NATURAL LANGUAGE PROCESSING (NLP) FUNDAMENTALS

TEXT PRE-PROCESSING

DR LEK HSIANG HUI

Introduction to Text Pre-processing

Text Pre-processing

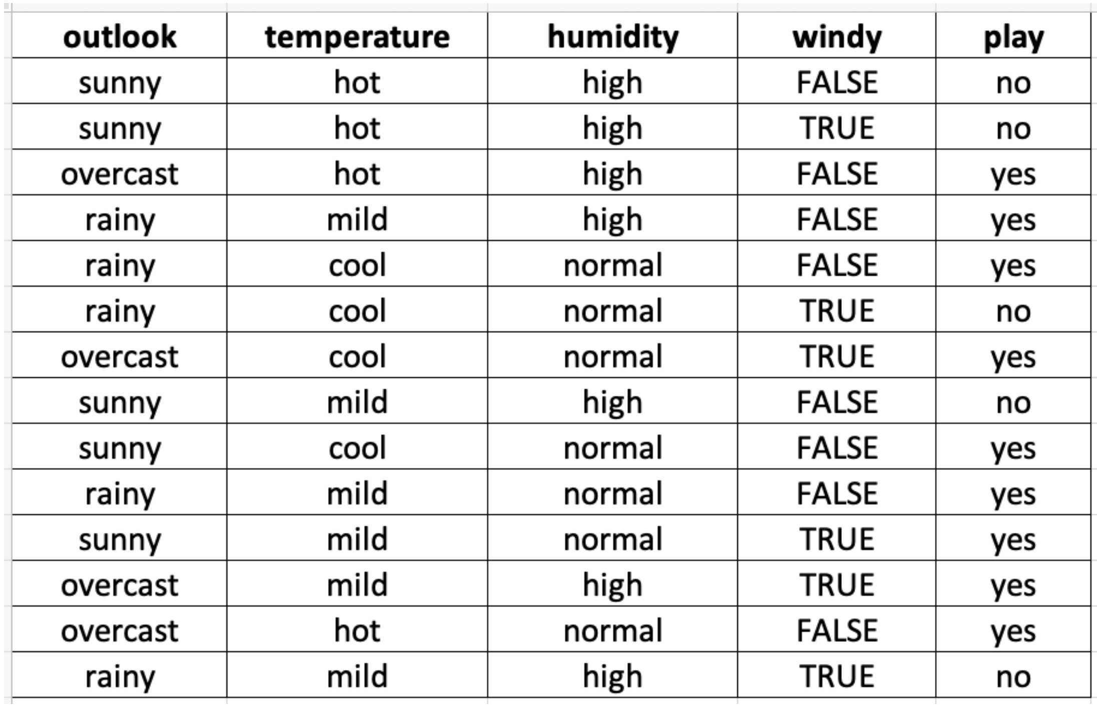
- Before we can build useful NLP application, one of the most important steps is to perform **text pre-processing**
- Like the **data pre-processing** step in data analytics workflow, text pre-processing is focused at preparing the text data into a more ideal form

Text Pre-processing

- High-level objectives of text pre-processing:
  - Convert **unstructured** text data into a **structured** form
  - Convert text data into a **more general form** so that **machine learning** algorithms can work better

