新聞真假辨識

Real News? Fake News?

黃鉦程

2022.01.07

Outline

01 背景介紹

02 資料前處理

03 模型建構

04 結論

01　背景介紹

背景介紹

科技網路蓬勃發展，使得現在成為了資訊爆炸的時代，很多的問題都能夠透過網路搜索，獲取最新的資訊，但也衍生了假新聞的問題，訊息可能是不實資訊故意誤導大眾，會使帶風向的人帶來政治、經濟、心理成就感的新聞，社交媒體上傳播錯誤訊息可能導致暴力、自殺等問題，成為一種負面循環。

What

假新聞時有所聞

Where

網路、報紙

When

閱讀新聞的時候

Who

收看新聞的大眾

Why

減少資訊誤導、暴力、自殺行為發生

How

運用LSTM辨別新聞真假

02　資料前處理

資料前處理- 資料集介紹

原始資料集	
真新聞	21417筆
假新聞	23481筆

欄位			
標題	內文	主題	日期

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

資料前處理- 出版社去除、空值移除

- 真新聞有出版社的資訊，而假新聞沒有此一資訊，將文字與出版社去除。
- 將新聞標題與內文有嚴重缺失的移除。

```
real.head()
```

	title	text
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...

```
real.head()
```

	title	text
0	As U.S. budget fight looms, Republicans flip t...	The head of a conservative Republican faction...
1	U.S. military to accept transgender recruits o...	Transgender people will be allowed for the fi...
2	Senior U.S. Republican senator: 'Let Mr. Muell...	The special counsel investigation of links be...

```
[index for index,text in enumerate(real.text.values) if str(text).strip() == '']
```

```
[8970]
```

```
real.iloc[8970]
```

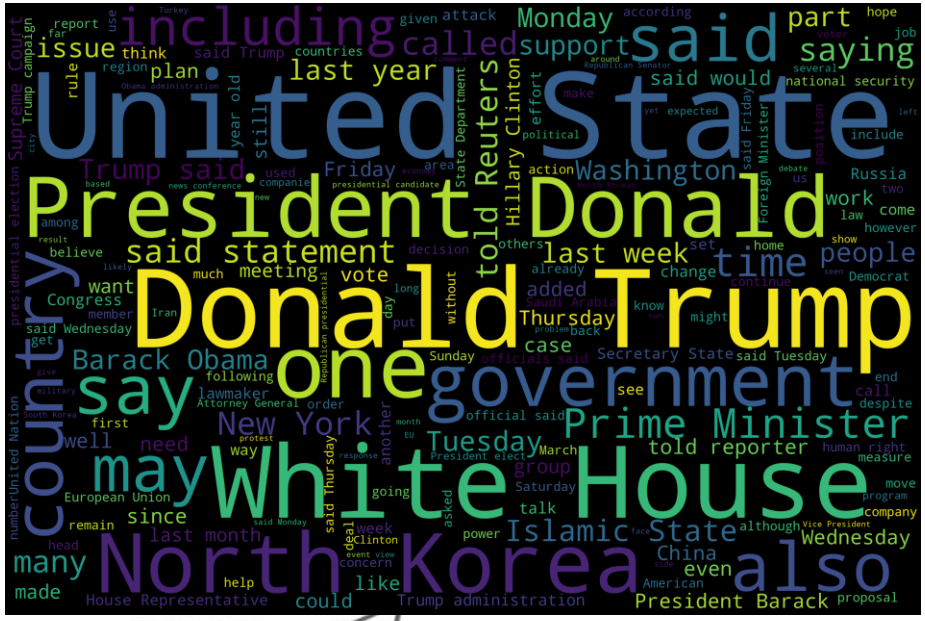
```
title      Graphic: Supreme Court roundup
text
subject                        politicsNews
date                         June 16, 2016
Name: 8970, dtype: object
```

```
real = real.drop(8970, axis=0)
```

