Introduction to NLP

SCHOOL OF INFOCOMM

Need to know

- **Trainers**
  - Ho Chee Wai (ho_chee_wai@rp.edu.sg)
  - Tan Poh Keam (tan_poh_keam@rp.edu.sg)

- **Assessment**
  - Written assignment, due 13 Mar 2020

Learning Objectives

- Describe the principles, concepts and usage of NLP
- Perform text pre-processing activities
- Discuss and use suitable methods for feature engineering of text
- Discuss and use suitable approaches to analysis, classify and compare text
- Explain how deep-learning techniques are used for text generation tasks
- Consume Machine Learning services to build NLP applications
- Use Azure Bot Services to build conversation systems

Where We’re Headed

When	Topics
27 Feb	NLP Introduction Regular Expression Text Pre-processing
28 Feb	Features Engineering Document Classification
02 Mar	Topic Modelling
03 Mar	Word Embedding Text Generation
04 Mar	Google Language Services
05 Mar	Azure Bot Services

What is Natural Language Processing (NLP)

By “natural language” we mean a language that is used for everyday communication by humans.

Two main components:

- Natural Language Understanding (NLU)
- Natural Language Generation (NLG)

NLP is an Intersection of several fields

- Computer Science
- Artificial Intelligence
- Linguistics

NLP is AI-Complete

- Requires all types of knowledge humans possess → It’s hard!

Ambiguity of Language

- Synonymy – different words, same meaning
- Polysemy – same word, different meaning
- Text and speech are unstructured data
- No fixed structure – Sentence format
- No fixed schema – Grammar
- Misspell, slang, abbreviations

History of NLP

- NLP has been through (at least) 3 major eras:
  - 1950s-1980s: Linguistics Methods and Rules
  - 1980s-Now: Statistical + Machine Learning Methods
  - Now - ???: Deep Learning
- You’re right near the start of a paradigm shift!

1950s - 1970s: Linguistics & Rules

- Approach focused on:
  - Linguistics: Grammar rules, sentence structure parsing
  - Handwritten Rules: Huge sets of logical (if/else) statements
  - Phase structure grammar : conversion of sentences into forms computers can understand
- Problems:
  - Too complex to maintain
  - Cannot scale
  - Cannot generalize

1980s - Now: Statistical / Machine Learning Methods

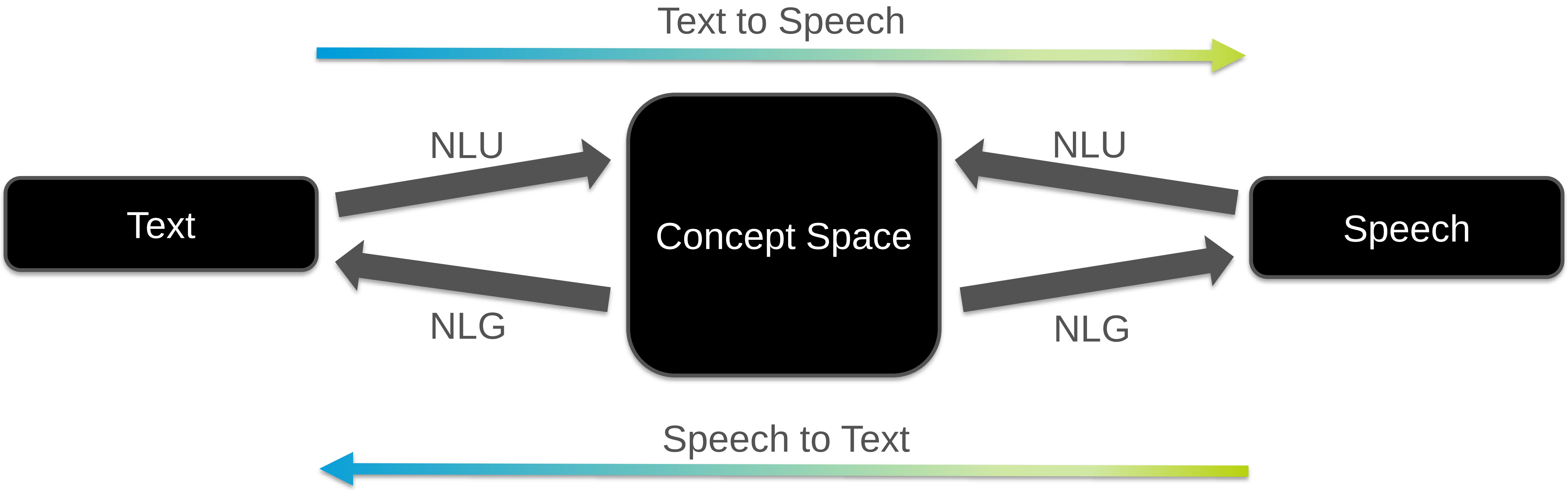
- Approach shifted from linguistics to data-driven
- Increasing computational power and ease to access of text
  - First web page (1991) → discussion forums, blogs, news portal
  - Digital archives
- NLP starts using statistical and probabilistic models
  - Data minding → text mining
- Generic machine learning algorithms applied to NLP tasks
  - Sentiment analysis using logistic regression
  - Language models with Markov models

Now and future: Deep Learning

- More advances in computing power with parallelization (GPU)
- Availability of large datasets becomes the norm
- CNN
  - Learnt word representation with finite dimensions
  - Capture semantic and relationships among words
- RNN / LSTM
  - Allows sequential processing and learning of text
  - Application into machine translation tasks and questions/answering systems
- Attention-based model
  - A way to place various degree of focus (attention) on different part of the text
  - Break-through in machine translation and text generation tasks

NLP: Speech vs Text

- Data source can refer to written text or speech
- Goal of both is the same: translate raw data (text or speech) into underlying concepts (NLU) then possibly into the other form (NLG)



NLU Applications

- Classification
- Natural Language Search
- Document Recommendation
- Topic Modeling
- Language Identification
- Intent matching

NLU Application: Document Classification

- Classify documents (discrete collections of text) - into categories
  - Classify emails as spam vs not spam
  - Classify product reviews as positive vs negative
  - Assign labels to documents

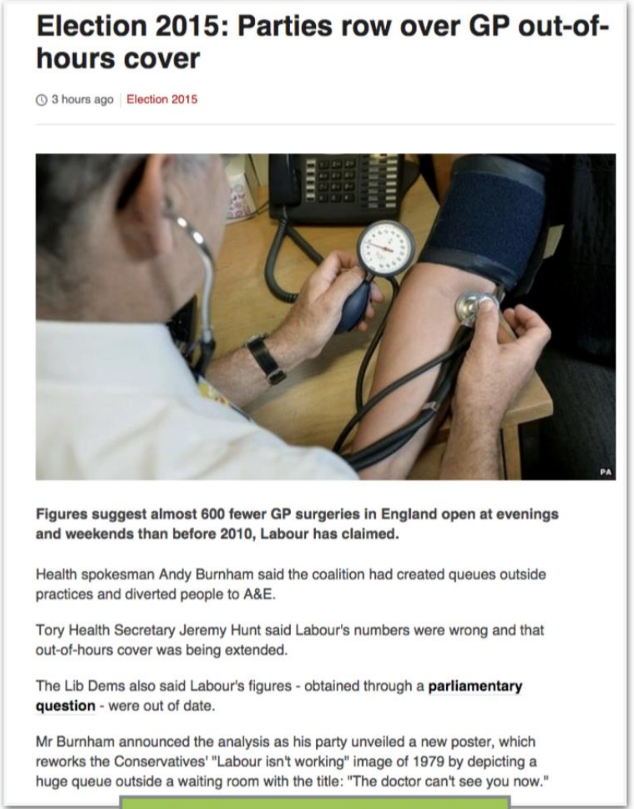
NLU Application: Document Recommendation

- Choosing the most relevant document based on some information or "finger print":
  - Show most relevant webpages based on query to search engine
  - Recommend news articles based on past articles liked or read
  - Recommend restaurants based on restaurant reviews

NLU Application: Topic Modeling

Breaking a set of documents into topics at the word level

- See how prevalence of certain topics covered in a magazine changes over time
- Find documents belonging to a certain topic



Election 2015: Parties row over GP out-of-hours cover

3 hours ago | Election 2015

Figures suggest almost 600 fewer GP surgeries in England open at evenings and weekends than before 2010, Labour has claimed.

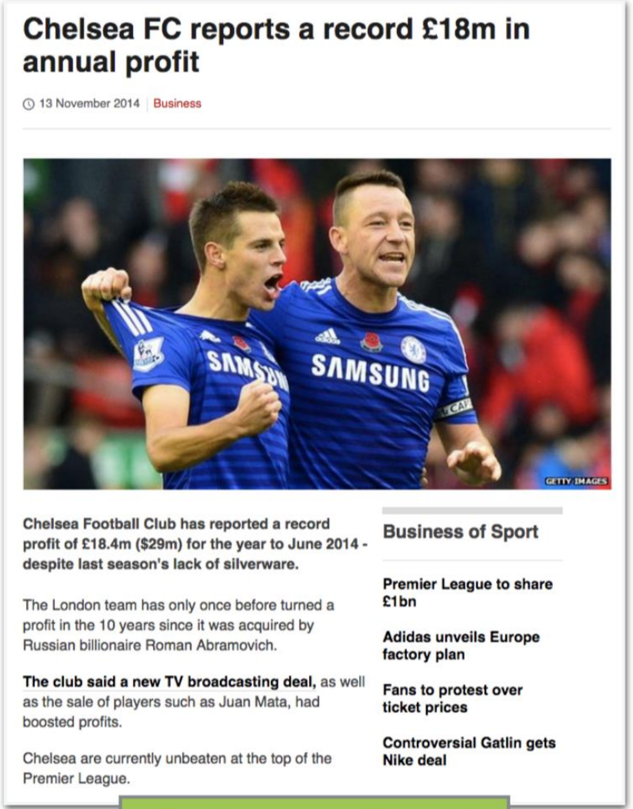
Health spokesman Andy Burnham said the coalition had created queues outside practices and diverted people to A&E.

