A Gentle Introduction to Natural Language Processing

Natalia Klyueva

The Hong Kong Polytechnic University
Chinese and Bilingual Studies

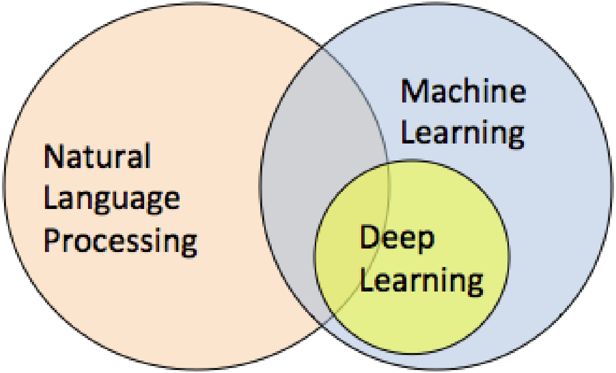
natalia.klyueva@polyu.edu.hk

June 25, 2018

Overview

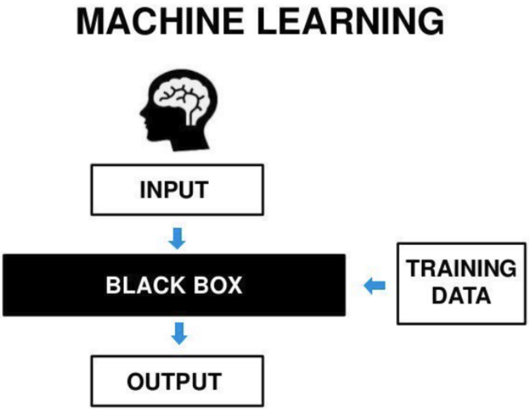
1. ML and NLP
2. From a sense to a word
3. Sentiment analysis

ML and NLP



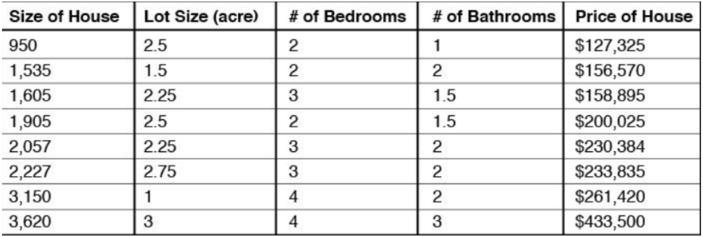
picture taken from https://rutumulkar.com/blog/2016/NLP-ML

What is ML: Recap

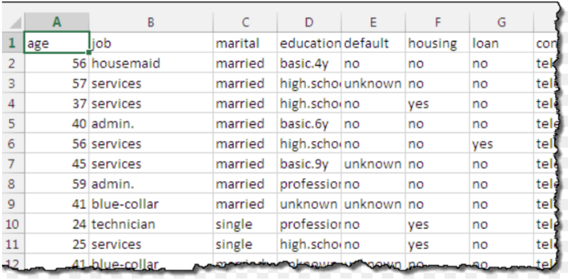


picture taken from https://twitter.com/KirkDBorne

Canonical Examples of ML data



Size of House	Lot Size (acre)	# of Bedrooms	# of Bathrooms	Price of House
950	2.5	2	1	$127,325
1,535	1.5	2	2	$156,570
1,605	2.25	3	1.5	$158,895
1,905	2.5	2	1.5	$200,025
2,057	2.25	3	2	$230,384
2,227	2.75	3	2	$233,835
3,150	1	4	2	$261,420
3,620	3	4	3	$433,500



	A	B	C	D	E	F	G	
1	age	job	marital	education	default	housing	loan	con
2	56	housemaid	married	basic.4y	no	no	no	tel
3	57	services	married	high.scho	unknown	no	no	tel
4	37	services	married	high.scho	no	yes	no	tel
5	40	admin.	married	basic.6y	no	no	no	tel
6	56	services	married	high.scho	no	no	yes	tel
7	45	services	married	basic.9y	unknown	no	no	tel
8	59	admin.	married	professio	no	no	no	tel
9	41	blue-collar	married	unknown	unknown	no	no	tel
10	24	technician	single	professio	no	yes	no	tel
11	25	services	single	high.scho	no	yes	no	tel
12	41	blue-collar	married	unknown	unknown	no	no	tel

