Week 8 Tutorial

Regular Expressions,
Nonregular Languages

Part 1: ***Designing Regular Expressions***

Designing Regexes

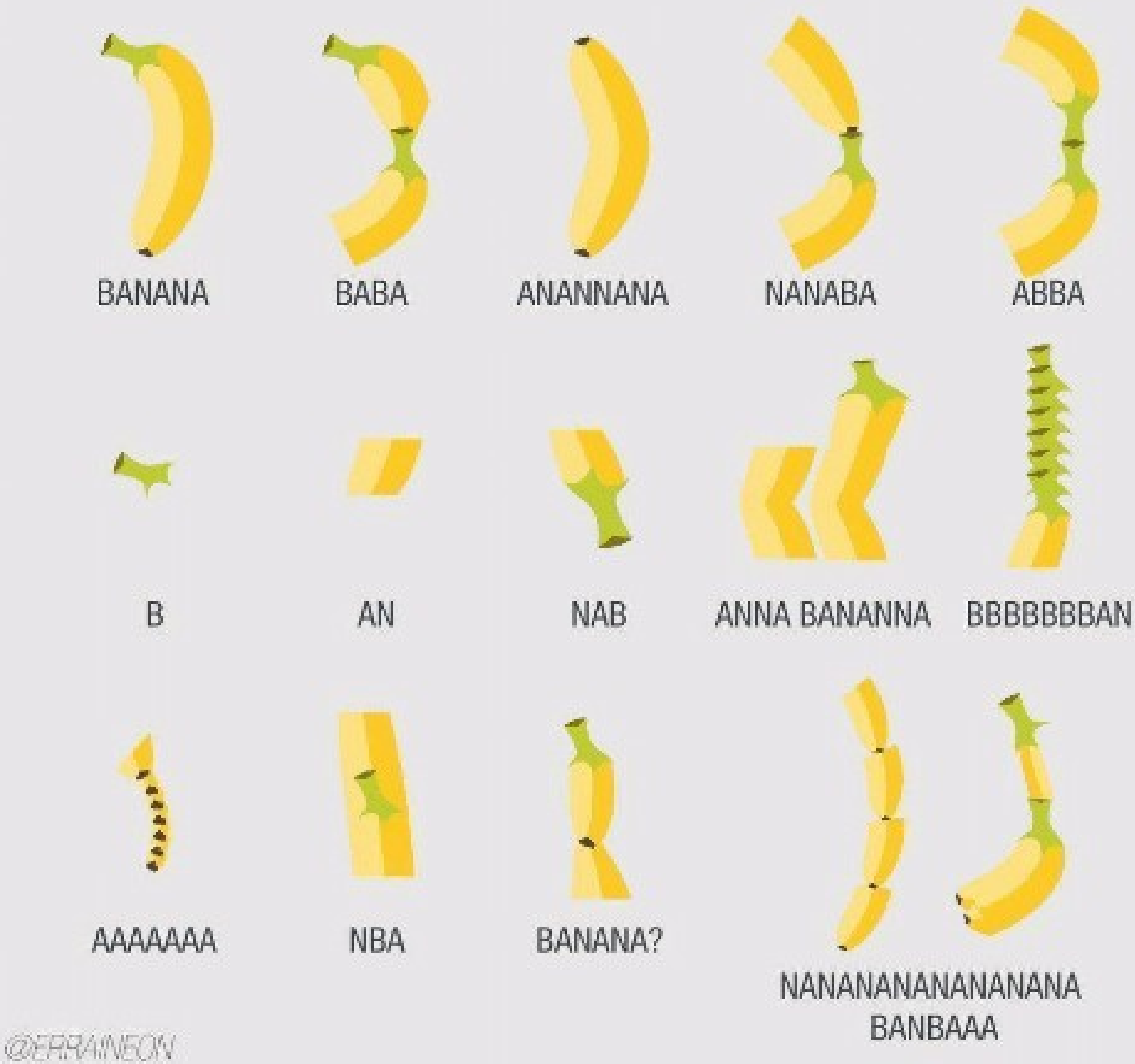
Write out some sample strings in the language and look for patterns:

- Can I separate out the strings into two (or more) categories?
  - ***Union*** – find the pattern for each category, then union together

- Can I break this problem down into solving some smaller subproblems?
  - ***Concatenation*** - find the pattern for each piece/subproblem, then concatenate together

- Is there some sort of repeating structure?
  - ***Kleene star*** – find smallest repeating unit, then star that pattern

BANANA GRAMMAR



Banana Languages

Let $\Sigma = \{$ **A**, **N** $\}$.

