Information Retrieval

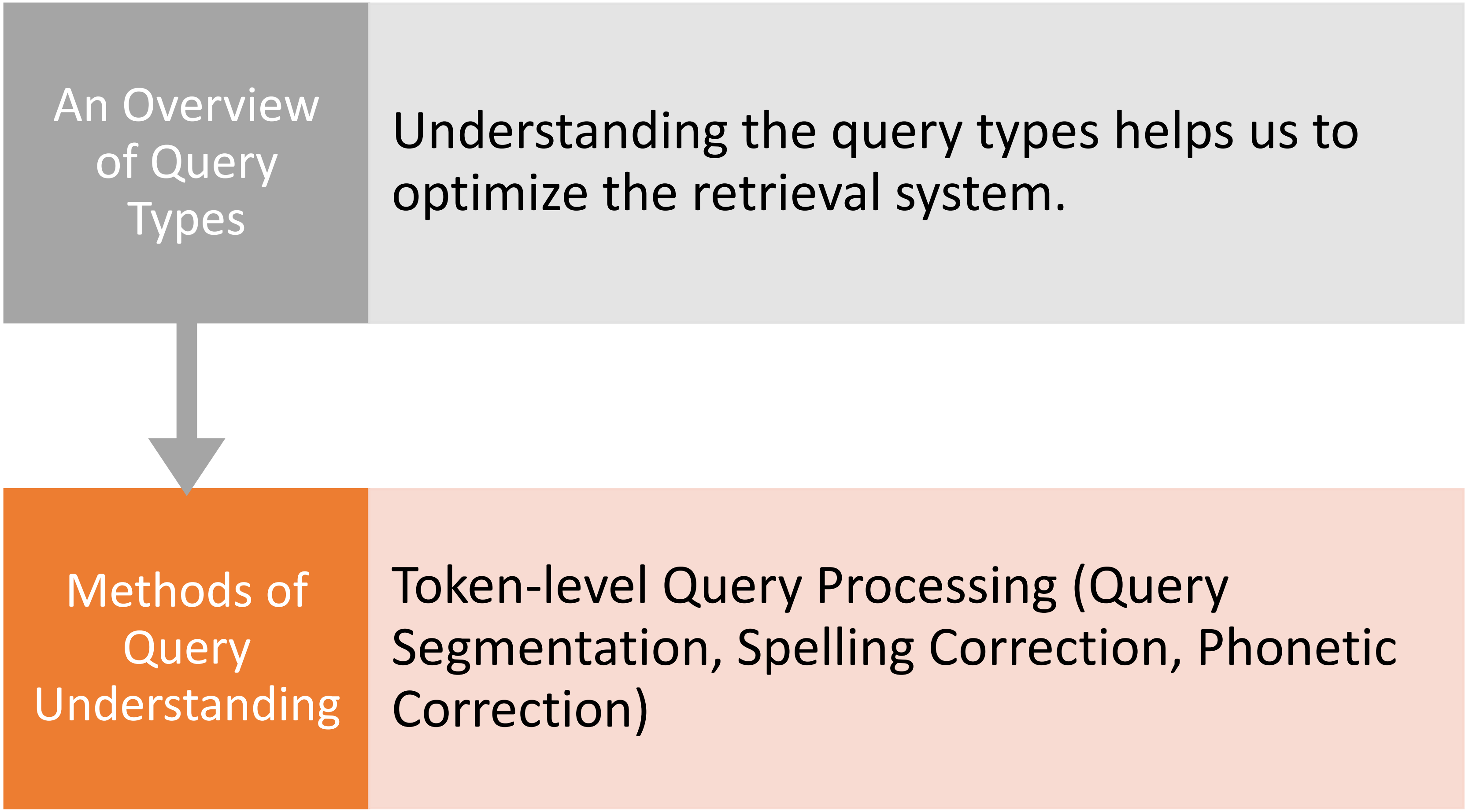
Venkatesh Vinayakarao

Term: Aug – Sep, 2019
Chennai Mathematical Institute

Mission defines strategy, and strategy defines structure.
– Peter Drucker.

