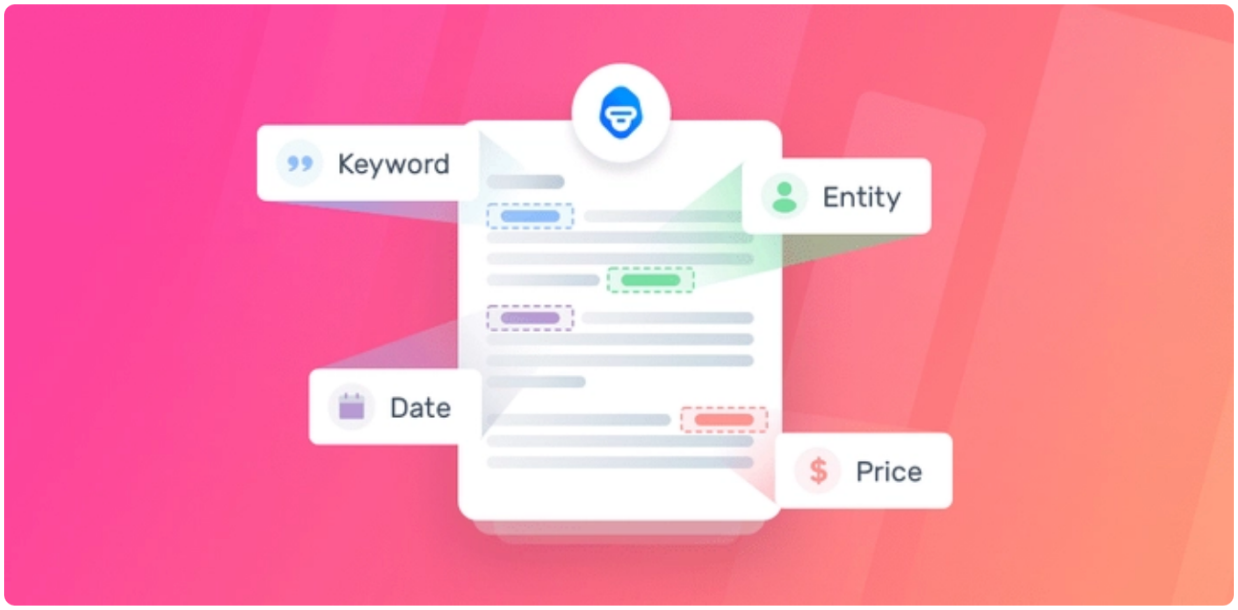
Entities, Keywords and Phrases Extracting Using Watson NLP

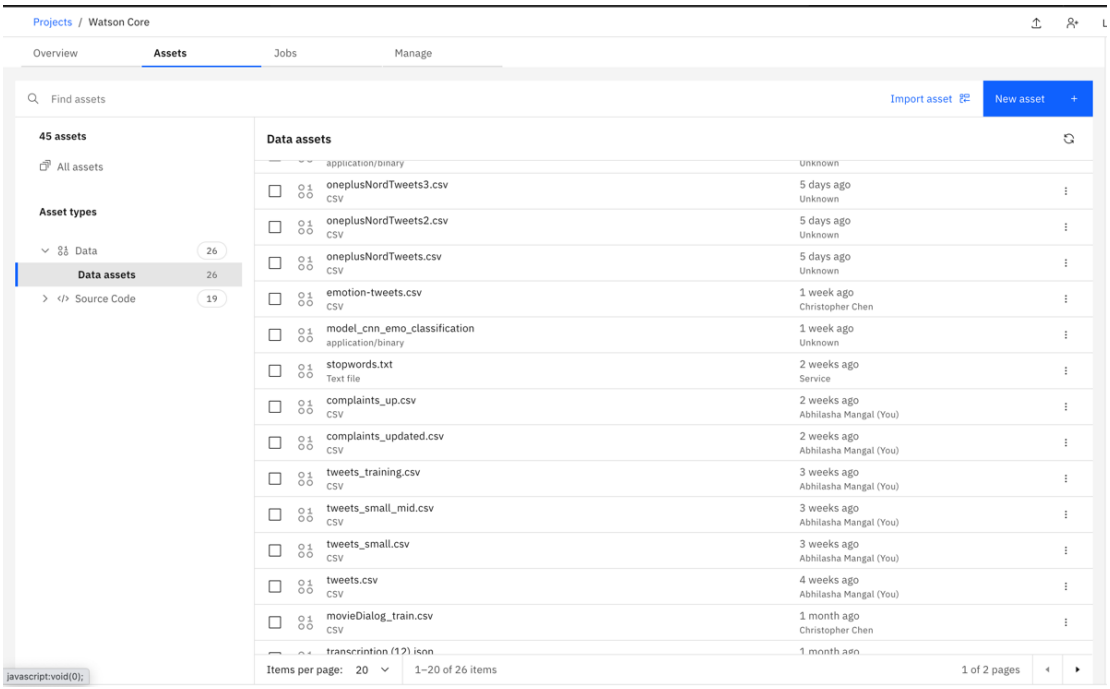
Entities, Keyword and Phrase Extraction play key roles to understanding unstructured text data. By using these techniques, we can find out from the text which entities, keywords and phrases are the most important. These entities can include **People name, Organization name, Date, Price, Facility** etc. it is also called **Named Entity Extraction**.