Tory Health Secretary Jeremy Hunt said Labour's numbers were wrong and that out-of-hours cover was being extended.

The Lib Dems also said Labour's figures - obtained through a **parliamentary question** - were out of date.

Mr Burnham announced the analysis as his party unveiled a new poster, which reworks the Conservatives' "Labour isn't working" image of 1979 by depicting a huge queue outside a waiting room with the title: "The doctor can't see you now."

Politics or Health?



Chelsea FC reports a record £18m in annual profit

13 November 2014 | Business

Chelsea Football Club has reported a record profit of £18.4m ($29m) for the year to June 2014 - despite last season's lack of silverware.

The London team has only once before turned a profit in the 10 years since it was acquired by Russian billionaire Roman Abramovich.

The club said a new TV broadcasting deal, as well as the sale of players such as Juan Mata, had boosted profits.

Chelsea are currently unbeaten at the top of the Premier League.

Business of Sport

Premier League to share £1bn

Adidas unveils Europe factory plan

Fans to protest over ticket prices

Controversial Gatlin gets Nike deal

Business or Sport?

NLU Application: Intent Matching

Understanding that there are many ways to say or ask for the same thing

- Use in conversation systems

May I know the opening time for the store?

Can I come to shop now?

What time do you close?

Is the mall still open?

I can come at 8 pm, is it okay ?

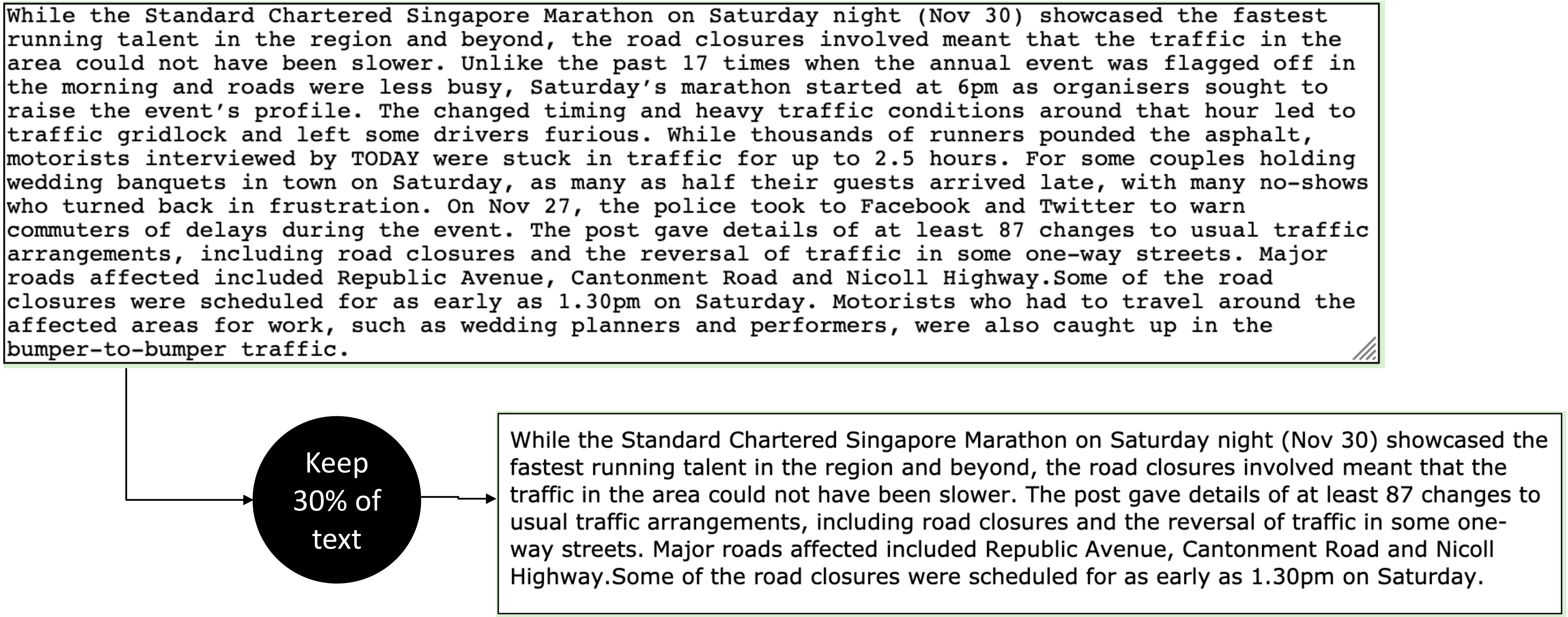
NLG Applications

- Machine Translation
- Document Summarization
- Text Generation
- Question Answering
- Image Captioning

Notice NLU is almost a prerequisite for NLG

NLG Application: Document Summarization

Read less



While the Standard Chartered Singapore Marathon on Saturday night (Nov 30) showcased the fastest running talent in the region and beyond, the road closures involved meant that the traffic in the area could not have been slower. Unlike the past 17 times when the annual event was flagged off in the morning and roads were less busy, Saturday’s marathon started at 6pm as organisers sought to raise the event’s profile. The changed timing and heavy traffic conditions around that hour led to traffic gridlock and left some drivers furious. While thousands of runners pounded the asphalt, motorists interviewed by TODAY were stuck in traffic for up to 2.5 hours. For some couples holding wedding banquets in town on Saturday, as many as half their guests arrived late, with many no-shows who turned back in frustration. On Nov 27, the police took to Facebook and Twitter to warn commuters of delays during the event. The post gave details of at least 87 changes to usual traffic arrangements, including road closures and the reversal of traffic in some one-way streets. Major roads affected included Republic Avenue, Cantonment Road and Nicoll Highway.Some of the road closures were scheduled for as early as 1.30pm on Saturday. Motorists who had to travel around the affected areas for work, such as wedding planners and performers, were also caught up in the bumper-to-bumper traffic.

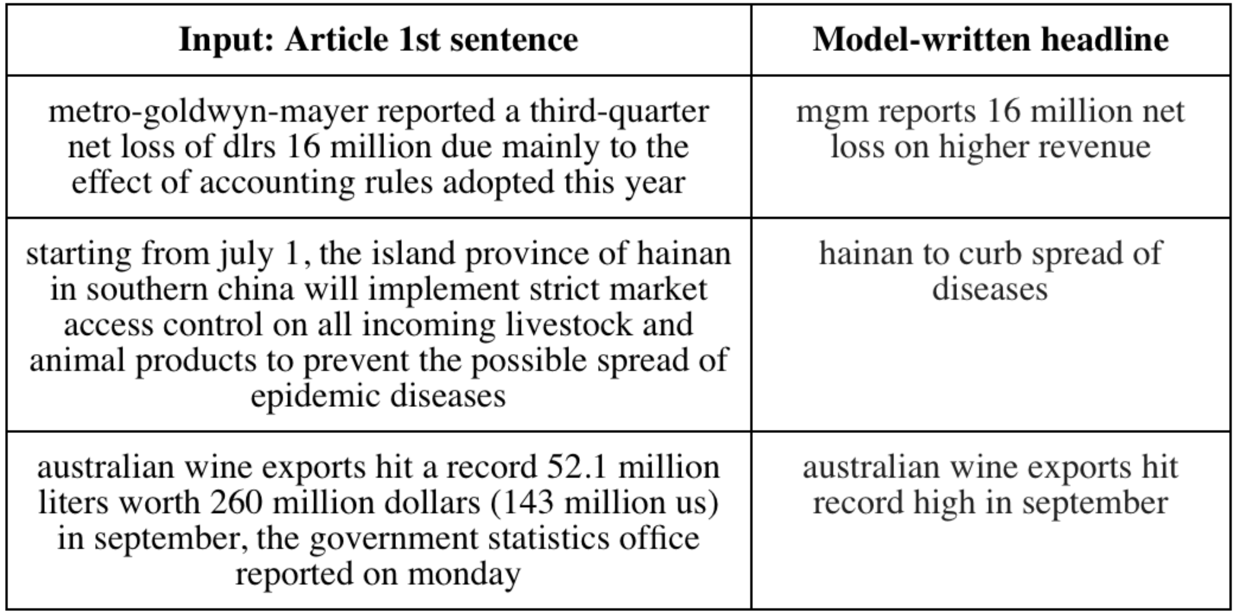
While the Standard Chartered Singapore Marathon on Saturday night (Nov 30) showcased the fastest running talent in the region and beyond, the road closures involved meant that the traffic in the area could not have been slower. The post gave details of at least 87 changes to usual traffic arrangements, including road closures and the reversal of traffic in some one-way streets. Major roads affected included Republic Avenue, Cantonment Road and Nicoll Highway.Some of the road closures were scheduled for as early as 1.30pm on Saturday.