Query Understanding

Agenda



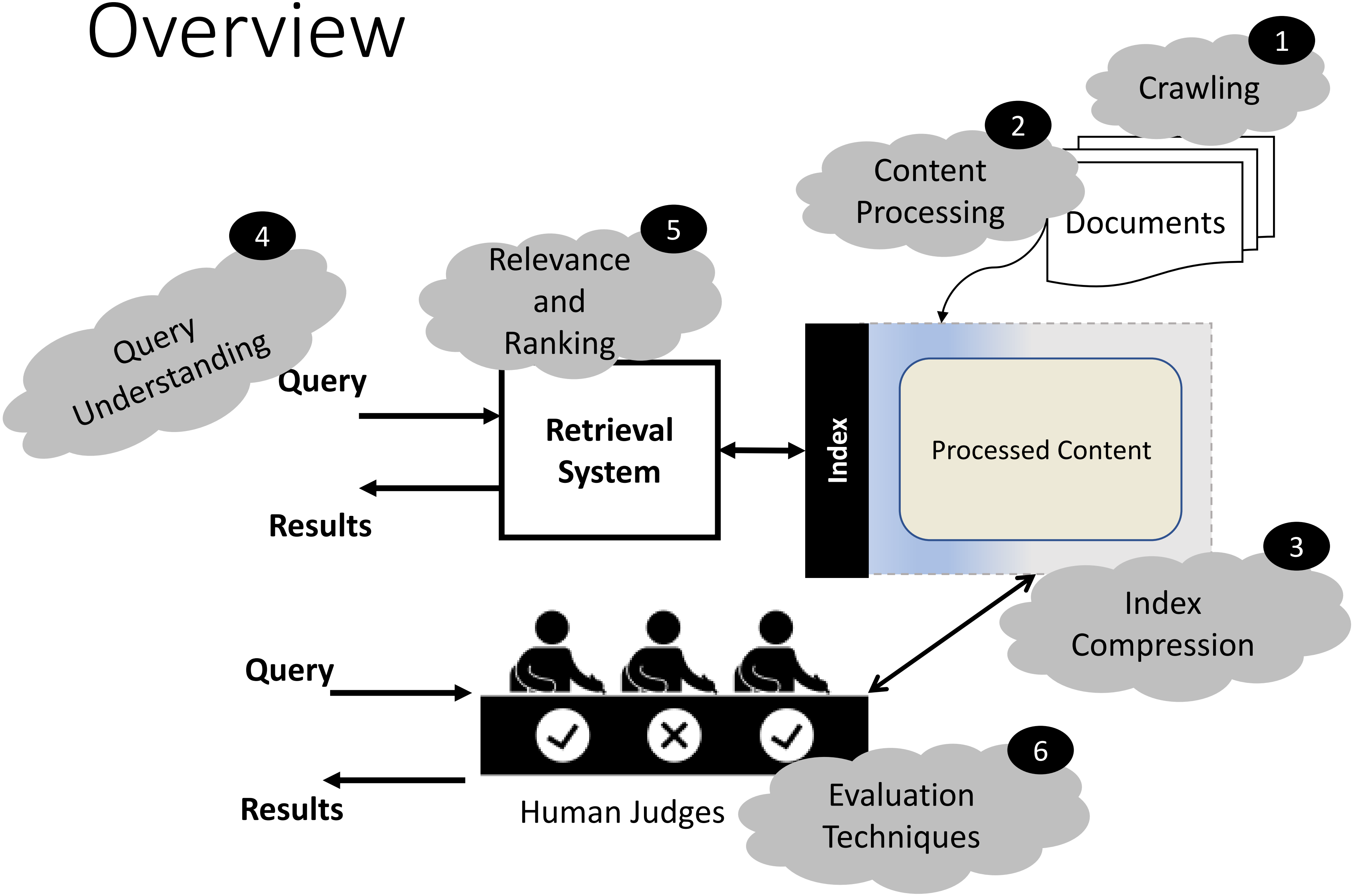
An Overview of Query Types

Understanding the query types helps us to optimize the retrieval system.

Methods of Query Understanding

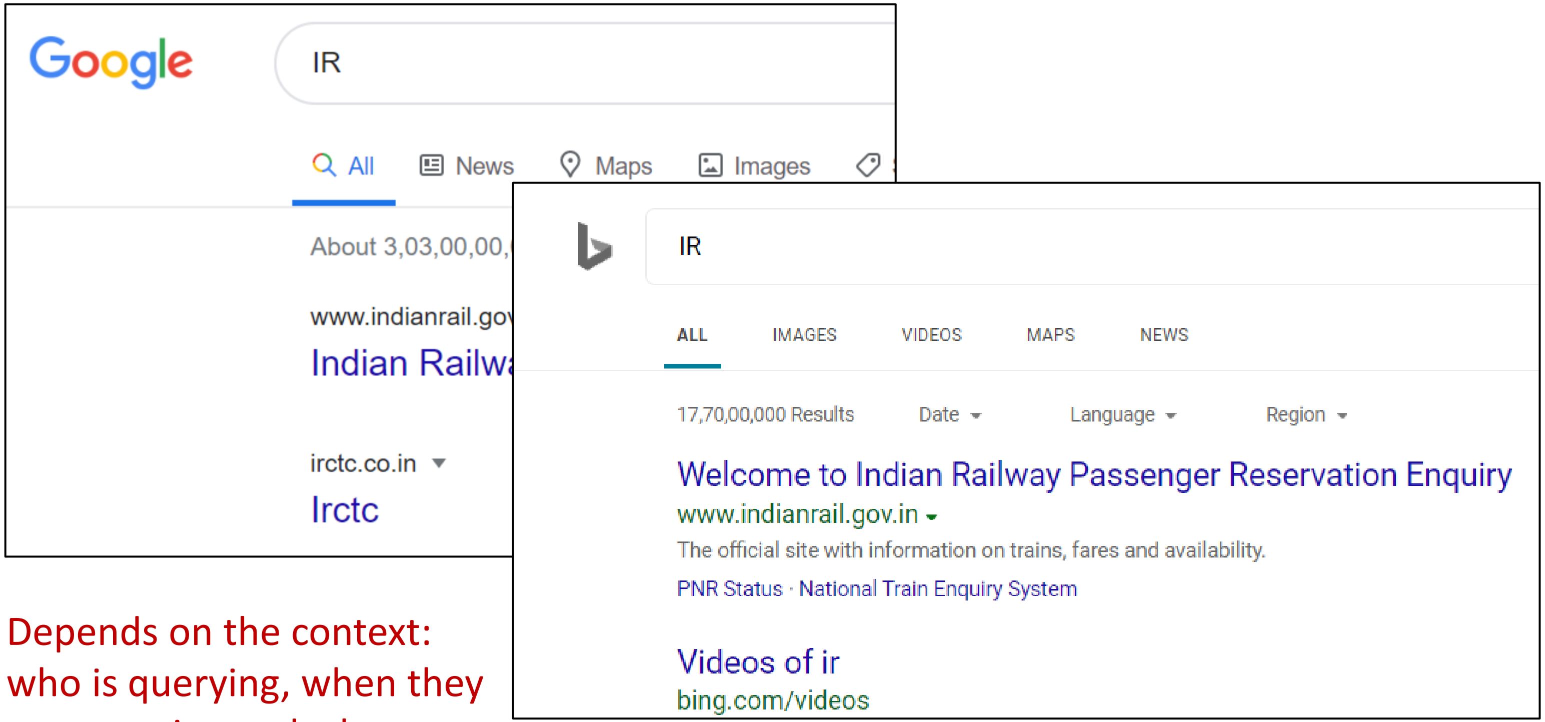
Token-level Query Processing (Query Segmentation, Spelling Correction, Phonetic Correction)

Overview


1
Crawling
2
Content Processing
Documents
4
Query Understanding
5
Relevance and Ranking
Query
Retrieval System
Results
Index
Processed Content
3
Index Compression
Query
Results
Human Judges
6
Evaluation Techniques

Some Queries are Hard to Understand!

- Guess, what should the query “IR” return?



About 3,03,00,00,

www.indianrail.gov

Indian Railw

irctc.co.in

Irctc

17,70,00,000 Results

Welcome to Indian Railway Passenger Reservation Enquiry

www.indianrail.gov.in

The official site with information on trains, fares and availability.

PNR Status · National Train Enquiry System

Videos of ir

bing.com/videos

Depends on the context: who is querying, when they are querying and what was queried before, popularity of keyword, etc.

Query Types: The N, I & T!

- Navigational
  - Example: “fb”
  - Say, a user wants to visit facebook.com. He might hit fb on the search bar and use the first result to go to the page.
- Informational
  - Example: “Amitabh bachchan”
  - Seeks information about Amitabh Bachchan.
- Transactional
  - Example: “Chennai to Delhi air ticket”
  - Say, the intent is to buy an air ticket and this query is the first step of searching for the best price/route/vendor.

