FRAUD DETECTION IN PYTHON

Using text data to detect fraud

Charlotte Werger
Data Scientist

You will often encounter text data during fraud detection

Types of useful text data:

1. Emails from employees and/or clients
2. Transaction descriptions
3. Employee notes
4. Insurance claim form description box
5. Recorded telephone conversations
6. ...

Text mining techniques for fraud detection

1. Word search
2. Sentiment analysis
3. Word frequencies and topic analysis
4. Style

Word search for fraud detection

Flagging suspicious words:

1. Simple, straightforward and easy to explain
2. Match results can be used as a filter on top of machine learning model
3. Match results can be used as a feature in a machine learning model

Word counts to flag fraud with pandas

```
# Using a string operator to find words
df['email_body'].str.contains('money laundering')
```

```
# Select data that matches
df.loc[df['email_body'].str.contains('money laundering', na=False)]
```

```
# Create a list of words to search for
list_of_words = ['police', 'money laundering']
df.loc[df['email_body'].str.contains('|'.join(list_of_words)
, na=False)]
```

```
# Create a fraud flag
df['flag'] = np.where((df['email_body'].str.contains('|'.join
(list_of_words)) == True), 1, 0)
```

FRAUD DETECTION IN PYTHON

Let's practice!

FRAUD DETECTION IN PYTHON

Text mining techniques for fraud detection

Charlotte Werger
Data Scientist

Cleaning your text data

Must do's when working with textual data:

1. Tokenization
2. Remove all stopwords
3. Lemmatize your words
4. Stem your words

Go from this...

	headline_text	index
0	aba decides against community broadcasting lic...	0
1	act fire witnesses must be aware of defamation	1
2	a g calls for infrastructure protection summit	2
3	air nz staff in aust strike for pay rise	3
4	air nz strike to affect australian travellers	4

To this...

```
0                [decid, communiti, broadcast, licenc]
1                                   [wit, awar, defam]
2               [call, infrastructur, protect, summit]
3                          [staff, aust, strike, rise]
4                [strike, affect, australian, travel]
5                   [ambiti, olsson, win, tripl, jump]
6               [antic, delight, record, break, barca]
7        [aussi, qualifi, stosur, wast, memphi, match]
8                [aust, address, secur, council, iraq]
9                             [australia, lock, timet]
Name: headline_text, dtype: object
```

Data preprocessing part 1

```
# 1. Tokenization
from nltk import word_tokenize

text = df.apply(lambda row: word_tokenize(row["email_body"]), axis=1)
text = text.rstrip()
text = re.sub(r'[^a-zA-Z]', ' ', text)
```

```
# 2. Remove all stopwords and punctuation
from nltk.corpus import stopwords
import string

exclude = set(string.punctuation)
stop = set(stopwords.words('english'))
stop_free = " ".join([word for word in text
           if((word not in stop) and (not word.isdigit()))])
punc_free = ''.join(word for word in stop_free
           if word not in exclude)
```

Data preprocessing part 2

```
# Lemmatize words
from nltk.stem.wordnet import WordNetLemmatizer
lemma = WordNetLemmatizer()
normalized = " ".join(lemma.lemmatize(word) for word in punc_free.split())
```

```
# Stem words
from nltk.stem.porter import PorterStemmer
porter= PorterStemmer()
cleaned_text = " ".join(porter.stem(token) for token in normalized.split())
```

```
print (cleaned_text)

['philip','going','street','curious','hear','perspective','may','wish',
'offer','trading','floor','enron','stock','lower','joined','company',
'business','school','imagine','quite','happy','people','day','relate',
'somewhat','stock','around','fact','broke','day','ago','knowing',
'imagine','letting','event','get','much','taken','similar',
'problem','hope','everything','else','going','well','family','knee',
'surgery','yet','give','call','chance','later']
```

FRAUD DETECTION IN PYTHON

Let's practice!

FRAUD DETECTION IN PYTHON

Topic modelling

Charlotte Werger
Data Scientist

Topic modelling: discover hidden patterns in text data

1. Discovering topics in text data
2. "What is the text about"
3. Conceptually similar to clustering data
4. Compare topics of fraud cases to non-fraud cases and use as a feature or flag
5. Or.. is there a particular topic in the data that seems to point to fraud?

