Text 101 with spaCy

Lisa Ong
Lecturer & Consultant
NUS ISS StackUp

Speaker Intro

bit.ly/iss-lisaong

linkedin.com/in/lisaong

github.com/lisaong

What I Teach

SOFTWARE SYSTEMS
NICF- Designing Intelligent Edge Computing (SF)

SOFTWARE SYSTEMS
NICF- Humanizing Smart Systems (SF)

STACKUP - STARTUP TECH TALENT DEVELOPMENT
NICF- Sequence Modeling with Deep Learning (SF)

STACKUP - STARTUP TECH TALENT DEVELOPMENT
NICF- Data and Feature Engineering for Machine Learning (SF)

STACKUP - STARTUP TECH TALENT DEVELOPMENT
NICF- Supervised and Unsupervised Modeling with Machine Learning (SF)

STACKUP - STARTUP TECH TALENT DEVELOPMENT
NICF- Feature Extraction and Supervised Modeling with Deep Learning (SF)

What are we doing today?

Part 1: Text Processing

- Tokenisation, Cleaning & Exploration
- Hands-on with spaCy

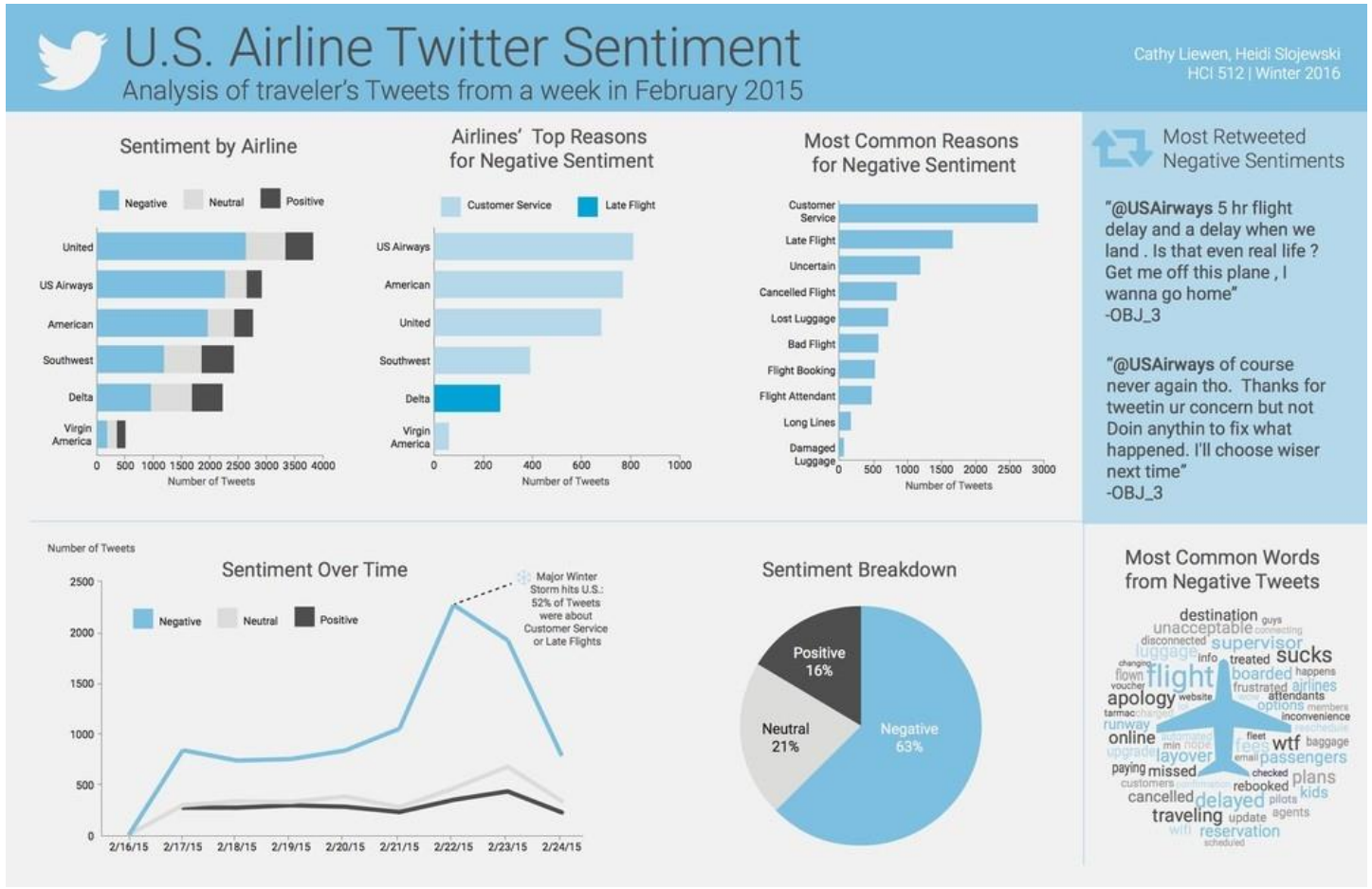
Part 2: Text Meaning

- Word Vectors & Similarity
- Hands-on with spaCy

Part 1: Text Processing

Tokenisation, Cleaning & Exploration

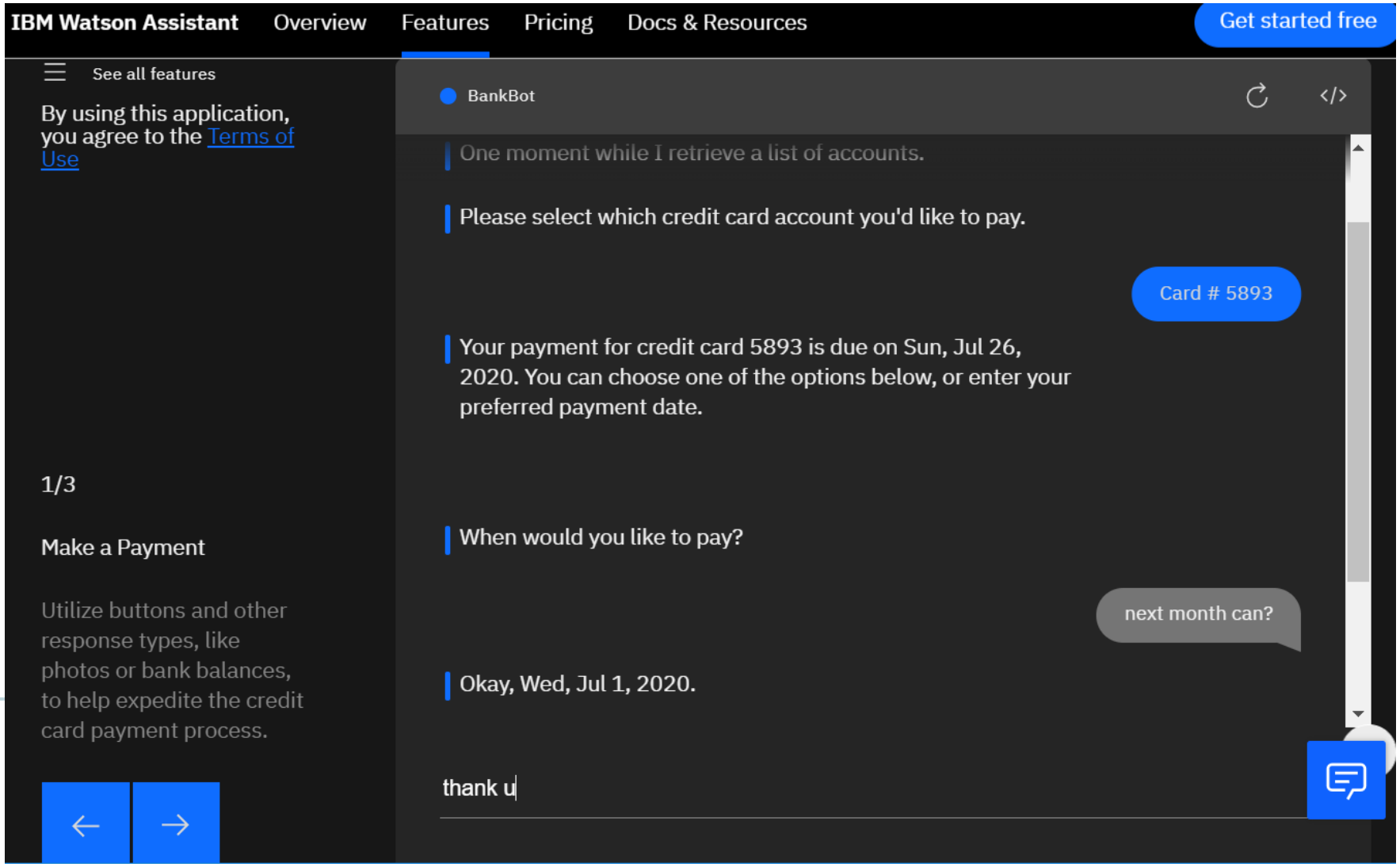
Sentiment Mining



www.wonderflow.co/blog/sentiment-analysis-examples

Chatbots (NLU)

NLU: Natural Language Understanding



www.ibm.com/cloud/watson-assistant

Workflow

Collect Raw Text

Process Text

Extract Meaning

Perform Analysis

Raw Text Examples

"To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them. To die: to sleep;"

Language is not static: models need to be created and refreshed periodically

