Handling unstructured data in real estate research projects

ESCP Lecture

Stephen Ryan 18 May 2021

Two recent research projects

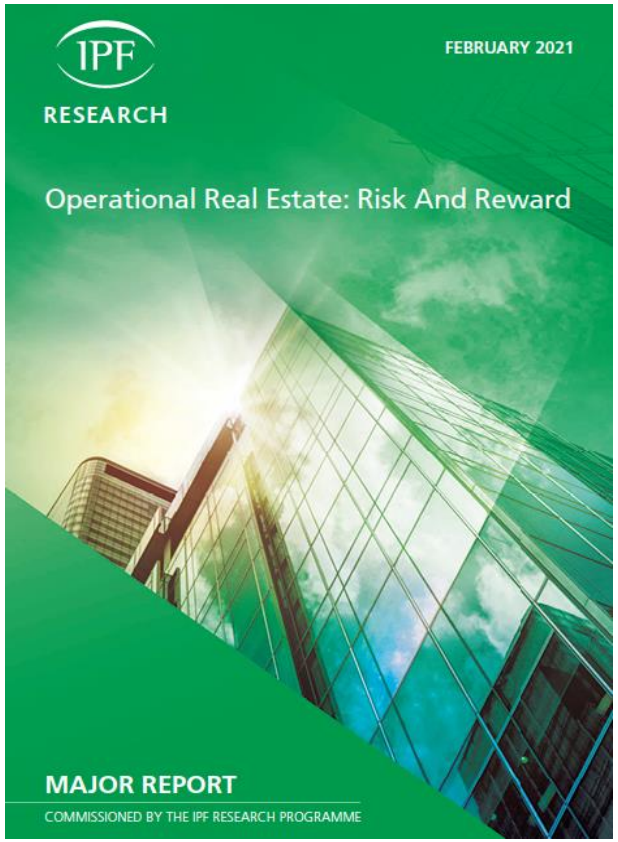
Very different topics but unstructured text data featured in both

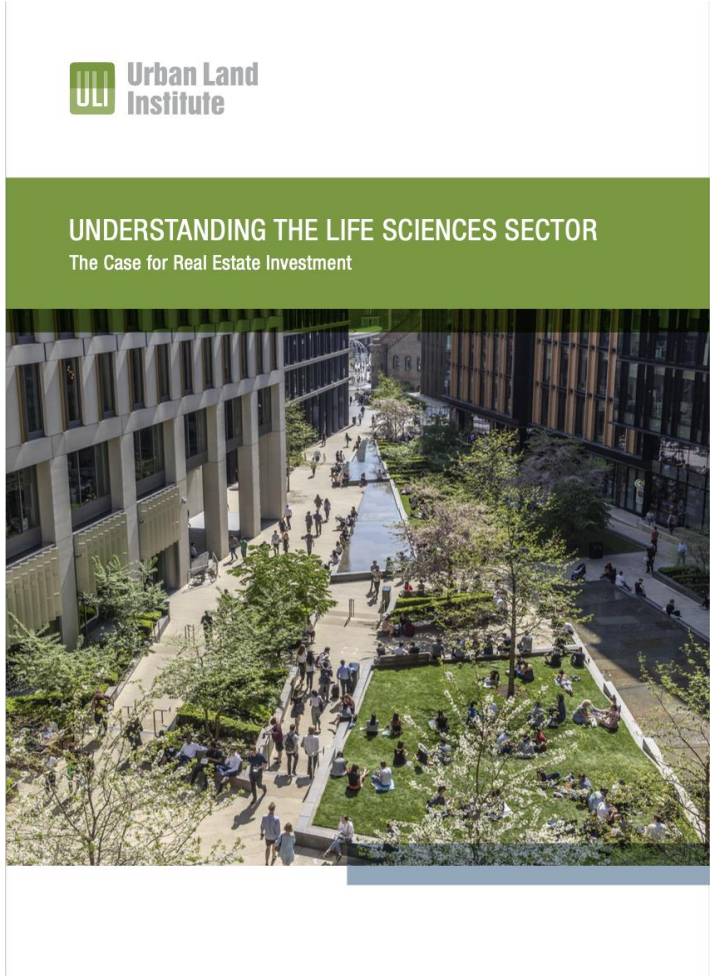
S RYAN INVEST

Project 1
Operational real estate: risk and reward

Project 2
Understanding the life sciences sector: the case for real estate investment