Latent Dirichlet Allocation (LDA)

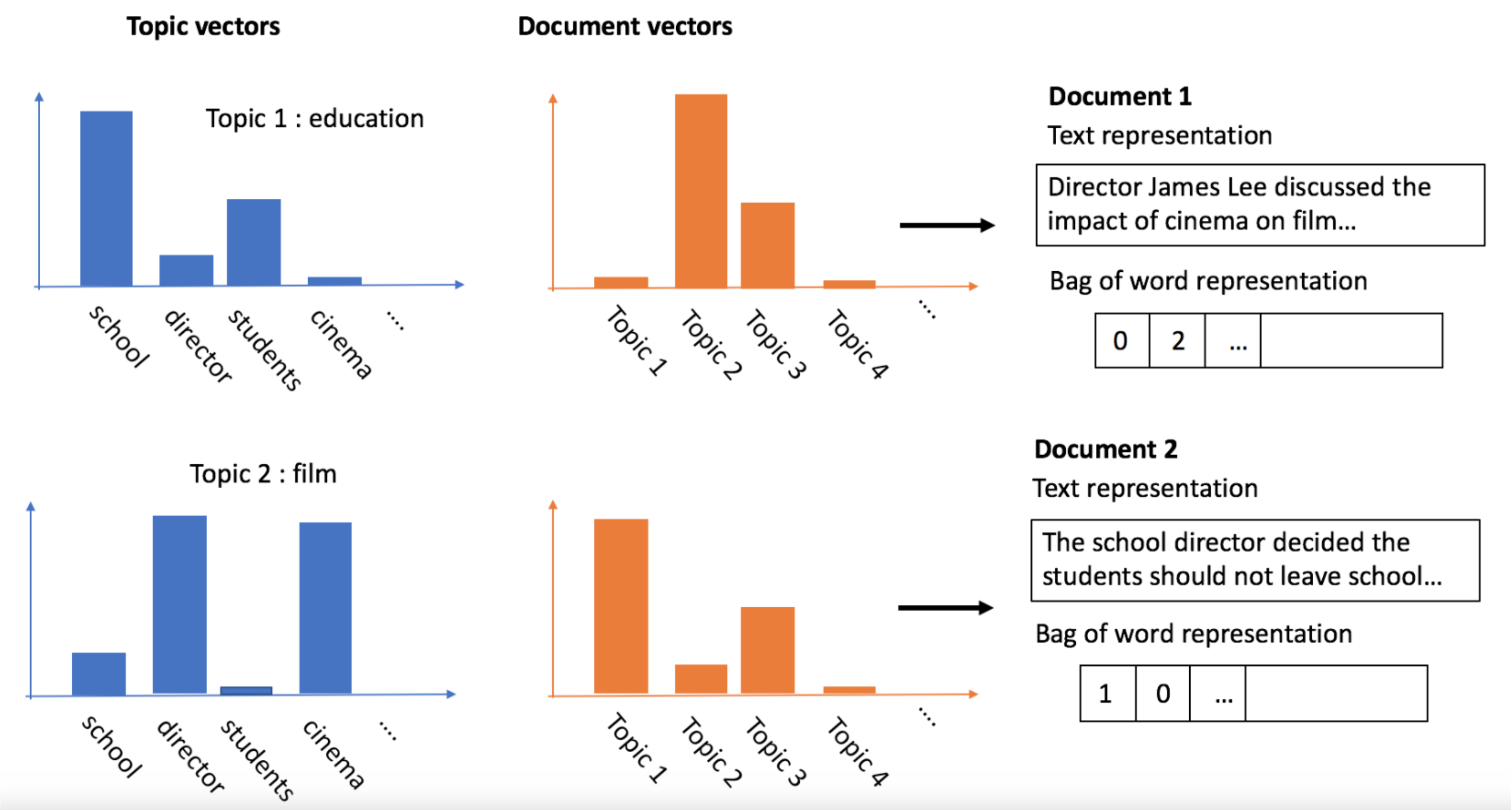
With LDA you obtain:

1. "topics per text item" model (i.e. probabilities)
2. "words per topic" model

Creating your own topic model:

1. Clean your data
2. Create a bag of words with dictionary and corpus
3. Feed dictionary and corpus into the LDA model

Latent Dirichlet Allocation (LDA)



Bag of words: dictionary and corpus

```
from gensim import corpora
```

```
# Create dictionary number of times a word appears
dictionary = corpora.Dictionary(cleaned_emails)
```

```
# Filter out (non)frequent words
dictionary.filter_extremes(no_below=5, keep_n=50000)
```

```
# Create corpus
corpus = [dictionary.doc2bow(text) for text in cleaned_emails]
```

Latent Dirichlet Allocation (LDA) with gensim

```
import gensim

# Define the LDA model
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = 3,
id2word=dictionary, passes=15)
```

```
# Print the three topics from the model with top words
topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)

(0, '0.029*"email" + 0.016*"send" + 0.016*"results" + 0.016*"invoice"')
(1, '0.026*"price" + 0.026*"work" + 0.026*"management" + 0.026*"sell"')
(2, '0.029*"distribute" + 0.029*"contact" + 0.016*"supply" + 0.016*"fast"')
```

FRAUD DETECTION IN PYTHON

Let's practice!

FRAUD DETECTION IN PYTHON

Flagging fraud based on topics

Charlotte Werger
Data Scientist

Using your LDA model results for fraud detection

1. Are there any suspicious topics? (no labels)
2. Are the topics in fraud and non-fraud cases similar? (with labels)
3. Are fraud cases associated more with certain topics? (with labels)

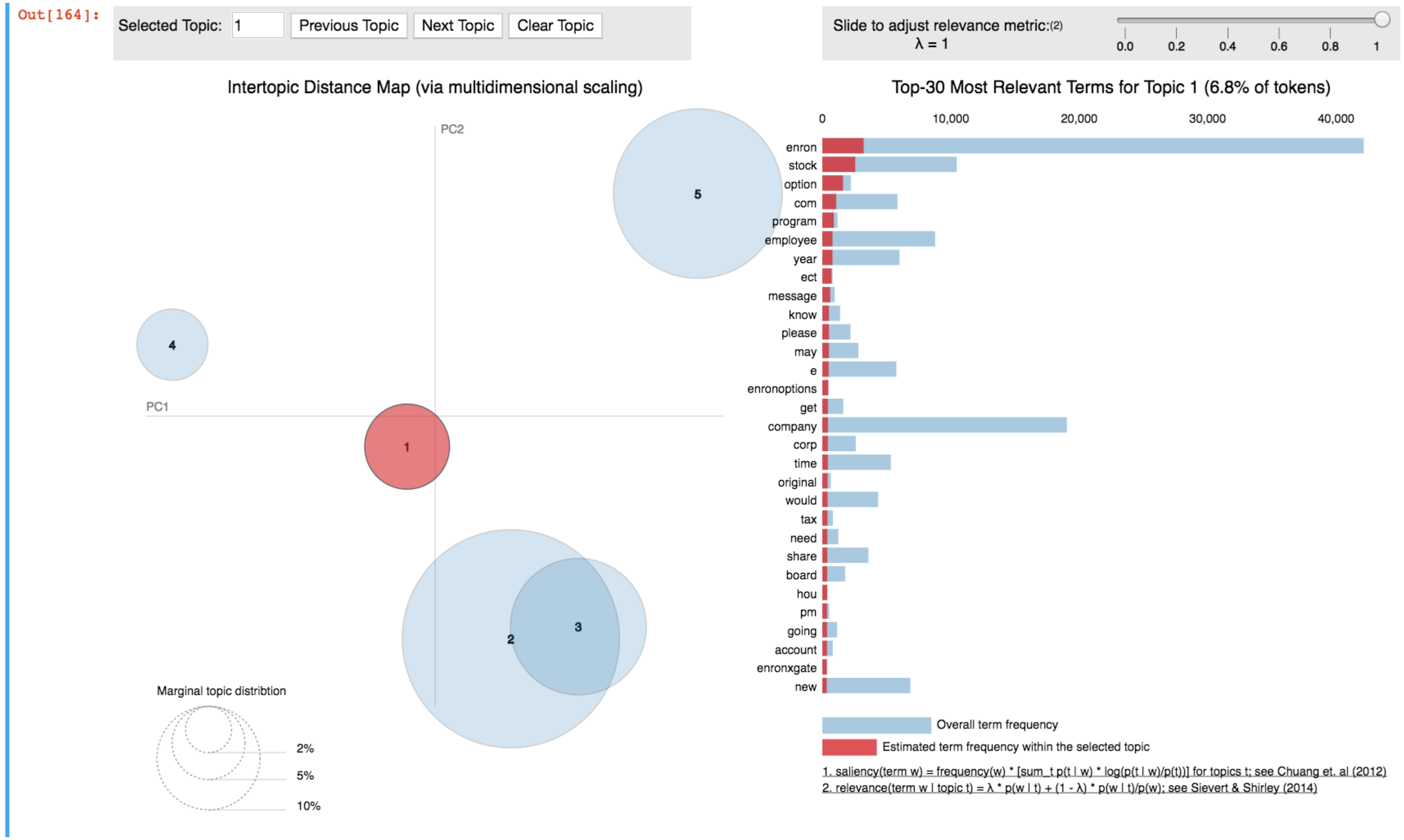
To understand topics, you need to visualize

```
import pyLDAvis.gensim
```

```
lda_display = pyLDAvis.gensim.prepare(ldamodel, corpus,
                                      dictionary, sort_topics=False)
```

```
pyLDAvis.display(lda_display)
```

