Web and Social Media Analytics

Social media (data collection and pre-processing)

Dr Philip Worrall

School of Computer Science and Engineering
115 New Cavendish Street
University of Westminster
London, W1W 6UW
worralph@westminster.ac.uk

LW8

Week 8

Plan for today

- **Programming with Python**
  - Functions, OOP, External Packages and File IO
- **Data collection with Python**
  - Demo using Twitter Streaming API
  - Demo using the Reddit API
  - Demo using the YouTube Data API
- **The text mining process**
  - Data preparation and pre-processing

Data mining / Text mining

Fundamentals of data mining

The computer assisted process of extracting meaningful information from large datasets. To date, it has been used to derive patterns in consumer behavior, prediction of a likely outcome and support various types of decision making.

Examples

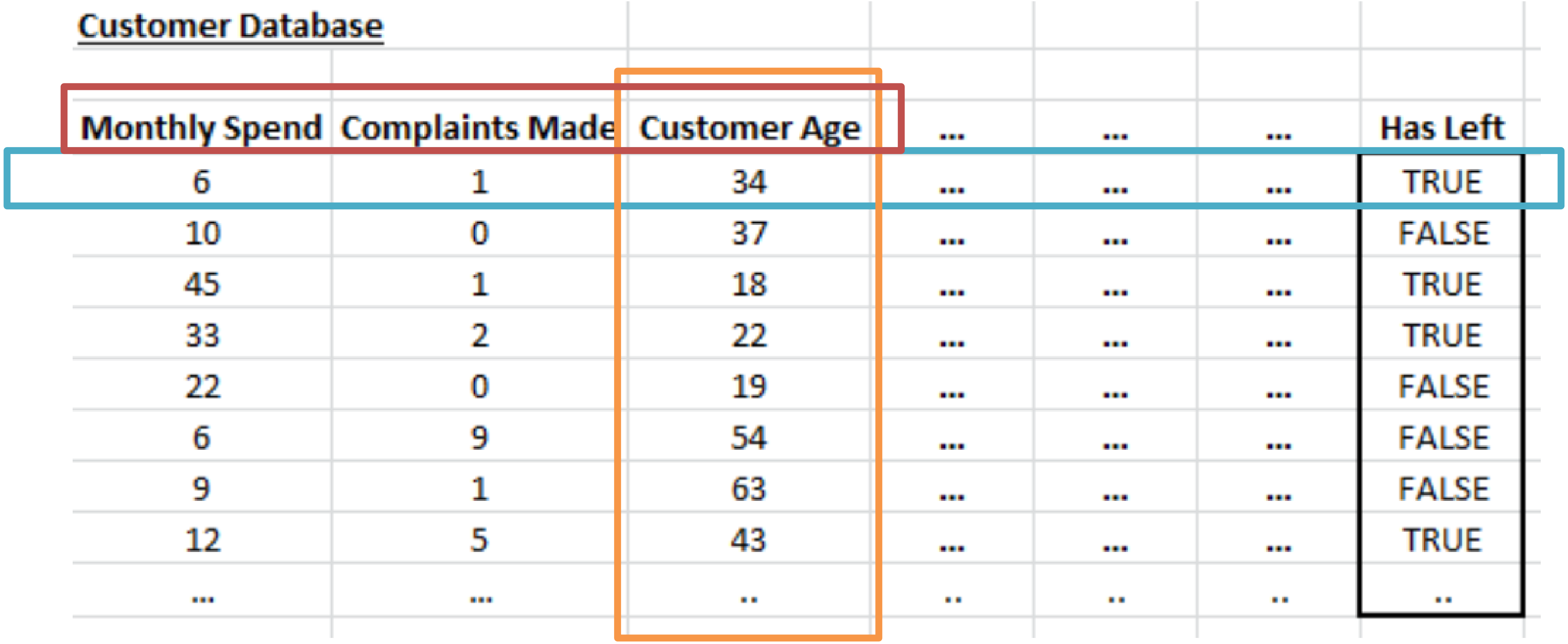
Prediction of weather, ability to pay back a loan, credit scoring, demographic profile.

Data mining / Text mining

Relationship with predictive modeling

Data mining is closely related to predictive modeling but the algorithms used in data mining are generally better suited to modeling situations in which highly complex non-linear relationships exist.