Natural Language Processing

- The same as in Machine Learning tasks above, but...
- features are: words, sentences, paragraphs, documents etc.
- Canonical NLP tasks:
  - search engines
  - Machine Translation
  - face recognition systems
  - chatbots
  - natural language generation
  - opinion mining (sentiment analysis)

Overview

1. ML and NLP
2. From a sense to a word
3. Sentiment analysis

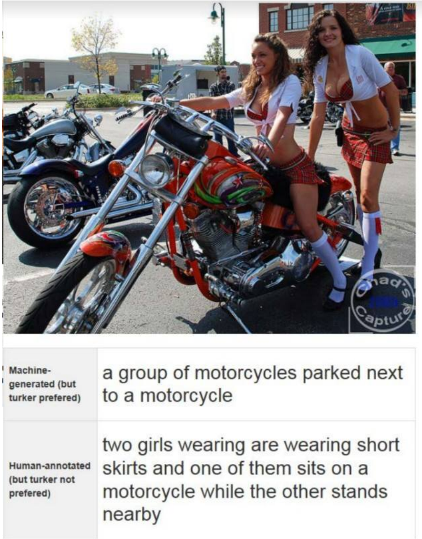
Word <-> Sense: overview

How can machine understand meaning? Let's start with the basic notion - 'sense'

- How can machine understand meaning of a word?
- image recognition task: how to relate from an image (real world object) to a word/sentence
    - "Is this a cat on the picture?"
    - FB: "Do you want to tag FRIENDS_NAME FRIENDS_SURNAME?"
- given a picture, give a description of an object

Image recognition: near-NLP task

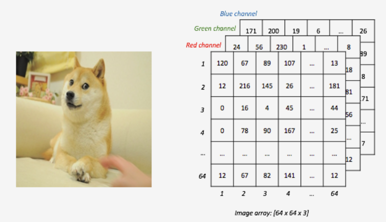
What object do we(humans) and they(machines) refer to?



Machine-generated (but turker prefered)	a group of motorcycles parked next to a motorcycle
Human-annotated (but turker not prefered)	two girls wearing are wearing short skirts and one of them sits on a motorcycle while the other stands nearby

How the machine sees an image: oversimplified

- We identify object by its features (is furry, has 4 paws, tail, special form of ears etc...)? Or by observations?
- Representation for a machine: all images are 64x64 pixels, and each pixel is a certain value of RGB
- Input: array of numbers



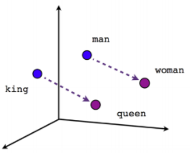
1	120	67	89	107	...	13
2	12	216	145	26	...	181
3	0	16	4	45	...	44
4	0	78	90	167	...	25
...	...	...	...	...	...	...
64	12	67	82	141	...	12
	1	2	3	4	...	64

Image array: [64 x 64 x 3]

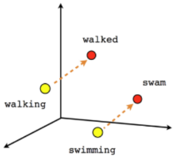
- Output: 97% it is a dog
- That is how to represent a picture.
- And how do we represent words?

Word representation

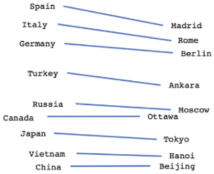
- Obviously, we need numbers to make calculations easier and faster
- We need to assign some numerical representation for each word
- I->id1; see->id2; a->id3 cat->id4 ... Is it a good one????
- Small revolution in NLP: word2vec https://www.tensorflow.org/tutorials/word2vec
- Input: words,sentences,text,documents. Output: vector



Male-Female



Verb tense

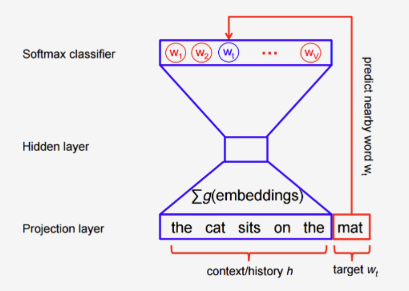


Country-Capital

- '2+2' of NLP: king−man+women=queen
- https://projector.tensorflow.org/

word2vec/doc2vec: what is inside?