More Raw Text Examples

(͡° ͜ʖ ͡°)

(ʊ_ʊ)

¯_(ツ)_/¯

(ﾉ °□°）ﾉ ︵ ┻━┻

(•_•) (•_•)>⌐■-■ (⌐■_■)

twistedsifter.files.wordpress.com

Article 1 All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

المادة 1 يولد جميع الناس أحرارًا متساوين في الكرامة والحقوق. وقد وهبوا عقلاً وضميرًا وعليهم أن يعامل بعضهم بعضًا بروح الإخاء.

第1条 すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心とを授けられており、互いに同胞の精神をもって行動しなければならない。

www.w3.org/International/articles/vertical-text/

Typically use a specific "emoticon" language model or custom rules

Typically handled by rendering attributes (not part of the text content)

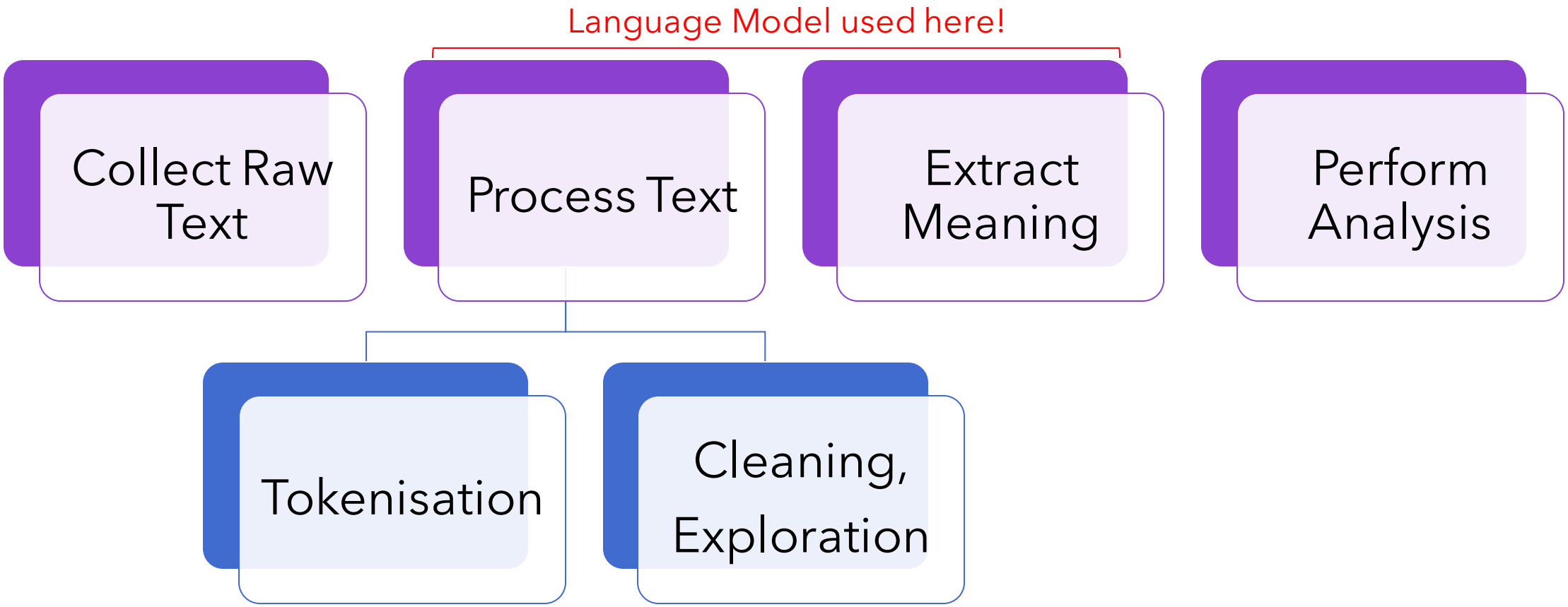
Processing Text

Select the language model

Split the text into tokens

Clean & Explore the tokenised text

Workflow: Process Text



Basic Text Processing

- **Language Model**
- Tokenisation
- Cleaning
- Extraction

Selecting Language Models

- What is the **primary language in use**?
  - English is the **default**, but not always **appropriate**
- What language is **supported by the tools**?
  - Not all languages have the same **quality of support**
  - If non ideal, may need to fine-tune or train own model (data- and time-investment)

Language support

spaCy currently provides support for the following languages. You can help by improving the existing language data and extending the tokenization patterns. See here for details on how to contribute to model development.