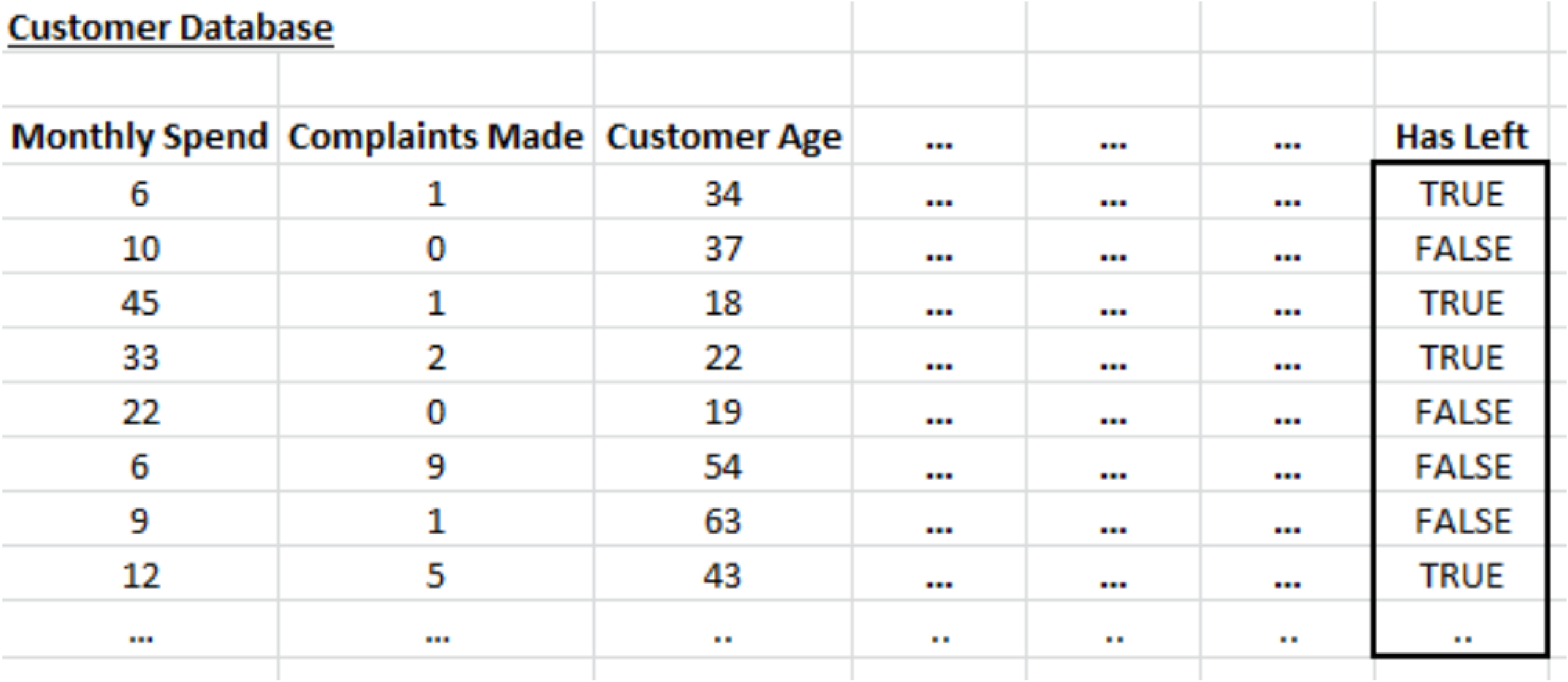
Typical scenario



Customer Database

Monthly Spend	Complaints Made	Customer Age	...	...	...	Has Left
6	1	34	...	...	...	TRUE
10	0	37	...	...	...	FALSE
45	1	18	...	...	...	TRUE
33	2	22	...	...	...	TRUE
22	0	19	...	...	...	FALSE
6	9	54	...	...	...	FALSE
9	1	63	...	...	...	FALSE
12	5	43	...	...	...	TRUE
...	...	..	..	..	..	..

Typical scenario



Customer Database

Monthly Spend	Complaints Made	Customer Age	...	...	...	Has Left
6	1	34	...	...	...	TRUE
10	0	37	...	...	...	FALSE
45	1	18	...	...	...	TRUE
33	2	22	...	...	...	TRUE
22	0	19	...	...	...	FALSE
6	9	54	...	...	...	FALSE
9	1	63	...	...	...	FALSE
12	5	43	...	...	...	TRUE
...	...	..	..	..	..	..

What kind of problems can we investigate with this dataset?