- Predict the word:



- For doc2vec - the same principle, with document information included https://cs.stanford.edu/~quocle/paragraph_vector.pdf

Overview

1. ML and NLP
2. From a sense to a word
3. Sentiment analysis

Hot NLP topic: Sentiment analysis (or opinion mining)

- Likes/dislikes on facebook, wechat; swipes right/left; comments on products/services on internet
- User generated content. Comes for free and in large quantity
- The questions are: "Do you like this or dislike this?", "What are your emotions"

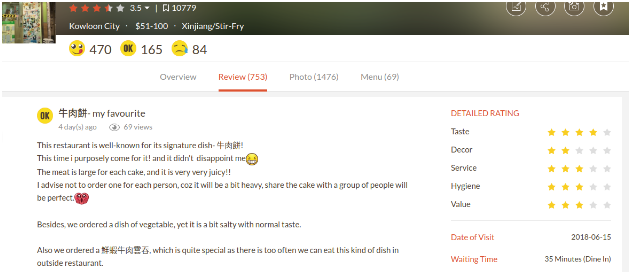
Training the model

- Sentiment analysis of the text
- On the user generated content, you can perform various machine learning experiments and model "sentiment"
- Input: review -> MODEL -> Output:sentiment/emotion

Our data: openrice

- Sentiment analysis of the text
- On the user generated content, you can perform various machine learning experiments and model "sentiment"
- Input: review -> MODEL -> Output:sentiment/emotion
- Openrice scrapped
- About 360,000 comments; 5,000 restaurant IDs

Data formatting



3.5 | 10779

Kowloon City · $51-100 · Xinjiang/Stir-Fry

470 165 84

Overview Review (753) Photo (1476) Menu (69)

牛肉餅- my favourite

4 day(s) ago 69 views

This restaurant is well-known for its signature dish- 牛肉餅!

This time i purposely come for it! and it didn't disappoint me

The meat is large for each cake, and it is very very juicy!!

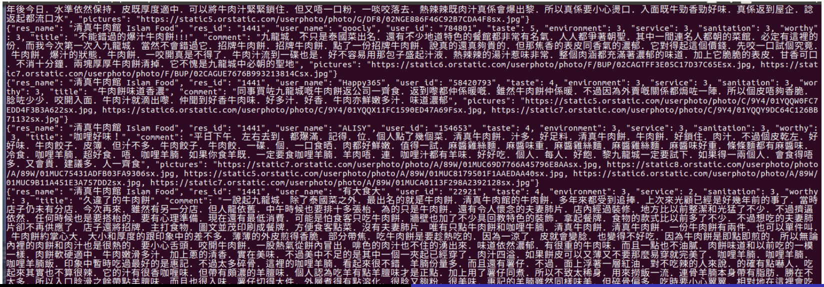
I advise not to order one for each person, coz it will be a bit heavy, share the cake with a group of people will be perfect.

Besides, we ordered a dish of vegetable, yet it is a bit salty with normal taste.

Also we ordered a 鮮蝦牛肉雲吞, which is quite special as there is too often we can eat this kind of dish in outside restaurant.

DETAILED RATING

Taste	★★★★☆
Decor	★★☆☆☆
Service	★★★☆☆
Hygiene	★★★☆☆
Value	★★★☆☆
Date of Visit	2018-06-15
Waiting Time	35 Minutes (Dine In)