Both co-written with Didobi and both published in Q1 2021

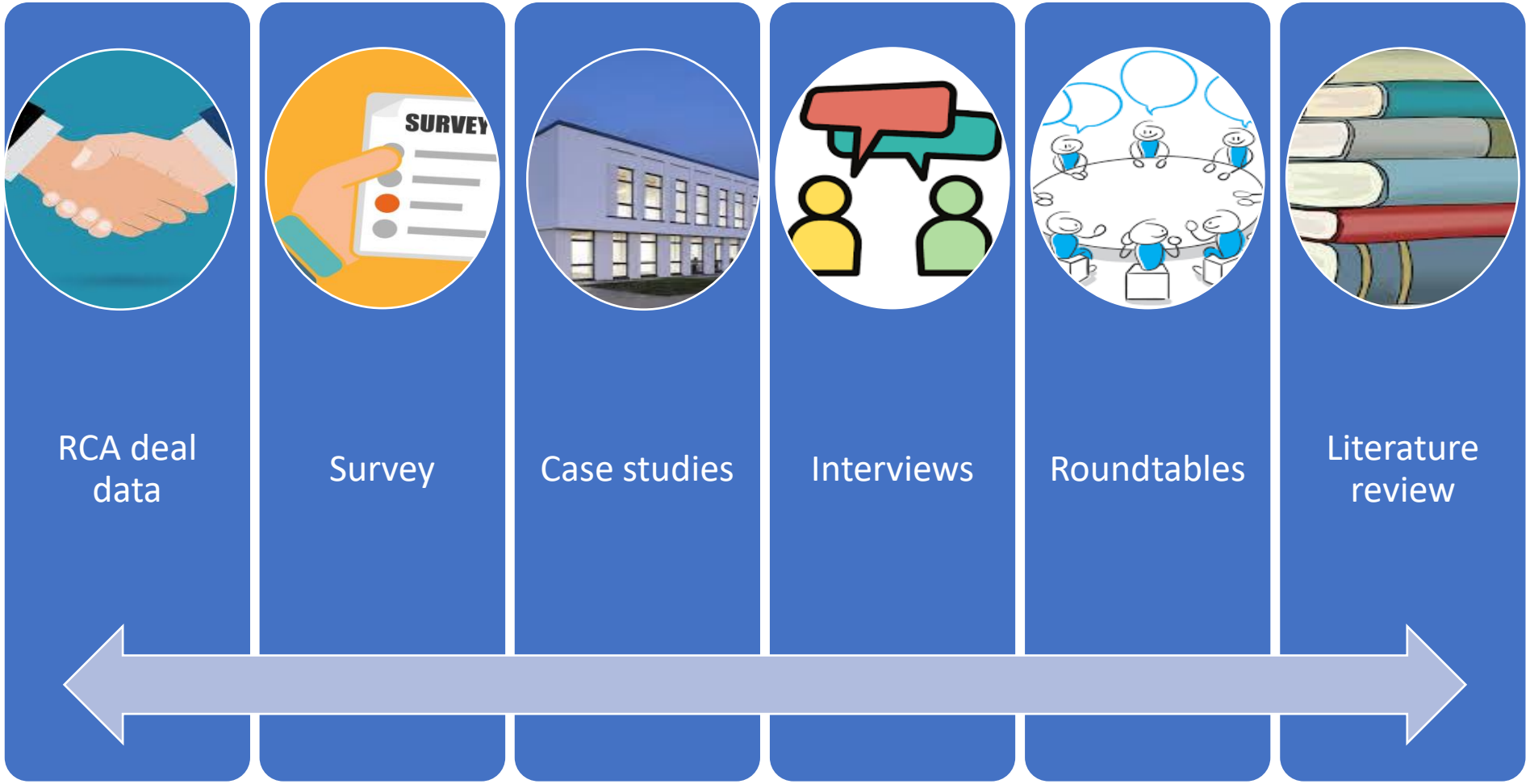




Our research methodology for the ULI project

Six data sets

S RYAN INVEST



Mix of structured and unstructured data

Neatly arranged in rows and columns v. words as they were spoken or written

S RYAN INVEST

	Transaction data from RCA	Survey	Case studies	Interviews	Roundtables	Literature review[1]
	STRUCTURED DATA			UNSTRUCTURED DATA		
FORMAT	Spreadsheet	Spreadsheet	Spreadsheet	Text	Text	Text
SIZE	272 rows * 32 columns	100 rows * 143 columns	16 rows * 2 columns	7,000 words	13,000 words	279,000 words
KB	81	59	54	43	68	2,600

We use NLP to handle unstructured text data

[1] 42 reports and articles converted from PDF

stephen@sryaninvest.com

Natural language processing

What it is and where you find it

S RYAN INVEST

What is natural language processing (NLP)?