LANGUAGE	CODE	LANGUAGE DATA	MODELS
Chinese	zh	lang/zh	3 models
Danish	da	lang/da	3 models
Dutch	nl	lang/nl	3 models
English	en	lang/en	3 models
French	fr	lang/fr	3 models
German	de	lang/de	3 models
Greek	el	lang/el	3 models

spacy.io/usage/models

What about text like this?

Kindle商店 › Kindle电子书 › 社会科学

在线试读

要领（斯坦福大学原校长、谷歌母公司Alphabet董事会主席、图灵奖得主、“硅谷教父”约翰·汉尼斯重磅力作，写给有为者而非有位者的领导指南）

Kindle电子书

[美] 约翰·汉尼斯(John L. Hennessy) (作者) | 格式: Kindle电子书

| 分享 <分享样章> | #1 亚马逊最畅销商品 在高等教育中

› 显示所有 格式和版本

Kindle电子书
￥25.99

使用我们的 免费Kindle阅读软件

【内容简介】

电子书定价: ~~￥80.99~~
Kindle电子书价格: ￥25.99

立即购买

或

一键下单

发送至您的Kindle设备或Kindle阅读软件

请输入促销代码或礼品卡

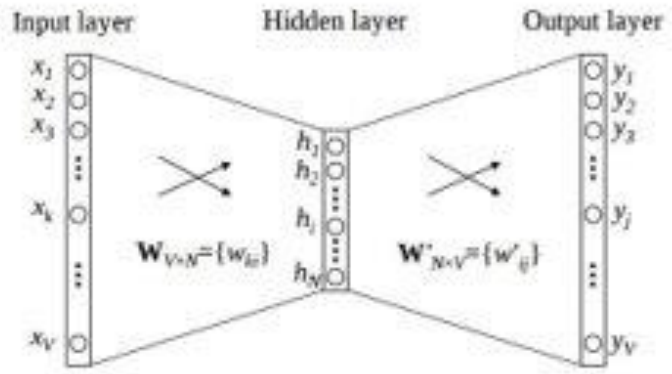
Kindle账户充值

amazon.cn

Most models support only **1 language per model**.
Limited functionality multi-language models do exist.
Some borrowed words (e.g. Kindle) do appear in the Chinese vocab, so can be **trained** or handled through **rules.**

How are Language Models trained?

1. Choose a large text corpus (ideally '00Ks or MMs of words)
2. Vocabulary: **encode words** into **numbers (vectors)**
   1. Probabilistic approach: GloVe
   2. Neural Network approach: Word2Vec
3. Linguistic aspects (e.g. nouns, verbs)
   1. Rule-based matching
   2. Neural Network Classifiers (e.g. https://spacy.io/usage/tra
4. Updating models
   - Add custom rules (spaCy: Matcher)
   - Fine-tune Neural Network (spaCy: nlp.update())



stackoverflow

Basic Text Processing

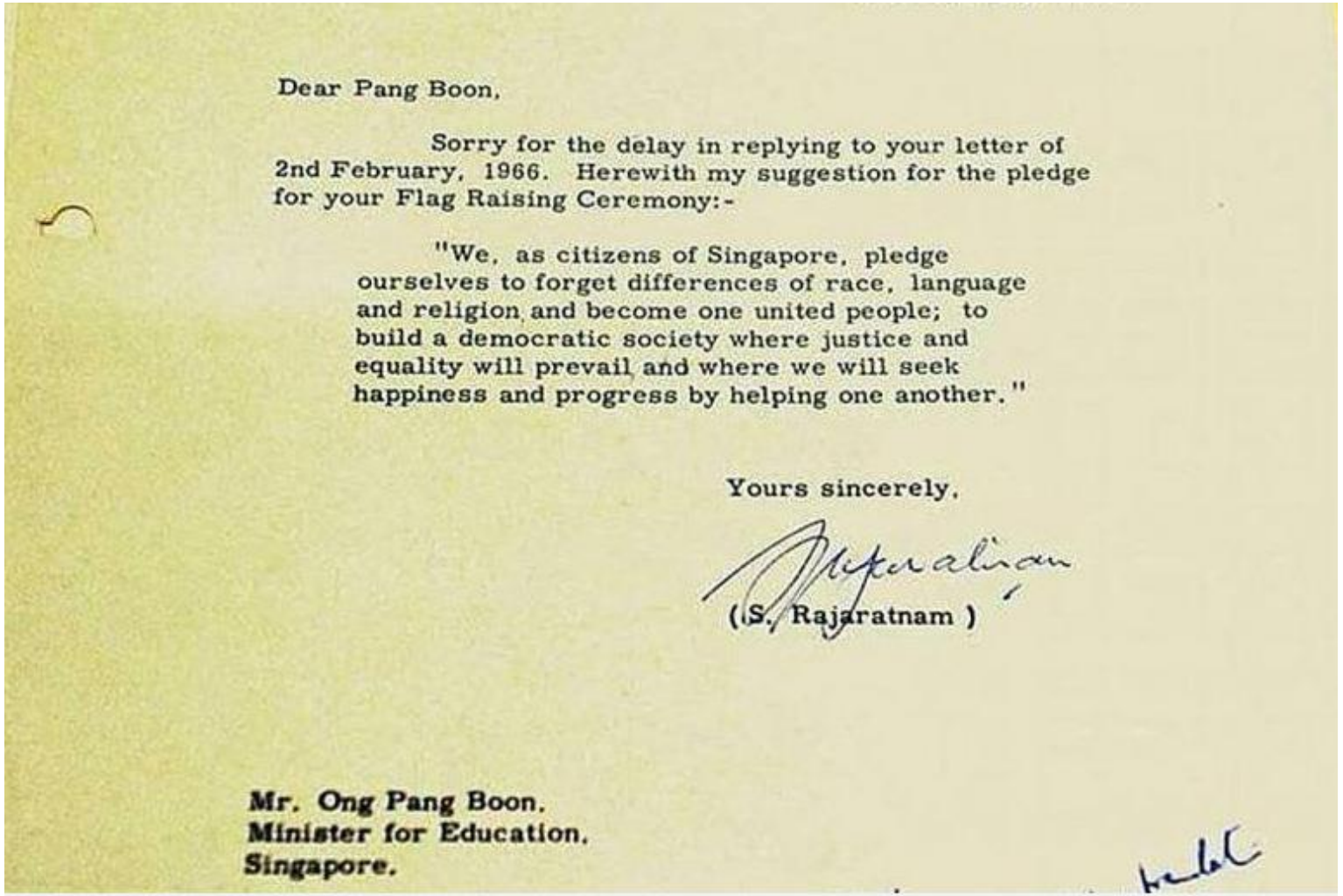
- Language Model
- **Tokenisation**
- Cleaning
- Extraction

Text Tokenization

The process of **splitting a text** into small **units**

The smallest unit is a **token**

What are the **token boundaries**?



