

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Approximate String Matching

COMP90049

Knowledge Technologies

<https://powcoder.com>

Jeremy Nicholson and Justin Zobel and Karin Verspoor

Add WeChat powcoder

Semester 2, 2017



THE UNIVERSITY OF
MELBOURNE

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Week 3:

- Approximate String Search and Matching

- Common Applications

- Methods:

- Neighbourhood Search

- Edit Distance

- N-Gram Distance

- [Phonetic methods]

- Evaluation

- [Genomics]

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Consider:

- Given a string, is some substring contained within it?
- Given a string (document), find all occurrences of some substring

<https://powcoder.com>

For example, find Exxon in:

In exes for foxes rex dux mixes a pox of waxed luxes.

An axe, and an axon, to exo Exxon max oxen.

Grexit or Brexit as quixotic farmers with burrom rex taxation.

Not (really) a Knowledge Technology!

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Find exon in:

In exes for foxes rex dux mixes a pox of waxed luxes.

An axe, and an axon, to exo Exxon max oxen.

Gxerit or Bxerit as qxirotic maxxers with buxom rex taxation.

Not present!

...But what is the “closest” or “best” match?

Add WeChat powcoder

This is a Knowledge Technology!

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Two main applications for Approximate String Search:

<https://powcoder.com>

- Spelling correction

- Computational Genomics

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

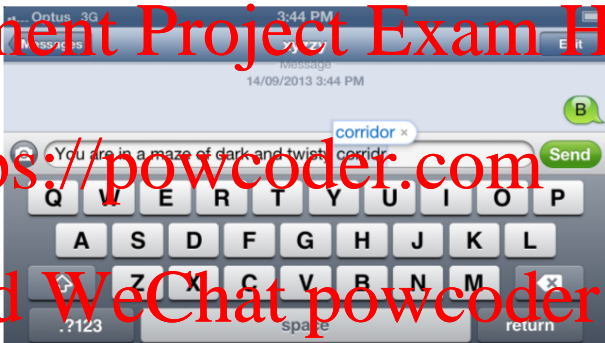
References

Genomics

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Need the notion of a **dictionary**:

- Here, a list of entries that are “correct”
- We can break our input into substrings that we wish to match, and compare each of them against the entries in the dictionary
- An item in the input which *doesn't* appear in the dictionary is *misspelled*
- An item in the input which *does* appear in the dictionary might be correctly spelled or misspelled (probably slightly beyond the scope of this subject)

<https://powcoder.com>

Add WeChat powcoder

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Therefore, the problem here:

<https://powcoder.com>
Given some item of interest — which does not appear in our dictionary
— which entry from the dictionary was truly intended?

Depends on the person who wrote the original string!

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

- Computational Genomics (later, if we have time)

- Name matching

- Query repair

- Phonetic matching (later, if we have time)

- Data cleaning

- ..

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Find approximate match(es) for exon in:

<https://powcoder.com>

`Il eyes for foxes rex dix nixes a pox of waxed luxes.`

`An axe, and an axon, to exo Exxon max oxen.`

`Grexit or Brexit as quixotic haxxers with buxom rex taxation.`

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

For a given string w of interest:

- Generate all variants of w that utilise at most k changes (Insertions/Deletions/Replacements) — **neighbours**
- Check whether generated variants exist in dictionary
- **All** results found in dictionary are returned

Unix command-line utility `fgrep` is an efficient mechanism for finding these.

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

For example:

... proceed if you can see no **ther** option ...

<https://powcoder.com>

Intended word: other

Requires 1 insertion (o) so intended word will be found using
neighbourhood search (and some unintended words...)

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Neighbourhood search is surprisingly fast!