Inspecting how topics differ



Assign topics to your original data

```
def get_topic_details(ldamodel, corpus):
    topic_details_df = pd.DataFrame()
    for i, row in enumerate(ldamodel[corpus]):
        row = sorted(row, key=lambda x: (x[1]), reverse=True)
        for j, (topic_num, prop_topic) in enumerate(row):
            if j == 0:  # => dominant topic
                wp = ldamodel.show_topic(topic_num)
                topic_details_df = topic_details_df.append(pd.Series([topic_num,
    topic_details_df.columns = ['Dominant_Topic', '% Score']
    return topic_details_df
```

```
contents = pd.DataFrame({'Original text':text_clean})
topic_details = pd.concat([get_topic_details(ldamodel,
                           corpus), contents], axis=1)
topic_details.head()
```

```
    Dominant_Topic     % Score     Original text
0     0.0              0.989108    [investools, advisory, free, ...
1     0.0              0.993513    [forwarded, richard, b, ...
2     1.0              0.964858    [hey, wearing, target, purple, ...
3     0.0              0.989241    [leslie, milosevich, santa, clara, ...
```

FRAUD DETECTION IN PYTHON

Let's practice!

FRAUD DETECTION IN PYTHON

Fraud detection in Python
Recap

Charlotte Werger
Data Scientist

Working with imbalanced data

- Worked with highly imbalanced fraud data
- Learned how to resample your data
- Learned about different resampling methods

Fraud detection with labeled data

- Refreshed supervised learning techniques to detect fraud
- Learned how to get reliable performance metrics and worked with the precision recall trade-off
- Explored how to optimise your model parameters to handle fraud data
- Applied ensemble methods to fraud detection

Fraud detection without labels

- Learned about the importance of segmentation
- Refreshed your knowledge on clustering methods
- Learned how to detect fraud using outliers and small clusters with K-means clustering
- Applied a DB-scan clustering model for fraud detection

Text mining for fraud detection

- Know how to augment fraud detection analysis with text mining techniques
- Applied word searches to flag use of certain words, and learned how to apply topic modelling for fraud detection
- Learned how to effectively clean messy text data

Further learning for fraud detection

- Network analysis to detect fraud
- Different supervised and unsupervised learning techniques (e.g. Neural Networks)
- Working with very large data

FRAUD DETECTION IN PYTHON

End of this course