Dear Pang Boon,

Sorry for the delay in replying to your letter of 2nd February, 1966. Herewith my suggestion for the pledge for your Flag Raising Ceremony:-

"We, as citizens of Singapore, pledge ourselves to forget differences of race, language and religion and become one united people; to build a democratic society where justice and equality will prevail and where we will seek happiness and progress by helping one another."

Yours sincerely,

(S. Rajaratnam)

Mr. Ong Pang Boon,
Minister for Education,
Singapore.

The third draft of the National Pledge went through another round of revisions by the Education Ministry and then Prime Minister Lee Kuan Yew before it was approved by the Cabinet in August 1966. PHOTO: FACEBOOK/K. SHANMUGAM

English: space, punctuation, …

Basic Text Processing

- Language Model
- Tokenisation
- **Cleaning**
- **Extraction**

Cleaning and Extraction

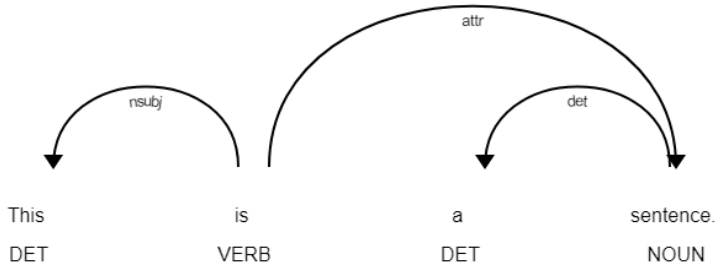


Image: spacy

Cleaning

- Remove **irrelevant** words/tokens
- E.g.: punctuation, extra spacing, symbols, "is, the, a, that, …"

Extraction

- **Part of speech**: noun, verb, determiner
- **Entity**: Place, Person, Date, Organisation, …
- **Syntactic relationship**: Words that come before (ancestor) or after (child) current word

Part of Speech Tagging

POS	DESCRIPTION	EXAMPLES
ADJ	adjective	big, old, green, incomprehensible, first
ADP	adposition	in, to, during
ADV	adverb	very, tomorrow, down, where, there
AUX	auxiliary	is, has (done), will (do), should (do)
CONJ	conjunction	and, or, but
CCONJ	coordinating conjunction	and, or, but
DET	determiner	a, an, the
INTJ	interjection	psst, ouch, bravo, hello
NOUN	noun	girl, cat, tree, air, beauty
NUM	numeral	1, 2017, one, seventy-seven, IV, MMXIV
PART	particle	’s, not,
PRON	pronoun	I, you, he, she, myself, themselves, somebody
PROPN	proper noun	Mary, John, London, NATO, HBO
PUNCT	punctuation	., (,), ?
SCONJ	subordinating conjunction	if, while, that
SYM	symbol	$, %, §, ©, +, −, ×, ÷, =, :), 😝
VERB	verb	run, runs, running, eat, ate, eating
X	other	sfpksdpsxmsa
SPACE	space	

spacy.io/api/annotation#named-entities

Named Entity Recognition

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

spacy.io/api/annotation#named-entities

Hands-on with spaCy

Load a Language Model

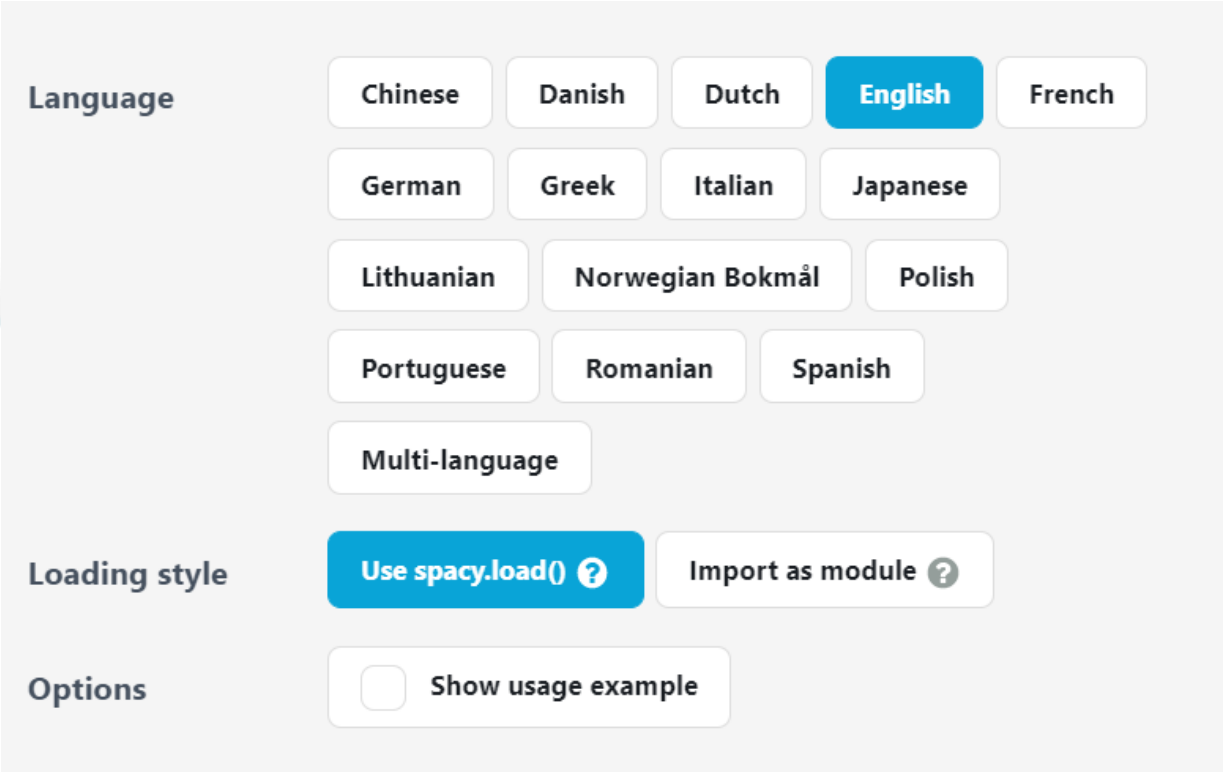
Tokenise the text

Cleaning, Exploration

The next 2 slides introduce the tools…

spaCy

- Popular Open Source NLP library, written in Python and Cython
- Good for beginners
- Extensible for advanced users through pipelines
- Supports English, Chinese, Japanese, French, Multi-language models, etc




```
$ python -m spacy download en_core_web_sm

>>> import spacy
>>> nlp = spacy.load("en_core_web_sm")
```

spacy.io

Logo: By Artichok3 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=54243937

Google Colab

- Run Python in the Browser
- Pre-installed libraries for statistics, plotting, NLP, ML, DL, …
- Linux cloud machine
- Free GPUs