Consider: alphabet size is Σ , length of string is $|w|$:

For k edits, roughly $O(\Sigma^k \cdot |w|^k)$ neighbours

...But Σ is a small constant, string of interest is usually short, and k is usually small

Add WeChat powcoder

For each neighbour, need a dictionary read (dict has D entries):
Binary search yields $O(|w|^k \log D)$ string comparisons

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

So, efficiency isn't our problem.
<https://powcoder.com>
(agrep example)

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Alternative methods:
<https://powcoder.com>

Scan through each dictionary entry looking for the “best” match

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Global Edit Distance:

<https://powcoder.com>
Transform the string of interest into each dictionary entry, using the operations Insert, Delete, Replace, and Match (character)

Each operation is associated with a score;

Best match is the dictionary entry with best aggregate **score**

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

For example:

Item of interest: `crat`

Dictionary: `cart`, `arts`

Score: Match +1, Insert -1, Delete -, Replace -1

`crat` → `cart`:

Match `c` (+1), Delete `r` (-1), Match `a` (+1), Insert `r` (-1), Match `t` (+1) = +1

`crat` → `arts`:

Replace `c` with `a` (-1), Match `r` (+1), Delete `a` (-1), Match `t` (+1), Insert `s` (-1) = -1

`cart` is the better match

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Confusingly, Global Edit Distance isn't a "distance"

.. But depends on parameter
<https://powcoder.com>

Match (0), Insert (+1), Delete (+1), Replace (+1)

This is the Levenshtein Distance (which is a "distance"): it counts the number of edits required to transform one string into the other

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Hypothetically, any parameter is possible!

But some choices make no sense, e.g.:

<https://powcoder.com>

aba: Which corresponds to best match?

- foo: Insert, Delete, Insert, Delete, Insert, Delete = +18
- aba: Match, Match, Match = +12
- cbc: Replace, Match, Replace = +4

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Often, “direction” doesn’t matter: Insert = Delete (“Indel”)

Sometimes, score of Replace depends on which character is being replaced.

Consider:

Is faxing more likely to be facing or faking?

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

From string f to string t , given array A of $|f| + 1$ columns and $|t| + 1$ rows, we can solve using the Needleman–Wunsch algorithm:

```
lf = strlen(f); lt = strlen(t);
A[0][0]=0;
for (j=1; j<=lt; j++) A[j][0] = j * i;
for (k=1; k<=lf; k++) A[0][k] = k * d;
for (j=1; j<=lt; j++)
    for (k=1; k<=lf; k++)
        A[j][k] = max3( //Or min3 if m<i,d,r
            A[j][k-1] + d, //deletion
            A[j-1][k] + i, //Insertion
            A[j-1][k-1] + equal(f[k-1],t[j-1])); //Replace or match
```

`equal()` returns m if characters match, r otherwise

Final score is at $A[lt][lf]$

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

In action: from erat to arts, Match (-1), Insert/Delete/Replace (-1)

<https://powcoder.com>

Add WeChat powcoder

| | ϵ | c | r | a | t |
|---|------------|----|----|----|----|
| e | 0 | -1 | -2 | -2 | -4 |
| a | -1 | -1 | -2 | -1 | -2 |
| r | -2 | -2 | 0 | -1 | -2 |
| t | -3 | -3 | -1 | -1 | 0 |
| s | -4 | -4 | -2 | -2 | -4 |

Global Edit Distance: -1 (Replace, Match, Delete, Match, Insert)

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Algorithm actually depends on parameter!

```
A[j][k] = max3(  
    A[f,j][k+1] + d, //Deletion  
    A[f,j-1][k] + i, //insertion  
    A[j-1][k-1] + equal(f[k-1],t[j-1])); //Replace or match
```

→ Match score greater than Insert/Delete/Replace

e.g. Match (+1), Insert/Delete/Replace (-1)

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help

Algorithm actually depends on parameter!

```
A[j][k] = min3(
    A[j][k-1] + d, //Deletion
    A[j-1][k] + i, //Insertion
    A[j-1][k-1] + equal(f[k-1],t[j-1])); //Replace or match
```