Demo: https://www.textcompactor.com/

NLG Application: Document Summarization

Automatically generate text summaries of documents

- generate **headlines of news articles**

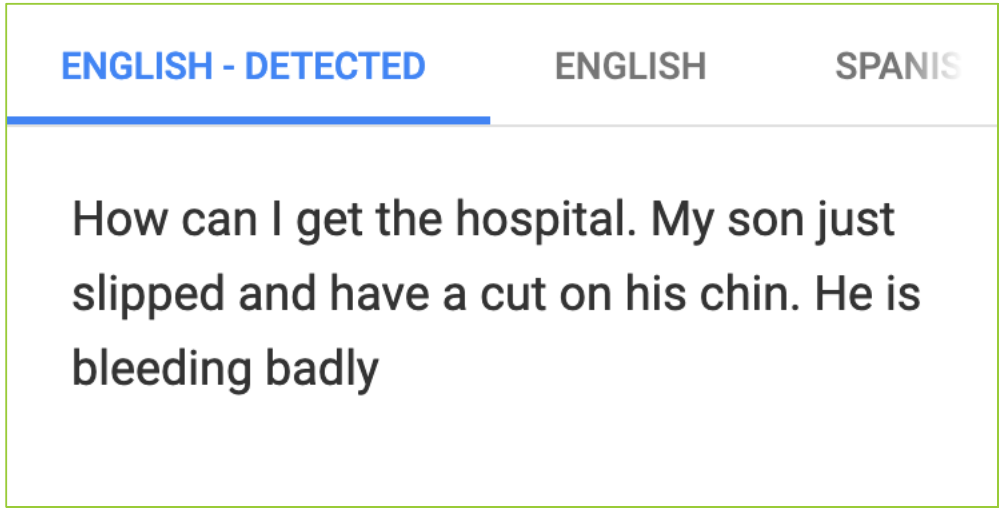


Input: Article 1st sentence	Model-written headline
metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year	mgm reports 16 million net loss on higher revenue
starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases	hainan to curb spread of diseases
australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday	australian wine exports hit record high in september

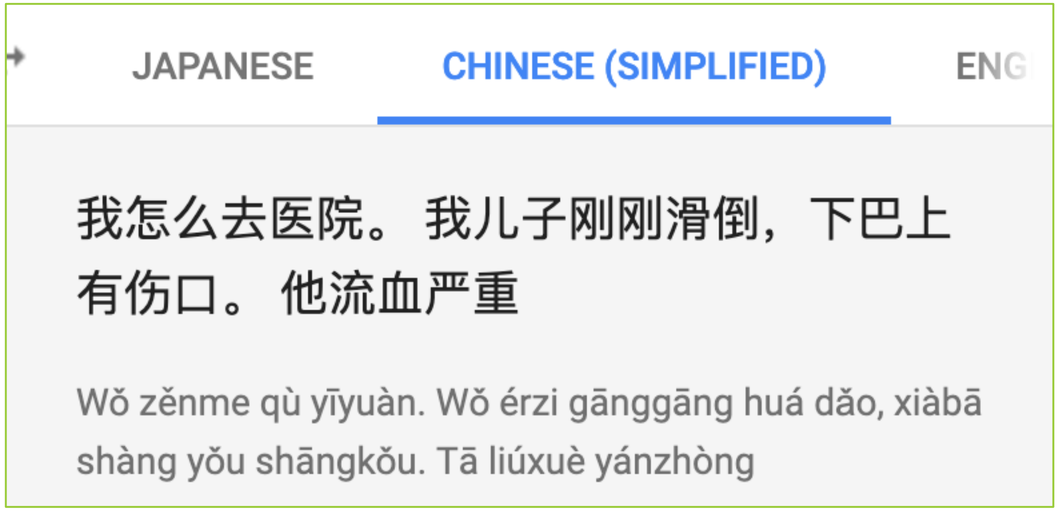
Source: https://ai.googleblog.com/2016/08/text-summarization-with-tensorflow.html

NLG Application: Machine Translation

Automatically translate text between language

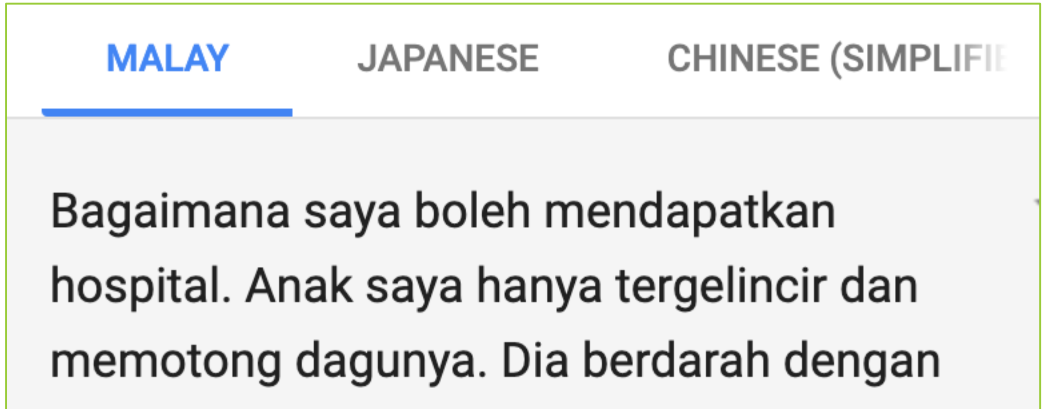


How can I get the hospital. My son just slipped and have a cut on his chin. He is bleeding badly



我怎么去医院。 我儿子刚刚滑倒，下巴上有伤口。 他流血严重

Wǒ zěnme qù yīyuàn. Wǒ érzi gānggāng huá dǎo, xiàbā shàng yǒu shāngkǒu. Tā liúxuè yánzhòng

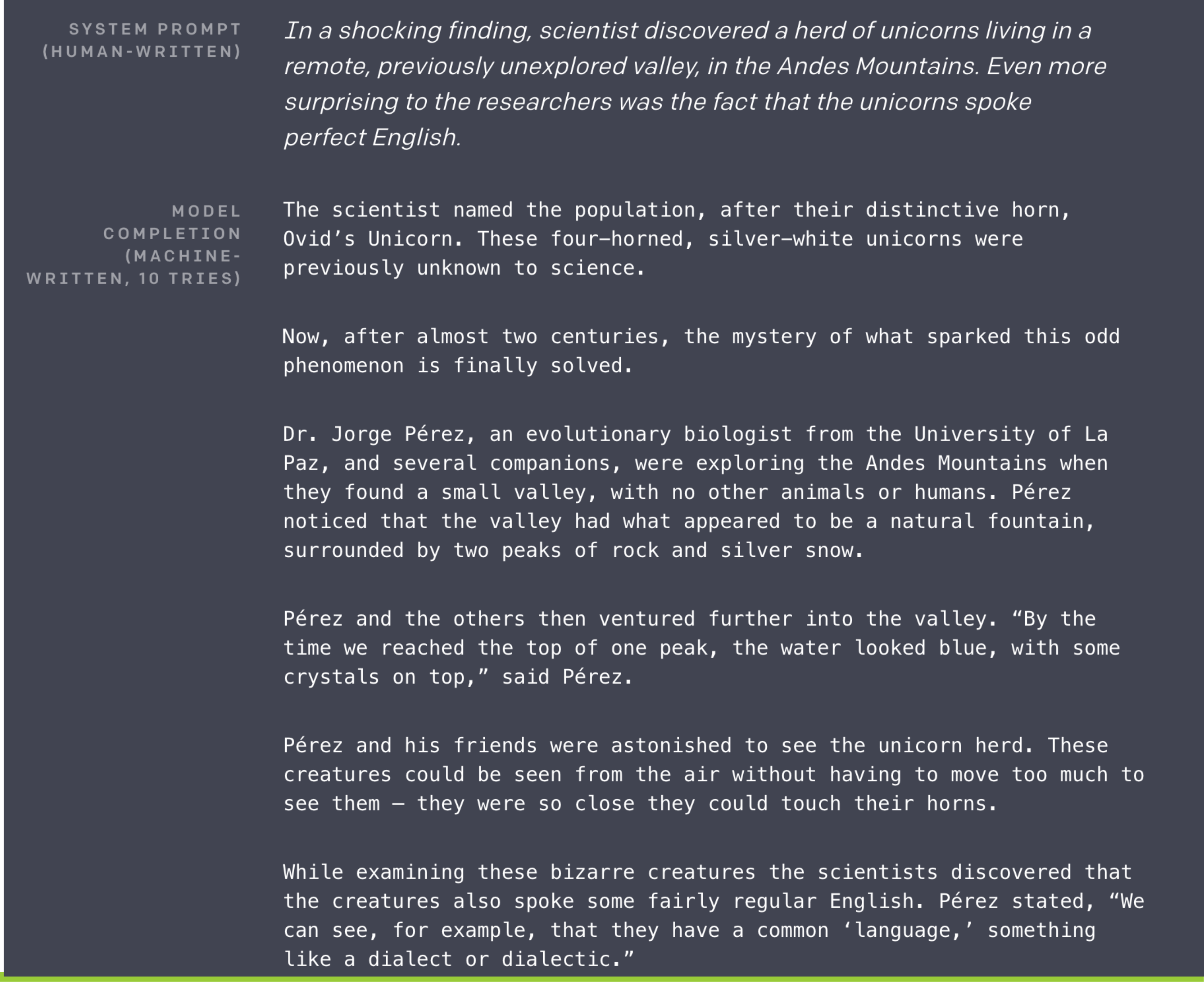


Bagaimana saya boleh mendapatkan hospital. Anak saya hanya tergelincir dan memotong dagunya. Dia berdarah dengan

