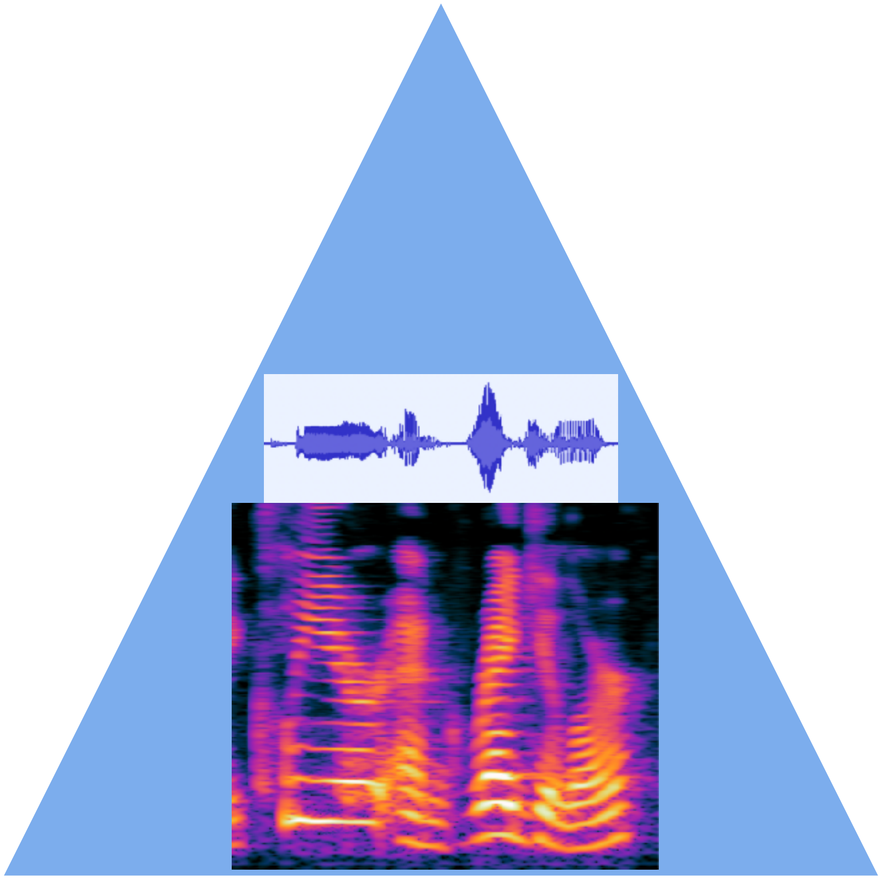
Lecture 6: Text processing and regular expression

Zhizheng Wu

Agenda

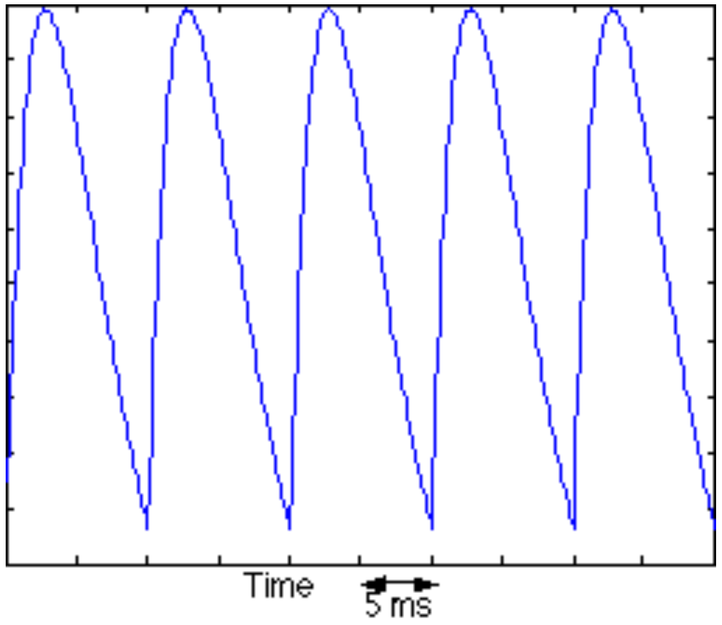
- Recap
- Text normalization
- Edit distance
- Regular expression