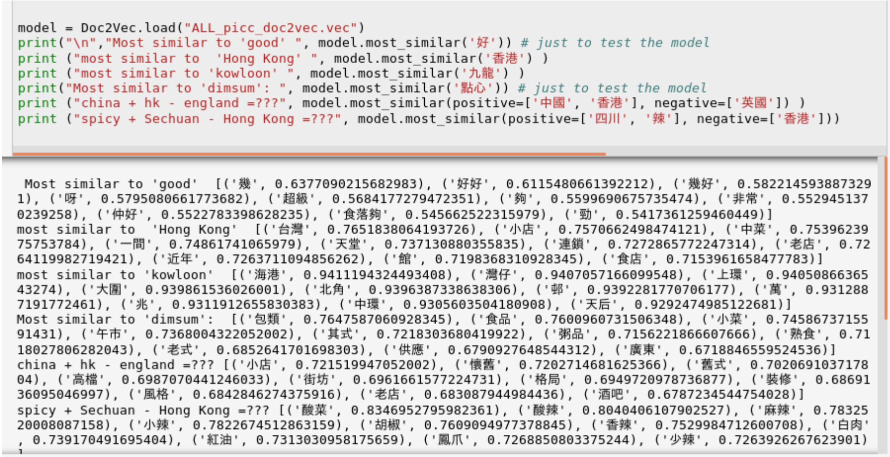


```
年後今日．水準依然保持．皮既厚度適中．可以將牛肉汁緊緊鎖住．但又唔一口粉．一啖咬落去．熱辣辣既肉汁真係會爆出黎．所以真係要小心燙口．入面既牛勁香勁好味．真係返到屋企．諗
返起都流口水", "pictures": "https://static5.orstatic.com/userphoto/photo/G/DF8/02NGE886F46C92B7CDA4F8sx.jpg"}
{"res_name": "清真牛肉館 Islam Food", "res_id": "1441", "user_name": "goocly", "user_id": "784801", "taste": 5, "environment": 3, "service": 3, "sanitation": 3, "worthy"
: 3, "title": "不能錯過的爆汁牛肉餅!!!", "comment": "九龍城．不只是泰國菜出名．還有不少地道特色的餐館都非常有名氣．人人都爭著朝聖．其中一間連名人都朝的菜館．必定有這裡的
份．而我今次第一次入九龍城．當然不會錯過它．招牌牛肉餅．招牌牛肉餅．點了一份招牌牛肉餅．說真的還真夠貴的．但那焦香的表皮同香氣的濃郁．它對得起這個價錢．先咬一口試個究竟．
 牛肉餅．爆汁的狀態．牛肉餅．一咬開真是不得了．牛肉汁流到一碟也是．好不容易用那包子盛起汁液．熱辣辣的湯汁惹味非常．整個肉涵都充滿著濃郁的味道．加上它脆脆的表皮．甘香可口
．不消十分鐘．兩塊厚厚牛肉餅清掉．它不愧是九龍城中必朝的聖地", "pictures": "https://static6.orstatic.com/userphoto/photo/F/BUP/02CAGTFF3E05C17D37C65Esx.jpg, https://stat
ic7.orstatic.com/userphoto/photo/F/BUP/02CAGUE7676B993213814Csx.jpg"}
{"res_name": "清真牛肉館 Islam Food", "res_id": "1441", "user_name": "Happy365", "user_id": "58420793", "taste": 4, "environment": 3, "service": 3, "sanitation": 3, "wor
thy": 3, "title": "牛肉餅味道香濃", "comment": "同事買咗九龍城嘅牛肉餅返公司一齊食．返到嚟都仲係暖嘅．雖然牛肉餅仲係暖．不過因為外賣嘅關係都焗咗一陣．所以個皮唔夠香脆．
腍咗少少．咬開入面．牛肉汁就滴出嚟．仲聞到好香牛肉味．好多汁．好香．牛肉亦鮮嫩多汁．味道濃郁", "pictures": "https://static5.orstatic.com/userphoto/photo/C/9Y4/01YQQW0FC7
EDD4F3B3A622sx.jpg, https://static6.orstatic.com/userphoto/photo/C/9Y4/01YQQX11FC1590ED47A69Fsx.jpg, https://static7.orstatic.com/userphoto/photo/C/9Y4/01YQQY9DC64C126BB
71132sx.jpg"}
{"res_name": "清真牛肉館 Islam Food", "res_id": "1441", "user_name": "ALISY", "user_id": "154653", "taste": 4, "environment": 3, "service": 3, "sanitation": 3, "worthy":
 3, "title": "咖哩好味！", "comment": "平日下午．左右去到．都爆滿．記得．位．個人點了幾個菜．清真牛肉餅．汁多．好足料．清真牛肉餅．牛肉餅．好鎖住．肉汁．不過個皮乾左．好