Structured Data

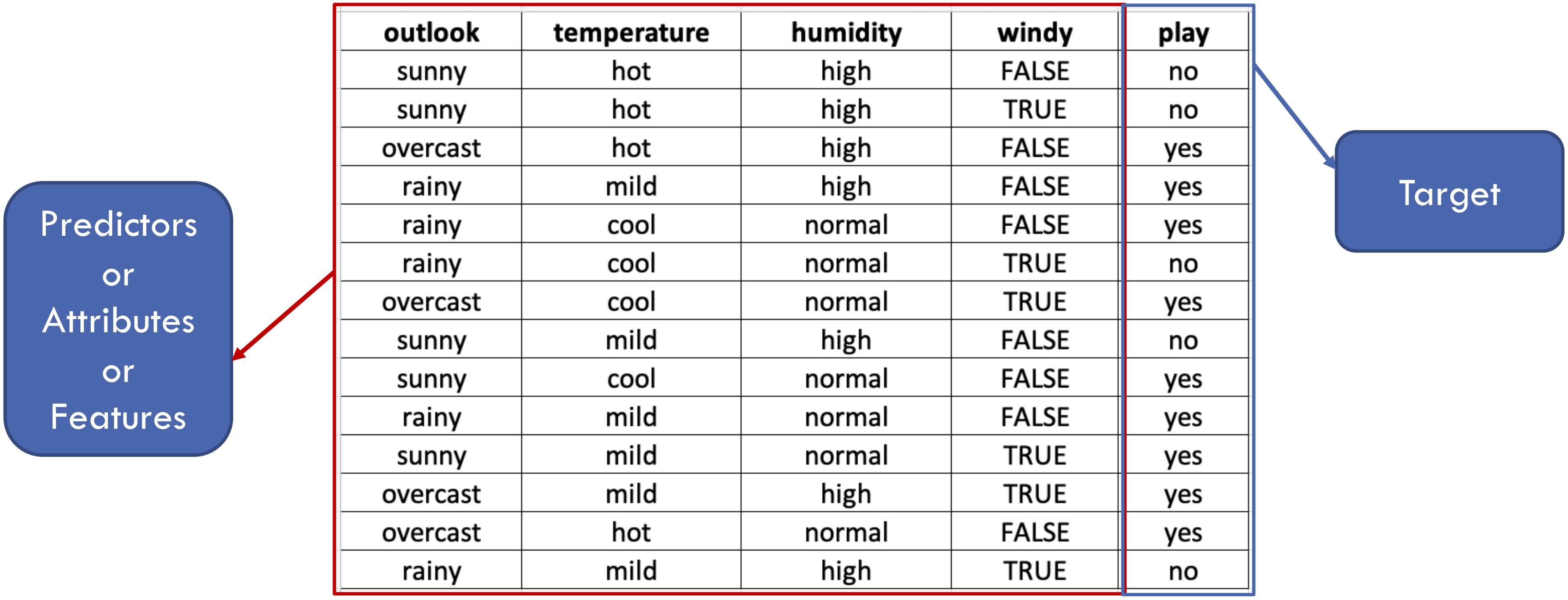
- □ Data useful for analytics are often in a structured form that looks like a spreadsheet



outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Structured Data

□ Data useful for analytics are often in a structured form that looks like a spreadsheet



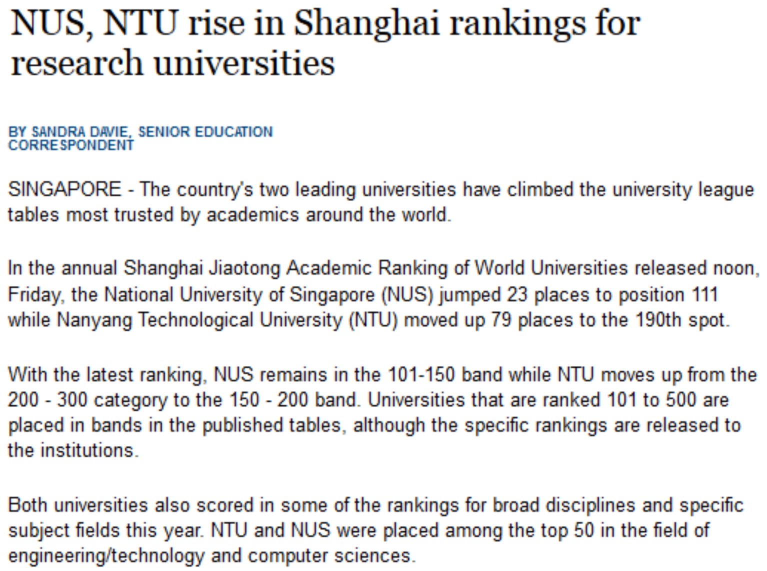
outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Predictors or Attributes or Features

Target

Unstructured Data

- □ Text data by default is unstructured
- □ So, there is a need to do some pre-processing to convert it into a structured form before performing data modeling



NUS, NTU rise in Shanghai rankings for research universities

BY SANDRA DAVIE, SENIOR EDUCATION CORRESPONDENT

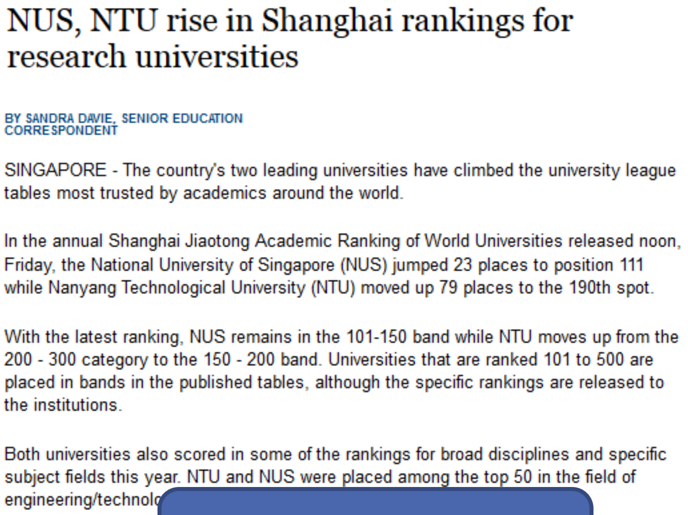
SINGAPORE - The country's two leading universities have climbed the university league tables most trusted by academics around the world.