Data mining approaches

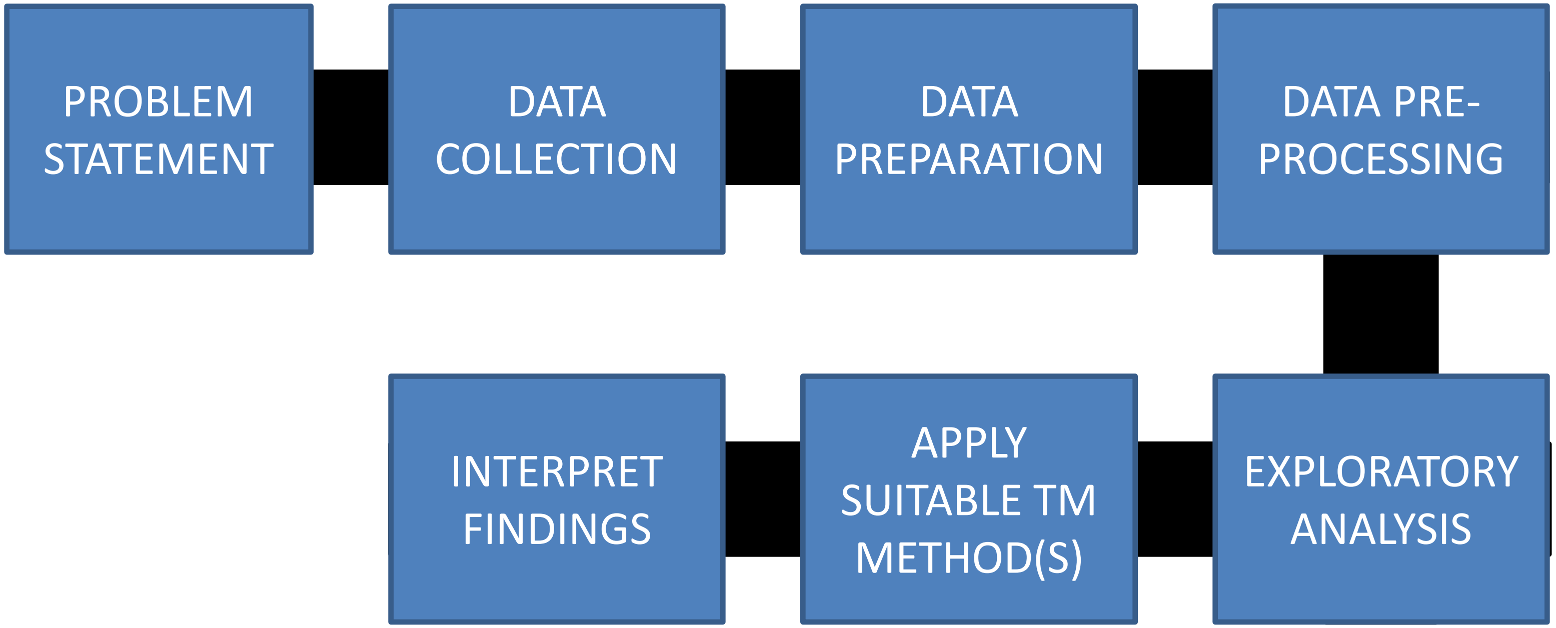
- 2 possible cases…
- Supervised (labelled)
  - class membership
  - level prediction
- Unsupervised (unlabelled)
  - clustering
  - feature extraction (e.g. PCA)

Text mining

Relationship with data mining

Text mining concerns the extraction of high quality information from text. Although the basic principles are similar, because we are dealing with natural language, a key aspect of text mining is the translation of text (and its associated meaning) into quantitative data so that we can apply mathematical models.

The text mining process…



PROBLEM STATEMENT

DATA COLLECTION

DATA PREPARATION

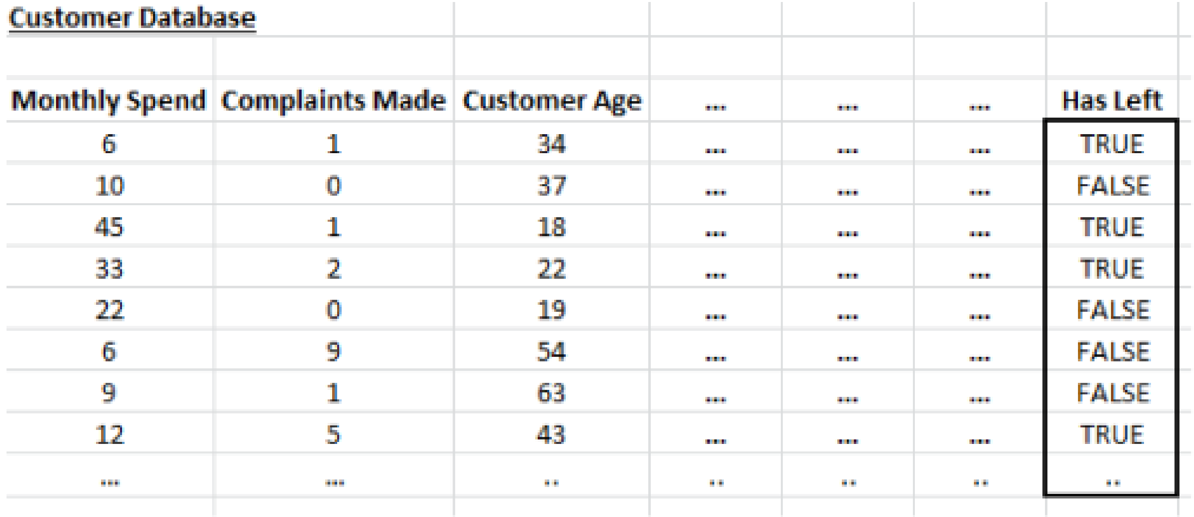
DATA PRE-PROCESSING

EXPLORATORY ANALYSIS

APPLY SUITABLE TM METHOD(S)