This blog demonstrates how to analyse **HOTEL Reviews** using Watson NLP step by step.

1. Collecting the dataset

The data used in this notebook is the data scraped from Booking.com and TripAdvisor. This dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe. A review contains the customer's narrative description of their experience. In this blog, you will focus on detecting entity mentions and phrases in this narrative description. The data is sourced from the publicly available Kaggle 515K Hotel Reviews Data in Europe and OpinRank Review Dataset. The dataset used in this notebook has combined and transformed both datasets for a cleaner and simpler approach to the data preparation step.



The data for each of the hotels can be downloaded from GitHub: Hotel 1, Hotel 2, Hotel 3. Since these CSVs should already be in the reserved environment, we will use the Project library to load all three files as Data Frames.

	date	text	website	hotel
0	2017-07-13	Although I appreciate this is a 4 star and not a 5 star it is nowhere near the standard of the other Firmdale hotels we have stayed at Room we had looked over a busy road not the leafy garden square and had just a shower no bath Very friendly staff good cocktails in bar	Booking.com	Dorset
1	2017-06-26	A night time bottle of water on the second night or something similar Beautiful furnishings comfortable bed gorgeous view of Dorset Square very pleasant staff	Booking.com	Dorset
2	2017-06-12	Our room was very small and had no view Tea and coffee facilities in the room would be an added bonus The hotel is beautifully decorated The staff was very friendly and accommodating The restaurant had a good breakfast selection I can definitely recommend the hotel to anybody who would like a boutique hotel experience The hotel is within walking distance from Lords cricket grounds The mini ...	Booking.com	Dorset
3	2017-06-06	The bed was a bit small and short given that we are rather tall people Even though we thoroughly enjoyed breakfast the selection could have benefitted from also some more savoury items next to the egg varieties such as some cheese cold cuts Nothing really to find fault with Overall a very well maintained property Beautiful room although a bit small with a nice view of the square Equally bea...	Booking.com	Dorset
4	2017-05-29	We didnt like our first room but the staff traded our room when we asked They were as helpful as they could be while being totally booked up We loved the location as well as the staff Lorenzo and Sarah from the front desk were incredibly helpful as well as Virgil and all the others we had a wonderful time	Booking.com	Dorset

2. Data processing and Exploratory Data Analysis

2.1 Text Pre-processing

Our first step is to pre-process the documents in a way that cleans distracting signals and makes them easier to process and analyse. This is a standard step in many NLP pipelines. Here we perform three types of pre-processing:

1. Stop-words filtering:
   To remove stop-words we used **Watson NLP pre-defined list**. We can remove and extend this stop-words list. You can download this list by using ‘download_and_load’method of Watson NLP library.

```
wnlp_stop_words = watson_nlp.download_and_load('text_stopwords_classification_ensemble_en_stock').stopwords
```

2. Remove some Patterns:
   This dataset has consumer personal information. That is hidden by pattern of 'XX/XX/XXXX' or another format. Replace all the patterns with blanks.

3. Entities Extraction

Entity extraction uses the entity-mentions block to encapsulate algorithms for the task of extracting mentions of entities (person, organizations, dates, locations,...) from the input text. The block offers implementations of strong entity extraction algorithms from each of the four families: **rule-based, classic ML, deep-learning,** and **transformers**.