In the annual Shanghai Jiaotong Academic Ranking of World Universities released noon, Friday, the National University of Singapore (NUS) jumped 23 places to position 111 while Nanyang Technological University (NTU) moved up 79 places to the 190th spot.

With the latest ranking, NUS remains in the 101-150 band while NTU moves up from the 200 - 300 category to the 150 - 200 band. Universities that are ranked 101 to 500 are placed in bands in the published tables, although the specific rankings are released to the institutions.

Both universities also scored in some of the rankings for broad disciplines and specific subject fields this year. NTU and NUS were placed among the top 50 in the field of engineering/technology and computer sciences.

Text Pre-processing



NUS, NTU rise in Shanghai rankings for research universities

BY SANDRA DAVIE, SENIOR EDUCATION CORRESPONDENT

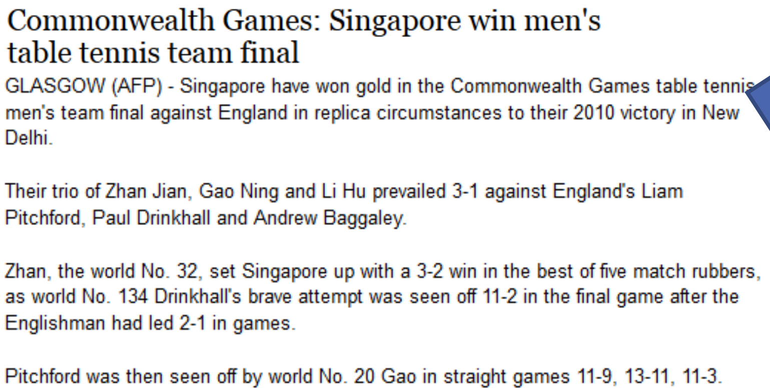
SINGAPORE - The country's two leading universities have climbed the university league tables most trusted by academics around the world.

In the annual Shanghai Jiaotong Academic Ranking of World Universities released noon, Friday, the National University of Singapore (NUS) jumped 23 places to position 111 while Nanyang Technological University (NTU) moved up 79 places to the 190th spot.

With the latest ranking, NUS remains in the 101-150 band while NTU moves up from the 200 - 300 category to the 150 - 200 band. Universities that are ranked 101 to 500 are placed in bands in the published tables, although the specific rankings are released to the institutions.

Both universities also scored in some of the rankings for broad disciplines and specific subject fields this year. NTU and NUS were placed among the top 50 in the field of engineering/technolo

Documents



Commonwealth Games: Singapore win men's table tennis team final

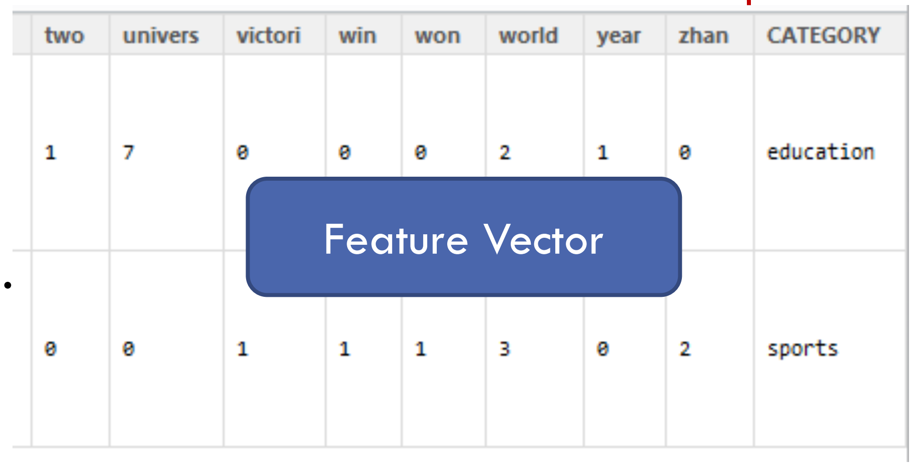
GLASGOW (AFP) - Singapore have won gold in the Commonwealth Games table tennis men's team final against England in replica circumstances to their 2010 victory in New Delhi.