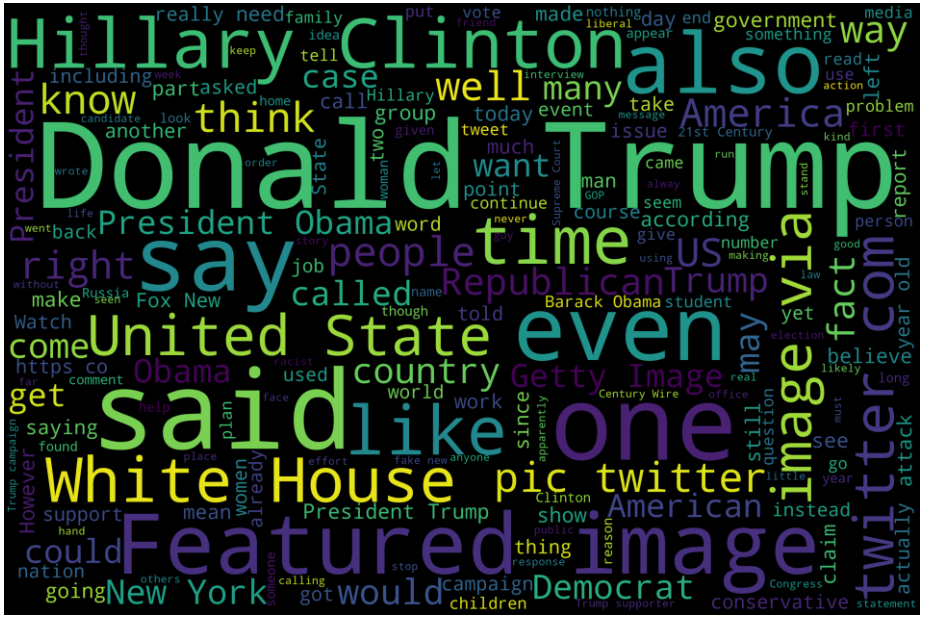
資料前處理- 字詞分析視覺化

利用文字雲了解資料集關鍵字，使用nltk的stopwords來去除停用詞。

真新聞



假新聞



```
text = ''
for news in fake.text.values:
        text += f" {news}"
wordcloud = WordCloud(
        width = 3000,
        height = 2000,
        background_color = 'black',
        stopwords = set(nltk.corpus.stopwords.words("english"))).generate(text)
fig = plt.figure(
        figsize = (40, 30),
        facecolor = 'k',
        edgecolor = 'k')
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```

資料前處理- 標籤化、分詞(Tokenization)、移除停用詞(stopwords)

- 資料標籤化。
- 標題與內文合併。
- 不需要的欄位移除。
- 將文字做分詞並且移除所有停用詞。

```
real["class"] = 1
fake["class"] = 0
```

```
real["text"] = real["title"] + " " + real["text"]
fake["text"] = fake["title"] + " " + fake["text"]
```

```
real = real.drop(["subject", "date","title", "publisher"], axis=1)
fake = fake.drop(["subject", "date", "title"], axis=1)
```

```
data = real.append(fake, ignore_index=True)
del real, fake
```

```
y = data["class"].values
X = []
stop_words = set(nltk.corpus.stopwords.words("english"))
tokenizer = nltk.tokenize.RegexpTokenizer(r'\w+')
for par in data["text"].values:
        tmp = []
        sentences = nltk.sent_tokenize(par)
        for sent in sentences:
                sent = sent.lower()
                tokens = tokenizer.tokenize(sent)
                filtered_words = [w.strip() for w in tokens if w not in stop_words and len(w) > 1]
                tmp.extend(filtered_words)
        X.append(tmp)
```

03　模型建構

模型建構- 向量化 (Word2Vec)

- One-hot encoding是資訊密度低、維度高的向量，word embedding 改成資訊密度高、維度低的向量來代表一個詞。

- Word2vec是一群用來產生詞向量的相關模型，可以訓練出自己的 word vectors，也可以用 pre-trained word vectors 查看 similar words。

