ALL VEHICLES SHOPPING TOOLS

2010 MUSTANG Gallery Features Specs Models & Options Pricing Reviews Build & Price

Thrill Machine, Pure and Simple.
Watch Videos of the New Mustang Unleashed

\$20,995 Starting MSRP
As shown \$32,740
Prices for Sep 25, 2009 and ZIP code Not your ZIP code?
Request a Local Quote
Let Us Find It for You
Search Dealer Inventory
More Shopping Tools
18 mpg^{city} 26 mpg^{hwy}

HIGHLIGHTS Exterior See the 360 View

New Exterior Design New Interior Design Shelby Mustang GT 500 The symbol of "care free" 4.6-liter SOHC V8 Engine Glass Roof

See All Mustang Features

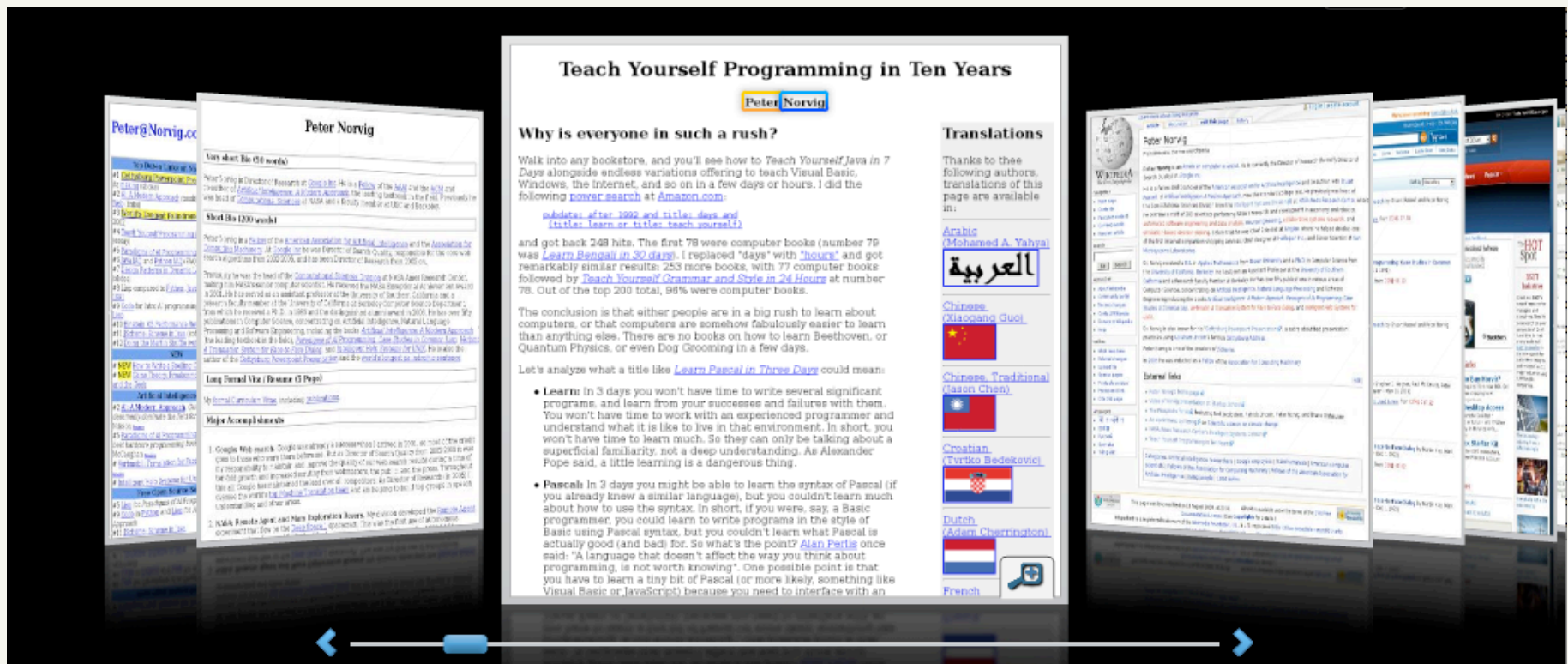
Local incentives and offers. See how tough goes smart. COMMERCIAL TRUCK SEASON Learn about our adaptive equipment programs. MOBILITY MOTORING

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



Spelling correction

Google

are you my moter?