Their trio of Zhan Jian, Gao Ning and Li Hu prevailed 3-1 against England's Liam Pitchford, Paul Drinkhall and Andrew Baggaley.

Zhan, the world No. 32, set Singapore up with a 3-2 win in the best of five match rubbers, as world No. 134 Drinkhall's brave attempt was seen off 11-2 in the final game after the Englishman had led 2-1 in games.

Pitchford was then seen off by world No. 20 Gao in straight games 11-9, 13-11, 11-3.

Usually large amount of features



two	univers	victori	win	won	world	year	zhan	CATEGORY
1	7	0	0	0	2	1	0	education
0	0	1	1	1	3	0	2	sports

Feature Vector

…

Basic Building Blocks of Text Pre-processing

- Some basic building blocks useful for text pre-processing includes:
  - **Sentence Segmentation**
  - **Tokenization**
  - **Text Normalization** *(will be covered in the next lecture)*
- We also need to be able to perform **string manipulation** effectively
  - Use **Regular Expression**!

Regular Expressions

Regular Expressions (Regex)

- What is Regular Expressions (Regex)?
  - A language for specifying text search pattern useful for performing string matching and processing

Regex – Simple Case

- The simple case is when we are searching for a predefined string
  - E.g.
    natural language processing
    is searching for the presence of the string *natural language processing*

Regex – Character Matching

- Possible to define regex to match characters using [ ]
  - E.g. In order to match the following:
    color, Color
  - We can use this Regex:
    [cC]olor

Pattern	Description	Match Example
[abcABC]	Matches a single character that is either a, b, c, A, B, C	A
[a-zA-Z]	Matches a single character in the range a to z or A to Z	C
[a-zABC]	Matches a single character that is in the range of a to z or A, B, C	b

Regex – Character Matching

Use \ to escape characters

Pattern	Description	Match Example
[a-z0-9\-]	Matches a single character in the range a to z or 0 to 9 or the ”-” character	-
[^a-zA-Z]	Matches a single character that is **not** in the range a to z and A to Z	0
[^abc]	Matches a single character that is **not** a, b, c	A
[a^]	Matches a single character that is a or the “^” character	^
[^a^]	Matches a single character that is not a or the “^” character	b

Negation

Regex – “Or”

Pattern	Description	Match Example
mouse|mice	Matches the word “mouse” or “mice” but does not match “mousemice”	mouse
a|b|c	Same as [abc]	a
[hH](ello|i)	Matches “hello” or “Hello” or “hi” or “Hi"	hello

Regex – Metacharacters

- **Metacharacters** are characters with special meaning

Pattern	Description	Match Equivalent
\w	Matches a word character	[a-zA-Z0-9_]
\W	Matches a non-word character	[^a-zA-Z0-9_]
\d	Matches a digit	[0-9]
\D	Matches a non-digit	[^0-9]
\s	Matches a whitespace character	[\n\r\t\f]
\S	Matches a non-whitespace	[^ \n\r\t\f]
\b	Matches boundary between word and non-word	
.	Matches any character	

Regex – Repetition

Pattern	Description	Match Example
[0-9]*	Matches a string with digits appearing **zero or more times**	1234
[0-9]+	Matches a string with digits appearing **one or more times**	[^a-zA-Z0-9_]
colou?r	? Makes the previous expression/character **optional**	color colour
[0-9]{3}	Matches a string with 3 digits	789
[0-9]{1,3}	Matches a string with 1 to 3 digits	1 12 512
[0-9]{2,}	Matches a string with 2 (or more) digits	1251

Regex – Repetition

- A common pattern is to match any character or any word

Pattern	Description	Match Example
.*	Matches any character zero or more times	(empty string) (space) hello world
.+	Matches any character one or more times	Similar to .* except that does not match (empty string)
\w+	Matches any word one or more times	hello world
(hello)+	Matches the string hello one or more times	hellohello

Regex – Greedy Matching

- Note that * and + performs **greedy** matching

  e.g.