```
#Dimension of vectors we are generating
EMBEDDING_DIM = 100
#Creating Word Vectors by Word2Vec Method (takes time...)
w2v_model = gensim.models.Word2Vec(sentences=X, size=EMBEDDING_DIM, window=5, min_count=1)
```

```
len(w2v_model.wv.vocab)
```

```
122248
```

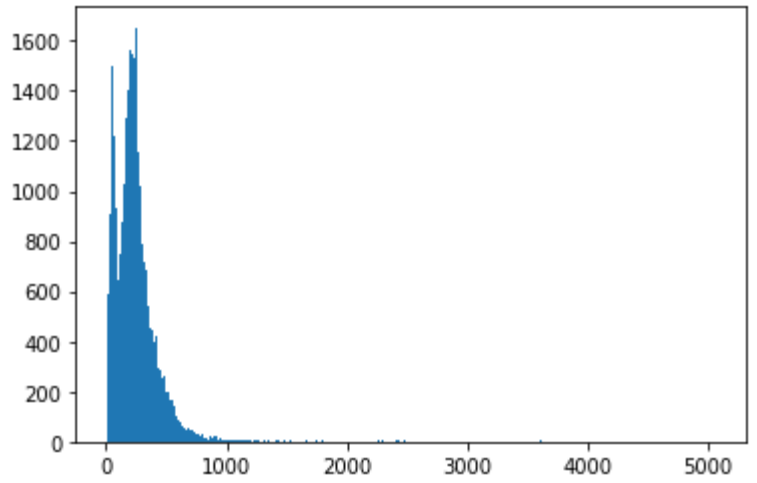
```
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X)

X = tokenizer.texts_to_sequences(X)
```

參數	
Size = 100,200	詞向量的維度大小，維度太小會無法有效表達詞與詞的關係。
Window = 5	決定Word2Vec一次取多少詞來預測中間詞。
Min_count = 1	出現次數大於等於min_count的詞，才會納入Word2Vec的詞典中。

模型建構-多對一模型(many to one) 輸入字數固定

- 利用pad_sequences()函式，使每個序列擁有相同的長度。
- 將每個序列都刪減或補值成700個字。

```
plt.hist([len(x) for x in X], bins=500)
plt.show()
```




```
nos = np.array([len(x) for x in X])
len(nos[nos < 700])
```

```
43982
```

```
maxlen = 700
#Making all news of size maxlen defined above
X = pad_sequences(X, maxlen=maxlen)
```

模型建構- LSTM model

```
#Defining  Neural  Network
model  =  Sequential()
#Non-trainable  embeddidng  layer
model.add(Embedding(vocab_size,  output_dim=EMBEDDING_DIM,  weights=[embedding_vectors],  input_length=maxlen,  trainable=False))
#LSTM
model.add(LSTM(units=128))
model.add(Dense(1,  activation='sigmoid'))
model.compile(optimizer='adam',  loss='binary_crossentropy',  metrics=['acc'])
```

```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 700, 100)          12224900

 lstm (LSTM)                 (None, 128)               117248

 dense (Dense)               (None, 1)                 129

=================================================================
Total params: 12,342,277
Trainable params: 117,377
Non-trainable params: 12,224,900
_________________________________________________________________
```

模型建構- LSTM model

參數	
Embedding	(122249,100,700)
LSTM	(128)
Dense	(1,sigmoid)
Loss_function	binary_crossentropy
optimizer	adam
epoch	6
Loss	0.0295
accuracy	0.99082

參數	
Embedding	(122249,200,700)
LSTM	(128)
Dense	(1,sigmoid)
Loss_function	binary_crossentropy
optimizer	adam
epoch	6
Loss	0.0295
accuracy	0.98788

04　結論

結論

- 透過模型應用，可以辨別新聞真假，讓人民能夠不被誤導。
- Word embedding 維度越高效率精度會越低。

未來展望

- 蒐集國內中文新聞，並透過自然語言處理，也能夠訓練出模型。