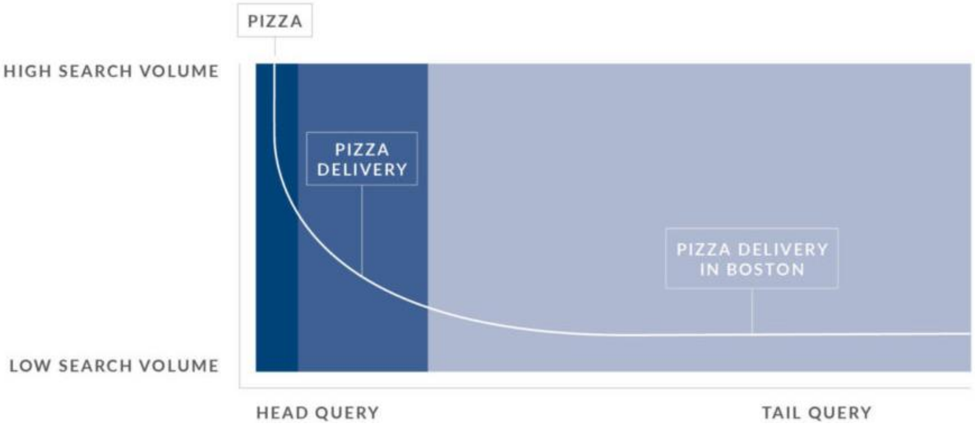
Query Types: Long and Short

- Typically, queries are **short phrases**
  - “data science degree india”
  - “Stanford semester start date”
- However, **long queries** are not uncommon
  - Example: “easter egg hunts in northeast columbus parks and recreation centers”
  - “Queries of length five words or more have increased at a year over year rate of 10%, while single word queries dropped 3%.” – Balasubramanian, Kumaran and Carvalho – 2010.

Query Types: Head and Tail Queries

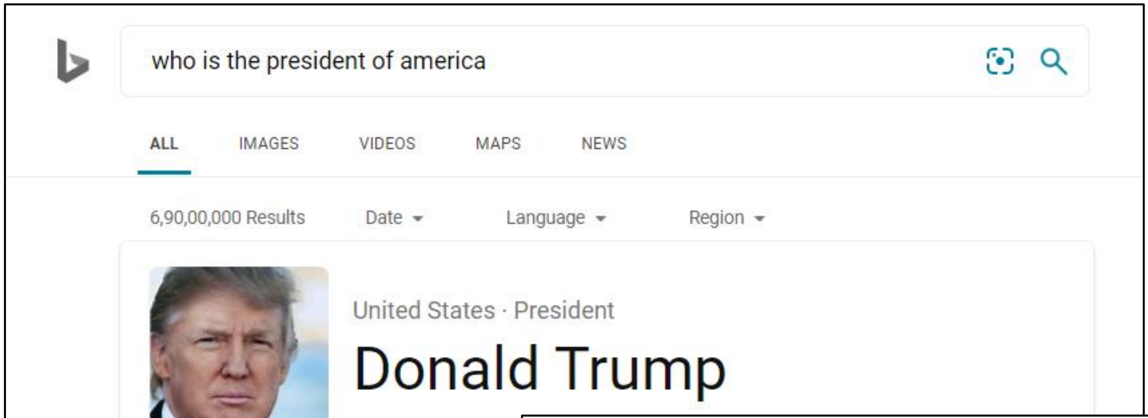
- Head Queries
  - Queries that appear very frequently
- Tail Queries
  - The “rare” queries

In a quest to improve overall performance, we often do not give the attention here that this deserves



Picture Source: https://lucidworks.com/ai-powered-search/head-tail-analysis/

Query Types: Question and Answers



who is the president of america

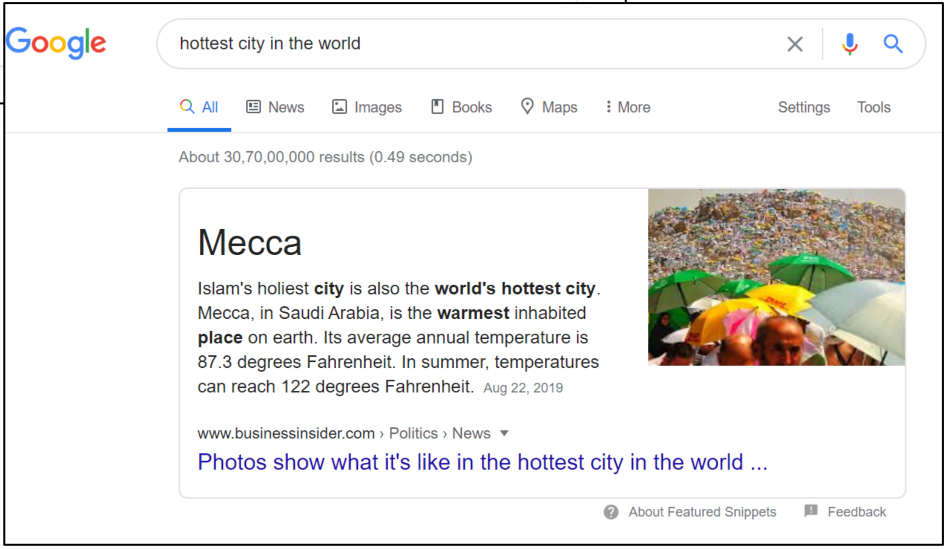
ALL IMAGES VIDEOS MAPS NEWS

6,90,00,000 Results Date Language Region

United States · President

Donald Trump

Wow! How did Bing understand this?



hottest city in the world

All News Images Books Maps More Settings Tools

About 30,70,00,000 results (0.49 seconds)

Mecca

Islam's holiest **city** is also the **world's hottest city**. Mecca, in Saudi Arabia, is the **warmest** inhabited **place** on earth. Its average annual temperature is 87.3 degrees Fahrenheit. In summer, temperatures can reach 122 degrees Fahrenheit. Aug 22, 2019

www.businessinsider.com › Politics › News

Photos show what it's like in the hottest city in the world ...

About Featured Snippets Feedback

Good Query Understanding

Methods for Query Understanding

- Token-level Query Processing
    - Spelling Errors
    - Query Segmentation
- Query Reduction
    - Remove less-important query tokens.
- Query Expansion
    - Add more terms to query to improve precision and recall.
- Query Rewriting
    - Transform the original query to a query friendlier to the retrieval system.

Token-Level Query Processing

Query Segmentation

- Users might miss spaces when they query. Consider, for example:
  - Statebankofindia for “State Bank of India”
  - Amazonprimevideo for “Amazon Prime Video”
- Can you give an algorithm to check if input can be split to arrive at dictionary terms?

A Recursive Algorithm

S	T	A	T	E	B	A	N	K	O	F	I	N	D	I	A

S	■	T	A	T	E	B	A	N	K	O	F	I	N	D	I	A

S not in our dictionary. Keep moving till we find a dictionary term.

S	T	A	T	E	■	B	A	N	K	O	F	I	N	D	I	A

Check if rest of the string “BANKOFINDIA“ can be split.

If “yes”, insert a space after STATE.