  String: ***hello how are you? nice to meet you***

  Regex: **hello.*you**

  Will match the entire string

  hello how are you? nice to meet you

Regex – Non-Greedy Matching

- □ If we want the matching to stop "as soon as we find a match", we can perform a **non-greedy** matching
  e.g.
  String: ***hello how are you? nice to meet you***

  Regex: **hello.*?you**

  Will match from "hello" until the first "you"

  hello how are you? nice to meet you

Regex – Anchor

- Often when matching regular expressions, we will get multiple matches
  e.g.
  String: ***hello how are you? nice to meet you***

  Regex: \w+

  8 matches: hello, how, are, you?, nice, to, meet, you

Regex – Anchor

- □ Sometimes we just want to match the beginning or end of the line

  e.g.

  String: ***hello how are you? nice to meet you***

  Regex (beginning of line): **^\w+**

  1 match: hello

  Regex (end of line): **\w+$**

  1 match: you

Regex – Capturing Group

- □ We have seen that ( ) can be used to group up characters

  e.g.

  [hH](ello|i)

- □ But sometimes instead of just finding a match, we want to extract certain parts of the match, we can also use ( ) to capture the pattern into **capturing groups**

Regex – Capturing Group

- e.g.
  <html><body><ul><li>**johndoe@gmail.com**</li><li>**janedoe@gmail.com**</li><li>**Robin**</li></ul></body></html>

- We want the bold section, can use this regex:
  <li>(.*?)</li>

Sentence Segmentation & Tokenization

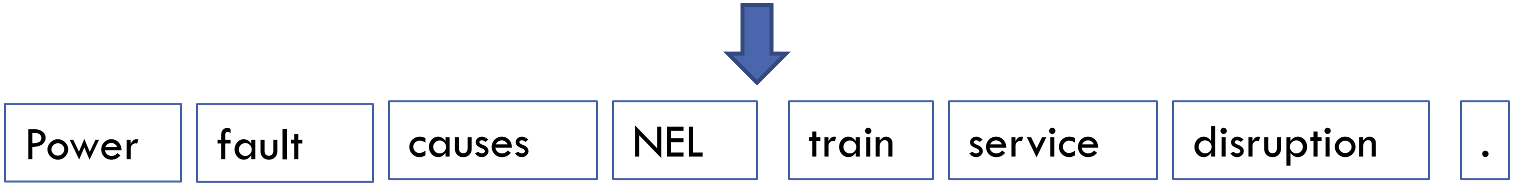
Sentence Segmentation

- Breaking passage into sentences
  - Using “!”, “?” as indicators of the end of a sentence
  - How about period “.”?
    - Sentence boundary
    - Used in abbreviation : “Ms.”
    - Used in numbers 0.12

Tokenization

- □ Process of breaking a sentence into different tokens

Power fault causes NEL train service disruption.



Power | fault | causes | NEL | train | service | disruption | .

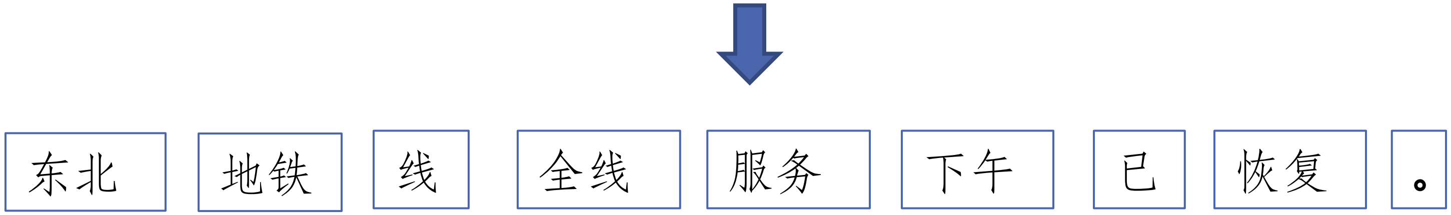
- □ Generally an easy task for English
  - Split the string by space and punctuation
- □ Some problems for hypenation, apostrophe, periods
  - aren’t → [“are”, “n’t”], o’neill → [“o’neill”]
  - Bishan-Toa Payoh→ [“Bishan”, “Toa Payoh”], co-education → [“co-education”]
  - Ph.D., Mr.

Tokenization

- □ More challenging for other languages (e.g. Chinese)

东北地铁线全线服务下午已恢复正常。

(The NEL train service has been fully restored in the afternoon.)



东北 地铁 线 全线 服务 下午 已 恢复 。

- □ Unlike English, there is no white space in Chinese

NATURAL LANGUAGE PROCESSING (NLP) FUNDAMENTALS

TEXT PRE-PROCESSING

DR LEK HSIANG HUI