NLP is a branch of artificial intelligence that helps computers understand and manipulate human language

NLP includes many different techniques, ranging from statistical and machine learning methods to rules-based and algorithmic approaches

NLP tasks break down language into shorter, elemental pieces. This enables, among other things:

Contextual extraction which means pulling structured information from text-based sources

(Source: SAS)

Where is it found?

- Predictive text
- Machine translation
- Chatbots
- Grammar checkers
- Spam filters
- Real estate research?

stephen@sryaninvest.com

How does NLP work?

Examples of NLP techniques

S RYAN INVEST

1. Part of speech (POS) tagging	• Marking a word as corresponding to a particular part of speech
2. Chunking	• Separating a sentence into its constituent non-overlapping phrases
3. N-grams	• All combinations of adjacent words of length *n* in a text • For example, bigrams = 2-word combinations
4. Sentiment analysis	• Determining if a chunk of text is negative, neutral or positive
5. Named entity recognition (NER)	• Locating and classifying entities mentioned in unstructured text into pre-defined categories

Technique 1: POS tagging

Identifying parts of speech

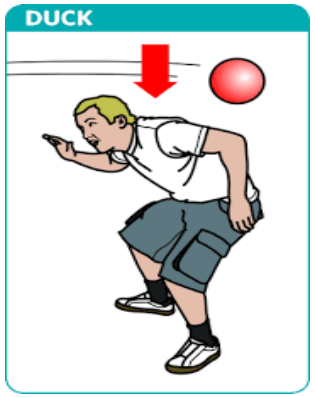
S Ryan Invest

1. Noun *(apartment, office, workshop)*
2. Verb *(build, demolish, wish)*
3. Adjective *(new, old, solid)*
4. Adverb *(early, too, very)*
5. Pronoun *(he, she, we)*
6. Preposition *(after, by, with)*
7. Conjunction *(and, but, neither)*
8. Determiner *(a, the, these)*
9. Exclamation *(aha, alas, hmmm)*

POS tagging is not straightforward

It depends on the context

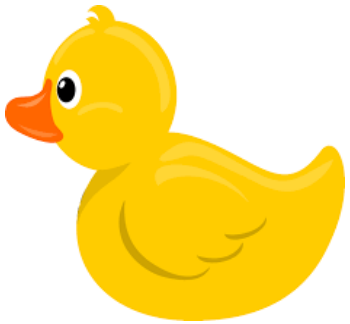
S RYAN INVEST



Verb

My left hand

Adjective



Noun

I left my bike

Verb

What about unknown words?

Machine learning comes to the rescue

S RYAN INVEST

The new chef made spaghetti

The brown dog chased cats

The small child ate sweets

Pattern is: determiner ⇨ adjective ⇨ noun ⇨ verb ⇨ noun

The mimsy butterfly touched hearts

Mimsy is an unknown word

Following the pattern, it is tagged as an adjective

Techniques 2 & 3: Chunking and N-grams

Breaking language down into smaller pieces

S RYAN INVEST

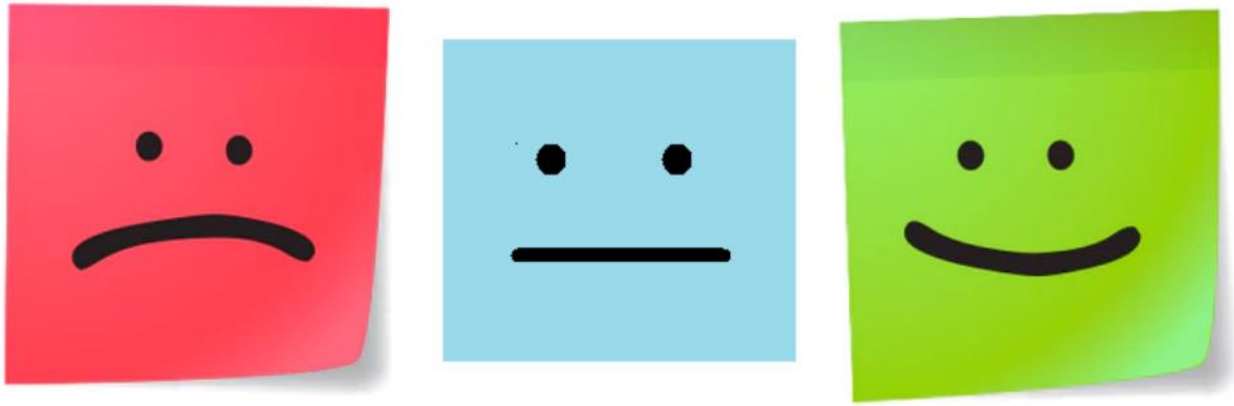
Sample sentence: “Institutional investors hesitate”