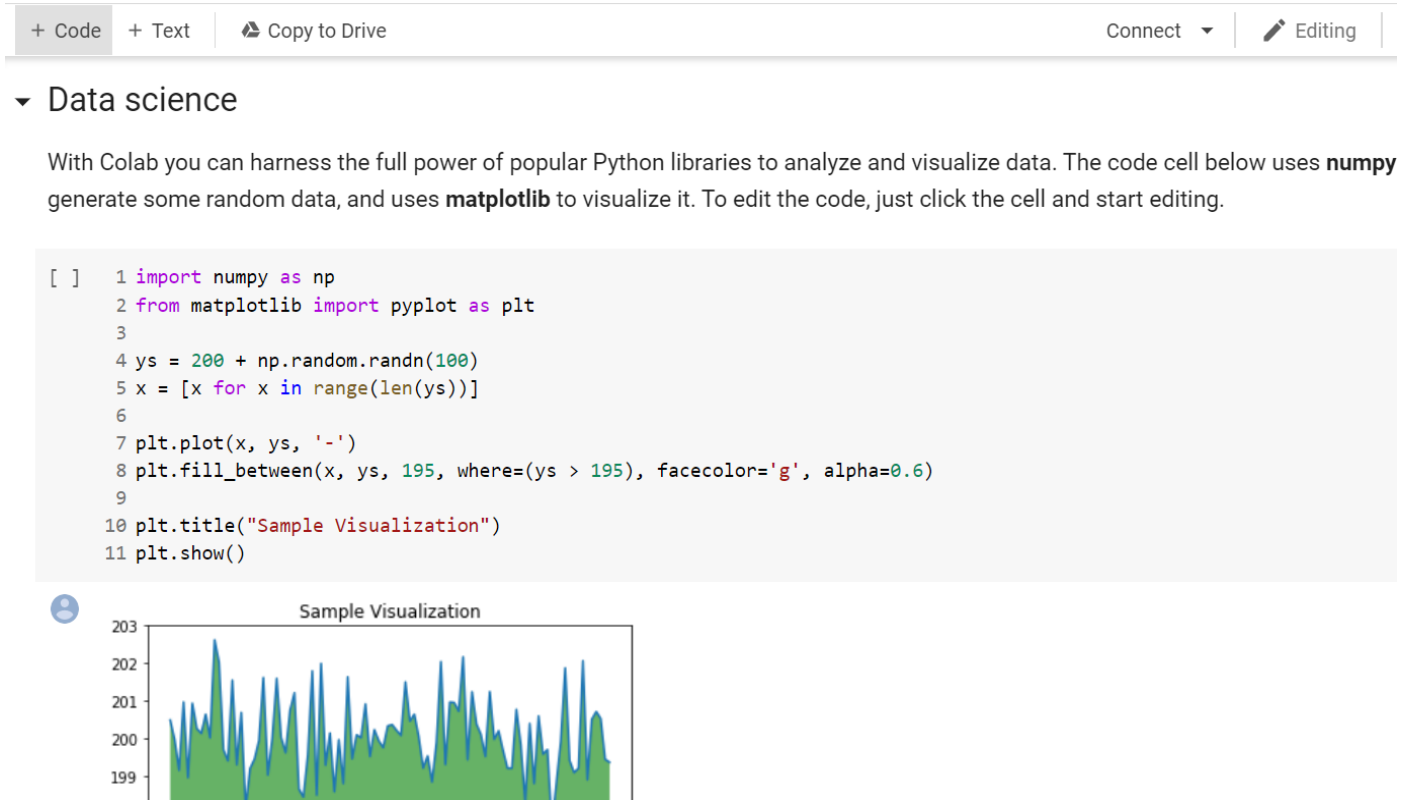


Image: colab.research.google.com

Go to: **bit.ly/iss-text101**

2890 lines (2890 sloc) | 161 KB

Raw Blame

Open in Colab

Text Processing 101 Demo Notebook

Useful references:

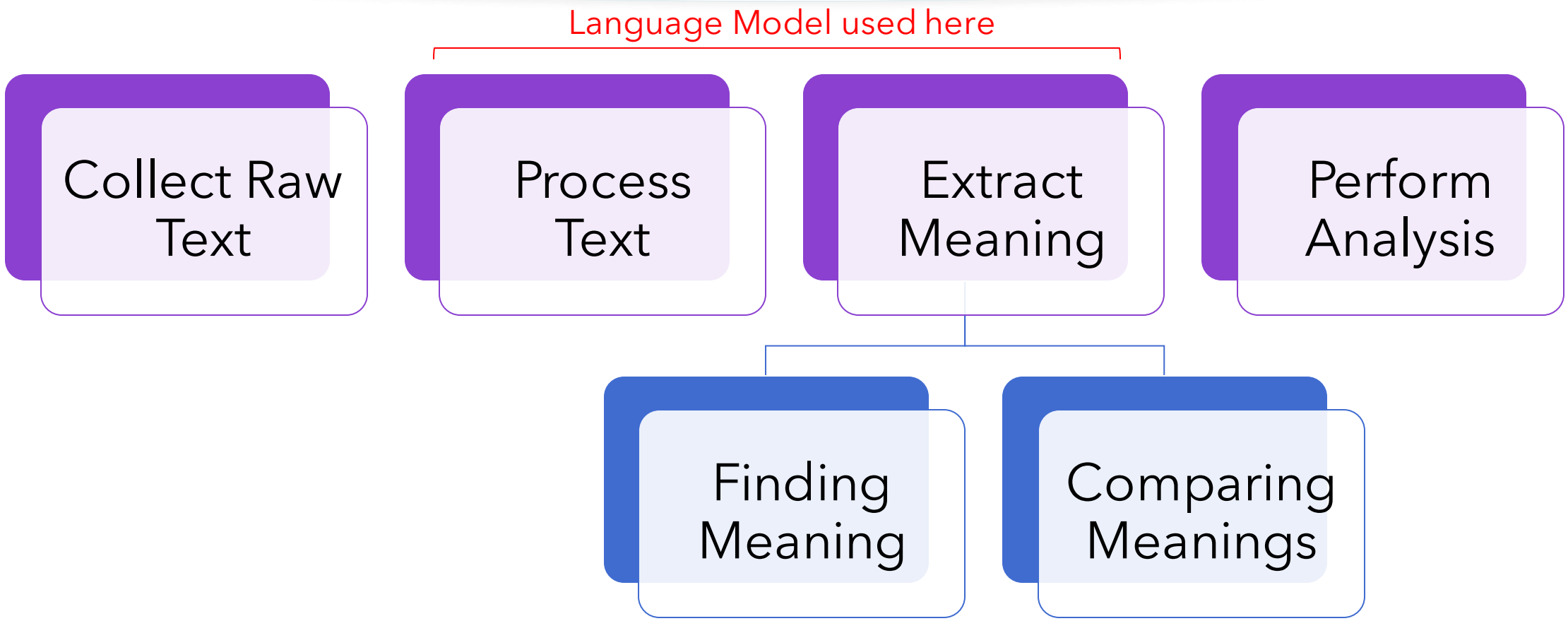
- spacy usage: https://spacy.io/usage
- spacy cheatsheet: https://www.datacamp.com/community/blog/spacy-cheatsheet
- matplotlib: https://www.datacamp.com/community/blog/python-matplotlib-cheat-sheet
- seaborn: https://www.datacamp.com/community/blog/seaborn-cheat-sheet-python

Click the highlighted button to launch in Google Colab.