Search

Web [Show options...](#) Results 1 - 10 of about 198,000,000 for are you my moter?. (0.27 seconds)

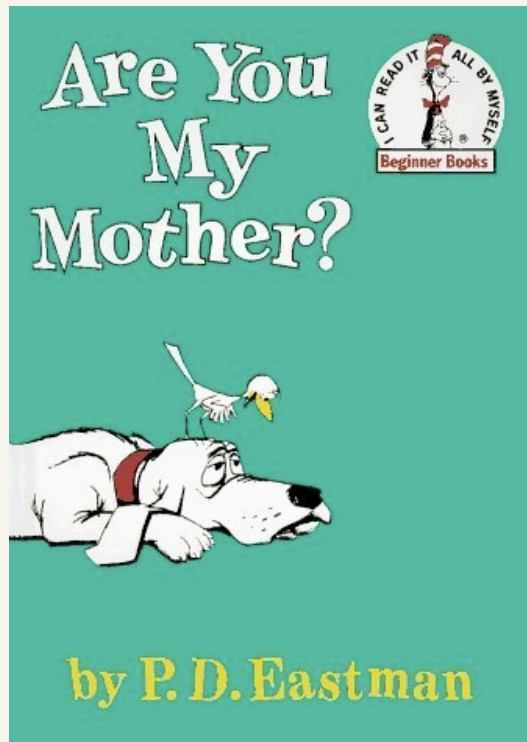
Did you mean: [are you my mother?](#)

[Amazon.com: Are You My Mother? \(9780394800189\): P.D. Eastman: Books](#)

Are You My Mother? follows a confused baby bird who's been denied the experience of

"Are You My Mother" is a fun little picture book with simple text, ...

www.amazon.com/Are-You-Mother-P-D.../0394800184 - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [✕](#)



Google?

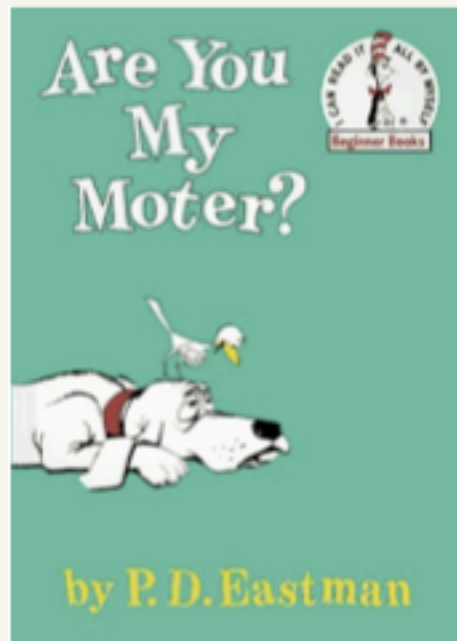
'Googlle': Google Logo Spelled Wrong To Celebrate 11th Anniversary (PHOTOS)

As you can see from a screen shot of [Google Trends](#), the search engine's new doodle seems to have created a lot of confusion. Two of the top ten queries on Google trends were "why does Google have two 'l's" and "why is Google spelled wrong?". Maybe "Goog11e" would have created less of a stir.

1. [roman polanski](#)
2. [tawny kitaen](#)
3. [samantha geimer](#)
4. [nephilim](#)
5. [why does google have two l s](#)

Spell correction

- How might we utilize spelling correction?
- Two principal 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 QWERTY keyboard, so more likely interchanged in typing).
- 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 Morisset***
- We can either
 - Retrieve documents indexed by the correct spelling
 - Return several suggested alternative queries with the correct spelling
 - *Did you mean ... ?*
 - **Advantages/disadvantages?**

Spell correction

- Two main flavors/approaches:
 - Isolated word
 - Check each word on its own for misspelling
 - Which of these is misspelled?
 - moter
 - from
 - Will not catch typos resulting in correctly spelled words
 - Context-sensitive
 - Look at surrounding words,
 - e.g., *I flew form Heathrow to Narita.*

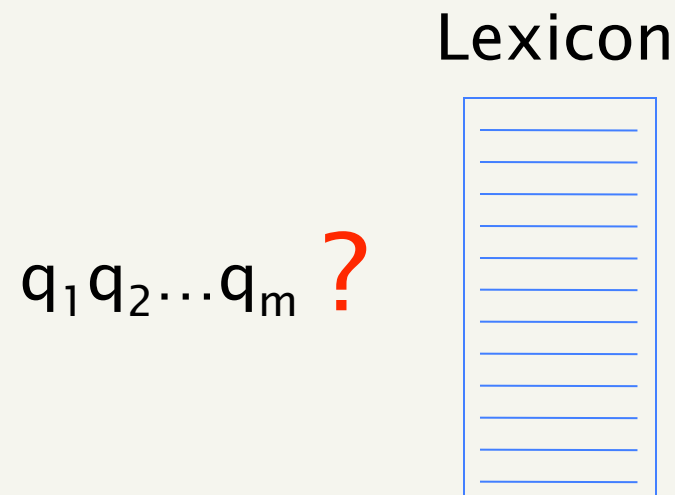
Isolated word correction

- Fundamental premise – there is a lexicon from which the correct spellings come
- Two basic choices for this
 - 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 mis-spellings)

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



- How might we measure “closest”?
 - Edit distance (Levenshtein distance)
 - Weighted edit distance
 - n -gram overlap

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 by dynamic programming
- See <http://www.merriampark.com/ld.htm> for a nice example plus an applet.