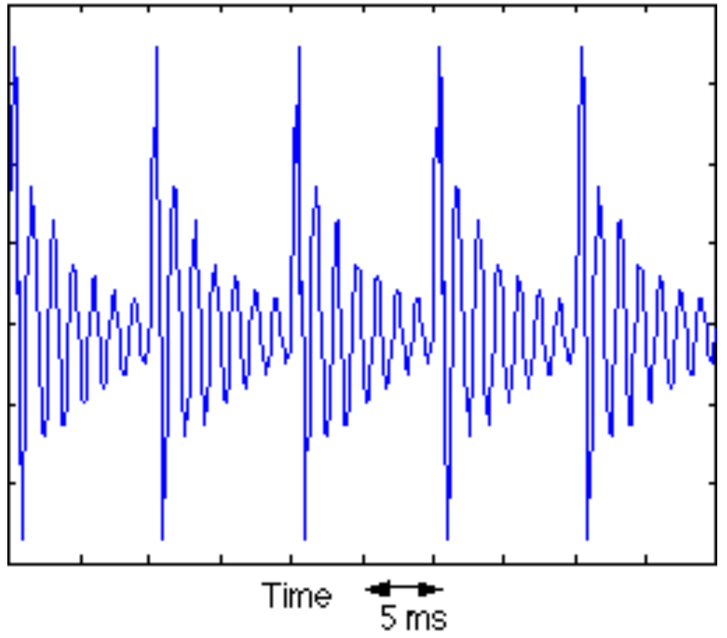
Content

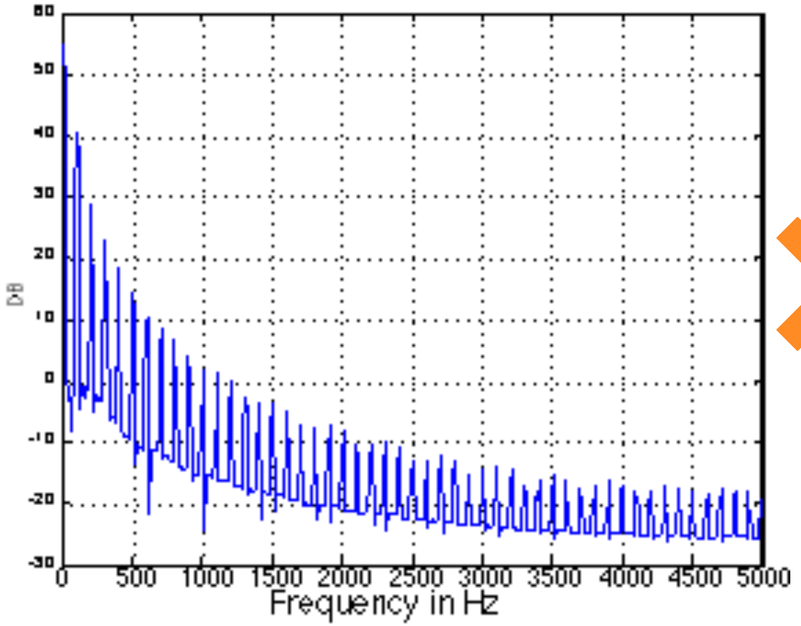


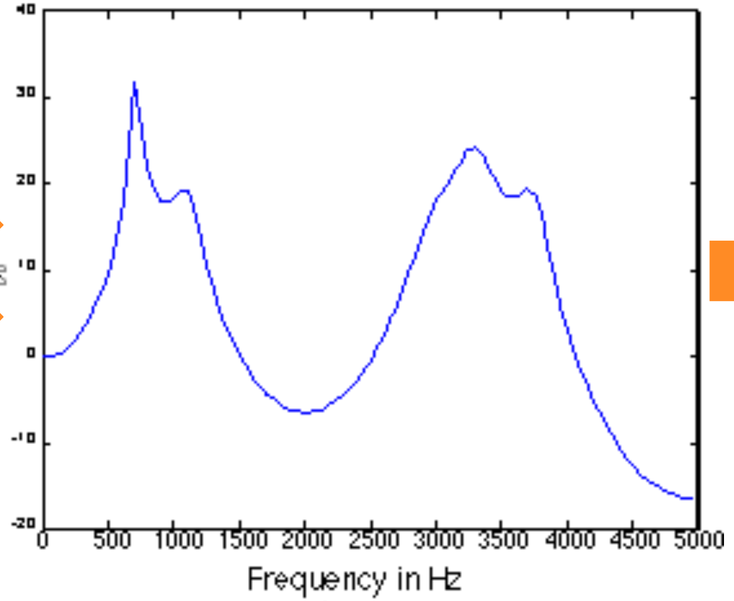
Timbre

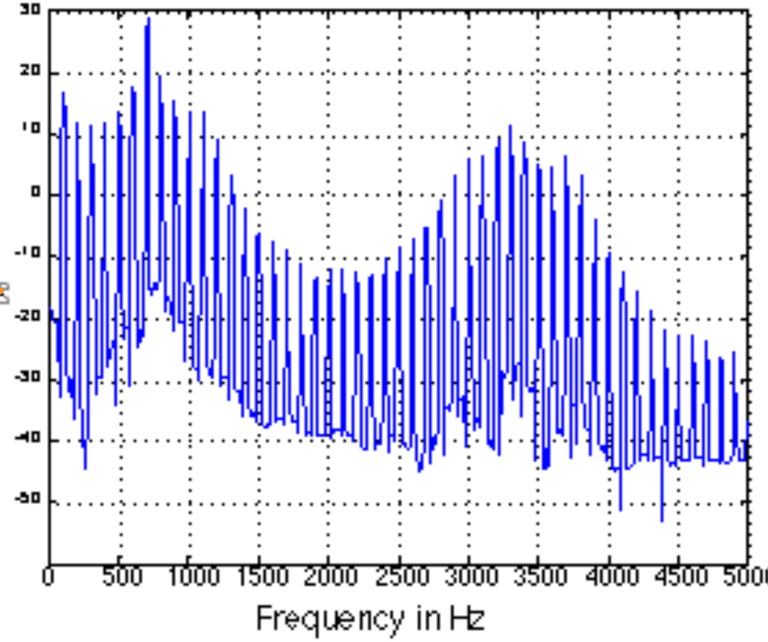
Prosody





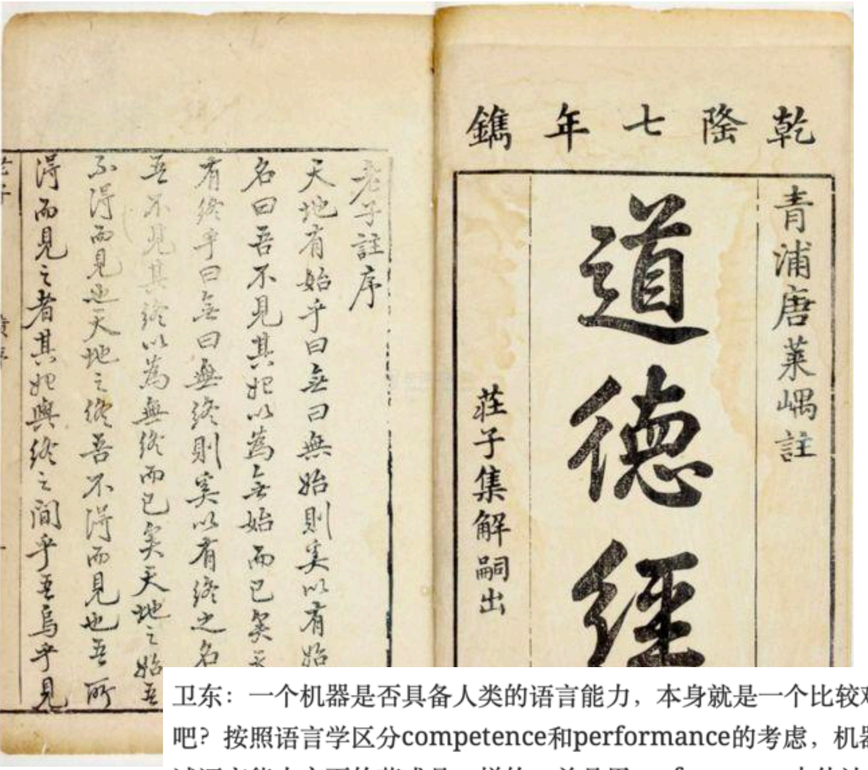


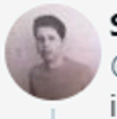




https://sail.usc.edu/~lgoldste/General_Phonetics/Source_Filter/MATLAB_demo/source-filter.html

From spoken language to written language





Sam Altman · Dec 4, 2022
@sama · Follow
interesting to me how many of the ChatGPT takes are either "this is AGI" (obviously not close, lol) or "this approach can't really go that much further".

trust the exponential. flat looking backwards, vertical looking forwards.

Elon Musk
@elonmusk · Follow

ChatGPT is scary good. We are not far from dangerously strong AI.

3:48 AM · Dec 4, 2022

25.3K Reply Share

卫东：一个机器是否具备人类的语言能力，本身就是一个比较难判断的问题吧？按照语言学区分competence和performance的考虑，机器和人，在测试语言能力方面的范式是一样的，总是用performance去估计competence。所以，真正的“语言能力”，大概也只能是一种“感觉”吧。chatgpt现在的表现，应该是让很多人“觉得”它掌握了语言。人们似乎还没有想出比图灵测试更高明的方法，来判断机器是否具有语言能力。