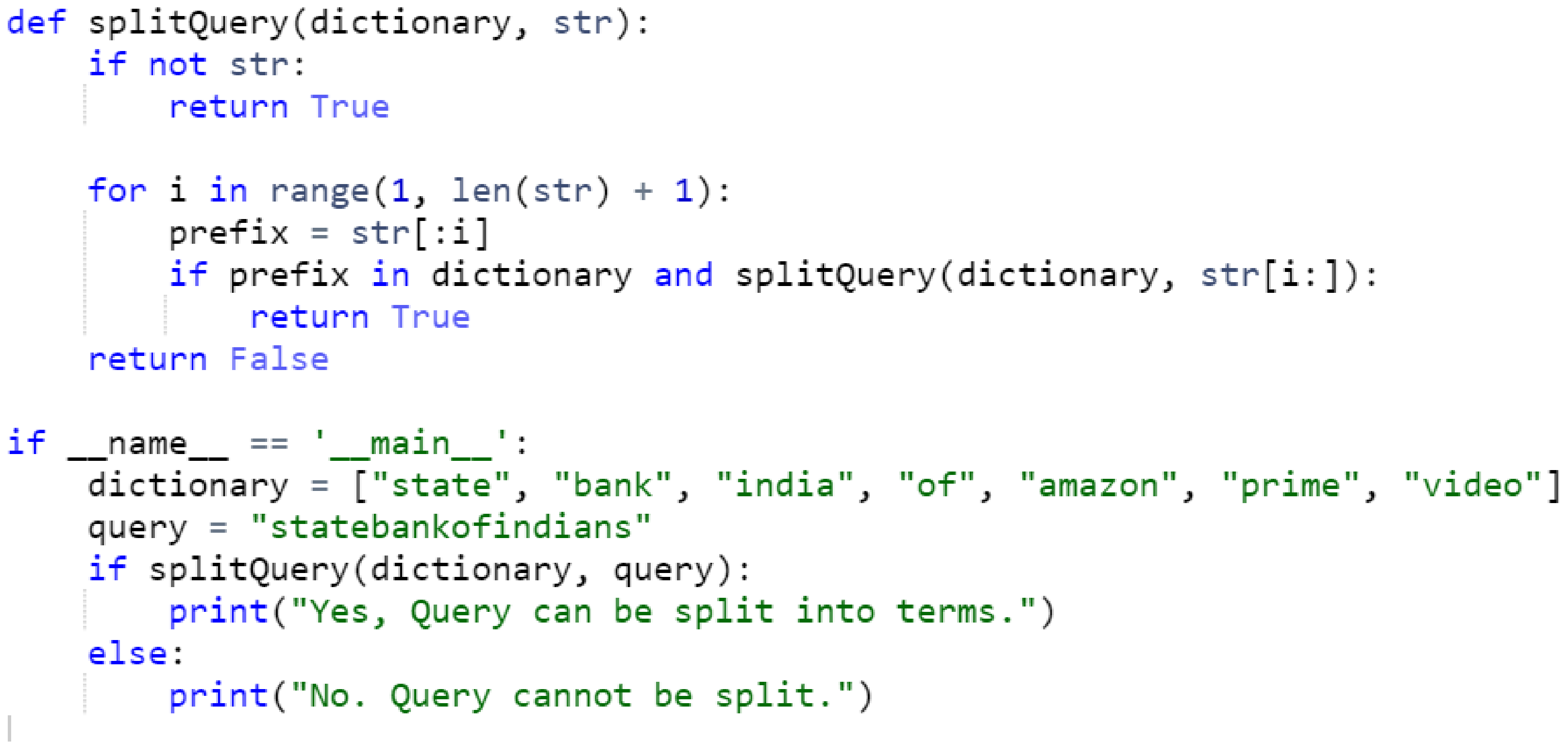
Else, continue with a longer term.

S	T	A	T	E	■	B	A	N	K	■	O	F	■	I	N	D	I	A

Recurse and backtrack till you find a split.

Dictionary
State
Bank
Of
India
Amazon
Prime
Video
…

A Python Solution

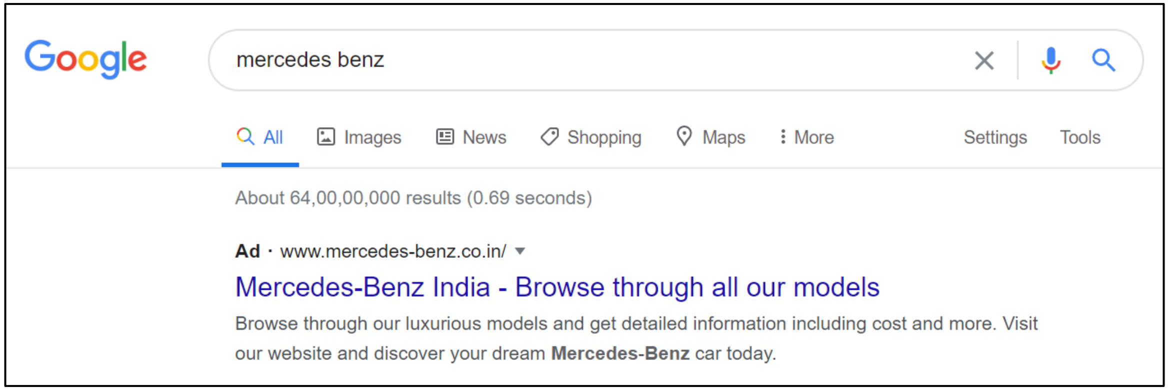


```python
def splitQuery(dictionary, str):
    if not str:
        return True

    for i in range(1, len(str) + 1):
        prefix = str[:i]
        if prefix in dictionary and splitQuery(dictionary, str[i:]):
            return True
    return False

if __name__ == '__main__':
    dictionary = ["state", "bank", "india", "of", "amazon", "prime", "video"]
    query = "statebankofindians"
    if splitQuery(dictionary, query):
        print("Yes, Query can be split into terms.")
    else:
        print("No. Query cannot be split.")
```

Spelling Errors



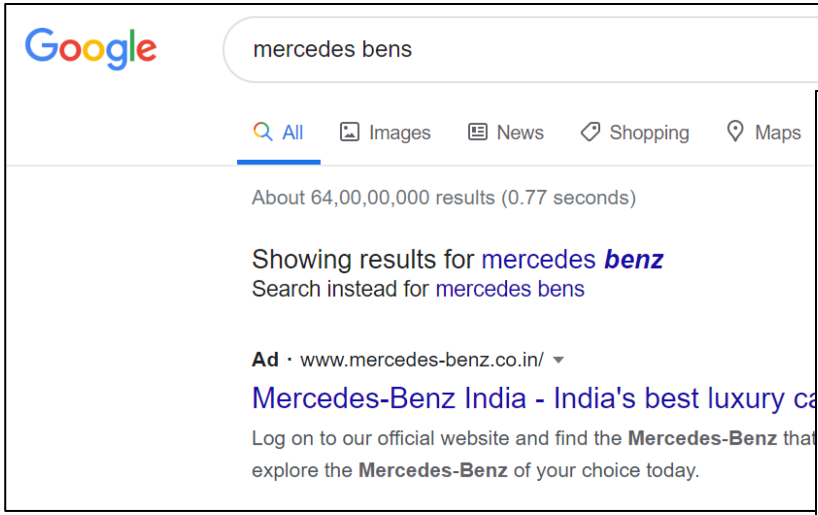
mercedes benz

About 64,00,00,000 results (0.69 seconds)