Design a regex for the language

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

Banana Languages

Let Σ = { **A**, **N** }.

Design a regex for the language

L = { $w \in \Sigma^*$ | $w \neq \varepsilon$ and the characters of w alternate between **A** and **N** }

N $\in L$	**AAN** $\notin L$
ANANA $\in L$	**NNNNN** $\notin L$
NANANANAN $\in L$	**ANAANA** $\notin L$

Banana Languages

Let Σ = { **A**, **N** }.

Design a regex for the language

L = { $w \in \Sigma^*$ | $w \neq \varepsilon$ and the characters of w alternate between **A** and **N** }

1) Create a regex for the language above.

Submit on Gradescope!

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

A

N

AN

NA

ANA

NAN

ANAN

NANA

ANANA

NANAN

...

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

A

N

AN

NA

ANA

NAN

ANAN

NANA

ANANA

NANAN

...

Can I separate out the strings into two (or more) categories?

- ***Union*** – find the pattern for each category, then union together

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

Starts with **A**	Starts with **N**
A	**N**
AN	**NA**
ANA	**NAN**
ANAN	**NANA**
ANANA	**NANAN**
...	**...**

Can I separate out the strings into two (or more) categories?

- ***Union*** – find the pattern for each category, then union together

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

Starts with **A**	Starts with **N**
A	**N**
AN	**NA**
ANA	**NAN**
ANAN	**NANA**
ANANA	**NANAN**
...	...

Banana Languages

L = { $w \in \Sigma^*$ | $w \neq \varepsilon$ and the characters of w alternate between **A** and **N** }

Starts with **A**	Starts with **N**
A	**N**
AN	**NA**
ANA	**NAN**
ANAN	**NANA**
ANANA	**NANAN**
. . .	**. . .**

Can I break this problem down into solving some smaller subproblems?

- ***Concatenation*** - find the pattern for each piece/subproblem, then concatenate together

Banana Languages

L = { $w \in \Sigma^*$ | $w \neq \varepsilon$ and the characters of w alternate between **A** and **N** }

Starts with **A**	Starts with **N**
A	**N**
AN	**NA**
ANA	**NAN**
ANAN	**NANA**
ANANA	**NANAN**
...	**...**

Can I break this problem down into solving some smaller subproblems?

- ***Concatenation*** - find the pattern for each piece/subproblem, then concatenate together

A(sequence of **NA**s)(possibly another **N**)

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

Starts with **A**	Starts with **N**
A	**N**
AN	**NA**
ANA	**NAN**
ANAN	**NANA**
ANANA	**NANAN**
...	**...**

Is there some sort of repeating structure?

- ***Kleene star*** – find smallest repeating unit, then star that pattern

A(sequence of **NA**s)(possibly another **N**)

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

Starts with **A**	Starts with **N**
A	**N**
AN	**NA**
ANA	**NAN**
ANAN	**NANA**
ANANA	**NANAN**
...	...

A(**NA**)***N**?

Is there some sort of repeating structure?

- ***Kleene star*** – find smallest repeating unit, then star that pattern

Banana Languages

$L = \{ w \in \Sigma^* \mid w \neq \varepsilon$ and the characters of w alternate between **A** and **N** $\}$

Starts with **A**	Starts with **N**
A	**N**
AN	**NA**
ANA	**NAN**
ANAN	**NANA**
ANANA	**NANAN**
...	...

A(**NA**)***N**? ∪ **N**(**AN**)***A**?

Part 2: ***Myhill Nerode***

Approaching Myhill-Nerode

In lecture we saw how to prove a language L is non-regular using the Myhill-Nerode theorem. To so do, we:

1) Find an infinite, distinguishing set S.
2) Prove that S is an infinite set.
3) Prove that S is a distinguishing set by picking two arbitrary strings from S and showing that they're distinguishable relative to L.