There are two types of models:

1. A **rule-based model** (the rbr models), which handles syntactically regular entity types such as number, email and phone.
2. A **model trained on labelled data** for the more complex entity types such as person, organization location.

   To extract the entities, we are using pre-trained Watson NLP models. You can download and use these models by using below steps:

```
# Load a syntax model to split the text into sentences and tokens
syntax_model = watson_nlp.load(watson_nlp.download('syntax_izumo_en_stock'))
# Load bilstm model in WatsonNLP
bilstm_model = watson_nlp.load(watson_nlp.download('entity-mentions_bilstm_en_stock'))
# Load rbr model in WatsonNLP
rbr_model = watson_nlp.load(watson_nlp.download('entity-mentions_rbr_en_stock'))
# Load bert model in WatsonNLP
bert_model = watson_nlp.load(watson_nlp.download('entity-mentions_bert_multi_stock'))
# Load transformer model in WatsonNLP
#transformer_model = watson_nlp.load(watson_nlp.download('entity-mentions_transformer_multi_stock'))
```

After loading the model, we can extract entities using **run()** method of Watson NLP. We are able to see different types of entities like Location, Organization, Facility etc.

|:

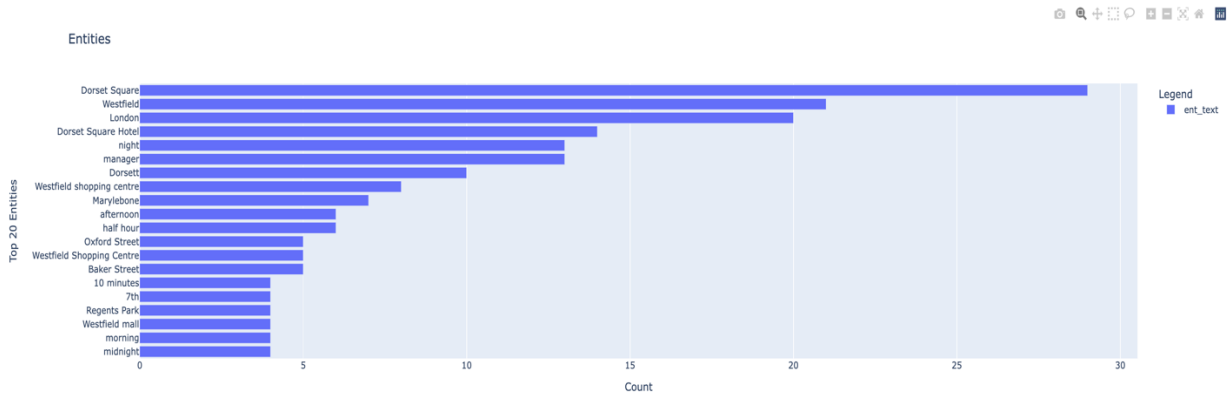
	Document	Hotel Name	Website	ent_type	ent_text
2578	Dorset Square Hotel conveniently located, moderately priced luxury hotel. blocks Wax museum London. street pleasant, peaceful park named (what else) Dorset Square, adds quiet ambiance. proximity busy streets tourist attractions.We nights April, 2000 enjoyed stay. restaurant served excellent breakfast, offered English European fare. Buses taxi's easy catch, great night spots short distance away...	Dorset	TripAdvisor	Location	street pleasant
2579	Dorset Square Hotel conveniently located, moderately priced luxury hotel. blocks Wax museum London. street pleasant, peaceful park named (what else) Dorset Square, adds quiet ambiance. proximity busy streets tourist attractions.We nights April, 2000 enjoyed stay. restaurant served excellent breakfast, offered English European fare. Buses taxi's easy catch, great night spots short distance away...	Dorset	TripAdvisor	Facility	Dorset Square
2580	Dorset Square Hotel conveniently located, moderately priced luxury hotel. blocks Wax museum London. street pleasant, peaceful park named (what else) Dorset Square, adds quiet ambiance. proximity busy streets tourist attractions.We nights April, 2000 enjoyed stay. restaurant served excellent breakfast, offered English European fare. Buses taxi's easy catch, great night spots short distance away...	Dorset	TripAdvisor	Date	April, 2000
2581	Dorset Square Hotel conveniently located, moderately priced luxury hotel. blocks Wax museum London. street pleasant, peaceful park named (what else) Dorset Square, adds quiet ambiance. proximity busy streets tourist attractions.We nights April, 2000 enjoyed stay. restaurant served excellent breakfast, offered English European fare. Buses taxi's easy catch, great night spots short distance away...	Dorset	TripAdvisor	Organization	English European
2582	Dorset Square Hotel conveniently located, moderately priced luxury hotel. blocks Wax museum London. street pleasant, peaceful park named (what else) Dorset Square, adds quiet ambiance. proximity busy streets tourist attractions.We nights April, 2000 enjoyed stay. restaurant served excellent breakfast, offered English European fare. Buses taxi's easy catch, great night spots short distance away...	Dorset	TripAdvisor	Facility	away.This hotel