	INSTITUTIONAL	INVESTORS	HESITATE
Chunking	Noun phrase		Verb
N-grams: Unigrams	① Institutional	② investors	③ hesitate
N-grams: Bigrams	① Institutional investors	② investors hesitate	
N-grams: Trigrams	① Institutional investors hesitate		

Technique 4: Sentiment analysis

Sometimes called opinion mining

S RYAN INVEST



- A procedure used to determine if a chunk of text is negative, neutral or positive
- Polarity ranges from most negative (-1) to most positive (+1), for example:
  - “Epic” = +1
  - “Slanderous” = -1
- Overall polarity calculated using rules-based approach or machine learning
- Other factors can be added, and scale is not always -1 to +1

Technique 5: Named entity recognition (NER)

Locating and classifying entities into pre-defined categories

S RYAN INVEST

NLP can recognise words that represent:

- Geopolitical entities (countries and cities)
- Organisations
- People
- Facilities (bridge, monument, road, airport)
- Events
- Laws
- Times, dates, numbers and more

JP Morgan

= ORGANISATION

Eoin Morgan

= PERSON

So how does it work in practice?

Example: unstructured datasets from the life sciences project

S RYAN INVEST

SAMPLE FROM INTERVIEWS

Q: In your firm, what do you consider the greatest challenge in life sciences real estate?

A: Top talent from abroad won't wish to work in a field in Kent

SAMPLE FROM ROUNDTABLES

Q: Are there examples of real estate playing a part in attracting and retaining top talent?

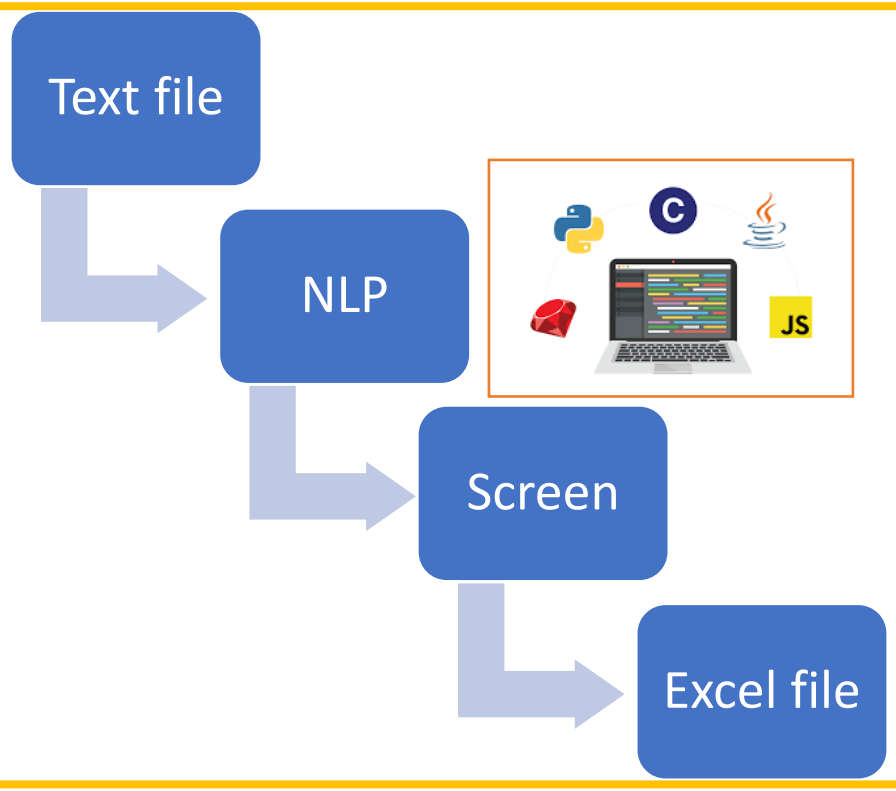
A: Spaces are being created to facilitate links between occupiers