Approaching Myhill-Nerode

- The challenge in using the Myhill-Nerode theorem is finding the right set of strings.
- ***General intuition:***
  - Start by thinking about what information a computer "must" remember in order to answer correctly.
  - Choose a group of strings that all require different information.
  - Prove that those strings are distinguishable relative to the language in question.

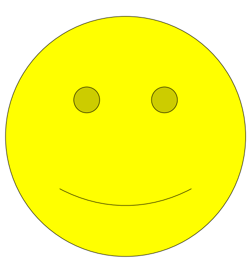
An Analogy

Imagine a scenario where Bob is thinking of a string and Alice has to figure out whether that string is in a particular language.

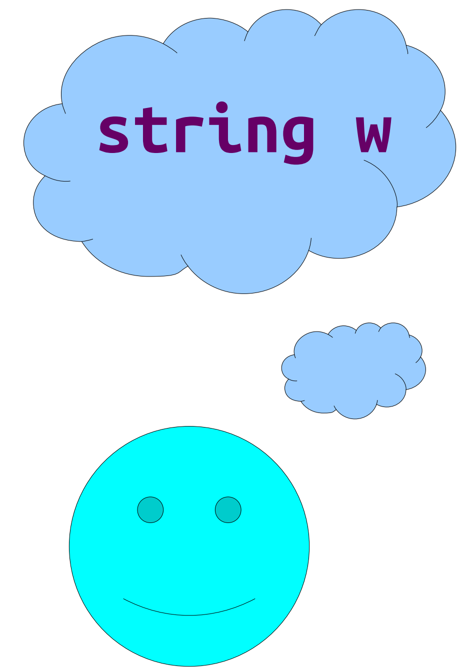
An Analogy

The catch: Bob can only send Alice one character at a time, and Alice doesn’t know how long the string is until Bob tells her that he’s done sending input.

language L



Alice

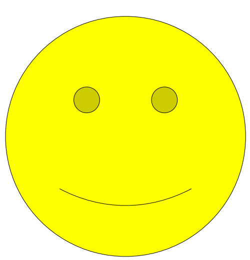


Bob

An Analogy

What does Alice need to remember about the characters she's receiving from Bob?

language L



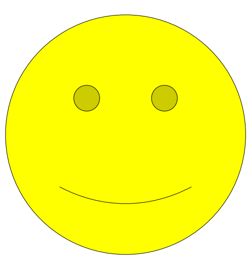
Alice

Bob

An Analogy

What does Alice need to remember about the characters she's receiving from Bob?

$L = \{ w \mid w$ is a natural number divisible by 5 $\}$



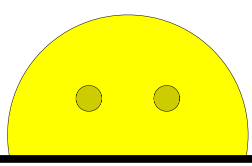
Alice

Bob

An Analogy

What does Alice need to remember about the characters she's receiving from Bob?

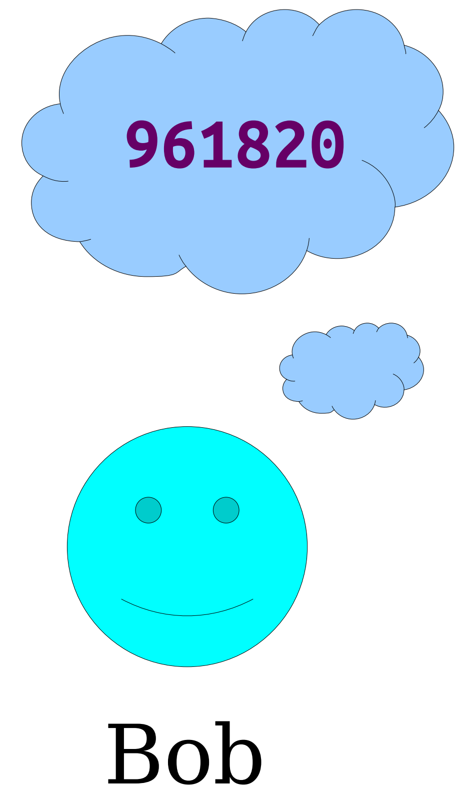
L = { w | w is a natural number divisible by 5 }