Part 2: Text Meaning

Word Vectors and Similarity

Workflow: Extract Meaning



Finding meaning

meaning [mee-ning] SHOW IPA

SEE SYNONYMS FOR *meaning* ON THESAURUS.COM

noun

1. what is intended to be, or actually is, expressed or indicated; signification; import:
   the three meanings of a word.
2. the end, purpose, or significance of something:
   What is the meaning of life? What is the meaning of this intrusion?
3. *Linguistics.*
   a. the nonlinguistic cultural correlate, reference, or denotation of a linguistic form; expression.
   b. linguistic content (opposed to expression).

adjective

4. intentioned (usually used in combination):
   She's a well-meaning person.
5. full of significance; expressive:

Image: dictionary.com

What is meaning (to machines)?

- **A bunch of numbers**, derived from **learning** on large volumes of text data
- Numbers that can be added, subtracted, multiplied, and **compared**

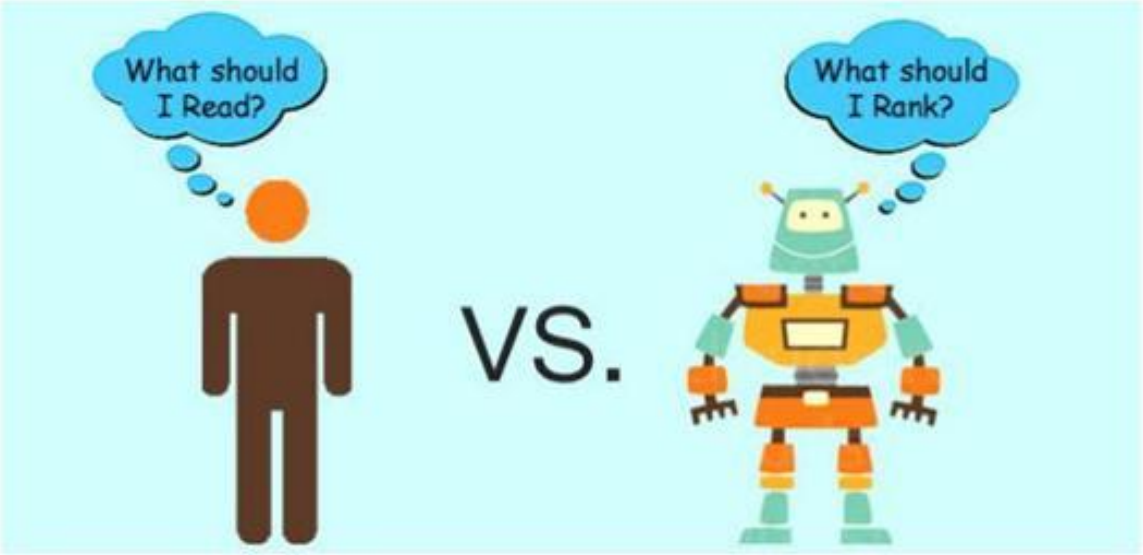


Image: www.techwyse.com

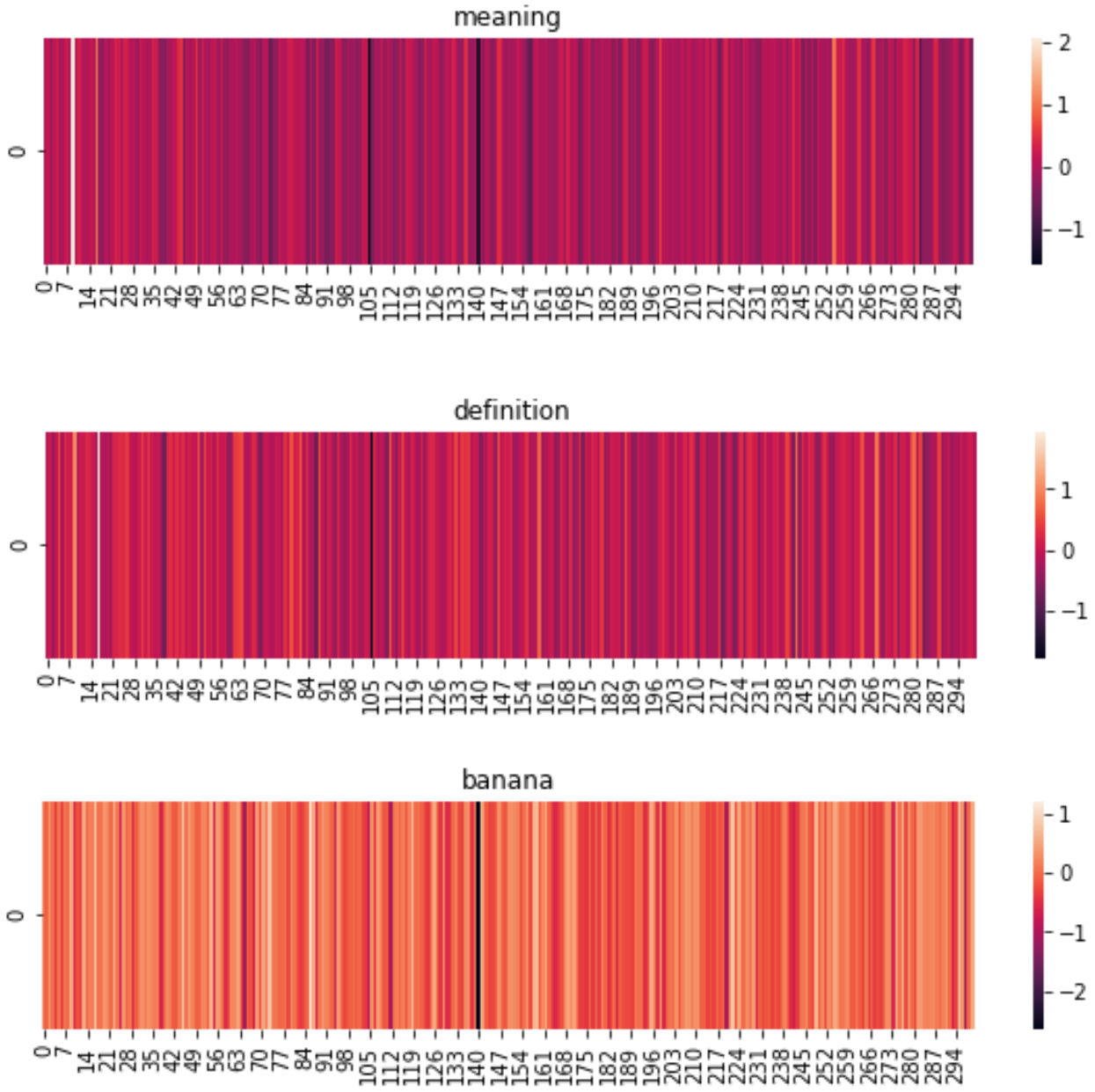
Part of the word vector for "meaning" (length = 300)

```
array([ 3.2456e-02,  1.5584e-01, -2.2763e-01,  1.2952e-01,  2.9447e-01,
       -3.1122e-02,  6.2653e-02,  2.3528e-01, -8.5213e-02,  2.0571e+00,
       -1.9374e-02,  3.3405e-01,  9.8610e-02,  2.7788e-02,  6.0454e-02,
        1.7122e-01, -4.9111e-02,  9.8979e-01, -2.1726e-01, -3.2660e-01,
        9.0827e-02, -1.0801e-01,  1.2777e-01,  3.8531e-01,  2.8327e-01,
       -6.9632e-02,  3.2282e-01,  2.6586e-01,  3.7181e-02,  1.2763e-01,
       -1.3892e-01,  2.2034e-01,  8.4188e-02, -9.5130e-03,  2.0808e-03,
        3.5423e-01,  2.6161e-01, -2.7047e-01, -3.2764e-01, -3.4673e-01,
       -7.1548e-02,  2.2396e-02,  1.3196e-01,  3.9390e-01,  4.9190e-01,
       -4.1159e-01,  4.4494e-02, -9.1078e-02,  5.4010e-02,  4.7170e-01,
```