霄云：图灵测试 is not for language only, it is end to end “common sense “ test, human intelligence via language.

卫东：是的。它包含了语言能力。

南山：所以纠结机器是否智能在可预见未来是无解的，相关的判别标准和概念大家都没有清晰、一致，对于chatgpt、alphzero这类，看疗效才是王道。

霄云：单独测 language 是不是 翻译 或者别的 normalization 就可以？ @詹卫东

[Assignment1 Q1 2]Some questions about concepts and parameters

For task1, I'm not sure about the meaning of frameshift and window size.
Is the hop_length matches the frameshift and the win_size matches the window size? And if the hop_length is in ms?
For task2...

Read more

↑ 1 ☺ 1 comment

WhiteEurya Collaborator posted in General · yesterday

How can I submit my assignment?

I know it is a dumb question :(
But I want to check again that, should I just replace the "assignment_1.ipynb" in my private project or something else?

↑ 1 ☺ 1 comment

Corpora

- Words don’t appear out of nowhere
- Any particular piece of text is produced
  - by one or more specific speakers or writers
  - in a specific dialect of a specific language
  - at a specific time
  - in a specific place
  - for a specific function

Corpora along multiple dimensions

- Language: English, Chinese, etc
- Genre: Fiction, Scientific articles, Twitter, etc
- Author Demographics: writer's age, gender, etc
- Code switching: e.g. English/Chinese
- Variety: organization vs organisation

Corpus: tokens vs vocabulary

- Type: an element of the vocabulary
- Token: an instance of that type in running text

GPT-3 training data

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

How many words in a sentence?

they lay back on the San Francisco grass and looked at the stars and their

How many?
Tokens: 15
Types: 13

Text normalization

- Normalizing text into standard format

- Every NLP task requires text normalization
  - Tokenizing (segmenting) words
  - Normalizing word formats
  - Segmenting sentences

Word tokenization

- Splitting a text into separate words, or tokens, while preserving the meaning of the text

- Examples
  - I can't believe it's 2023 already!
    - Tokens: ["I", "can't", "believe", "it's", "2023", “already!"]
  - Let's meet at 7 PM at the café.
    - Tokens: ["Let's", "meet", "at", "7", "PM", "at", "the", “café."]