好味．牛肉餃子．皮簿．但汁不多．牛肉餃子．牛肉餃．一碟．個．一口食晒．肉都好鮮嫩．值得一試．麻醬雞絲麵．麻醬味重．麻醬雞絲麵．麻醬雞絲麵．麻醬味好重．條條麵都有麻醬味．
冷食．咖哩羊腩．超好食．唔．咖哩羊腩．如果你食羊既．一定要食咖哩羊腩．羊肉唔．連．咖哩汁都有羊味．好好吃．個人．每人．好飽．黎九龍城一定要試下．如果得一兩個人．會食得唔
多．又會貴．建議多．人一齊食", "pictures": "https://static7.orstatic.com/userphoto/photo/A/89W/01MUC69D7766A45796E8AAsx.jpg, https://static8.orstatic.com/userphoto/photo
/A/89W/01MUC75431ADFB03FA9306sx.jpg, https://static5.orstatic.com/userphoto/photo/A/89W/01MUC8179501F1AAEDAA40sx.jpg, https://static6.orstatic.com/userphoto/photo/A/89W/
01MUC9B11A451E3A757DD2sx.jpg, https://static7.orstatic.com/userphoto/photo/A/89W/01MUCA0113F298A2392128sx.jpg"}
{"res_name": "清真牛肉館 Islam Food", "res_id": "1441", "user_name": "有大食大", "user_id": "22921", "taste": 4, "environment": 3, "service": 2, "sanitation": 3, "worthy
": 3, "title": "久違了的牛肉餅", "comment": "一說起九龍城．除了泰國菜之外．最出名的就是牛肉餅．清真牛肉館的牛肉餅．多年來都受到追捧．上次來光顧已經是好幾年前的事了．當時
店子仍未有分店．今次再來．雖然有另一分店．但人龍依舊．中午時候也要排十多張枱．為的只是牛肉餅．還有令人懷念的夫妻肺片．店內經過裝修．地方比以前整潔和光猛了不少．不過擠逼
依然．任何時候也是要搭枱的．要有心理準備．現在還有最低消費．可能是怕食客只吃牛肉餅．牆壁也加了不少具回教特色的裝飾．拿起餐牌．食物的款式比以前多了不少．不過想吃的夫妻肺
片卻不再供應了．店子還將招牌．主打食物．圖文並茂印刷成餐牌．方便食客點菜．沒有夫妻肺片．唯有只點牛肉餅和咖哩牛腩．清真牛肉餅．清真牛肉餅．一份牛肉餅有兩件．也可以單件叫．
 牛肉餅約掌心大．大小和厚度的跟印象中的差不多．薄薄的外皮煎得香脆．部分帶焦．吃牛肉餅是要趁熱吃的．因為一涼了．皮就會變腍．也變得不好吃．因為牛肉餅是即點即煎的．所以無論
內裡的肉餅和肉汁也是很熱的．要小心舌頭．咬開牛肉餅．一股熱氣從餅內冒出．啡色的肉汁也不住的湧出來．味道依然濃郁．有很重的牛肉味．而且一點也不油膩．肉餅味道和以前吃的一模
一樣．肉餅軟硬適中．牛肉嫩滑多汁．加上蔥的清香．實在美味．不過美中不足的是其中一個一夾起已經穿了．肉汁四溢．如果餅皮可以又薄又不要那麼易穿就完美了．咖哩羊腩．咖哩羊腩．
咖哩羊腩飯．印象中暫時吃過最好的是惠記．不過太多碎骨．這裡的咖哩羊腩．看起來很不錯．羊腩份量多．而且還有薯仔．不過．面上浮著一層紅油．對不吃辣的人來說．的確有點嚇人．吃
起來其實也不算很辣．它的汁有很香咖喱味．但帶有頗濃的羊膻味．個人認為吃羊有點羊膻味才是正點．加上用了薯仔同煮．所以不致太稀身．用來撈飯一流．連骨羊腩本身帶有脂肪．勝在不
太多．所以入口腍滑之餘帶點羊膻味．而且也很入味．薯仔切得大件．外層煮得有點溶化．很腍又夠粉．很美味．惠記的羊腩雖然同樣味美．但碎骨偏多．吃時要小心翼翼．相對地在這裡會吃
```