Example from Google Translate

Reading : https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

NLG Application: Text Generation



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

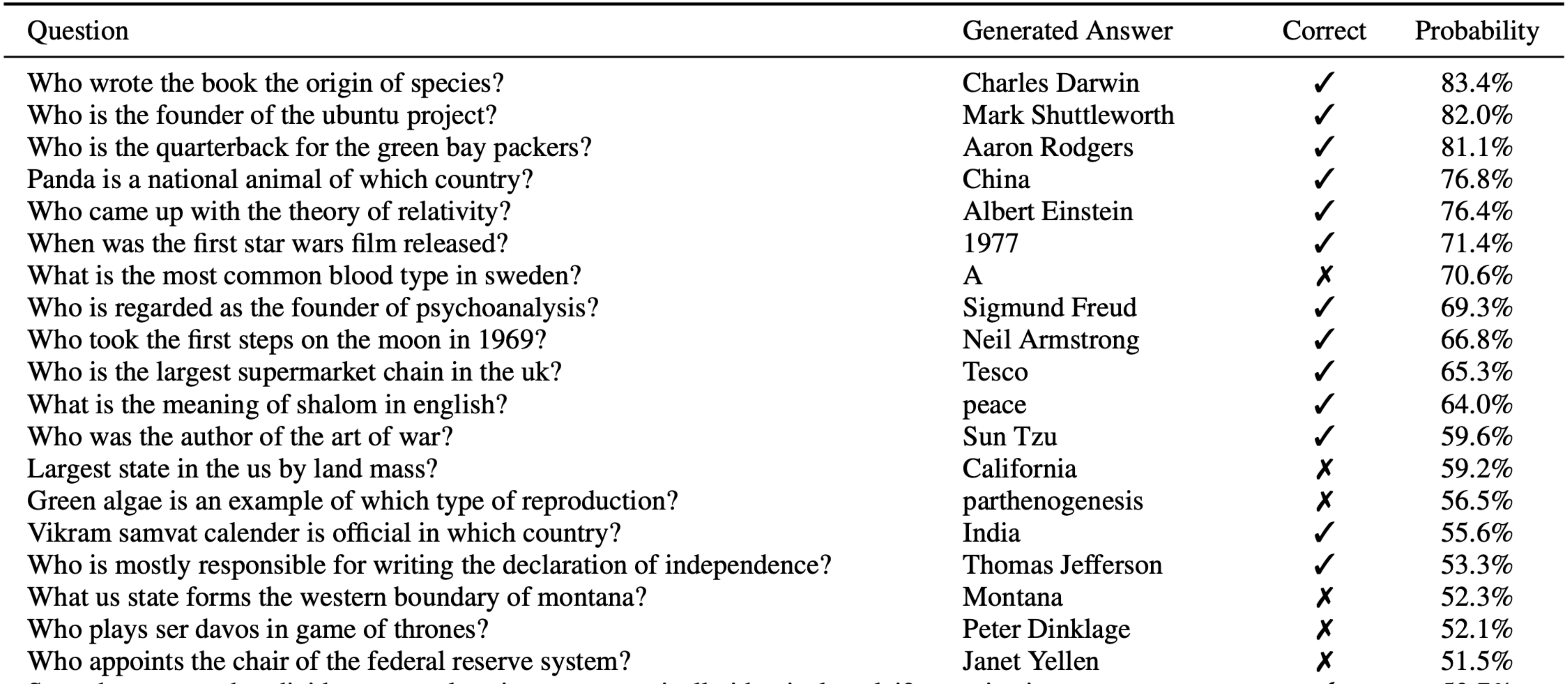
Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, “We can see, for example, that they have a common ‘language,’ something like a dialect or dialectic.”

Source: https://openai.com/blog/better-language-models/

NLG Application: Question Answering



Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%

Source: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

NLG Application: Question Answering

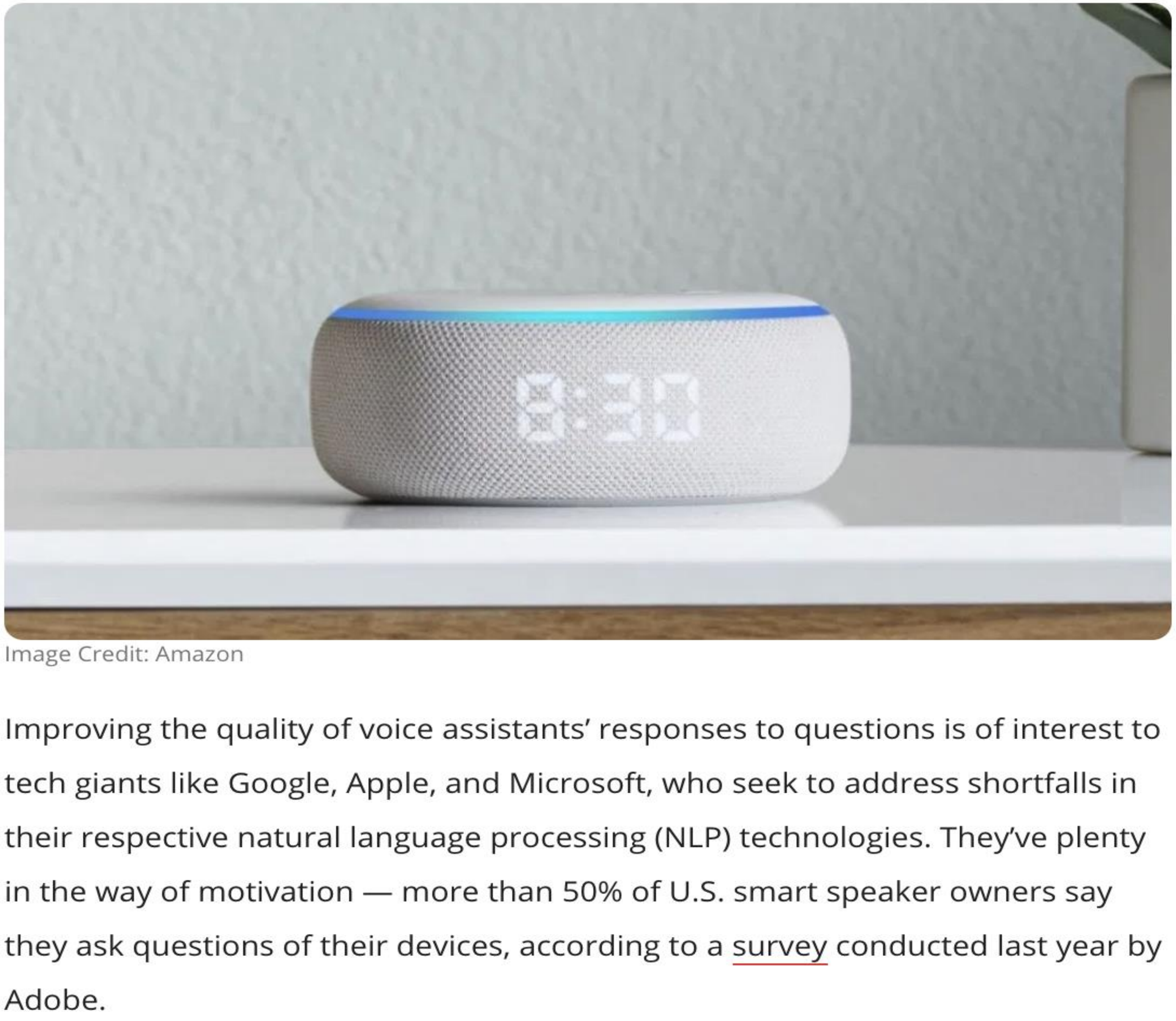
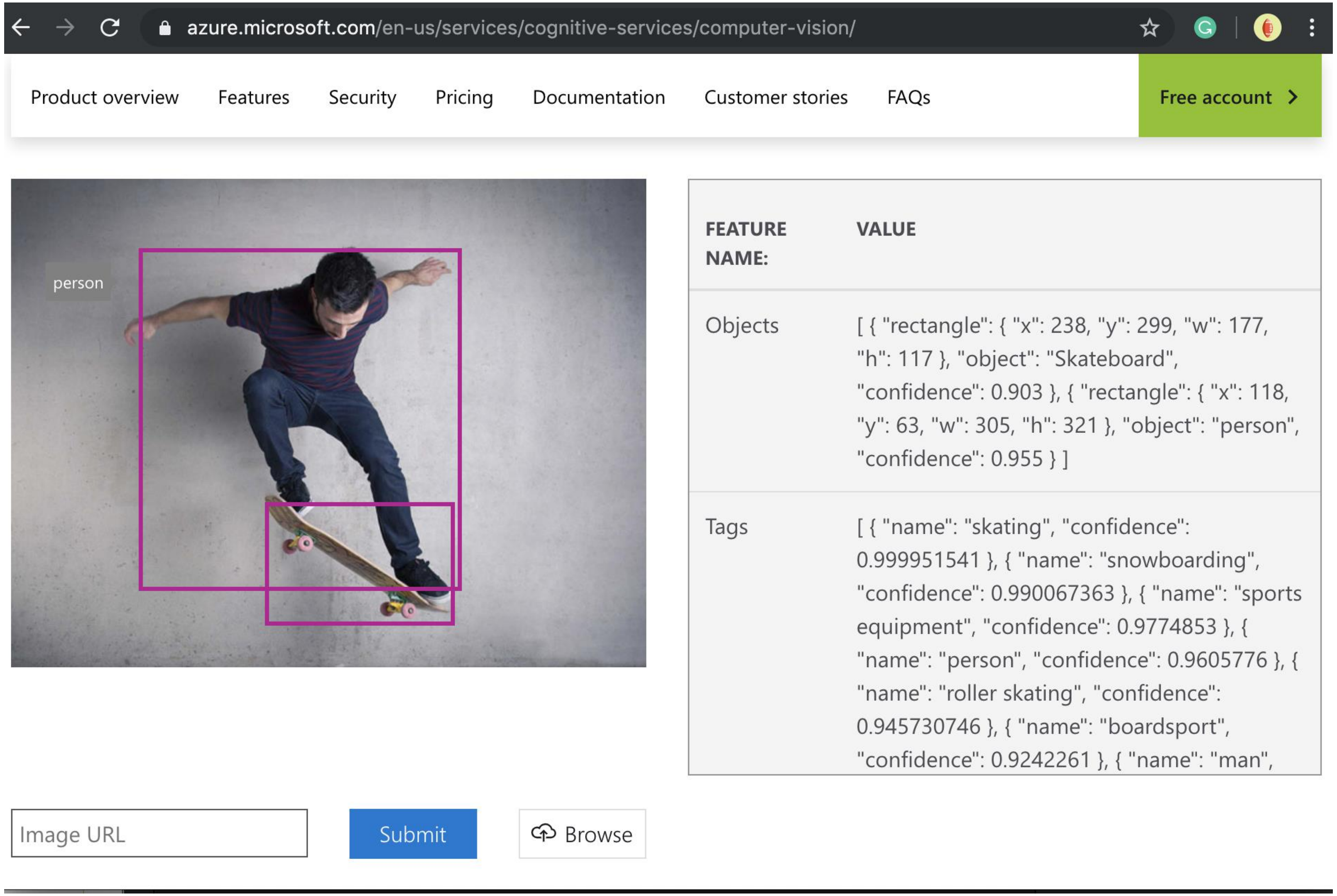


Image Credit: Amazon

Improving the quality of voice assistants' responses to questions is of interest to tech giants like Google, Apple, and Microsoft, who seek to address shortfalls in their respective natural language processing (NLP) technologies. They've plenty in the way of motivation — more than 50% of U.S. smart speaker owners say they ask questions of their devices, according to a survey conducted last year by Adobe.