Word tokenization

the Rock 'n' Roll Brooklyn Half Marathon course in Brooklyn, New York

["the", "Rock", "'n'", "Roll", "Brooklyn", "Half", "Marathon", "course", "in", "Brooklyn,", "New", "York"]

["the", “Rock ’n' Roll", "Brooklyn", "Half", "Marathon", "course", "in", "Brooklyn,", “New York"]

Tokenization in languages without spaces

- Many languages (e.g. Chinese) don’t use spaces to separate words
- How do we decide where the token boundaries should be?

- Chinese as an example
  - 乒乓球拍卖完了

Chinese word segmentation

乒乓球拍卖完了

乒乓球拍/卖完了

乒乓球/拍卖/完了

Chinese word segmentation

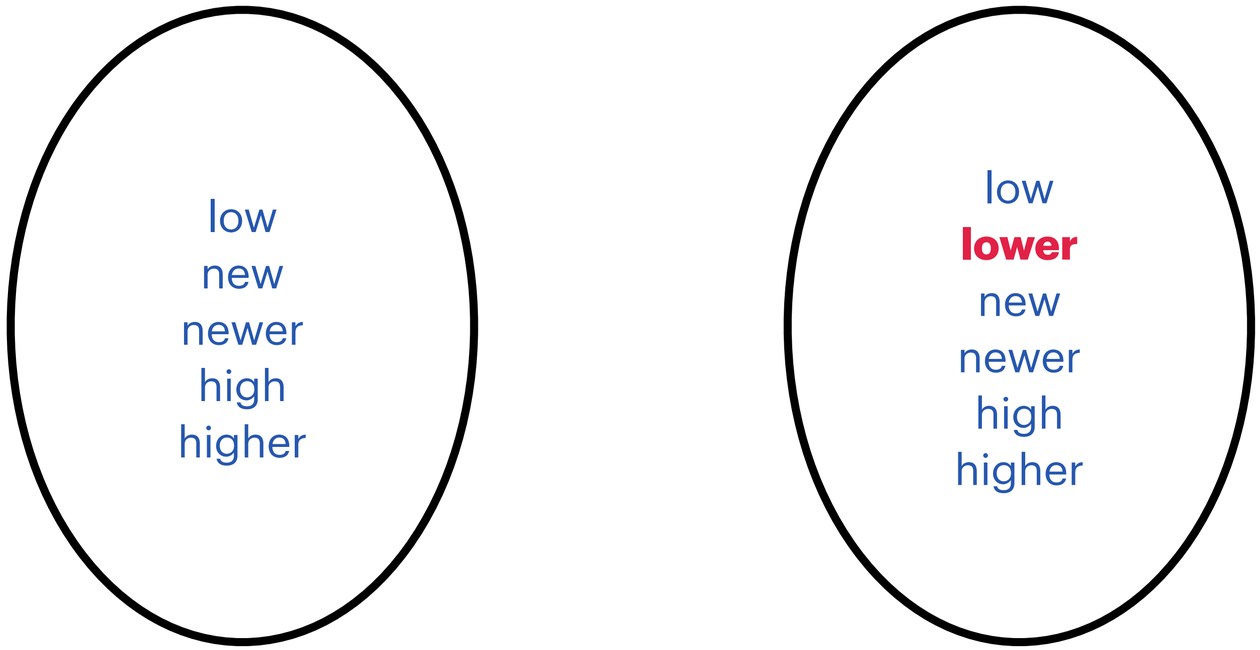
姚明进入总决赛

姚明 进入 总决赛

姚 明 进入 总 决赛

姚 明 进 入 总 决 赛

Word tokenization: Out-Of-Vocabulary



Subword tokenization

- Definition: tokens are smaller than words. Subwords can be arbitrary substrings