Ad · www.mercedes-benz.co.in/

Mercedes-Benz India - Browse through all our models

Browse through our luxurious models and get detailed information including cost and more. Visit our website and discover your dream **Mercedes-Benz** car today.



mercedes bens

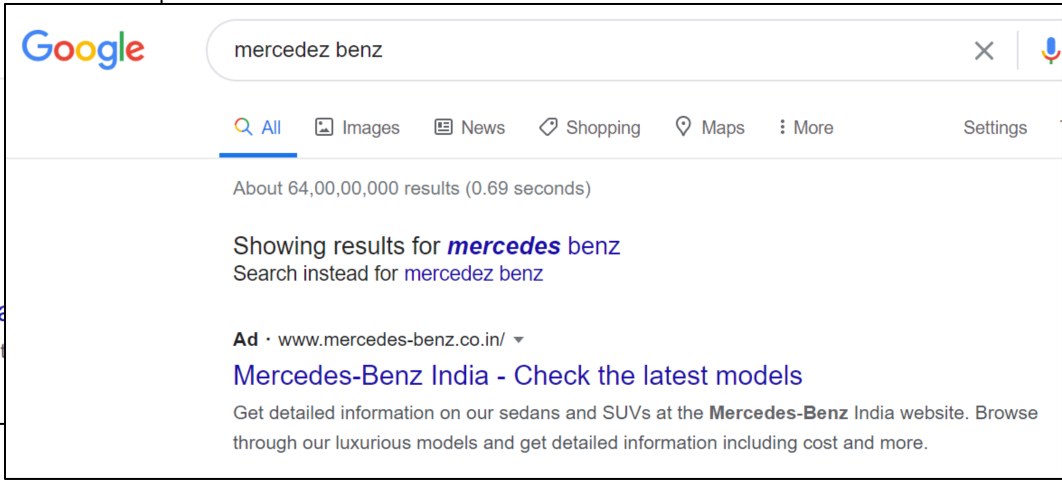
About 64,00,00,000 results (0.77 seconds)

Showing results for mercedes ***benz***
Search instead for mercedes bens

Ad · www.mercedes-benz.co.in/

Mercedes-Benz India - India's best luxury ca

Log on to our official website and find the **Mercedes-Benz** that explore the **Mercedes-Benz** of your choice today.



mercedez benz

About 64,00,00,000 results (0.69 seconds)

Showing results for ***mercedes*** benz
Search instead for mercedez benz

Ad · www.mercedes-benz.co.in/

Mercedes-Benz India - Check the latest models

Get detailed information on our sedans and SUVs at the **Mercedes-Benz** India website. Browse through our luxurious models and get detailed information including cost and more.

Some Types of Misspellings

Cause	Misspelling	Correction
Typing quickly	exxit mispell	exit misspell
Keyboard adjacency	importamt	important
Inconsistent rules	concieve conceirge	conceive concierge
Ambiguous word breaking	silver light	silverlight
New words	kinnect	kinect

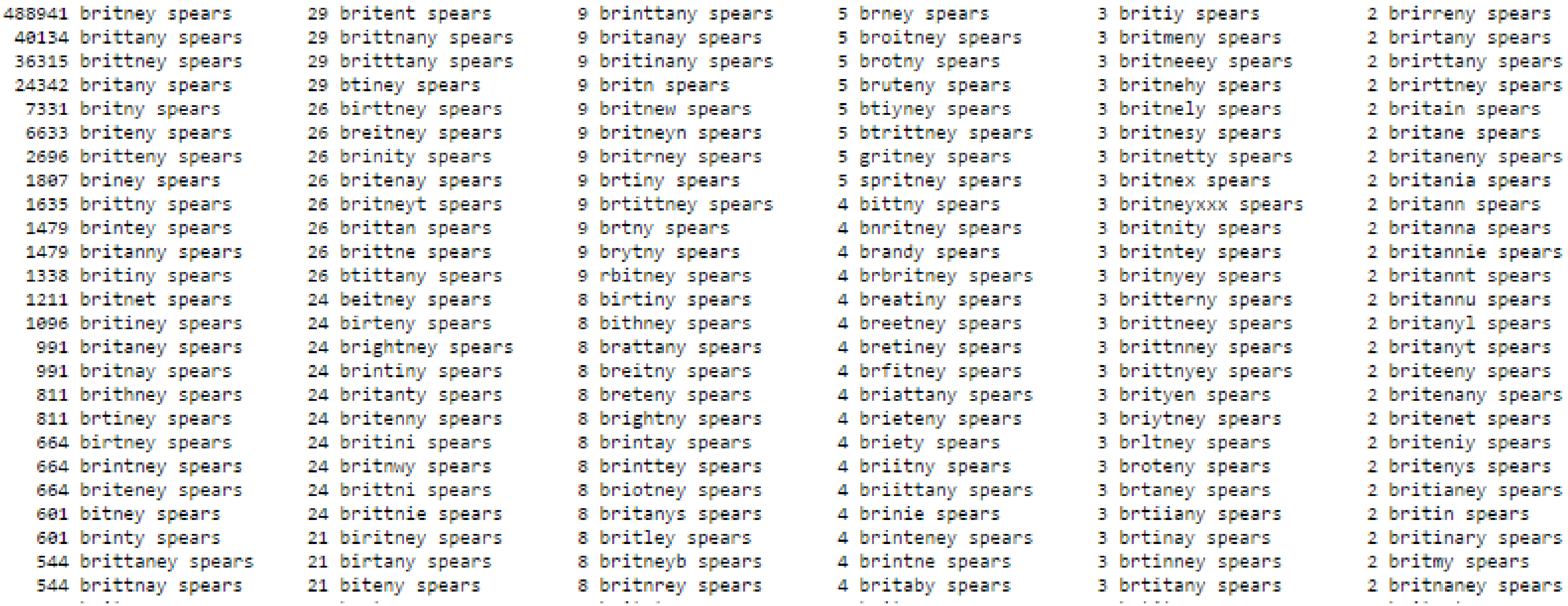
Spelling Errors in Query

- 10% to 20% of queries carry misspelt words[1].
- English is not 100% phonetic (e.g., colonel, read vs dead).
- How many of these phrases contain spelling errors?
  - cigarete lighter
  - fourty dollars
  - going to libary today
  - unforgetable holiday
  - successful businessman

Duan and Hsu, WWW 2011.