Source: https://venturebeat.com/2019/11/25/amazon-researchers-use-nlp-data-set-to-improve-alexas-answers/

NLG Application: Image Captioning

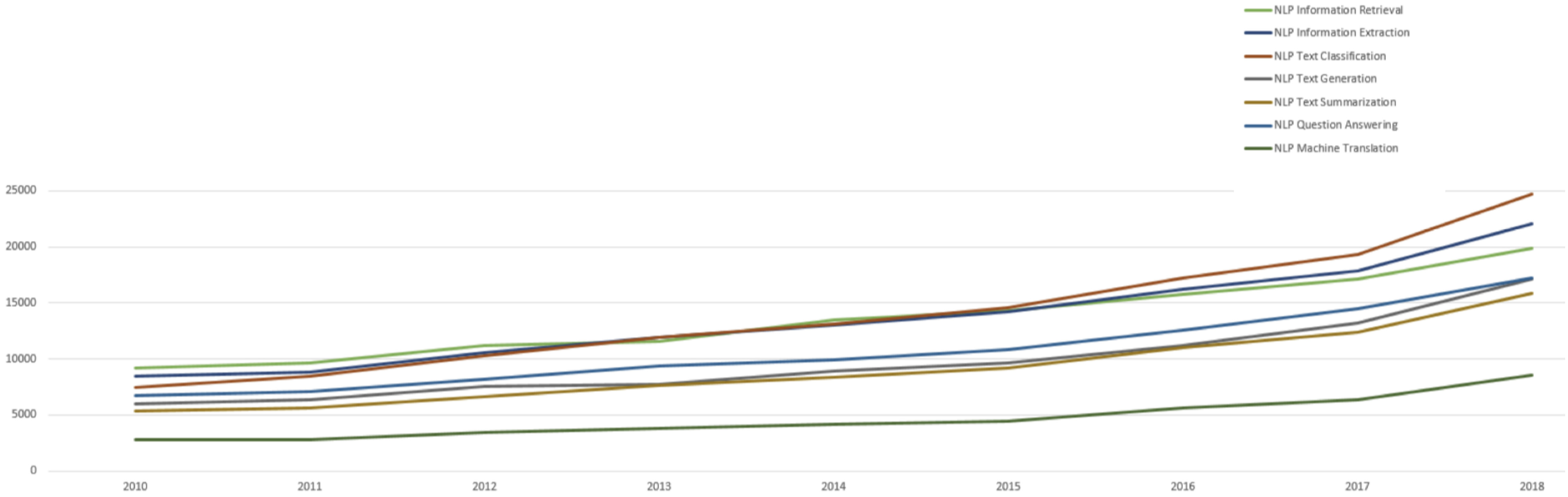


FEATURE NAME:	VALUE
Objects	[{ "rectangle": { "x": 238, "y": 299, "w": 177, "h": 117 }, "object": "Skateboard", "confidence": 0.903 }, { "rectangle": { "x": 118, "y": 63, "w": 305, "h": 321 }, "object": "person", "confidence": 0.955 }]
Tags	[{ "name": "skating", "confidence": 0.999951541 }, { "name": "snowboarding", "confidence": 0.990067363 }, { "name": "sports equipment", "confidence": 0.9774853 }, { "name": "person", "confidence": 0.9605776 }, { "name": "roller skating", "confidence": 0.945730746 }, { "name": "boardsport", "confidence": 0.9242261 }, { "name": "man",

Demo:

- https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

Publications on Applied Areas of NLP



Credit: A Survey of the Usages of Deep Learning for Natural Language Processing , Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita (https://arxiv.org/pdf/1807.10854.pdf)

NLP Benchmark

The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analysing natural language understanding systems

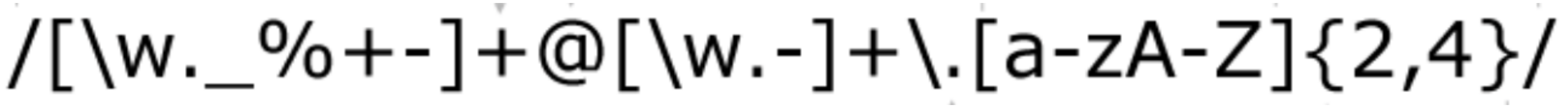
- Language understanding tasks built on established existing datasets and selected to cover a diverse range **of dataset sizes, text genres, and degrees of difficulty**

- Diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of **linguistic phenomena** found in natural language

- Leaderboard for **tracking performance** on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set (https://gluebenchmark.com/leaderboard)

Further reading: https://openreview.net/pdf?id=rJ4km2R5t7

Regular Expression



```
/[\w._%+-]+@[\w.-]+\.[a-zA-Z]{2,4}/
```

What are Regular Expressions (“RegEx”)?

- To understand text, we need mechanisms for analysing its structure and identifying particular types of constituents

- Regular expressions is the starting point

- A regular expression is a pattern which can match one or more characters strings

- One common outcome of using regular expression is to find some data in text and take actions

Use Cases for RegEx

Parsing documents with structured layout

- HTML, find all text within <p> tags
- Find body sections of emails

Handling personal identifiable information (PII)

- Is there something in the text that looks like NRIC or password?

  e.g. Series of digits + character

Find logic "cut point" of huge corpus of document

Simple examples of regular expressions (1)

Matching characters

- app - match text containing **app,** such as pineapple, apple, applications
- 1800 – Match to the text 1800

Character sets [   ]

- [amk] – Match to either a , m or k
- [a-z] – Match any alphabet from a to z
- [7-9] – Match text containing characters with 7, 8, 9
- [A-Z0-9] – Match characters A to Z or 0 to 9
- \d – numbers class (short cut for [0-9])
- \w – alpha numberic character class (e.g. digits, letters but exclude special symbols)

Simple examples of regular expressions (2)

OR |

- apple | orange | pineapple – match to either apple, orange or pineapple

Groups ( )

- Group symbols in the parathesis
- (Mon | Tues | Wednes)day – match to Monday, Tuesday or Wednesday

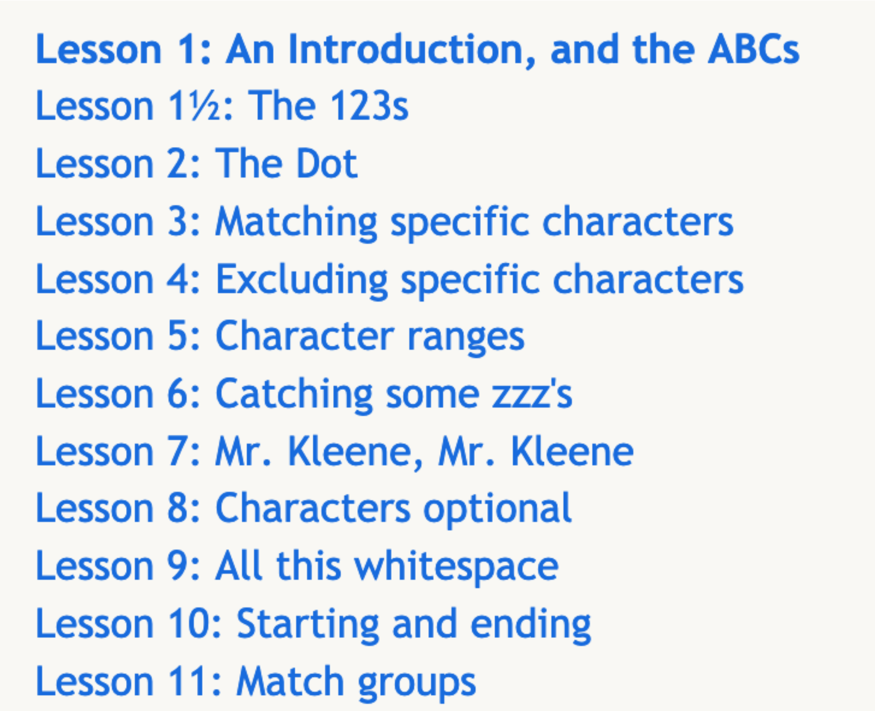
Quantifiers. *. ? + {n , m }

- * : match zero or more of the previous expression
- ? : Optional match
- + : match 1 or more of previous expression
- {n, m} : match n to m occurence of previous expression

