How do you implement a "Did you mean"? [duplicate]

Asked 16 years, 3 months ago Modified 12 years, 1 month ago

Viewed 33k times Part of NLP Collective

118

This question already has answers here:

Closed 11 years ago.

> **Possible Duplicate:**
> How does the Google “Did you mean?” Algorithm work?

Suppose you have a search system already in your website. How can you implement the "Did you mean: `<spell_checked_word>`" like Google does in some search queries?

NLP nlp

Share

Improve this question

Follow

edited May 23, 2017 at 12:09

Community Bot

1 • 1

asked Sep 3, 2008 at 10:36

pek
18k ● 28 ● 88 ● 99

@pek: I had the same thought a while ago... Have you thought of using an HTML scruber and using Google as the source of the corrections? – Ande Turner Nov 3, 2008 at 9:53

See stackoverflow.com/questions/3763640/… – John Sep 23, 2010 at 10:56

17 Answers

Sorted by: Highest score (default)

87

Actually what Google does is very much non-trivial and also at first counter-intuitive. They don't do anything like check against a dictionary, but rather they make use of statistics to identify "similar" queries that returned more results than your query, the exact algorithm is of course not known.

There are different sub-problems to solve here, as a fundamental basis for all Natural Language Processing statistics related there is one must have book: Foundation of Statistical Natural Language Processing.

Concretely to solve the problem of word/query similarity I have had good results with using Edit Distance, a mathematical measure of string similarity that works surprisingly well. I used to use Levenshtein but the others may be worth looking into.

Soundex - in my experience - is crap.

Actually efficiently storing and searching a large dictionary of misspelled words and having sub second retrieval is again non-trivial, your best bet is to make use of existing full text indexing and retrieval engines (i.e. not your database's one), of which Lucene is currently one of the best and coincidentally ported to many many platforms.

Share Improve this answer

Follow

answered Sep 3, 2008 at 10:55

Boris Terzic

10.9k ● 8 ● 46 ● 60

35

Google's Dr Norvig has outlined how it works; he even gives a 20ish line Python implementation:

http://googlesystem.blogspot.com/2007/04/simplified-version-of-googles-spell.html

http://www.norvig.com/spell-correct.html

Dr Norvig also discusses the "did you mean" in this excellent talk. Dr Norvig is *head of research* at Google - when asked how "did you mean" is implemented, his answer is ***authoritive***.

So its spell-checking, presumably with a dynamic dictionary build from other searches or even actual internet phrases and such. But that's still *spell checking*.

SOUNDEX and other guesses don't get a look in, people!

Share Improve this answer Follow

edited Oct 27, 2011 at 5:04

answered Nov 3, 2008 at 10:33

Will

75.6k ● 43 ● 174 ● 255

4 Dr. Norvig provided a toy example of the concept; it's not nearly accurate enough to provide 'did you mean' for web. For example: "barak" does not produce a suggestion; "barak obama" does (since they know "barack" occurs often with obama, and can infer the likely correction – SquareCog Jan 16, 2009 at 22:27

2 it isn't hard to go from his toy spell checker to something that does handle your example and that works well. An important thing to remember is that he is showing a spell checker which is subtly but significantly different from a query suggester. Training it with previous queries instead of english text is a good place to start. – jshen Jul 17, 2009 at 17:52

There's definitely more to it than just spell-checking. For one thing, I've seen cases where neither the thing I typed nor the suggested replacement are "dictionary words". – Ryan Lundy Mar 30, 2010 at 20:57

1 @Kyralessa: do you think their dictionary is static words in some language, or dynamic and based on the words on the internet and common search terms? It doesn't it isn't still a dictionary check. Dr Novig is after all head of research at google - when asked how "did you mean" is implemented, his answer is authoritive. – Will Mar 31, 2010 at 6:52

Check this article on wikipedia about the Levenshtein distance. Make sure you take a good look at Possible

13

improvements.