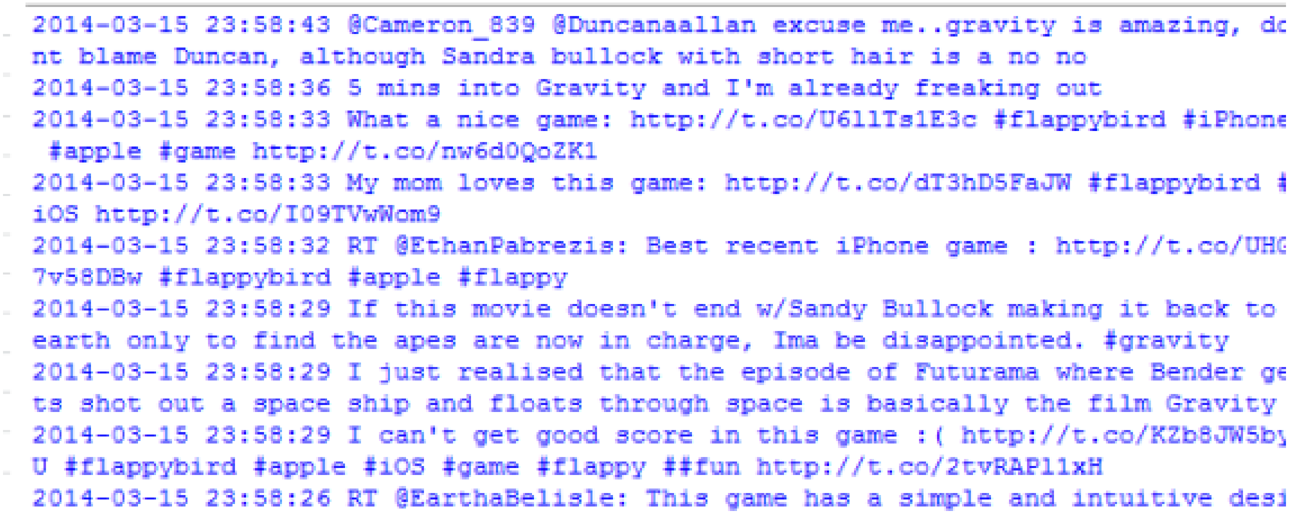
INTERPRET FINDINGS

Spot the difference…



Customer Database

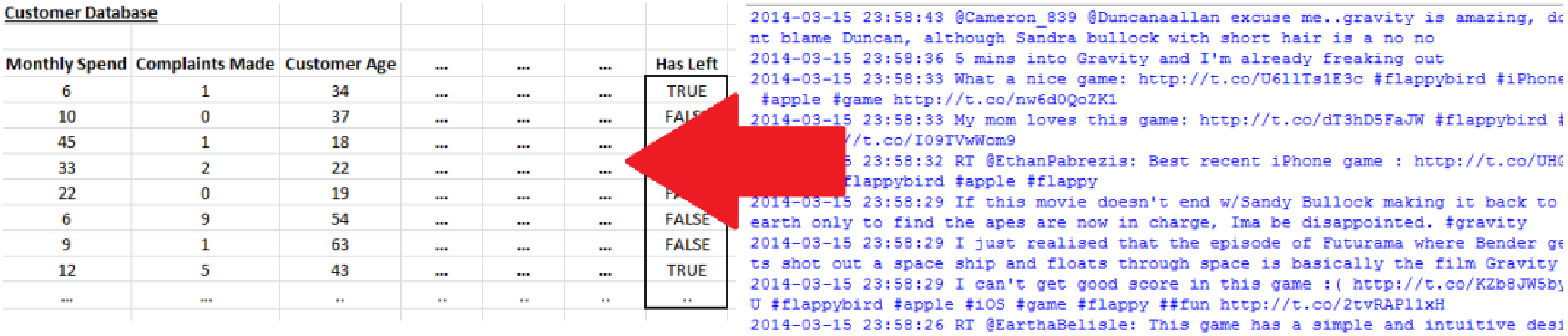
Monthly Spend	Complaints Made	Customer Age	....	....	....	Has Left
6	1	34	....	....	....	TRUE
10	0	37	....	....	....	FALSE
45	1	18	....	....	....	TRUE
33	2	22	....	....	....	TRUE
22	0	19	....	....	....	FALSE
6	9	54	....	....	....	FALSE
9	1	63	....	....	....	FALSE
12	5	43	....	....	....	TRUE
...	...	..	..	..	..	..



```
2014-03-15 23:58:43 @Cameron_839 @Duncanaallan excuse me..gravity is amazing, do
nt blame Duncan, although Sandra bullock with short hair is a no no
2014-03-15 23:58:36 5 mins into Gravity and I'm already freaking out
2014-03-15 23:58:33 What a nice game: http://t.co/U611Ts1E3c #flappybird #iPhone
 #apple #game http://t.co/nw6d0QoZK1
2014-03-15 23:58:33 My mom loves this game: http://t.co/dT3hD5FaJW #flappybird #
iOS http://t.co/I09TVwWom9
2014-03-15 23:58:32 RT @EthanPabrezis: Best recent iPhone game : http://t.co/UHG
7v58DBw #flappybird #apple #flappy
2014-03-15 23:58:29 If this movie doesn't end w/Sandy Bullock making it back to
earth only to find the apes are now in charge, Ima be disappointed. #gravity
2014-03-15 23:58:29 I just realised that the episode of Futurama where Bender ge
ts shot out a space ship and floats through space is basically the film Gravity
2014-03-15 23:58:29 I can't get good score in this game :( http://t.co/KZb8JW5by
U #flappybird #apple #iOS #game #flappy ##fun http://t.co/2tvRAP11xH
2014-03-15 23:58:26 RT @EarthaBelisle: This game has a simple and intuitive desi
```