Our data: openrice

- The simple (baseline) task: given a comment, predict the 'value' (1-5)
- Input: document embeddings, classifier from sklearn
- Demo: https://github.com/natalink/openrice_annotations/blob/master/clf_openrice.ipynb
- For the small sample (12506, on github) that I put online, the accuracy of the model was about 0.4
- If we select only very positive(rank 5) or negative (rank 1) comments, the accuracy was 0.7.
- For the whole data with rank value [1, 5] (26549 reviews) accuracy was around 0.850

Operations with vectors: spicy+sechuan-Hong Kong=???



```
model = Doc2Vec.load("ALL_picc_doc2vec.vec")
print("\n","Most similar to 'good' ", model.most_similar('好')) # just to test the model
print ("most similar to  'Hong Kong' ", model.most_similar('香港') )
print ("most similar to 'kowloon' ", model.most_similar('九龍') )
print("Most similar to 'dimsum': ", model.most_similar('點心')) # just to test the model
print ("china + hk - england =???", model.most_similar(positive=['中國', '香港'], negative=['英國']) )
print ("spicy + Sechuan - Hong Kong =???", model.most_similar(positive=['四川', '辣'], negative=['香港']))
```

```
 Most similar to 'good'  [('幾', 0.6377090215682983), ('好好', 0.6115480661392212), ('幾好', 0.582214593887329
1), ('呀', 0.5795080661773682), ('超級', 0.5684177279472351), ('夠', 0.5599690675735474), ('非常', 0.552945137
0239258), ('仲好', 0.5522783398628235), ('食落夠', 0.545662522315979), ('勁', 0.5417361259460449)]
most similar to  'Hong Kong'  [('台灣', 0.7651838064193726), ('小店', 0.7570662498474121), ('中菜', 0.75396239
75753784), ('一間', 0.74861741065979), ('天堂', 0.737130880355835), ('連鎖', 0.7272865772247314), ('老店', 0.72
64119982719421), ('近年', 0.7263711094856262), ('館', 0.7198368310928345), ('食店', 0.7153961658477783)]
most similar to 'kowloon'  [('海港', 0.9411194324493408), ('灣仔', 0.9407057166099548), ('上環', 0.94050866365
43274), ('大圍', 0.939861536026001), ('北角', 0.9396387338638306), ('邨', 0.9392281770706177), ('萬', 0.931288
7191772461), ('兆', 0.9311912655830383), ('中環', 0.9305603504180908), ('天后', 0.9292474985122681)]
Most similar to 'dimsum':  [('包類', 0.7647587060928345), ('食品', 0.7600960731506348), ('小菜', 0.74586737155
91431), ('午市', 0.7368004322052002), ('其式', 0.7218303680419922), ('粥品', 0.7156221866607666), ('熟食', 0.71
18027806282043), ('老式', 0.6852641701698303), ('供應', 0.6790927648544312), ('廣東', 0.6718846559524536)]
china + hk - england =??? [('小店', 0.721519947052002), ('懷舊', 0.7202714681625366), ('舊式', 0.7020691037178
04), ('高檔', 0.6987070441246033), ('街坊', 0.6961661577224731), ('格局', 0.6949720978736877), ('裝修', 0.68691
36095046997), ('風格', 0.6842846274375916), ('老店', 0.683087944984436), ('酒吧', 0.6787234544754028)]
spicy + Sechuan - Hong Kong =??? [('酸菜', 0.8346952795982361), ('酸辣', 0.8040406107902527), ('麻辣', 0.78325
20008087158), ('小辣', 0.7822674512863159), ('胡椒', 0.7609094977378845), ('香辣', 0.7529984712600708), ('白肉'
, 0.739170491695404), ('紅油', 0.7313030958175659), ('鳳爪', 0.7268850803375244), ('少辣', 0.7263926267623901)
]
```