Notorious Britney

The data below shows some of the misspellings detected by our spelling correction system for the query [britney spears], and the count of how many different users spelled her name that way. -- Google.



```
488941 britney spears
 40134 brittany spears
 36315 brittney spears
 24342 britany spears
  7331 britny spears
  6633 briteny spears
  2696 britteny spears
  1807 briney spears
  1635 brittny spears
  1479 brintey spears
  1479 britanny spears
  1338 britiny spears
  1211 britnet spears
  1096 britiney spears
   991 britaney spears
   991 britnay spears
   811 brithney spears
   811 brtiney spears
   664 birtney spears
   664 brintney spears
   664 briteney spears
   601 bitney spears
   601 brinty spears
   544 brittaney spears
   544 brittnay spears

29 britent spears
29 brittnany spears
29 britttany spears
29 btiney spears
26 birttney spears
26 breitney spears
26 brinity spears
26 britenay spears
26 britneyt spears
26 brittan spears
26 brittne spears
26 btittany spears
24 beitney spears
24 birteny spears
24 brightney spears
24 brintiny spears
24 britanty spears
24 britenny spears
24 britini spears
24 britnwy spears
24 brittni spears
24 brittnie spears
21 biritney spears
21 birtany spears
21 biteny spears

9 brinttany spears
9 britanay spears
9 britinany spears
9 britn spears
9 britnew spears
9 britneyn spears
9 britrney spears
9 brtiny spears
9 brtittney spears
9 brtny spears
9 brytny spears
9 rbitney spears
8 birtiny spears
8 bithney spears
8 brattany spears
8 breitny spears
8 breteny spears
8 brightny spears
8 brintay spears
8 brinttey spears
8 briotney spears
8 britanys spears
8 britley spears
8 britneyb spears
8 britnrey spears

5 brney spears
5 broitney spears
5 brotny spears
5 bruteny spears
5 btiyney spears
5 btrittney spears
5 gritney spears
5 spritney spears
4 bittny spears
4 bnritney spears
4 brandy spears
4 brbritney spears
4 breatiny spears
4 breetney spears
4 bretiney spears
4 brfitney spears
4 briattany spears
4 brieteny spears
4 briety spears
4 briitny spears
4 briittany spears
4 brinie spears
4 brinteney spears
4 brintne spears
4 britaby spears

3 britiy spears
3 britmeny spears
3 britneeey spears
3 britnehy spears
3 britnely spears
3 britnesy spears
3 britnetty spears
3 britnex spears
3 britneyxxx spears
3 britnity spears
3 britntey spears
3 britnyey spears
3 britterny spears
3 brittneey spears
3 brittnney spears
3 brittnyey spears
3 brityen spears
3 briytney spears
3 brltney spears
3 broteny spears
3 brtaney spears
3 brtiiany spears
3 brtinay spears
3 brtinney spears
3 brtitany spears

2 brirreny spears
2 brirtany spears
2 brirttany spears
2 brirttney spears
2 britain spears
2 britane spears
2 britaneny spears
2 britania spears
2 britann spears
2 britanna spears
2 britannie spears
2 britannt spears
2 britannu spears
2 britanyl spears
2 britanyt spears
2 briteeny spears
2 britenany spears
2 britenet spears
2 briteniy spears
2 britenys spears
2 britianey spears
2 britin spears
2 britinary spears
2 britmy spears
2 britnaney spears
```

Source: https://archive.google.com/jobs/britney.html

Two Major Approaches

- Two major approaches exist for spelling correction:
  - finding “**nearest**” dictionary term.
  - finding “**most commonly used**” dictionary term when there are multiple “nearest terms”.
- Two major kinds:
  - Isolated-Term Correction
    - Correct one word at a time.
  - Context-Sensitive Correction
    - “flew ***form*** New York” – Note that form is a dictionary term. Yet, this requires to be corrected to “flew from New York”.

Edit Distance

Spelling Correction: 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, (and perhaps Transposition*)
- E.g., the edit distance from ***dof*** to ***dog*** is 1
  - From ***cat*** to ***act*** is 2 (Just 1 with transpose.)
  - from ***cat*** to ***dog*** is 3.

*In this course, we do not consider transposition.

Quiz

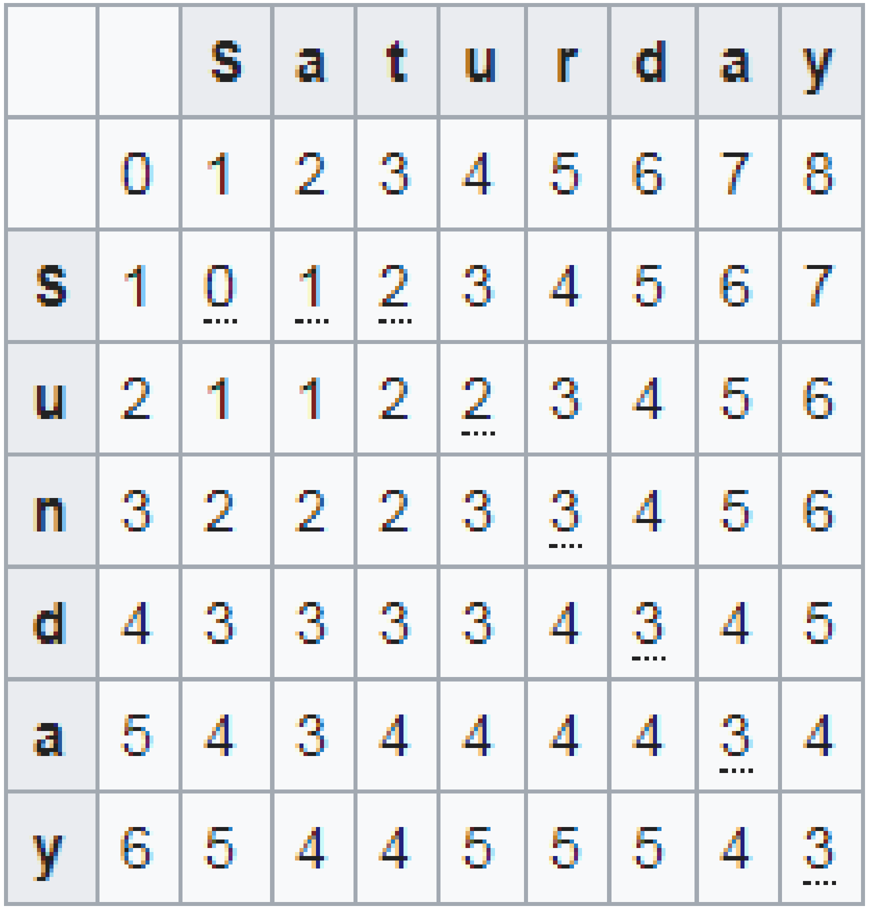
What is the edit distance between Sunday and Saturday?