Quantitative modelling

- DM/PM algorithms require numerical inputs
- We need fields, columns, data instances and (potentially) labels
- The first step involves **data encoding**



Customer Database

Monthly Spend	Complaints Made	Customer Age	...	...	...	Has Left
6	1	34	...	...	...	TRUE
10	0	37	...	...	...	FALS
45	1	18	...	...	...	
33	2	22	...	...	...	
22	0	19	...	...	...	F
6	9	54	...	...	...	FALSE
9	1	63	...	...	...	FALSE
12	5	43	...	...	...	TRUE
...	...	..	..	..	..	..

```
2014-03-15 23:58:43 @Cameron_839 @Duncanaallan excuse me..gravity is amazing, do
nt blame Duncan, although Sandra bullock with short hair is a no no
2014-03-15 23:58:36 5 mins into Gravity and I'm already freaking out
2014-03-15 23:58:33 What a nice game: http://t.co/U611Ts1E3c #flappybird #iPhone
 #apple #game http://t.co/nw6d0QoZK1
2014-03-15 23:58:33 My mom loves this game: http://t.co/dT3hD5FaJW #flappybird #
                 //t.co/I09TVwWom9
                5 23:58:32 RT @EthanPabrezis: Best recent iPhone game : http://t.co/UHG
                flappybird #apple #flappy
2014-03-15 23:58:29 If this movie doesn't end w/Sandy Bullock making it back to
earth only to find the apes are now in charge, Ima be disappointed. #gravity
2014-03-15 23:58:29 I just realised that the episode of Futurama where Bender ge
ts shot out a space ship and floats through space is basically the film Gravity
2014-03-15 23:58:29 I can't get good score in this game :( http://t.co/KZb8JW5by
U #flappybird #apple #iOS #game #flappy ##fun http://t.co/2tvRAP11xH
2014-03-15 23:58:26 RT @EarthaBelisle: This game has a simple and intuitive desi
```

Possible encoding strategies…

- *Any suggestions?*
  - What impact could the encoding have?

FROM:	TO:

Addressing dimensionality...

Preprocessing

An important first step is to consider applying forms of preprocessing to the textual data. This helps to reduce the dimensionality of the text mining problem.