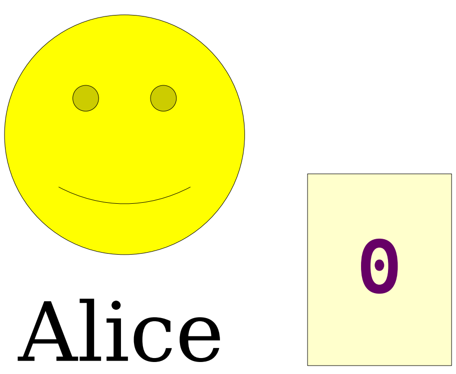
Initially, it seems like Alice has to remember the whole number that Bob is sending to her. But we only care about divisibility by 5 here, so we can get away with remembering a lot less!

An Analogy

Key insight: Alice only needs to remember ***the last character*** she received from Bob

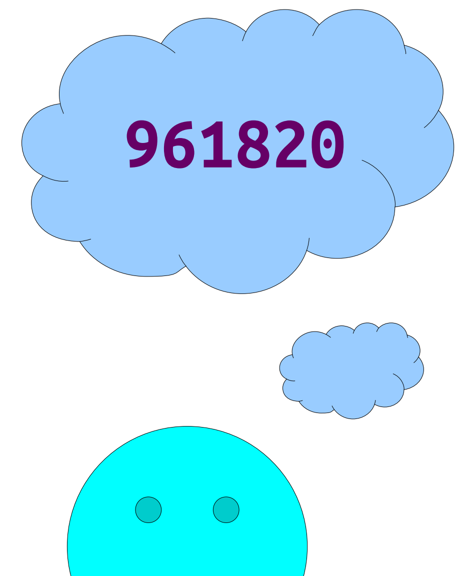


L = { w | w is a natural number divisible by 5 }

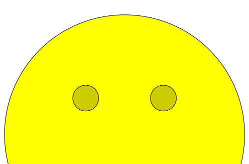


An Analogy

Key insight: Alice only needs to remember ***the last character*** she received from Bob



$L = \{ w \mid w$ is a natural number divisible by 5 $\}$

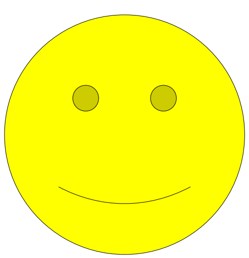


The number that Bob is thinking of could get unboundedly large, but the size of what Alice needs to remember remains constant (finite).

An Analogy

Let's contrast this with one of the non-regular languages we saw in class:

$$L = \{ \mathbf{a}^n\mathbf{b}^n \mid n \in \mathbb{N} \}$$



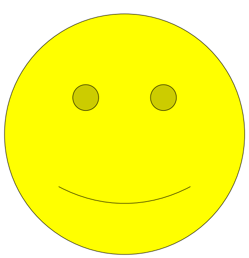
Alice

Bob

An Analogy

Alice needs to remember how many **a**'s she's seen so far, since she needs to verify that the number of **b**'s matches.

$$L = \{ \mathbf{a}^n \mathbf{b}^n \mid n \in \mathbb{N} \}$$



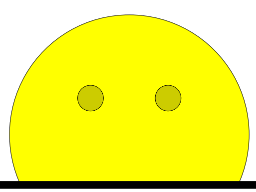
Alice

Bob

An Analogy

Alice needs to remember how many **a**'s she's seen so far, since she needs to verify that the number of **b**'s matches.