*You are allowed to perform only Insert, Delete, and Replace operations.

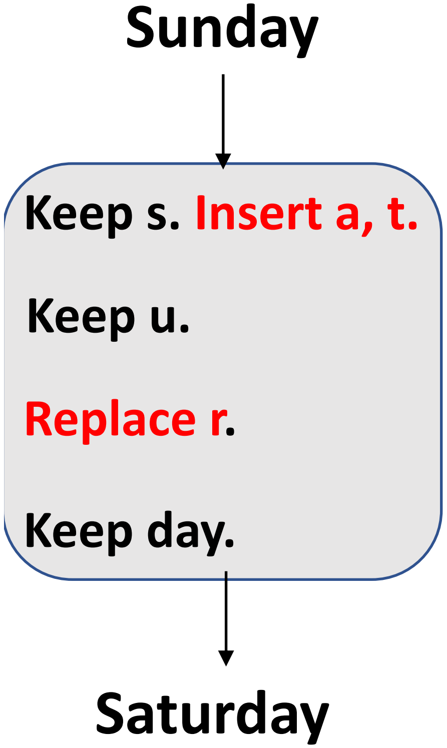
Answer

- Saturday = Sunday = S*day
- Problem is same as
  - What is the edit distance between atur and un?
  - Answer
    - Delete a,t. Replace r with n.
    - 3.

Levenshtein Example



		s	a	t	u	r	d	a	y
	0	1	2	3	4	5	6	7	8
s	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3



Sunday

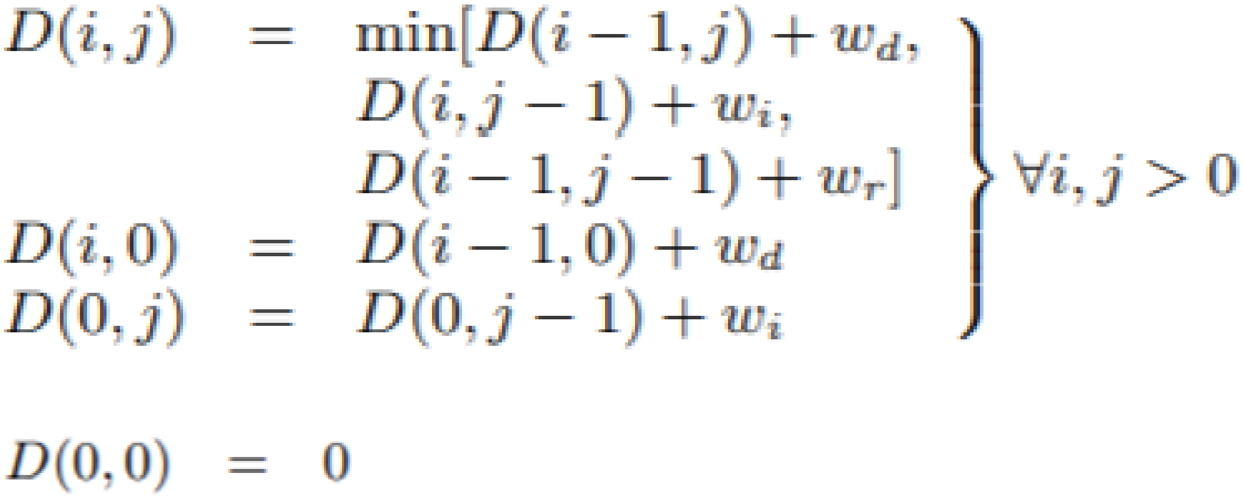
Keep s. Insert a, t.

Keep u.

Replace r.

Keep day.

Saturday

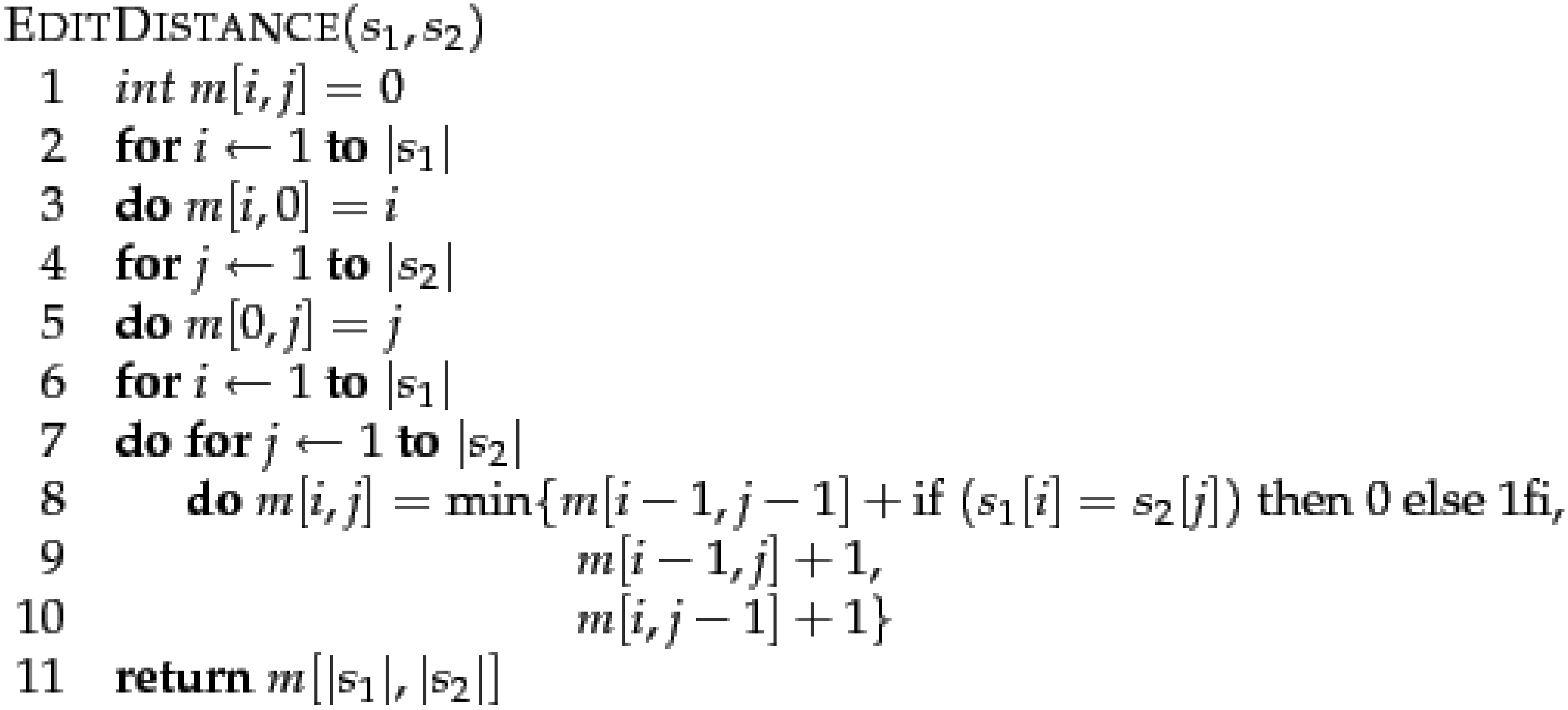


$$\left.\begin{array}{rcl} D(i,j) & = & \min[D(i-1,j)+w_d, \\ & & D(i,j-1)+w_i, \\ & & D(i-1,j-1)+w_r] \\ D(i,0) & = & D(i-1,0)+w_d \\ D(0,j) & = & D(0,j-1)+w_i \end{array}\right\} \forall i,j>0$$