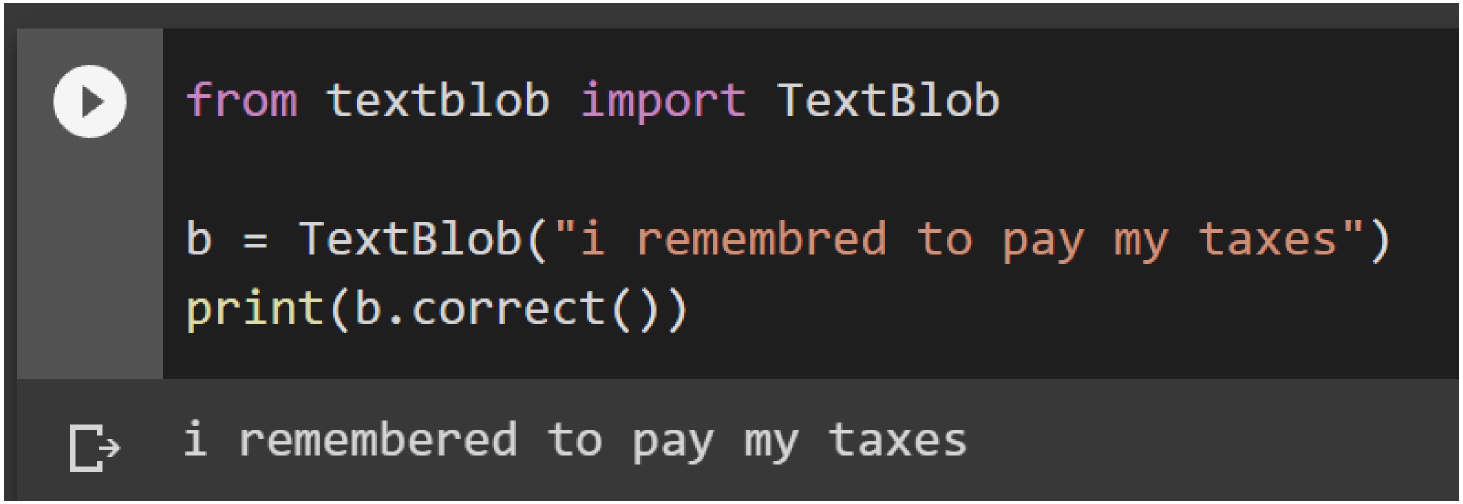
Dimensionality

In the data mining context, it refers to the number of different possible combinations of different data values in the dataset. Higher dimensionality leads to greater processing time and potentially weaker associations. In our text mining example, more words corresponds with greater dimensionality.

Too many different spellings...

Spelling

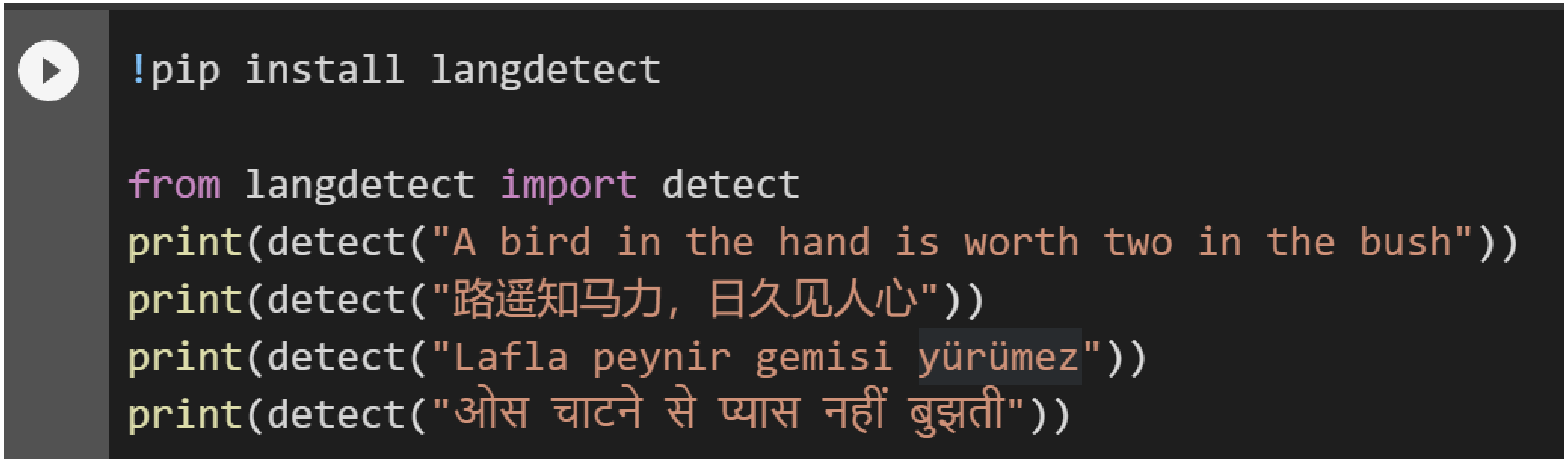
Reduce *colour* to *color*, *clothess* to *clothes*. A number of tools exist to help us correct spellings, including the python package *nltk* and *textblob*.



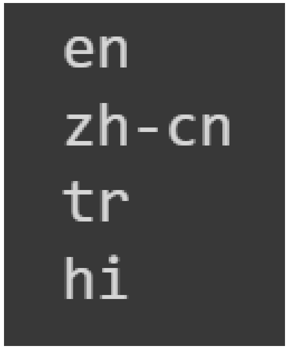
```
from textblob import TextBlob

b = TextBlob("i remembred to pay my taxes")
print(b.correct())
```

```
i remembered to pay my taxes
```

Data is in the same frame...