Word Vectors: Semantic Encoding

Notice how **synonym words** have **similar-looking vectors** when displayed as a heatmap.

These vectors **encode semantic meaning** about a word (derived from training data).



How are Word Vectors generated?

Common Contextual Methods:

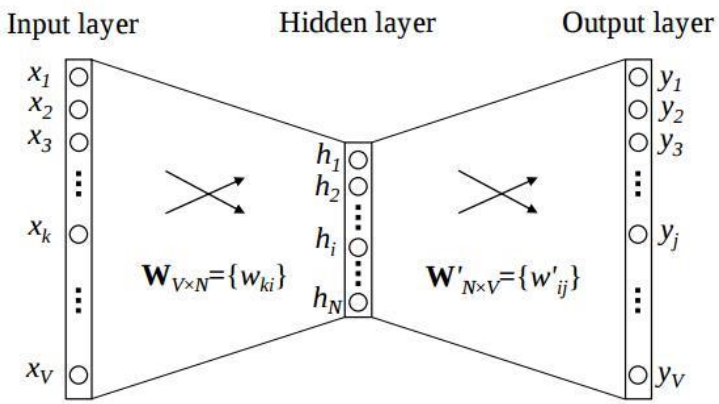
- Maximum likelihood estimation of **word co-occurrence probabilities (how often words appear together)**, and using the **<u>estimated parameters as word vectors</u>**

or

- Neural Network training to **predict words from nearby words**, and **extracting the <u>hidden layer as word vectors</u>**

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

nlp.stanford.edu/projects/glove



stackoverflow.com/questions/42281078/word2vec-output-vectors

How are Word Vectors used?

Comparing meanings with other text

Passed into Neural Networks to perform classification or regression tasks

E.g. US airline sentiment, document classification

Grouping similar documents (lists of word vectors) to find patterns

E.g. Clustering Amazon reviews

Comparing meanings

meaning [mee-ning]

SEE DEFINITION OF *meaning*

noun **message, signification** | *noun* **intention, aim**

SYNONYMS FOR *meaning*

connotation	nuance	bearing	pith	worth
content	sense	denotation	point	bottom line
context	significance	drift	purport	name of the game
definition	spirit	force	stuff	nature of beast
effect	subject	gist	suggestion	nuts and bolts
essence	substance	heart	symbolization	subject matter
explanation	understanding	import	tenor	
hint	value	intimation	thrust	
implication	acceptation	meat	upshot	
interpretation	allusion	nitty-gritty	use	

Image: dictionary.com

How do machines compare meanings?

- A meaning (vector) is an arrow from **the origin** with a **direction** and **length**
- Two meanings can be compared by computing the **distance between the tips of the arrows (= Euclidean distance)**
- Alternatively, you can **measure the angle between the arrows.** This compares **direction, but not length (=Cosine distance)**

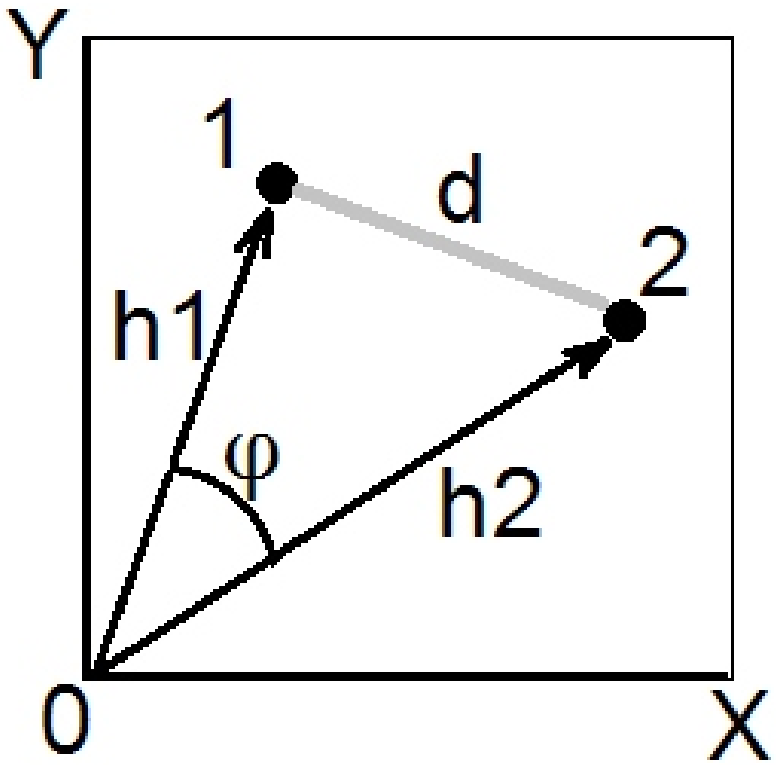
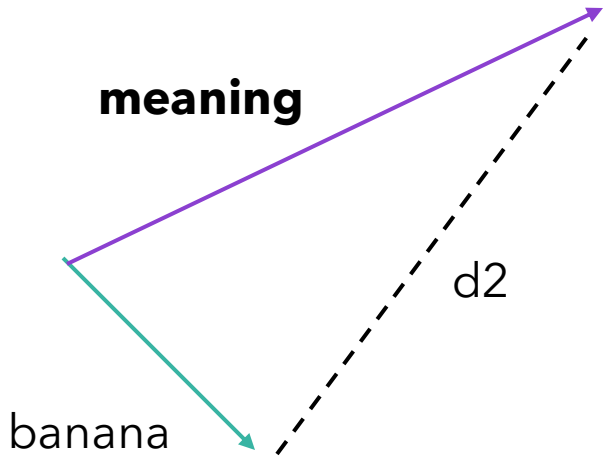


Image: stackexchange.com

Distance is relative to...?