$$D(0,0) = 0$$

V. I. Levenshtein, Binary codes capable of correcting deletions insertions and reversals. Soviet Physics. 10, 707-710, 1966.

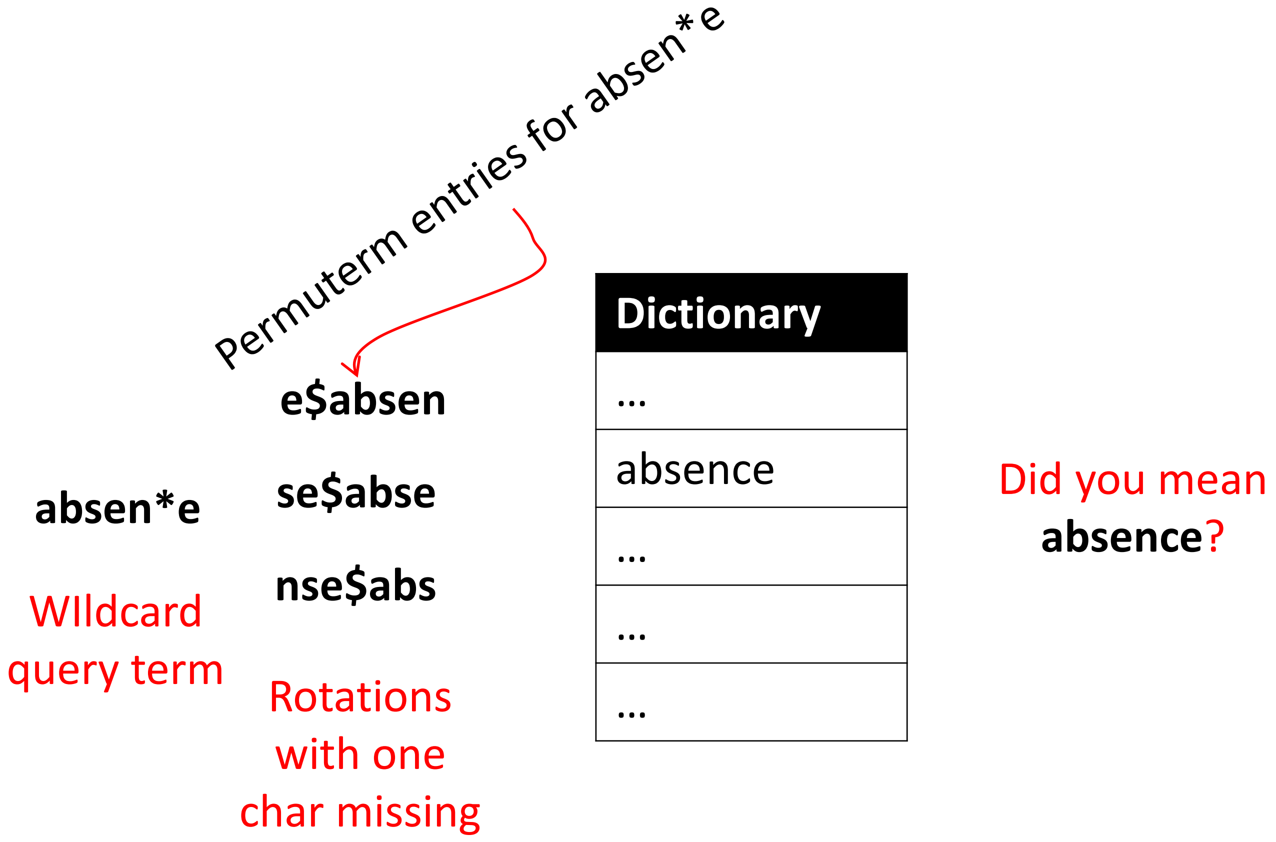
Levenshtein Algorithm



```
EDITDISTANCE($s_1, s_2$)
int $m[i, j] = 0$
for $i \leftarrow 1$ to $|s_1|$
do $m[i, 0] = i$
for $j \leftarrow 1$ to $|s_2|$
do $m[0, j] = j$
for $i \leftarrow 1$ to $|s_1|$
do for $j \leftarrow 1$ to $|s_2|$
    do $m[i, j] = \min\{m[i-1, j-1] +$ if $(s_1[i] = s_2[j])$ then 0 else 1fi,
                          $m[i-1, j] + 1,$
                          $m[i, j-1] + 1\}$
return $m[|s_1|, |s_2|]$
```

Can we use Levenshtein's Distance to answer wildcard queries?

Permuterm Index



absen*e

WIldcard query term

Permuterm entries for absen*e

e$absen

se$abse

nse$abs

Rotations with one char missing

Dictionary
…
absence
…
…
…

Did you mean **absence**?

Compute edit distance for each query term with each of its permuterm based matches. Very expensive!
Assume first letter will be correct. Apply such heuristics.

K-grams for Spelling Correction

k-gram Idea for Spelling Correction

- Many heuristics lead to poor matches.
- For example, “bored” misspelt as “bord” may match “boardroom” if the heuristic is
  - Match any two bigrams
    - and we matched “bo” and “rd”
- Potential Solution
  - Compute Jaccard Similarity between k-grams of matched term and that of the query term.

Jaccard Coefficient

- Jaccard Coefficient of two sets A and B

  = |A⋂B|/|AUB|

- Example: JS on bigrams of (“bord”, “boardroom”)

  = |{$b, bo, rd}|/|{$b, bo, or, rd, d$, oa, ar, dr, ro, oo, om, m$}|

  = 3/12.

*If you do not use end markings, we get 2/9.

Context-Sensitive Spelling Correction

- Our heuristics may lead to
  - "flew form Delhi" → "flew fore Delhi", "flew from Delhi"
- Surrounding words may determine the correction
- Potential Solution
  - Use query log frequency or collection frequency of these phrases to choose the best.

Thank You