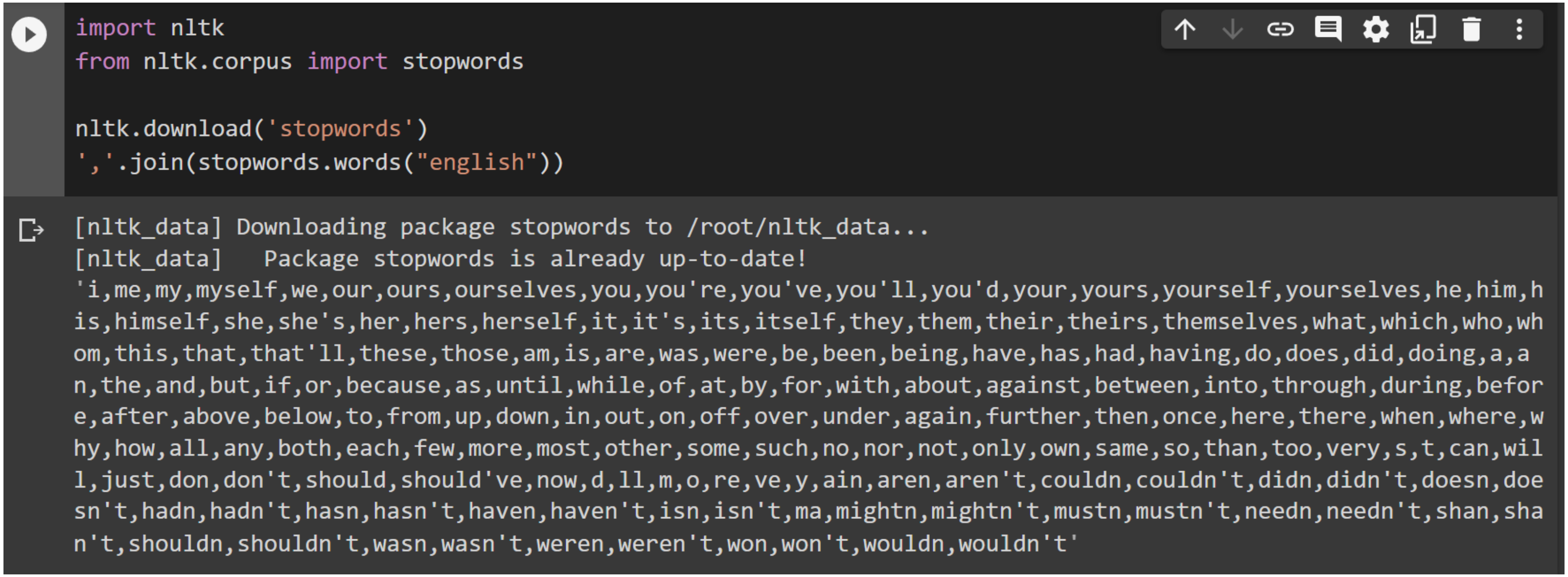


```
!pip install langdetect

from langdetect import detect
print(detect("A bird in the hand is worth two in the bush"))
print(detect("路遥知马力，日久见人心"))
print(detect("Lafla peynir gemisi yürümez"))
print(detect("ओस चाटने से प्यास नहीं बुझती"))
```



```
en
zh-cn
tr
hi
```