→ Match score less than Insert/Delete/Replace

e.g. Match (0), Insert/Delete/Replace (1)
 Add WeChat powcoder
 (Levenshtein Distance)

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Local Edit Distance is like Global Edit Distance, but we are searching for the best substring match

<https://powcoder.com>

Particularly suitable when comparing two strings of very different lengths, e.g. a word and a sentence

Add WeChat powcoder

Local Edit Distance Algorithm

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

From string f to string t , given array A of $|f| + 1$ columns and $|t| + 1$ rows, we can solve using the Smith–Waterman algorithm:

```
lf = strlen(f), lt = strlen(t);
```

```
A[0][0]=0;
```

```
for (j=1; j<=lt; j++) A[j][0] = 0;
```

```
for (k=1; k<=lf; k++) A[0][k] = 0;
```

```
for (j=1; j<=lt; j++)
```

```
    for (k=1; k<=lf; k++)
```

```
        A[j][k] = max4( //Or min4 if m<i,d,r
```

```
            0,
```

```
            A[j][k-1] + d, //deletion
```

```
            A[j-1][k] + i, //Insertion
```

```
            A[j-1][k-1] + equal(f[k-1],t[j-1])); //Replace or match
```

`equal()` returns m if characters match, r otherwise

Final score is greatest value in the entire table (or least value, if $m < i, d, r$)

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

In action: from eart to arts, Match (+1), Insert/Delete/Replace (-1)

| | ϵ | c | a | r | t |
|------------|------------|---|---|---|---|
| ϵ | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 1 | 0 | 0 |
| r | 0 | 0 | 0 | 2 | 1 |
| t | 0 | 0 | 0 | 1 | 3 |
| s | 0 | 0 | 0 | 0 | 2 |

<https://powcoder.com>

Add WeChat powcoder

Best match: art with art (+3); ties are possible.

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

For strings f and t , Both algorithms above are $\mathcal{O}(|f||t|)$ in both space and time. (Space can be improved, but time (probably) cannot.)

When approximate matching, we have a constant string f which we want to compare to each string t in the dictionary D :

$$\mathcal{O}(|f| \sum_{t \in D} |t|)$$

Hence, integer comparisons are roughly the number of characters in the dictionary. Whether this is feasible depends on the size of the dictionary.

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

N-Gram Distance has same goal as Edit Distance: compare two strings to determine “best” match

(character) n -gram: substring of length n

2-grams of crat: #c, cr, ra, at, t#

2-grams of cart: #c, ca, ar, rt, t#

2-grams of arts: #a, ar, rt, ts, s#

N-Gram Distance between n -grams of strings s ($G_n(s)$) and t ($G_n(t)$):
 $|G_n(s)| + |G_n(t)| - 2 \times |G_n(s) \cap G_n(t)|$

<https://powcoder.com>

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

n -gram: substring of length n

2-grams of crat: #c, cr, ra, at, t#

2-grams of cart: #c, ca, ar, rt, t#

2-grams of arts: #a, ar, rt, ts, s#

2-Gram Distance between crat and cart:

$$|G_2(\text{crat})| + |G_2(\text{cart})| - 2 \times |G_2(\text{crat}) \cap G_2(\text{cart})|$$

$$= 5 + 5 - 2 \times 2 = 6 \text{ (better)}$$

2-Gram Distance between crat and arts:

$$|G_2(\text{crat})| + |G_2(\text{arts})| - 2 \times |G_2(\text{crat}) \cap G_2(\text{arts})|$$

$$= 5 + 5 - 2 \times 0 = 10$$

<https://powcoder.com>

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Occasionally useful as a simpler variant of Edit Distance

More sensitive to long substring matches, less sensitive to relative ordering of strings (matches can be anywhere!)
<https://powcoder.com>

Despite its simplicity, takes roughly the same time to compare entire dictionary

Quite useless for very long strings and/or very small alphabets (Why?)
Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact
Approximate
Application

Methods

Neighbourhood
Edit Distance
N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

In English (and some other languages), **orthography** (spelling) isn't a good predictor of **phonetics** (sounds)

Salient concern in speech-to-text systems e.g.:
Georgia Conal

George O'Connell

Also relevant in spelling correction (English can be very difficult to spell correctly!)