The model will output a text's entity mention as well as its category of entity. For example, "london" mention is a **Location type** and "good soundproof rooms" is a **Facility type.**

3.1 Analysis on Each Hotel Data

By using the above model output, we can compare, amongst hotels, which type of facilities are most frequently used by customers:

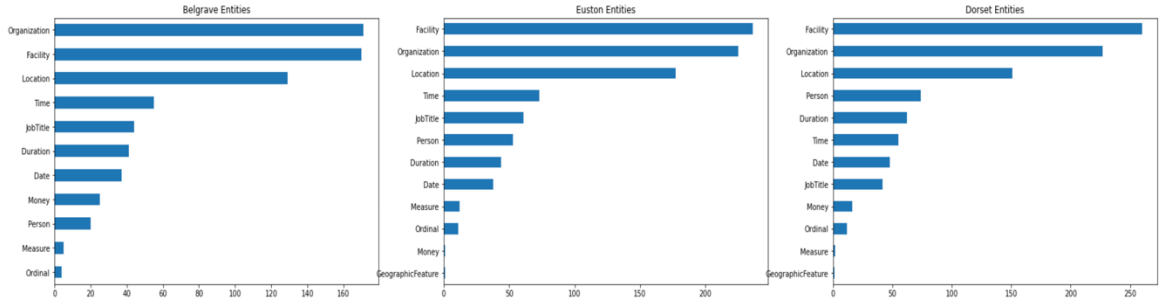
3.1.1 Top 20 Entities Values Hotel



It can be observed that most people are talking about '**Dorset Square'** and **'West Field'**.

3.1.2 Comparison in between 3 Hotels Top 20 Entities:

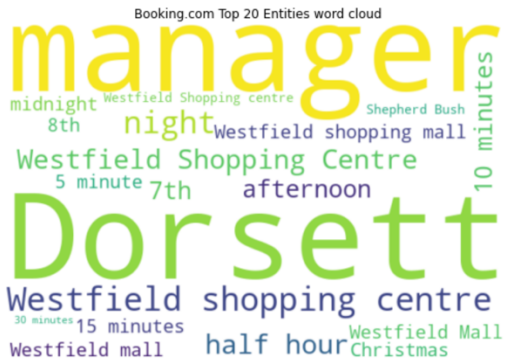
```
[19]: fig, axs = plt.subplots(nrows=1, ncols=3, figsize=(30, 6))
entities_df[entities_df['Hotel Name'] == 'Belgrave']['ent_type'].value_counts().head(20).sort_values().plot(kind='barh', ax=axs[0], title='Belgrave Entities')
entities_df[entities_df['Hotel Name'] == 'Euston']['ent_type'].value_counts().head(20).sort_values().plot(kind='barh', ax=axs[1], title='Euston Entities')
entities_df[entities_df['Hotel Name'] == 'Dorset']['ent_type'].value_counts().head(20).sort_values().plot(kind='barh', ax=axs[2], title='Dorset Entities')

plt.show()
```

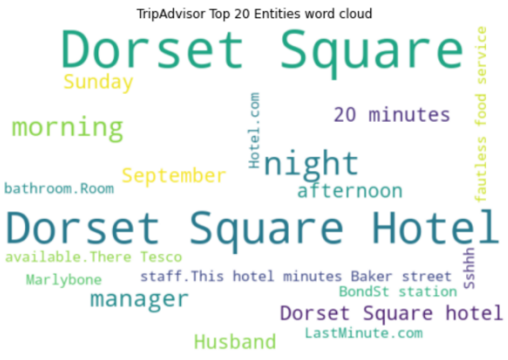


We can see that the most importantly reviewed attributes of a hotel are related to the facility, location, and organization. These are areas that management can look to target in more detail to understand what can be improved.