Stopwords



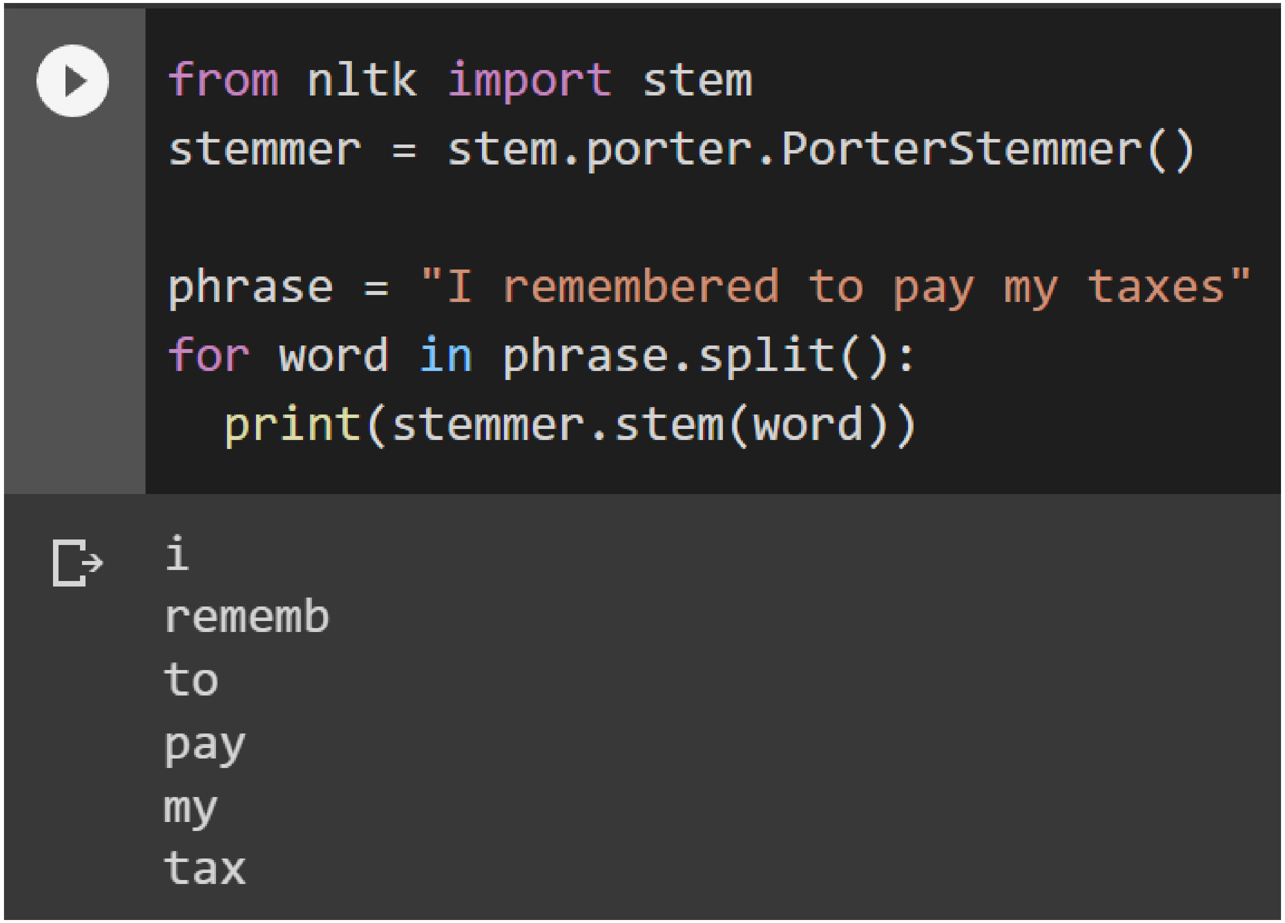
```python
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')
','.join(stopwords.words("english"))
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
'i,me,my,myself,we,our,ours,ourselves,you,you're,you've,you'll,you'd,your,yours,yourself,yourselves,he,him,h
is,himself,she,she's,her,hers,herself,it,it's,its,itself,they,them,their,theirs,themselves,what,which,who,wh
om,this,that,that'll,these,those,am,is,are,was,were,be,been,being,have,has,had,having,do,does,did,doing,a,a
n,the,and,but,if,or,because,as,until,while,of,at,by,for,with,about,against,between,into,through,during,befor
e,after,above,below,to,from,up,down,in,out,on,off,over,under,again,further,then,once,here,there,when,where,w
hy,how,all,any,both,each,few,more,most,other,some,such,no,nor,not,only,own,same,so,than,too,very,s,t,can,wil
l,just,don,don't,should,should've,now,d,ll,m,o,re,ve,y,ain,aren,aren't,couldn,couldn't,didn,didn't,doesn,doe
sn't,hadn,hadn't,hasn,hasn't,haven,haven't,isn,isn't,ma,mightn,mightn't,mustn,mustn't,needn,needn't,shan,sha
n't,shouldn,shouldn't,wasn,wasn't,weren,weren't,won,won't,wouldn,wouldn't'
```

Stemming and lemmatization…

- Stemming:
  - Determine the root or base form of a word.
- Lemmatization:
  - Findings words that all belong in the same group or lemma.
- For instance:
  - examples => example
  - running => run

Using the Porter Stemmer...



```
from nltk import stem
stemmer = stem.porter.PorterStemmer()

phrase = "I remembered to pay my taxes"
for word in phrase.split():
  print(stemmer.stem(word))
```

```
i
rememb
to
pay
my
tax
```

Different stemming techniques produce contrasting results
In the TM field the Porter Stemmer is one of
the least aggressive English language approaches.

In Summary

- Text mining and data mining are too closely related mathematical approaches used to derive insight from data.
- Text mining (TM) concerns the use of unstructured text as a data source, which will require suitable encoding to derive a quantitative representation.
- Preprocessing concerns reducing the number of data dimensions which we typically observe in conversational datasets.
- Preprocessing can involve steps such as stopword removal, correction of spelling, word stemming and lemmatization.
- High dimensionality can lead to increased model complexity and dampen associations.

End