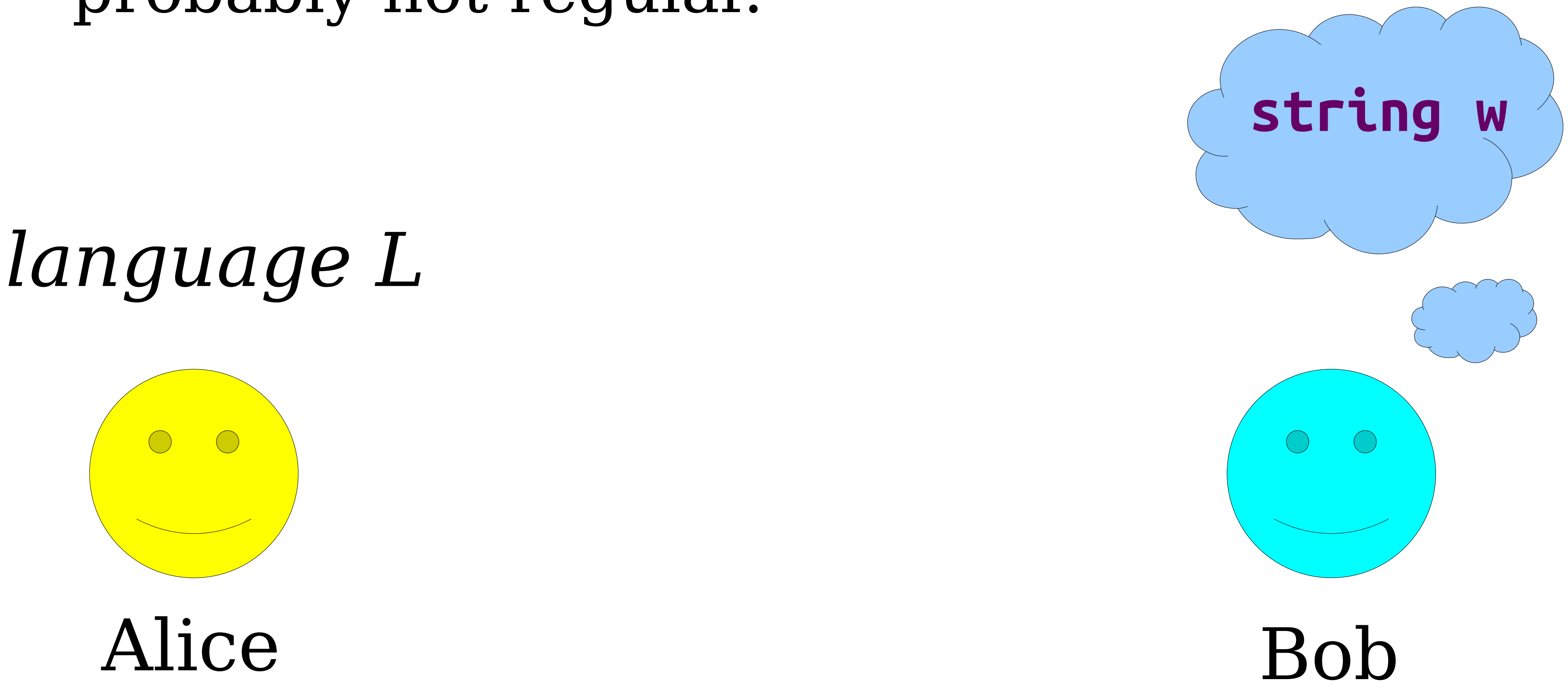
$$L = \{ \mathbf{a}^n \mathbf{b}^n \mid n \in \mathbb{N} \}$$



As the size of Bob's string gets larger, the amount of memory Alice needs also increases. Since Bob's string could get unboundedly large, we need infinite memory.

An Analogy

Key insight: if Alice has to remember ***infinitely*** many things, or one of ***infinitely*** many possibilities, the language is probably not regular.



More Banana Languages

Let Σ = { **A**, **N** }.

Consider the language

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$$

More Banana Languages

Let Σ = { **A**, **N** }.

Consider the language

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$$

N $\in L$

ANANA $\in L$

AAAA $\in L$

AN $\notin L$

NNNNAN $\notin L$

NAAA $\notin L$

More Banana Languages

Let $\Sigma = \{$ **A**, **N** $\}$.

Consider the language

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$$

2a) Which of the following sets would be a suitable choice to show that L is non-regular? For choices that don't work, explain why not.

i. { **A**, **AA**, **AAA**, **AAAA**, **AAAAA** }
ii. $\{ \mathtt{A}^n \mid n \in \mathbb{N} \}$
iii. $\{ \mathtt{A}^n\mathtt{N} \mid n \in \mathbb{N} \}$
iv. $\{ \mathtt{A}^n\mathtt{N}^n \mid n \in \mathbb{N} \}$

As a reminder, a distinguishing set $S \subseteq \Sigma^*$ is a set such that:

$$\forall x \in S.\ \forall y \in S.\ (x \neq y \rightarrow \exists w \in \Sigma^*.\ (xw \in L \leftrightarrow yw \notin L))$$

Submit on Gradescope!