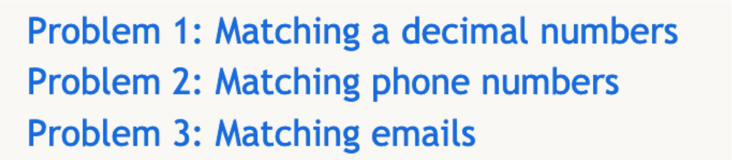
Exercise 1

Appreciation of ReEx syntax

Visit https://regexone.com/



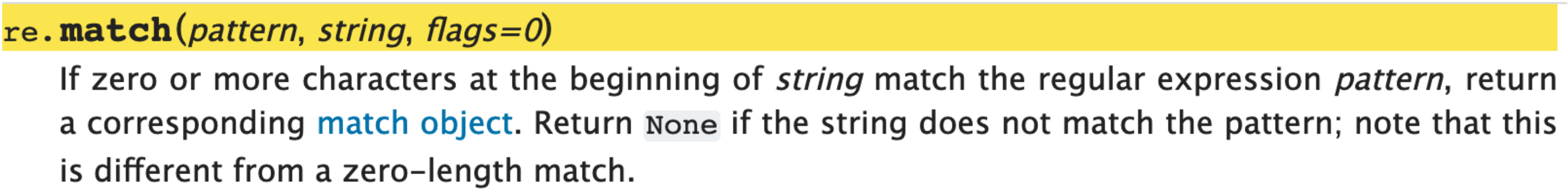
Lesson 1: An Introduction, and the ABCs
Lesson 1½: The 123s
Lesson 2: The Dot
Lesson 3: Matching specific characters
Lesson 4: Excluding specific characters
Lesson 5: Character ranges
Lesson 6: Catching some zzz's
Lesson 7: Mr. Kleene, Mr. Kleene
Lesson 8: Characters optional
Lesson 9: All this whitespace
Lesson 10: Starting and ending
Lesson 11: Match groups



Problem 1: Matching a decimal numbers
Problem 2: Matching phone numbers
Problem 3: Matching emails

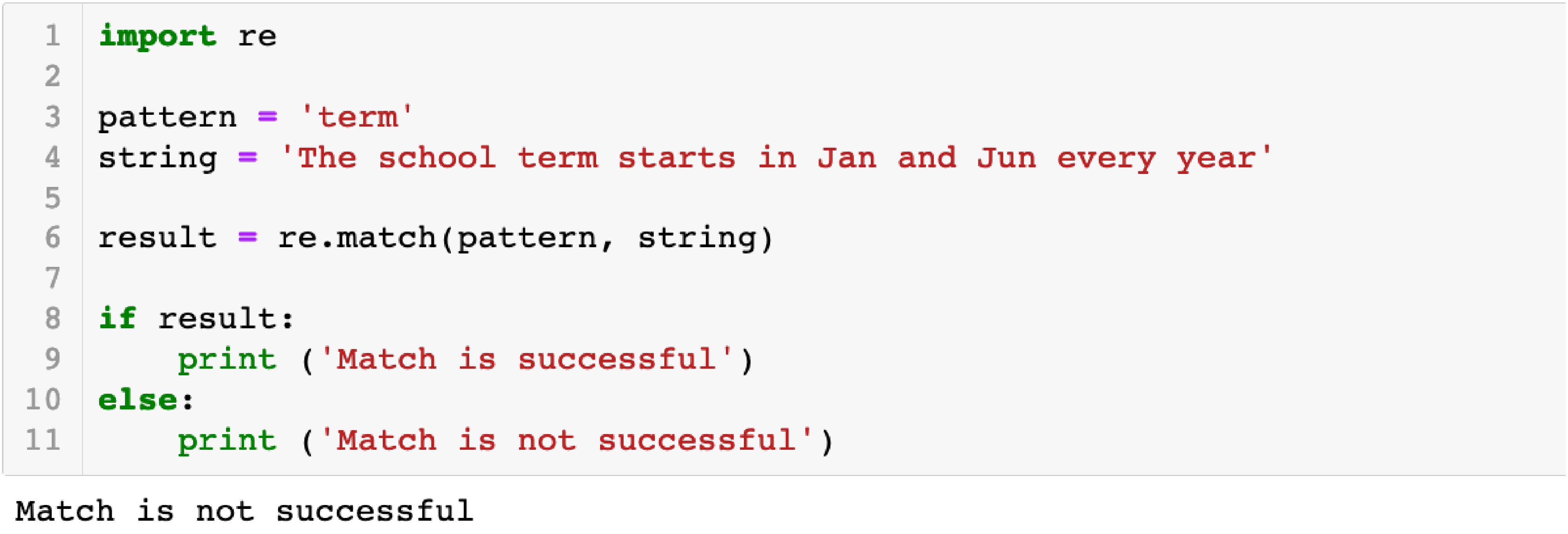
Instructions:
Work in pairs.
Complete Lesson 1 to Lesson 12.
Complete Problem 1 to Problem 3.

Regex Methods in Python



re.**match**(*pattern*, *string*, *flags=0*)

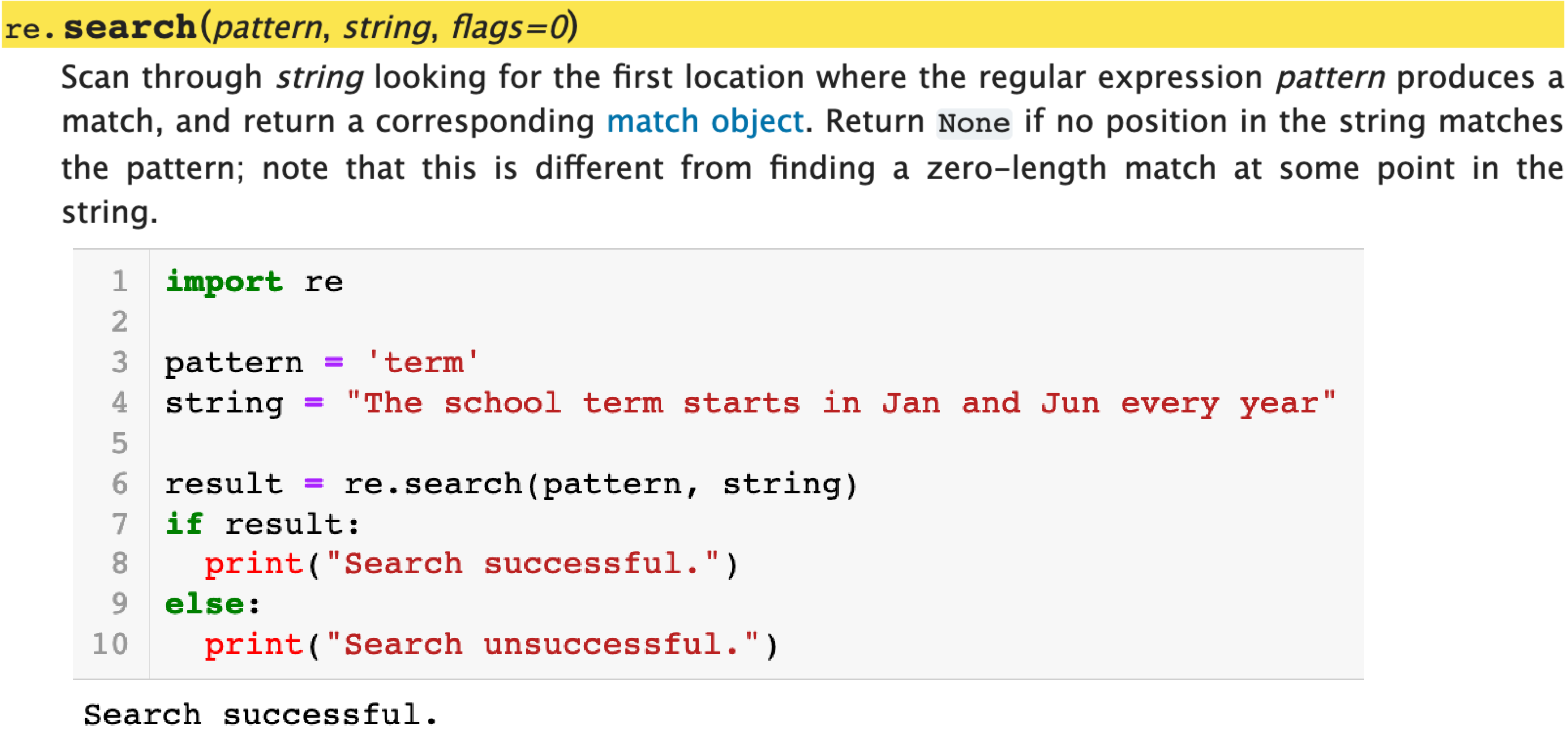
If zero or more characters at the beginning of *string* match the regular expression *pattern*, return a corresponding match object. Return `None` if the string does not match the pattern; note that this is different from a zero-length match.



```python
import re

pattern = 'term'
string = 'The school term starts in Jan and Jun every year'

result = re.match(pattern, string)

if result:
    print ('Match is successful')
else:
    print ('Match is not successful')
```

```
Match is not successful
```

Regex Methods in Python



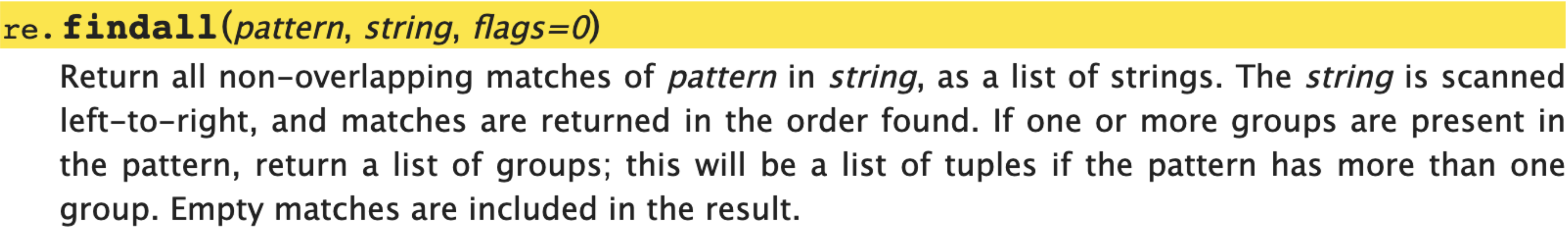
re.**search**(*pattern*, *string*, *flags=0*)