- Tokenization schemes:
  - Token learning
  - token segmenter

- Three algorithms
  - Byte-pair encoding
  - Unigram language modeling
  - Wordpiece

Byte-pair encoding

- Originally proposed for lossless data compression

aaabdaaabac

aaabdaaabac Replace aa with Z

ZabdZabac Replace ab with Y

ZabdZabac Replace ab with Y

ZYdZYac

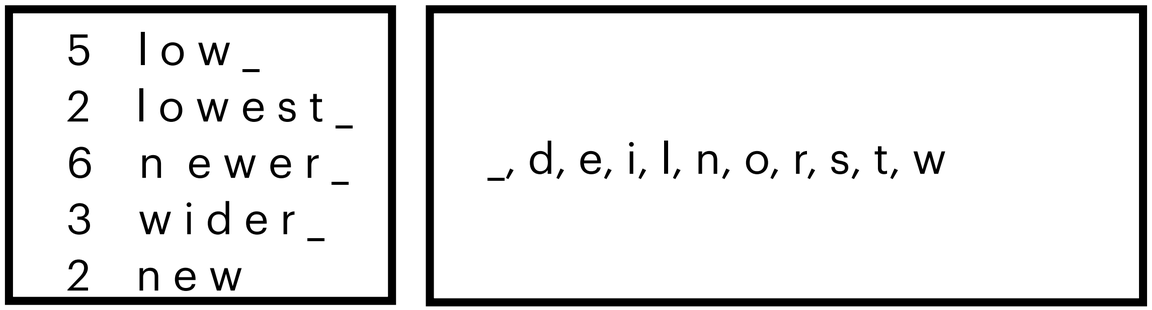
...

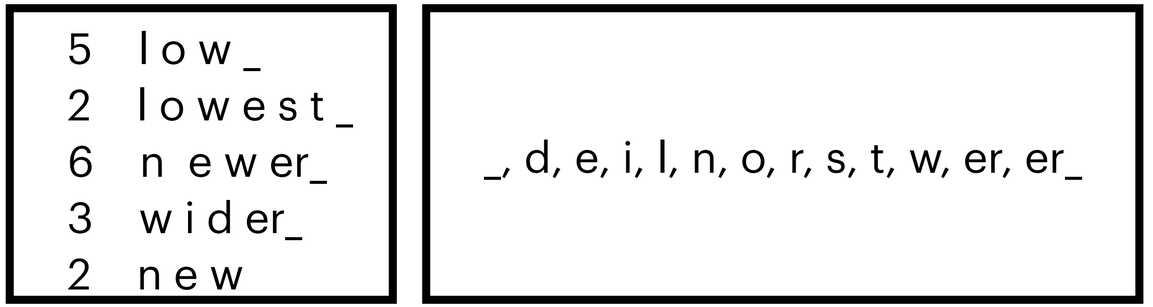
BPE algorithm

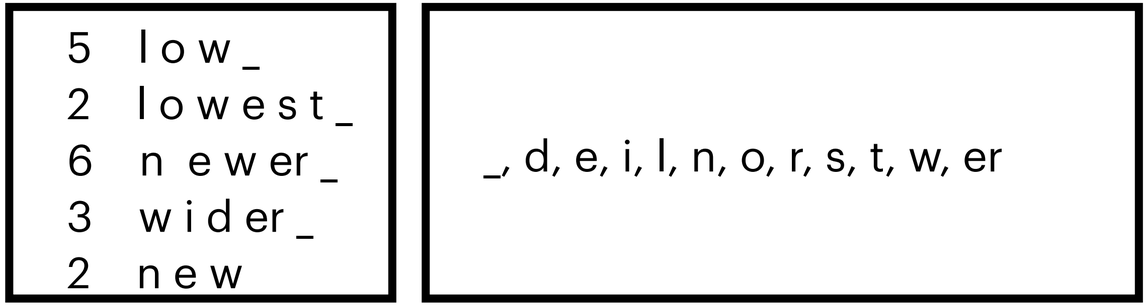
function BYTE-PAIR ENCODING(strings C, number of merges k) **returns** vocab V