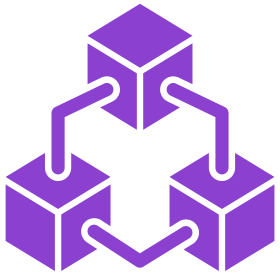
In NLP, we compute "similarity" or "most similar words" from a **reference word or vector**.

Uses of Semantic Similarity

Extracting meaning of a phrase, sentence, paragraph	By taking the average of the word vectors and finding similar words
Word suggestion	By suggesting topmost similar words
Content comparison	By comparing the similarity scores of two documents

Hands-on with spaCy

Getting the Word Vectors



Comparing the Word Vectors

Image:kimaverycoaching

Part 1: Text Processing

- Tokenisation, Cleaning & Exploration
- Hands-on with spaCy

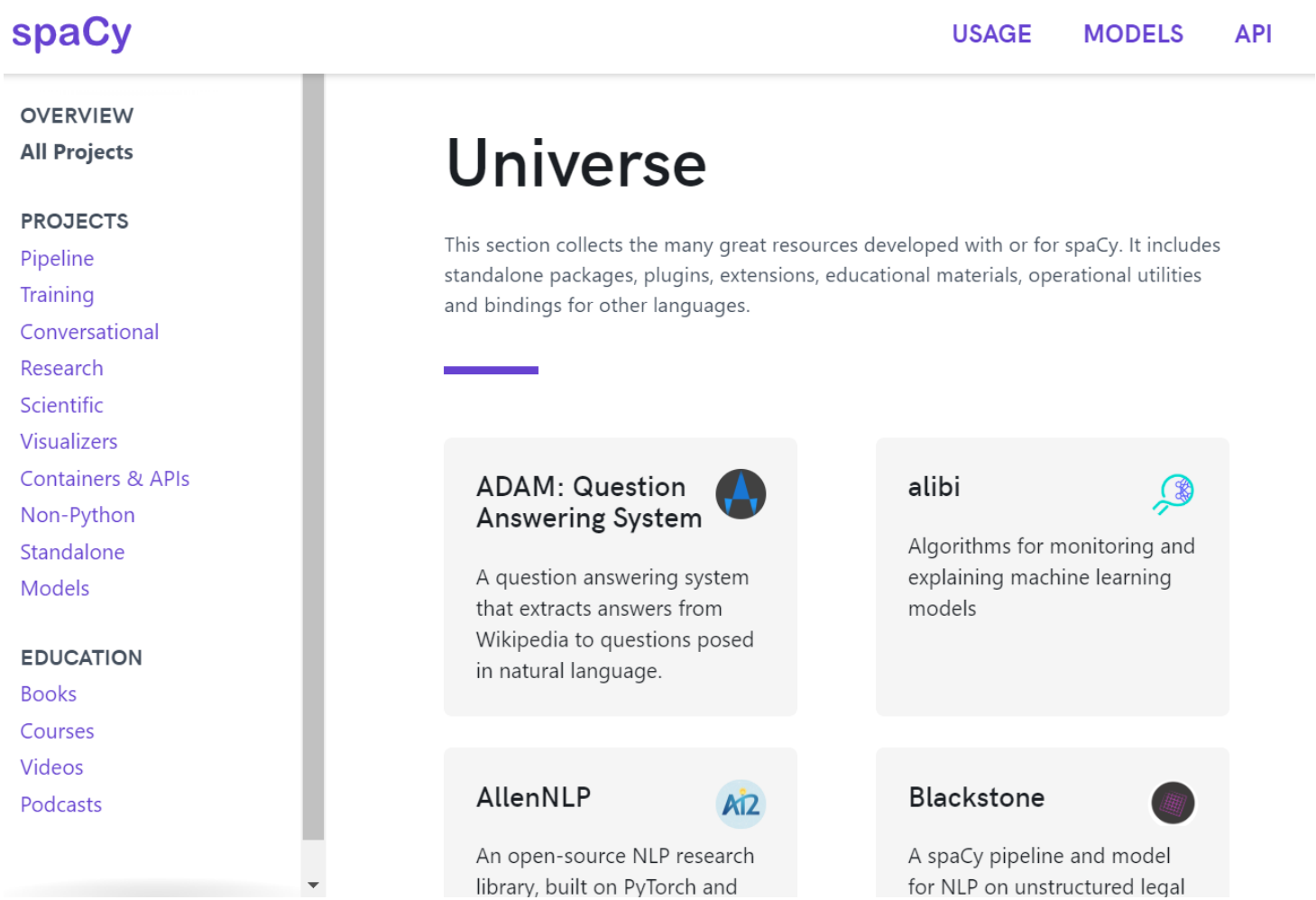
Part 2: Text Meaning

- Word Vectors & Similarity
- Hands-on with spaCy

Next steps...

Explore the **spaCy Universe**: spacy.io/universe

Tools, libraries, projects built using spaCy



Want to learn more?

spaCy online course (free)

course.spacy.io

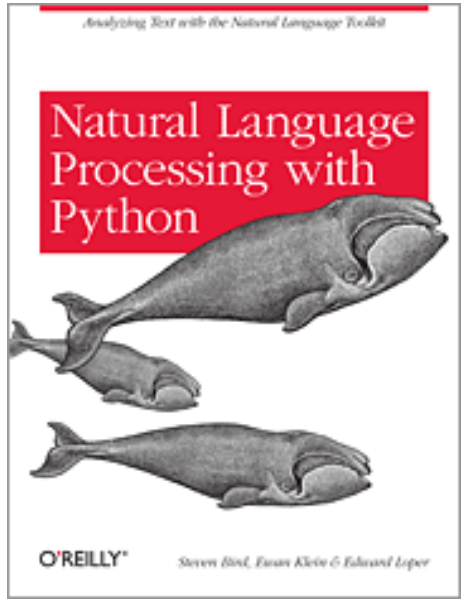
NLP with Python e-book (free)

www.nltk.org/book_1ed

NUS ISS **Graduate Certificate** in **Practical Language Processing**

Text Analytics, Sentiment Mining, Text Processing & ML, Chatbots.

27 days. Counts towards **MTech in Data Science**.



ADVANCED NLP with spaCy

Upcoming StackUp Courses

4 July

4-Course Series: Machine Learning-Driven Data Science: Part Time

13 July

Short Course: Autonomous Decision-making with Reinfocement Learning

27 July

Short Course: Containers for Deploying and Scaling Apps (SF)

bit.ly/iss-stackup

Thank you!!

- Help us improve your experience
- Scan the QR code or go to https://nus.edu/3eynlym
- Presentation slides and recording will be shared after completion & successful submission of feedback
- Allow 3-4 working days to receive the slides, upon successful submission.
- Alternatively, write to us at issmarketing@nus.edu.sg