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

One mechanism: Soundex

| | | |
|-----------------|---|------------------------------------|
| a e i o u y | → | 0 (vowels) |
| b p f v | → | 1 (labials) |
| c g j k q s x z | → | 2 (misc: fricatives, velars, etc.) |
| d t | → | 3 (dentals) |
| l | → | 4 (lateral) |
| m n | → | 5 (nasals) |
| r | → | 6 (rhotic) |

Translation table:

<https://powcoder.com>

Four step process:

Add WeChat powcoder

- 1 Except for initial character, translate string characters according to table
- 2 Remove duplicates (e.g. 4444 → 4)
- 3 Remove 0s
- 4 Truncate to four symbols

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

One mechanism: Soundex

| | | |
|----------|---|------------------------------------|
| aehiouwy | → | 0 (vowels) |
| bpfv | → | 1 (labials) |
| cgjksxz | → | 2 (misc: fricatives, velars, etc.) |
| dt | → | 3 (dentals) |
| l | → | 4 (lateral) |
| mn | → | 5 (nasals) |
| r | → | 6 (rhotic) |

Translation table:

Assignment Project Exam Help

<https://powcoder.com>

Four step process:

| | |
|------|--------|
| king | kyngge |
| k052 | k05220 |
| k052 | k0520 |
| k52 | k52 |

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact
Approximate
Application

Methods

Neighbourhood
Edit Distance
N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Better phonetic methods make use of the fact that some letters sounds alike in certain contexts, and different in other contexts

Editex uses the Edit Distance to compare strings based on a similar transition table to Soundex

lpadist uses a text-to-sound algorithm to represent tokens according to the International Phonetic Alphabet (but context matters a lot)

Add WeChat powcoder

There are also worse variants, like Phonix.

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Evaluation: consider whether the system is effective at solving the user's problem

In this case: for a misspelled word, does the system identify the correct word?

<https://powcoder.com>

To evaluate, we need:

- A number of cases of misspelled words
- The intended (correct) word for each case
- An **evaluation metric**

Add WeChat powcoder

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

We have some cases:

| Misspelled Word | Correct Word | Predicted Word | Right/Wrong? |
|-----------------|--------------|----------------|--------------|
| ther | other | ther | × |
| corridr | corridor | corridor | ✓ |
| cracheyt | crotchety | cachet | × |
| ... | ... | ... | ... |

Add WeChat powcoder

Accuracy: fraction of correct responses ($\frac{1}{3}$)

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

More realistic situation:

| Misspelled Word | Correct Word | Predicted Word | Right/Wrong? |
|-----------------|--------------|----------------|--------------|
| | | there | × |
| ther | other | other | ✓ |
| | | thei | × |
| corridr | corridor | corridor | ✓ |
| | | carrier | × |
| cracheyt | crotchety | ??? | — |

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Precision: fraction of correct responses among attempted responses
 $\left(\frac{2}{5}\right)$

Recall: proportion of words with a correct response (somewhere) $\left(\frac{2}{3}\right)$

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Typically, the value of the evaluation metric has little intrinsic meaning

“This system gets 81% accuracy” — useful for users, or not?

“The system based on the Global Edit Distance gets 81% accuracy, whereas the system based on the N-Gram Distance gets 84% accuracy”

“The basic system gets 81% accuracy, but after making some changes, the accuracy becomes 74%”

“System A gets 45% precision and 80% recall;
System B gets 95% precision and 10% recall”

— Which one should we use? (Also: why?)

The answer depends on the problem (and the user)!