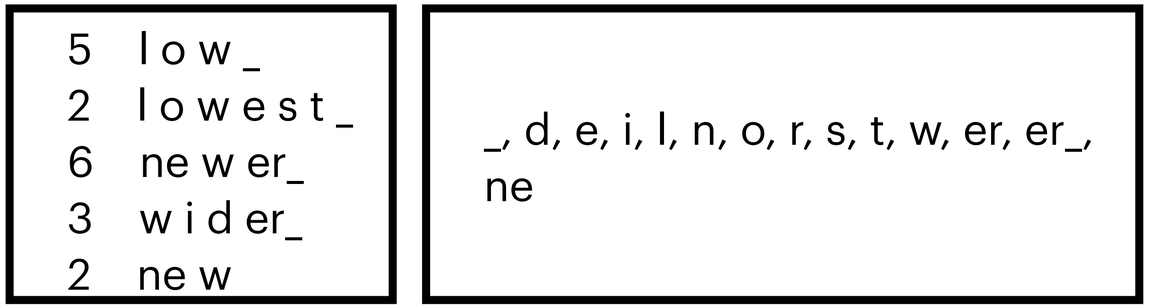
$V \leftarrow$ all unique characters in C # initial set of tokens is characters
for $i = 1$ **to** k **do** # merge tokens k times
  $t_L, t_R \leftarrow$ Most frequent pair of adjacent tokens in C
  $t_{NEW} \leftarrow t_L + t_R$ # make new token by concatenating
  $V \leftarrow V + t_{NEW}$ # update the vocabulary
  Replace each occurrence of t_L, t_R in C with t_{NEW} # and update the corpus
return V

BPE for subword tokenization









BPE for subword tokenization

Merge	Current vocabulary
(ne, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new
(l, o)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo
(lo, w)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low
(new, er_)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_
(low, _)	_, d, e, i, l, n, o, r, s, t, w, er, er_, ne, new, lo, low, newer_, low_

Applying BPE

- The word: ‘lower’

l o w e r _

l o w er _

l o w er_

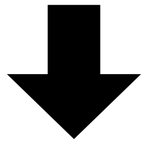
lo w er_

low er_

Word normalization

- A task to put word into a standard format, choosing a single normal form for words with multiple forms like USA and US.

CUHK-SZ, CUHK(SZ), CUHKSZ, CUHK-Shenzhen



CUHK-Shenzhen

Sentence segmentation

- Cut long text into individual sentences
- The most useful cues:
  - Punctuation (e.g. periods, question marks, and exclamation points)
  - The period character “.” is ambiguous between a sentence boundary marker and a marker of abbreviations like Mr. or Inc.

How similar are two strings?

- Given a word ‘*coleague*’, which is the closest?

  - Colleague
  - College
  - Colegio
  - …

Minimum Edit distance

- Edit distance gives us a way to quantify string similarity
- Edit operations
  - Insertion
  - Deletion
  - Substitution

- Minimum edit distance
  - the minimum number of editing operations (operations like insertion, deletion, substitution) needed to transform one string into another

Alignment

- An alignment is a correspondence between substring of two sequences
- The minimum edit distance can be represented as an alignment