Scan through *string* looking for the first location where the regular expression *pattern* produces a match, and return a corresponding match object. Return `None` if no position in the string matches the pattern; note that this is different from finding a zero-length match at some point in the string.

```python
import re

pattern = 'term'
string = "The school term starts in Jan and Jun every year"

result = re.search(pattern, string)
if result:
  print("Search successful.")
else:
  print("Search unsuccessful.")
```

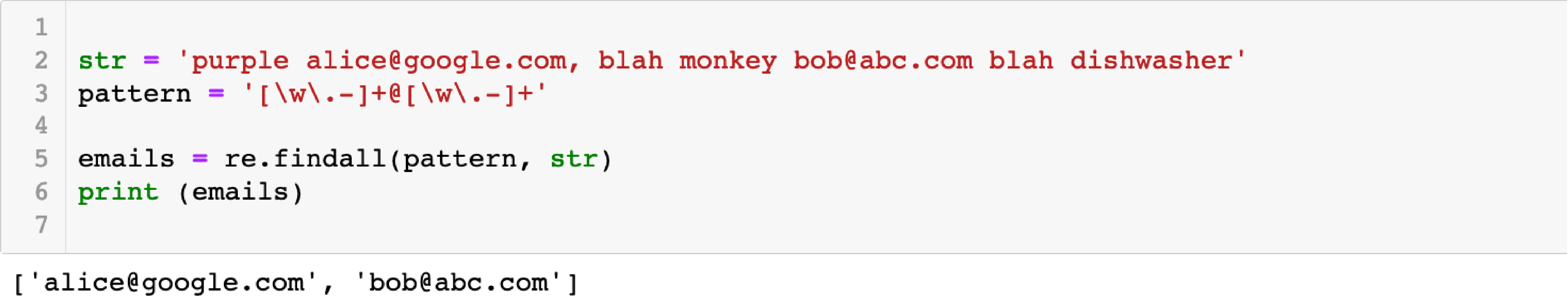
```
Search successful.
```

Regex Methods in Python



`re.findall(pattern, string, flags=0)`

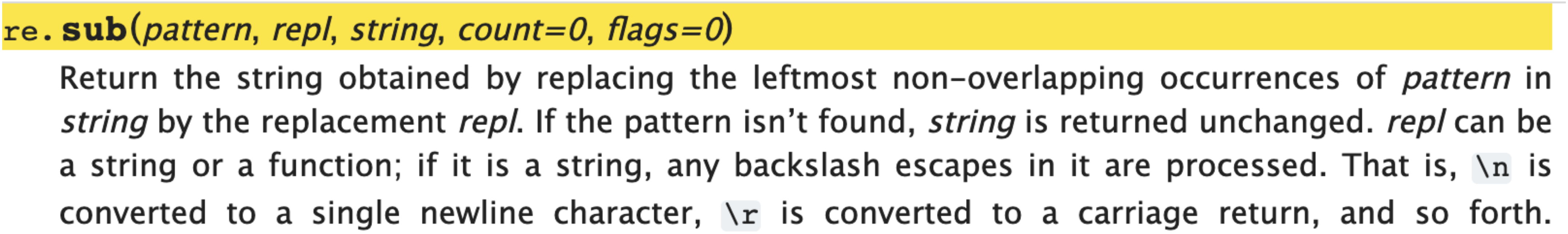
Return all non-overlapping matches of *pattern* in *string*, as a list of strings. The *string* is scanned left-to-right, and matches are returned in the order found. If one or more groups are present in the pattern, return a list of groups; this will be a list of tuples if the pattern has more than one group. Empty matches are included in the result.



```

str = 'purple alice@google.com, blah monkey bob@abc.com blah dishwasher'
pattern = '[\w\.-]+@[\w\.-]+'

emails = re.findall(pattern, str)
print (emails)

```

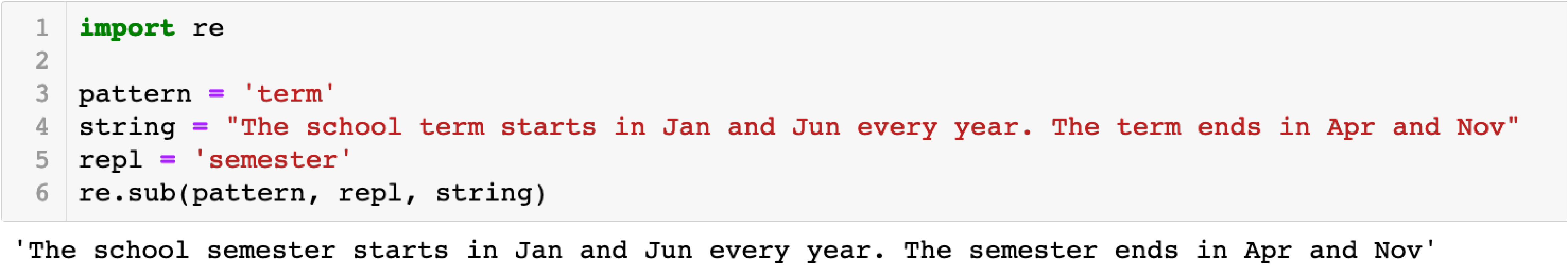
```
['alice@google.com', 'bob@abc.com']
```

Regex Methods in Python



re.**sub**(*pattern*, *repl*, *string*, *count=0*, *flags=0*)

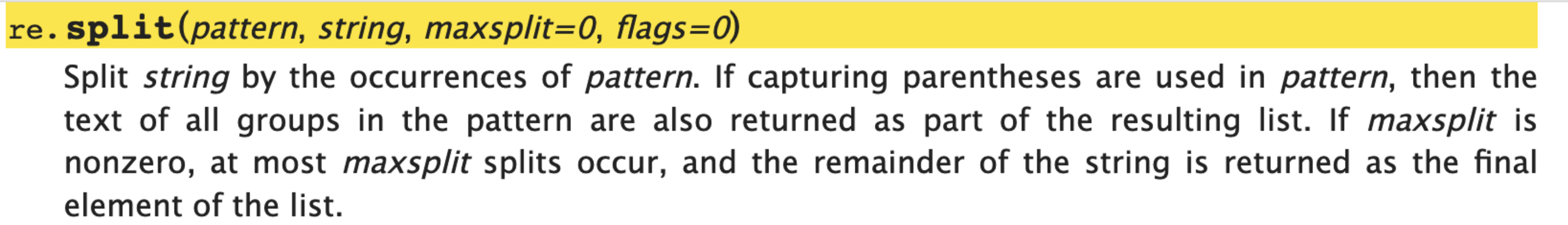
Return the string obtained by replacing the leftmost non-overlapping occurrences of *pattern* in *string* by the replacement *repl*. If the pattern isn't found, *string* is returned unchanged. *repl* can be a string or a function; if it is a string, any backslash escapes in it are processed. That is, `\n` is converted to a single newline character, `\r` is converted to a carriage return, and so forth.



```python
import re

pattern = 'term'
string = "The school term starts in Jan and Jun every year. The term ends in Apr and Nov"
repl = 'semester'
re.sub(pattern, repl, string)
```

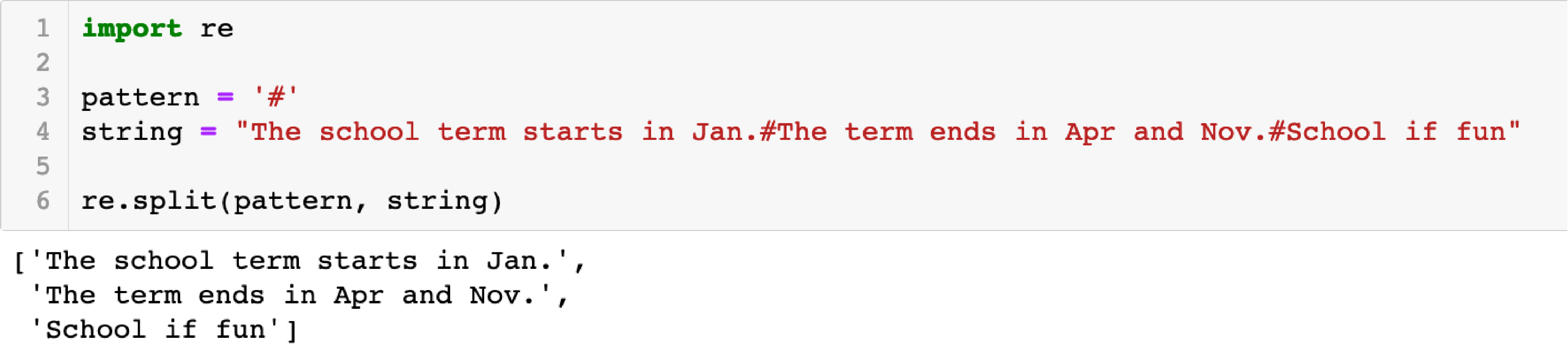
```
'The school semester starts in Jan and Jun every year. The semester ends in Apr and Nov'
```

Regex Methods in Python



`re.`**`split`**(*pattern*, *string*, *maxsplit=0*, *flags=0*)