Weighted edit distance

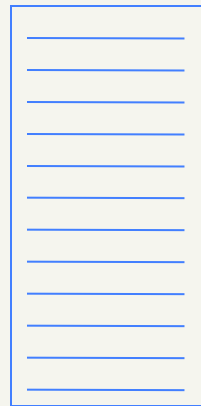
- As above, but the weight of an operation depends on the character(s) involved
 - Meant to capture OCR or keyboard errors, e.g. *m* more likely to be mis-typed as *n* than as *q*
 - Therefore, replacing *m* by *n* is a smaller edit distance than by *q*
 - This may be formulated as a probability model
- 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

$q_1 q_2 \dots q_m$?

Lexicon



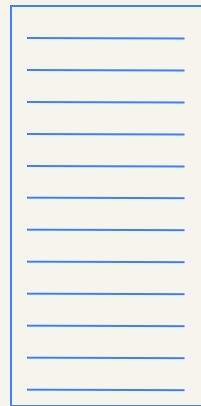
now what?

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

$q_1 q_2 \dots q_m$?

Lexicon

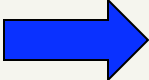


We need to reduce the candidate lexicon set

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

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

Character n -gram overlap

- Enumerate all n -grams in the query string
 - Identify all lexicon terms matching any of the query n -grams
 - Threshold by number of matching n -grams
 - Weight by keyboard layout, etc.
-
- What is the trigram overlap between “november” and “december”?

Example

- What is the trigram overlap between “november” and “december”?

november

nov
ove
vem
emb
mbe
ber

december

dec
ece
cem
emb
mbe
ber

Example

- What is the trigram overlap between “november” and “december”?

november

nov

ove

vem

emb

mbe

ber

december

dec

ece

cem

emb

mbe

ber

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

One option – Jaccard coefficient

- A commonly-used measure of overlap
- Let X and Y be two sets; then the J.C. is

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

- What does this mean?

$|X \cap Y|$ number of overlapping ngrams

$|X \cup Y|$ total n-grams between the two

Example

november

nov
ove
vem
emb
mbe
ber

december

dec
ece
cem
emb
mbe
ber

$$|X \cap Y| \quad 3$$

$$|X \cup Y| \quad 9$$

$$JC = 1/3$$

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 of 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

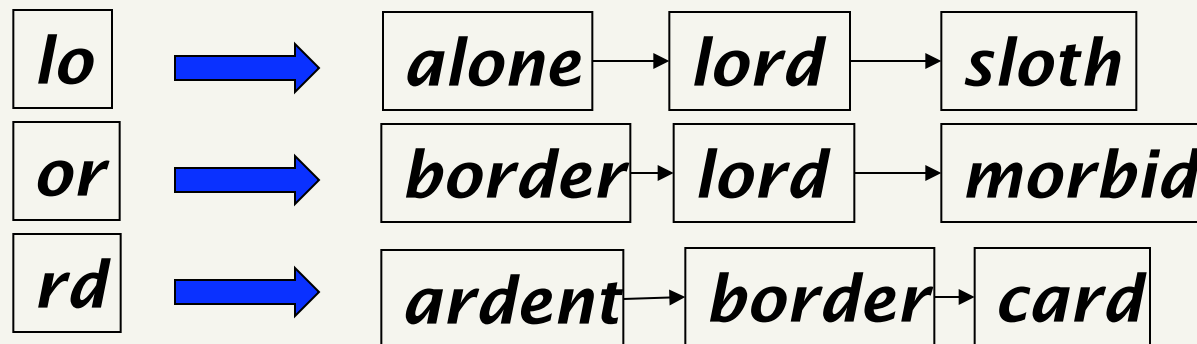
Efficiency

- 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?
- Index the words by n-grams!

lo → alone lord sloth

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 from Heathrow to Narita.*
- Consider the phrase query
“flew form 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

- How might we use query logs to assist in spelling correction?
 - Find similar queries, e.g. “flew form heathrow” and “flew from heathrow”
 - Query logs contain a temporal component!
 - User will type “flew form heathrow”
 - Not get any results (or any relevant results)
 - User will issue another query “flew from heathrow”
 - We can make this connection!

General issues in spell correction

- 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

Dictionaries

- How do new ways to get information change our information needs?
- How does our reliance on Google for all different types of information affect its ability to provide specific useful information?
- How does the undermining of traditional sources of rules of English (and other languages) affect society's use of English?