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

- What is approximate string search?
- What are some common applications of approximate string search; why are they hard?
- What are some methods for finding an approximate match to a string? What do we need to generate them?
- How can we evaluate a typical approximate matching system?

<https://powcoder.com>

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Needleman, Saul B. and Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48 (3): 44353.

doi:10.1016/0022-2836(70)90057-4

(Originally in Russian, published in English as:) Levenshtein, Vladimir I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady* 10 (8): 707710.

Smith, Temple F. and Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences". *Journal of Molecular Biology* 147: 195197. doi:10.1016/0022-2836(81)90087-5

Kondrak, Grzegorz. (2005). "N-Gram Similarity and Distance". In *Proceedings of the 12th international conference on String Processing and Information Retrieval (SPIRE'05)*, pp. 115-126, Buenos Aires, Argentina.

Zobel, Justin and Dart, Philip (1996). "Phonetic String Matching: Lessons from Information Retrieval". In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'96)*, pp. 166-172, New York, USA.

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Whitelaw, Casey and Hutchison, Ben and Chung, Grace Y and Ellis, Gerard (2009). "Using the Web for Language Independent Spellchecking and Autocorrection". In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), pp. 890-899, Singapore, Singapore.

<https://powcoder.com>

Ahmad, Farooq and Kondrak, Grzegorz (2005). "Learning a Spelling Error Model from Search Query Logs". In Proceedings of the Human Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 955-962, Vancouver, Canada.

Add WeChat powcoder

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Typical Genomics problem:

<https://powcoder.com>

- Given a nucleotide/amino acid sequence (substring)
- Find whether the sequence occurs within a larger sequence (string)
- Possibly with “errors” (nucleotide/amino acid changes)

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Typical Genomics problem:

- Given a substring, find whether the sequence occurs within a larger string, possibly with “errors”
- Almost the same as spelling correction
- But **much** larger strings: a small genomics problem might involve comparing perhaps 1K character sequence against several 100K character sequences; alphabet is smaller

<https://powcoder.com>

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact
Approximate
Application

Methods

Neighbourhood
Edit Distance
N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Recall: we have a “short” ($\sim 1\text{K}$ character) nucleotide/amino acid sequence to compare against many long ($\sim 100\text{K}$ character) chromosomes/genes/proteins/etc.

<https://powcoder.com>

For example, if some member of the population has 99% of the sequence of interest, they might be susceptible to some medical condition

Add WeChat powcoder

We're allowed ~ 10 errors; alphabet is ~ 4 or ~ 20 characters

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Neighbourhood search:

<https://powcoder.com>

Roughly $4^{10} \times 10^{30}$ possible neighbours.

... Forget it.

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Global Edit Distance:

One string is $\sim 1\text{K}$ characters other is $\sim 100\text{K}$ characters.

... Every string comparison involves $\sim 99\text{K}$ insertions.

→ Prefers shorter chromosomes (not intended behaviour)

Add WeChat powcoder

Approximate String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Local Edit Distance:

<https://powcoder.com>
One string is ~1k characters, other is ~100K characters.

... Seems like the right idea.

Add WeChat powcoder

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

Local Edit Distance:

<https://powcoder.com>

One string is ~ 100 K characters, other is ~ 1 G characters

... Can't fit table into memory.

... Requires approximate solutions with heuristics, e.g. BLAST, FASTA

Add WeChat powcoder

Approximate
String Matching

COMP90049
Knowledge
Technologies

String Search

Exact

Approximate

Application

Methods

Neighbourhood

Edit Distance

N-Gram Distance

Phonetics

Evaluation

References

Genomics

Assignment Project Exam Help

N-Gram Distance:

With huge n (e.g. 80% of length of shorter string) can almost work!

Tends to prefer shorter chromosomes like Global Edit Distance

But better methods for using n -gram information, e.g. de Bruijn graphs

Add WeChat powcoder