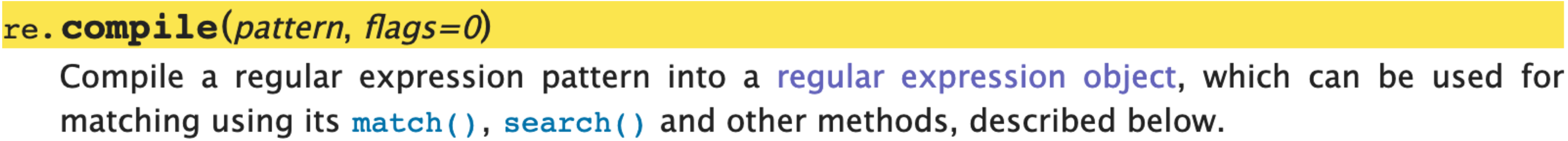
Split *string* by the occurrences of *pattern*. If capturing parentheses are used in *pattern*, then the text of all groups in the pattern are also returned as part of the resulting list. If *maxsplit* is nonzero, at most *maxsplit* splits occur, and the remainder of the string is returned as the final element of the list.



```python
import re

pattern = '#'
string = "The school term starts in Jan.#The term ends in Apr and Nov.#School if fun"

re.split(pattern, string)
```

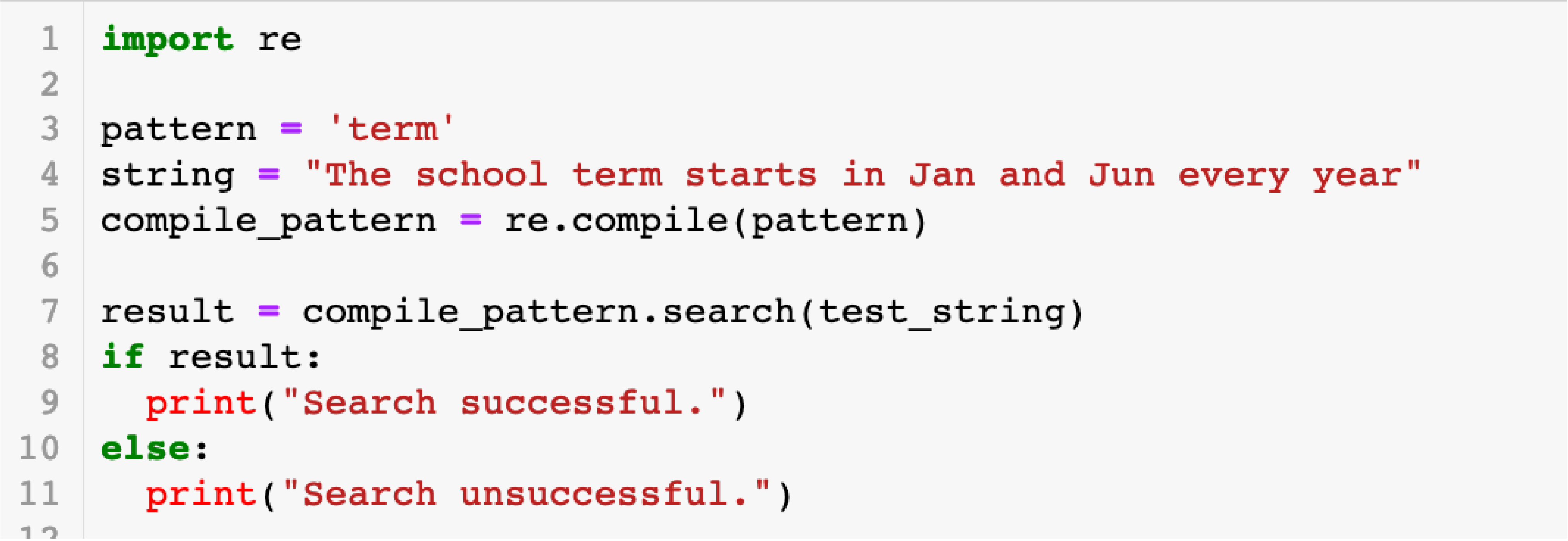
```
['The school term starts in Jan.',
 'The term ends in Apr and Nov.',
 'School if fun']
```

Alternatives: Using RE objects



re.compile(*pattern*, *flags=0*)

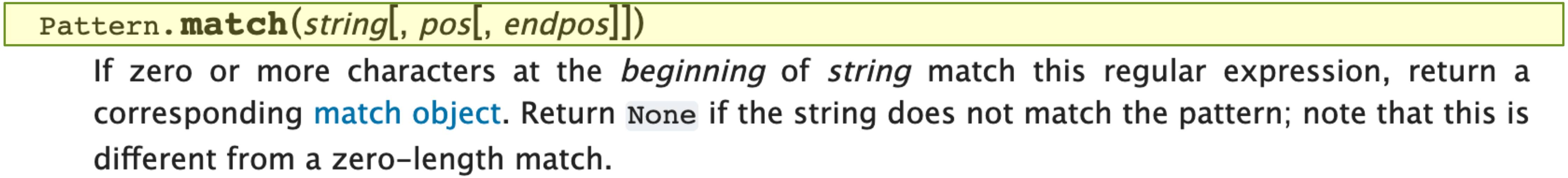
Compile a regular expression pattern into a regular expression object, which can be used for matching using its `match()`, `search()` and other methods, described below.



```python
import re

pattern = 'term'
string = "The school term starts in Jan and Jun every year"
compile_pattern = re.compile(pattern)

result = compile_pattern.search(test_string)
if result:
  print("Search successful.")
else:
  print("Search unsuccessful.")
```

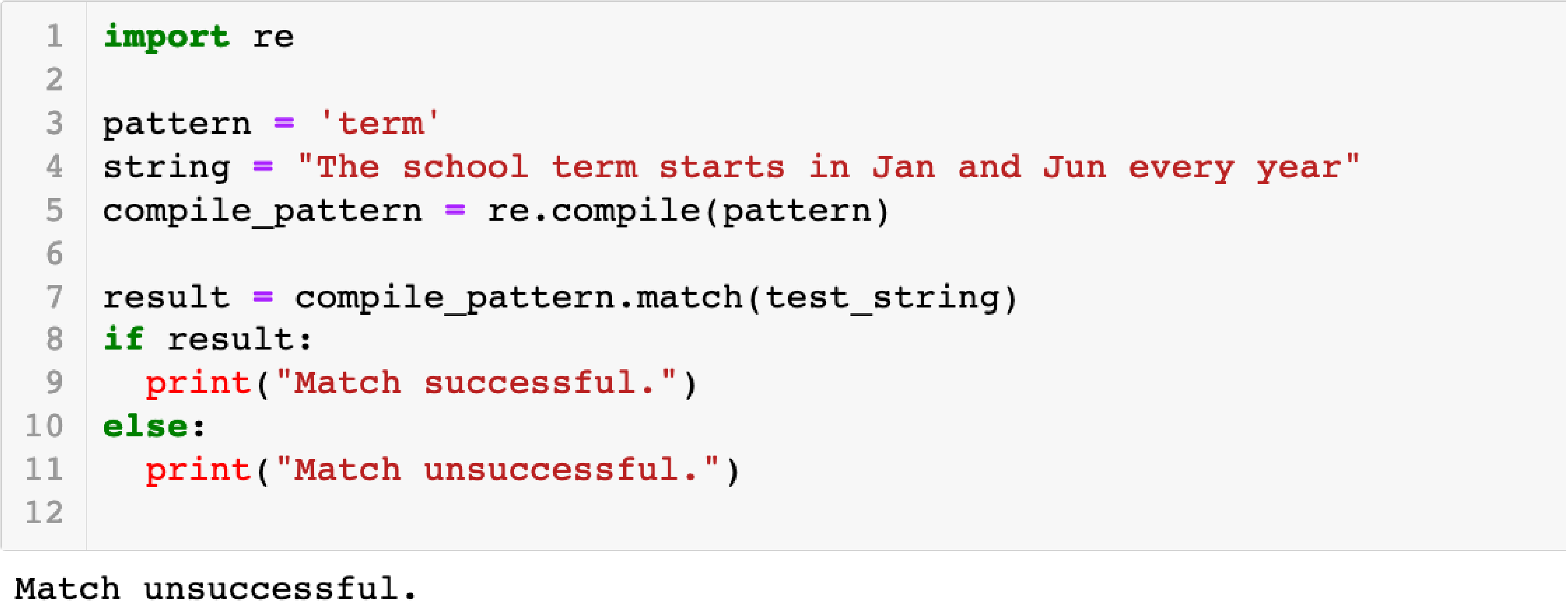
Compiled object can be reused and is more efficient when the expression is used multiple times

Alternatives: Using RE objects



`Pattern.match(string[, pos[, endpos]])`

If zero or more characters at the *beginning* of *string* match this regular expression, return a corresponding match object. Return `None` if the string does not match the pattern; note that this is different from a zero-length match.



```python
import re

pattern = 'term'
string = "The school term starts in Jan and Jun every year"
compile_pattern = re.compile(pattern)

result = compile_pattern.match(test_string)
if result:
  print("Match successful.")
else:
  print("Match unsuccessful.")

```

```
Match unsuccessful.
```

Exercise 2

Using RE library in Python

Proceed with Jupyter Notebook

ex2-regex.ipynb

Ref: https://www.w3schools.com/python/python_regex.asp

Exercise 3

Use regular expression to gain preliminary insights into a large corpus of data

Proceed with Jupyter Notebook

ex3-regex.ipynb

References

- Regular expression in Python, https://www.w3schools.com/python/python_regex.asp
- Python Regular Expressions, https://developers.google.com/edu/python/regular-expressions