SAMPLES OF LITERATURE

Step-by-step process looks like this

Text data is cleaned, processed and given some structure

S RYAN INVEST



- NLP techniques are implemented in code
- Python and Ruby (other languages work too)
- Output is checked on the screen
- Then onwards to an Excel file

Looking for patterns and topics using NLP

For the life sciences project NLP techniques 1-3 were used

S RYAN INVEST

CHUNKS
a forward-looking five-year business plan
a good long lead indicator
a more traditional office type lease
a two or three-year burn rate
a two or three-year lease term
a typical rent plus yield basis
an office and flex space
an office or residential project
an urban or CBD location
an urban real estate typology
as bigger more mature corporates
as higher and better use
at least nine bioscience hubs
between 1 to 10 or 15 people
it is such a broad spectrum
medical and high school students
pretty much fully fitted flexible space
six or seven or eight stories
some very interesting biotech start-ups
sort of 100% office occupancy
that not much natural air intake
the John Lewis department store
the most advanced real estate providers
the New York City life science cluster
their extractor, air or lavatory requirements

BIGRAMS
Lab space, 8
Ceiling heights, 7
Wet lab, 6
Floor ceiling, 6
City centre, 5
Office building, 4
Lease structures, 4
Science Park, 4
Ultimate flexibility, 4
Start ups, 3
Lab building, 3
Triple net, 3
Urban, setting, 3
Talent, 3
Dry labs, 3
Spin outs, 3
Lab office, 3
Office space, 3
Venture capitalists, 3
Hospitals - universities, 3
Building specifications, 3

1. Labs
2. Leases
3. Locations
4. Lifecyle (start-up, spin out, venture capital)
5. Transparency
6. Talent

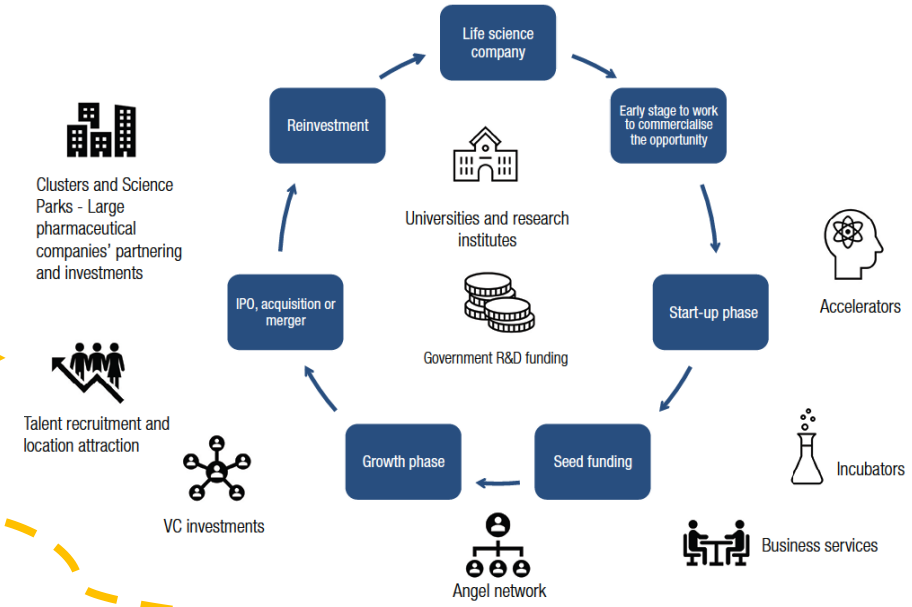
NLP shaped the report

Topics extracted by chunking and n-grams feature heavily in final report

S RYAN INVEST

1. Labs
2. Leases
3. Locations
4. Lifecyle
5. Transparency
6. Talent

Detailed descriptions



Specific recommendations

Occupational property types

- **Clean room/lab:** a room specifically designed to limit the number of airborne contaminants. Special air filters and air distribution systems keep the environment clean.

- **Overcome the lack of reliable data.** Investment research companies should collect and incorporate medical offices and laboratory space in their quarterly index reporting from specialist investors

stephen@sryaninvest.com

Conclusion

S RYAN INVEST

- Handling unstructured data in research projects can be challenging. But:
    1. NLP can quickly extract meaning from unstructured data
    2. Text is interrogated from different angles
    3. POS tagging, chunks and n-grams identify recurring topics

- Researcher knows which topics to emphasise (and which to avoid/downplay)
- For researchers, NLP offers *confidence* and *efficiency*

stephen@sryaninvest.com