Share Improve this answer

Follow

answered Sep 3, 2008 at 10:49

Ionut Anghelcovici

> The most common edit distance calculation. A common way to do this is the Wagner-Fischer algorithm. – Giuliano Apr 14, 2012 at 2:50

11

I was pleasantly surprised that someone has asked how to create a state-of-the-art spelling suggestion system for search engines. I have been working on this subject for more than a year for a search engine company and I can point to information on the public domain on the subject.

As was mentioned in a previous post, Google (and Microsoft and Yahoo!) do not use any predefined dictionary nor do they employ hordes of linguists that ponder over the possible misspellings of queries. That would be impossible due to the scale of the problem but also because it is not clear that people could actually correctly identify when and if a query is misspelled.

Instead there is a simple and rather effective principle that is also valid for all European languages. Get all the unique queries on your search logs, calculate the edit distance between all pairs of queries, assuming that the reference query is the one that has the highest count.

This simple algorithm will work great for many types of queries. If you want to take it to the next level then I suggest you read the paper by Microsoft Research on that subject. You can find it here

The paper has a great introduction but after that you will need to be knowledgeable with concepts such as the Hidden Markov Model.

Share Improve this answer

Follow

answered May 5, 2009 at 7:06

Costas Boulis

301 ● 2 ● 3

6

You may want to look at Peter Norvig's "How to Write a Spelling Corrector" article.

Share Improve this answer

Follow

answered Nov 1, 2008 at 6:45

FA.

97 ● 3 ● 4

6

I believe Google logs all queries and identifies when someone makes a spelling correction. This correction may then be suggested when others supply the same first query. This will work for any language, in fact any string of any characters.

Share Improve this answer

answered Nov 3, 2008 at 9:41

Follow

Liam
20.9k ● 24 ● 90 ● 129

They do indeed. This helps them learn new words easily -- they have the help of millions. – A. Rex Jan 16, 2009 at 7:32

2 Yes, this is actually the correct answer. According to the book "In the Plex", Google looks for cases where someone searches for something, gets results, then immediately adjusts their search terms a little bit. – Joel Spolsky Jun 21, 2011 at 3:21

5

I would suggest looking at SOUNDEX to find similar words in your database.

You can also access google own dictionary by using the Google API spelling suggestion request.

Share Improve this answer

Follow

edited Sep 3, 2008 at 10:46

answered Sep 3, 2008 at 10:39

Espo
41.9k ● 21 ● 136 ● 161

1 +1 for the link to the Google API which seems to be exactly what the asker was looking for, even if the chosen answer is more in depth and answers the 'why' and 'how' of Google's implementation. – dimo414 Jun 15, 2009 at 7:59

http://en.wikipedia.org/wiki/N-gram#Google_use_of_N-gram

4

Share Improve this answer

Follow

answered Sep 3, 2008 at 11:00

robaker

1,029 ● 7 ● 11

> Could you expand on this, in case your link dies of link-rot or Rampant Deletionism? The anchor is already dead...
> – Michael Paulukonis Apr 9, 2012 at 19:34

4

I think this depends on how big your website it. On our local Intranet which is used by about 500 member of staff, I simply look at the search phrases that returned zero results and enter that search phrase with the new suggested search phrase into a SQL table.

I them call on that table if no search results has been returned, however, this only works if the site is relatively small and I only do it for search phrases which are the most common.

You might also want to look at my answer to a similar question:

- "Similar Posts" like functionality using MS SQL Server?

Share Improve this answer

Follow

edited May 23, 2017 at 12:34

Community Bot

1 ● 1

answered Sep 3, 2008 at 13:11

GateKiller

75.8k ● 75 ● 175 ● 204

2

If you have industry specific translations, you will likely need a thesaurus. For example, I worked in the jewelry industry and there were abbreviate in our descriptions such as kt - karat, rd - round, cwt - carat weight... Endeca (the search engine at that job) has a thesaurus that will translate from common misspellings, but it does require manual intervention.