3.1.3 Comparison between Booking.com vs TripAdvisor for one hotel:

The below word cloud is created for '**Dorset Hotel':**





We can use this collective information to give priority to the website with reviews that better align with our own preferences about choosing a hotel. Do we care more about the convenience of the location of a hotel or do we care about the hotel's ambience, reception, perks?

4. Keyword and Noun Phrase Extraction:

To extract the noun phrase and key word extraction, we use the pre-trained models **noun-phrases_rbr_en_stock** and **'keywords_text-rank_en_stock'** of Watson NLP for English language**.** You can download and load these models by using below steps:

```
noun_phrases_model = watson_nlp.load(watson_nlp.download('noun-phrases_rbr_en_stock'))

keywords_model = watson_nlp.load(watson_nlp.download('keywords_text-rank_en_stock'))
```

By using above model, we are collecting all Noun and Keyword-phrases for each document. Extract keywords and noun phrases by passing the output of syntax model into the keywords and noun phrases model as shown below:

```
# Run the Syntax and Noun Phrases models
    syntax_prediction = syntax_model.run(text, parsers=('token', 'lemma',
'part_of_speech'))
    noun_phrases = noun_phrases_model.run(text)
    # Run the keywords model
    keywords = keywords_model.run(syntax_prediction, noun_phrases, limit=
5)
    keywords_list =keywords.to_dict()['keywords']
```

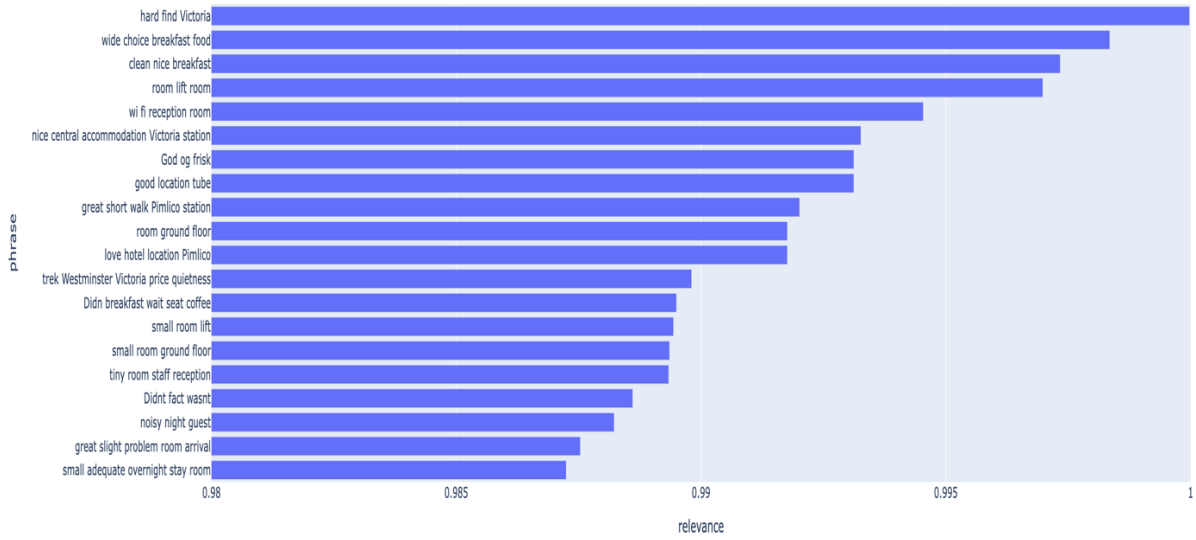
After collecting all Noun-phrases and Keywords, create a dataframe to observe the most frequent phrases in the dataset.