$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome} \}$

$S = \{$ **A**, **AA**, **AAA**, **AAAA**, **AAAAA** $\}$

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$$

$$S = \{ \mathtt{A}^n \mid n \in \mathbb{N} \}$$

$$L = \{\ w \in \Sigma^* \mid w \text{ is a palindrome}\ \}$$

$$S = \{\ \mathtt{A}^n\mathtt{N} \mid n \in \mathbb{N}\ \}$$

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome} \}$$

$$S = \{ \texttt{A}^n\texttt{N}^n \mid n \in \mathbb{N} \}$$

More Banana Languages

Let Σ = { A, N }.

Consider the language

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome} \}$$

2b) Let's say we choose S to be { $\mathtt{A}^n \mid n \in \mathbb{N}$ }. Identify the error in each of the following *incorrect* ways of proving that S is a distinguishing set.

i. Consider any two strings $\mathtt{A}^n, \mathtt{A}^m \in S$ where $m \neq n$. Then $\mathtt{A}^n\mathtt{A}^n \in L$ but $\mathtt{A}^{n+1}\mathtt{A}^n \notin L$.

ii. Consider any two strings $\mathtt{A}^n, \mathtt{A}^{n+1} \in S$. Then $\mathtt{A}^n\mathtt{NA}^n \in L$ but $\mathtt{A}^{n+1}\mathtt{NA}^n \notin L$.

iii. Consider any two strings $\mathtt{A}^n, \mathtt{A}^m \in S$ where $m \neq n$. Then $\mathtt{A}^n\mathtt{NA}^n \in L$ but $\mathtt{A}^n\mathtt{NA}^m \notin L$.

Submit on Gradescope!

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$$

$$S = \{ \texttt{A}^n \mid n \in \mathbb{N} \}$$

Consider any two strings $\texttt{A}^n$, $\texttt{A}^m \in S$ where $m \neq n$.
Then $\texttt{A}^n\texttt{A}^n \in L$ but $\texttt{A}^{n+1}\texttt{A}^n \notin L$.

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$$

$$S = \{ \texttt{A}^n \mid n \in \mathbb{N} \}$$

Consider any two strings $\texttt{A}^n$, $\texttt{A}^{n+1} \in S$.
Then $\texttt{A}^n\texttt{NA}^n \in L$ but $\texttt{A}^{n+1}\texttt{NA}^n \notin L$.

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome} \}$$

$$S = \{ \mathtt{A}^n \mid n \in \mathbb{N} \}$$

Consider any two strings $\mathtt{A}^n$, $\mathtt{A}^m \in S$ where $m \neq n$.
Then $\mathtt{A}^n\mathtt{NA}^n \in L$ but $\mathtt{A}^n\mathtt{NA}^m \notin L$.

More Banana Languages

Let Σ = { **A**, **N** }.

Consider the language

$$L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$$

2c) Write up a proof that this language is non-regular. You may use any distinguishing set you'd like, just make sure you haven't fallen into any of the common errors we saw in parts a) and b).

Submit on Gradescope!

Theorem: The language $L = \{ w \in \Sigma^* \mid w \text{ is a palindrome } \}$ is non-regular.

Proof: Let $S = \{ \mathtt{A}^n \mid n \in \mathbb{N} \}$. We will prove that S is infinite and that S is a distinguishing set for L.

To see that S is infinite, note that S contains one string for each natural number.

To see that S is a distinguishing set for L, consider any strings $\mathtt{A}^m, \mathtt{A}^n \in S$ where $m \neq n$. Note that $\mathtt{A}^m\mathtt{NA}^m \in L$ but $\mathtt{A}^n\mathtt{NA}^m \notin L$. Therefore, we see that $\mathtt{A}^m \not\equiv_L \mathtt{A}^n$, as required.

Since S is infinite and is a distinguishing set for L, by the Myhill-Nerode theorem we see that L is not regular. ■

Thanks for Calling In!

Stay safe, stay healthy,
and have a good week!

See you next time.