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

d: deletion

s: substitution

i: insertion

Minimum edit distance

- Initialization

  D(i, 0) = i

  D(0, j) = j

- Recurrence relation

  For i = 1…M

  For j = 1…N

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} \end{cases}$$

- Termination

  D(N, M) is distance

Edit distance table

	M	**O**	**N**	**K**	**E**	**Y**
M	0	1	2	3	4	5
O	1	0	1	2	3	4
N	2	1	0	1	2	3
E	3	2	1	2	1	2
Y	4	3	2	3	2	1

Regular expression

- A sequence of characters that specifies a ***pattern*** in text

Someone@cuhk.edu.cn

Someone@stanford.edu

Someone@mit.edu

Someone@ntu.edu.tw

Someone@ntu.edu.sg

Regular expression

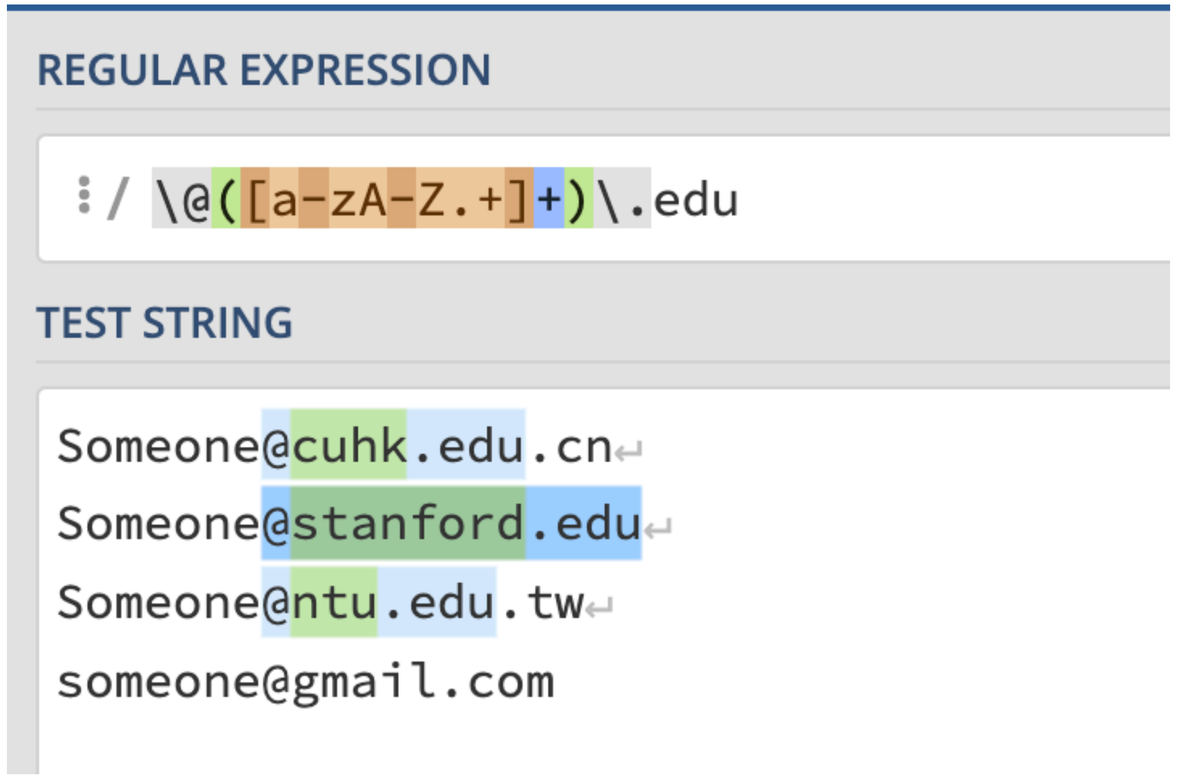
Someone@cuhk.edu.cn

Someone@stanford.edu

Someone@mit.edu

Someone@ntu.edu.tw

Someone@ntu.edu.sg



To practice: https://regex101.com/

Summary

- Every NLP task requires text normalization
  - Tokenizing (segmenting) words
  - Normalizing word formats
  - Segmenting

- Minimum edit distance
- Regular expression