7]:

	Complaint data	phrase	relevance	phrase_length
0	noisy property Workmen drilling renovations breakfast room sound made leave enjoy breakfast offerings loud hotel asked workmen begin work area breakfast service understand renovations undertaken care relation guest comfort disappointing listed booking website Hilton website renovations taking place bed comfortable room clean wifi strong staff members lovely welcoming location close Euston Stat...	renovation breakfast room sound	0.871282	4
0	noisy property Workmen drilling renovations breakfast room sound made leave enjoy breakfast offerings loud hotel asked workmen begin work area breakfast service understand renovations undertaken care relation guest comfort disappointing listed booking website Hilton website renovations taking place bed comfortable room clean wifi strong staff members lovely welcoming location close Euston Stat...	noisy property workman	0.721332	3
0	noisy property Workmen drilling renovations breakfast room sound made leave enjoy breakfast offerings loud hotel asked workmen begin work area breakfast service understand renovations undertaken care relation guest comfort disappointing listed booking website Hilton website renovations taking place bed comfortable room clean wifi strong staff members lovely welcoming location close Euston Stat...	book website Hilton website renovation	0.657340	5
0	noisy property Workmen drilling renovations breakfast room sound made leave enjoy breakfast offerings loud hotel asked workmen begin work area breakfast service understand renovations undertaken care relation guest comfort disappointing listed booking website Hilton website renovations taking place bed comfortable room clean wifi strong staff members lovely welcoming location close Euston Stat...	work area breakfast service	0.594659	4
0	noisy property Workmen drilling renovations breakfast room sound made leave enjoy breakfast offerings loud hotel asked workmen begin work area breakfast service understand renovations undertaken care relation guest comfort disappointing listed booking website Hilton website renovations taking place bed comfortable room clean wifi strong staff members lovely welcoming location close Euston Stat...	care relation guest comfort	0.588937	4
...	...	...	...	...
591	stayed hotel 3 nights October paid advance stayed 3 minutes. stayed hotels places years honestly worst. room small, clean (bedcover blanket stained carpet dirty) sheet pillow clean changed daily. shower batroom reasonable corridors stairs dirty badly coat paint. night change room 3.00am couldnot wink sleep due noise radiator. argument room bit left lot desired -there chair room. Breakfast poor...	argument room bit	0.681963	3
591	stayed hotel 3 nights October paid advance stayed 3 minutes. stayed hotels places years honestly worst. room small, clean (bedcover blanket stained carpet dirty) sheet pillow clean changed daily. shower batroom reasonable corridors stairs dirty badly coat paint. night change room 3.00am couldnot wink sleep due noise radiator. argument room bit left lot desired -there chair room. Breakfast poor...	night change room	0.679218	3
592	selected place hotel listing web site, describing modern haven relaxation enticing terms. arrived July 11th weary journey forward nice relaxing evening. felt uneasy mopment touched front door streaked dirt. reception gloomythought young staff pleasant enough. literally utter reservation credit card requested swiped full amount 4 stay. stage smelt rat pay advance this. horror discovered reason ...	place hotel listing web site	0.626401	5
592	selected place hotel listing web site, describing modern haven relaxation enticing terms. arrived July 11th weary journey forward nice relaxing evening. felt uneasy mopment touched front door streaked dirt. reception gloomythought young staff pleasant enough. literally utter reservation credit card requested swiped full amount 4 stay. stage smelt rat pay advance this. horror discovered reason ...	utter reservation credit card	0.591723	4
592	selected place hotel listing web site, describing modern haven relaxation enticing terms. arrived July 11th weary journey forward nice relaxing evening. felt uneasy mopment touched front door streaked dirt. reception gloomythought young staff pleasant enough. literally utter reservation credit card requested swiped full amount 4 stay. stage smelt rat pay advance this. horror discovered reason ...	refund internet company	0.587138	3

1116 rows × 4 columns

We applied some pre-processing steps to the above dataframe to calculate the length of each phrase and keyword. We removed 1-gram and bi-grams from this dataframe and we had collected Top 20 Keywords and Phrases to get the most frequent keywords and phrases. Here, we can say customers mostly talked about “hard find victoria” and “wide choice breakfast food”.



5. Conclusion

We have seen how easily we are able to analyse Hotel Reviews dataset by using Watson NLP. This entities and keywords extraction exercise can be used to understand most frequently discussed aspects (in terms of keywords and phrases) in their reviews. Based on this information, they can target the areas where they need to improve or work around. The next blog will cover model deployment to show how easily you can apply these models anywhere.