Share Improve this answer

Follow

answered Sep 3, 2008 at 13:04

oglester

6,655 ● 8 ● 44 ● 64

1

I do it with Lucene's Spell Checker.

Share Improve this answer

Follow

answered May 5, 2009 at 6:27

cherouvim

31.9k ● 15 ● 105 ● 155

0

Soundex is good for phonetic matches, but works best with peoples' names (it was originally developed for census data)

Also check out Full-Text-Indexing, the syntax is different from Google logic, but it's very quick and can deal with similar language elements.

Share Improve this answer

Follow

answered Sep 3, 2008 at 10:41

Keith

155k ● 82 ● 306 ● 446

one of the bad things of soundex is that it's too english-centric – Javier Oct 22, 2008 at 13:19

It was developed to Anglisize names, so Smith and Schmidt are suppose to match in it. Metaphone is better but has a similar problem. Any phonetic algorithm is going to be language dependant. – Keith Oct 22, 2008 at 15:40

0

Soundex and "Porter stemming" (soundex is trivial, not sure about porter stemming).

Share Improve this answer

Follow

answered Sep 3, 2008 at 10:46

Michael Neale

19.5k ● 19 ● 78 ● 111

1 Information (including implementations in 19 different coding languages) on Porter stemming can be found at

tartarus.org/~martin/PorterStemmer/index.html – msanders
Nov 3, 2008 at 11:24

0

There's something called aspell that might help: http://blog.evanweaver.com/files/doc/fauna/raspell/classes/Aspell.html

There's a ruby gem for it, but I don't know how to talk to it from python http://blog.evanweaver.com/files/doc/fauna/raspell/files/README.html

Here's a quote from the ruby implementation

> Usage
>
> Aspell lets you check words and suggest corrections. For example:
>
> ```
>   string = "my haert wil go on"
>
>   string.gsub(/[\w\']+/) do |word|
>     if !speller.check(word)
>       # word is wrong
>       puts "Possible correction for #
> {word}:"
>       puts speller.suggest(word).first
>     end
>   end
> ```

This outputs:

Possible correction for haert: heart Possible correction for wil: Will

Share Improve this answer

Follow

edited Jun 20, 2020 at 9:12

Community Bot

1 • 1

answered Nov 19, 2008 at 17:37

Vishu

0

Implementing spelling correction for search engines in an effective way is not trivial (you can't just compute the edit/levenshtein distance to every possible word). A solution based on k-gram indexes is described in Introduction to Information Retrieval (full text available online).

Share Improve this answer

Follow

answered Jan 16, 2009 at 22:20

Fabian Steeg

45.6k • 7 • 87 • 113

0

U could use ngram for the comparisment:
http://en.wikipedia.org/wiki/N-gram

Using python ngram module:
http://packages.python.org/ngram/index.html

```
import ngram

G2 = ngram.NGram([  "iis7 configure ftp 7.5",
```

```
                      "ubunto configre 8.5",
                      "mac configure ftp"])

print "String", "\t", "Similarity"
for i in G2.search("iis7 configurftp 7.5",
threshold=0.1):
    print i[1], "\t", i[0]
```

U get:

```
>>>
String  Similarity
0.76    "iis7 configure ftp 7.5"
0.24    "mac configure ftp"
0.19    "ubunto configre 8.5"
```

Share Improve this answer

Follow

answered Oct 8, 2010 at 7:35

hugo24

1,109 ● 13 ● 22

0

Why not use google's did you mean in your code.For how see here http://narenonit.blogspot.com/2012/08/trick-for-using-googles-did-you-mean.html

Share Improve this answer

Follow

answered Aug 20, 2012 at 12:30

Narendra Rajput

711 ● 9 ● 29

1 Page is not found anymore... :'(– Lauro Nov 1, 2016 at 11:35