University of BRISTOL

Weeks 8 & 9: Information Extraction

Edwin Simpson

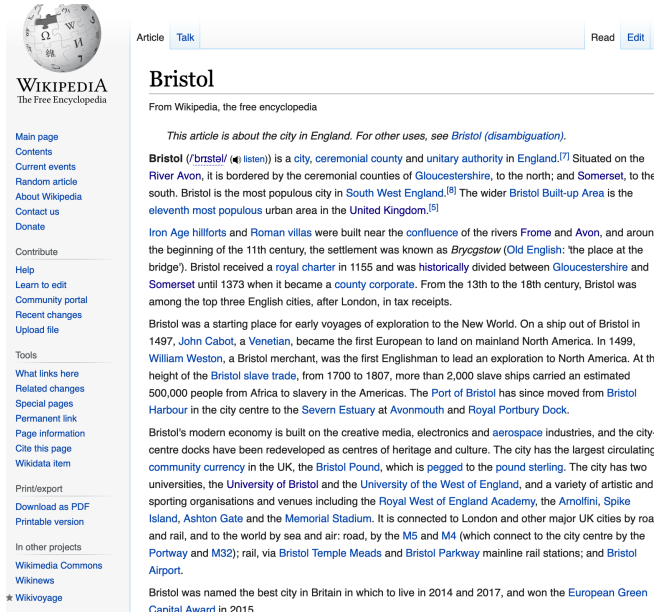
Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

Information Extraction

What steps are needed to go from the text to the info-box?



Wikipedia
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikinews
Wikivoyage

Article Talk Read Edit

Bristol

From Wikipedia, the free encyclopedia

This article is about the city in England. For other uses, see Bristol (disambiguation).

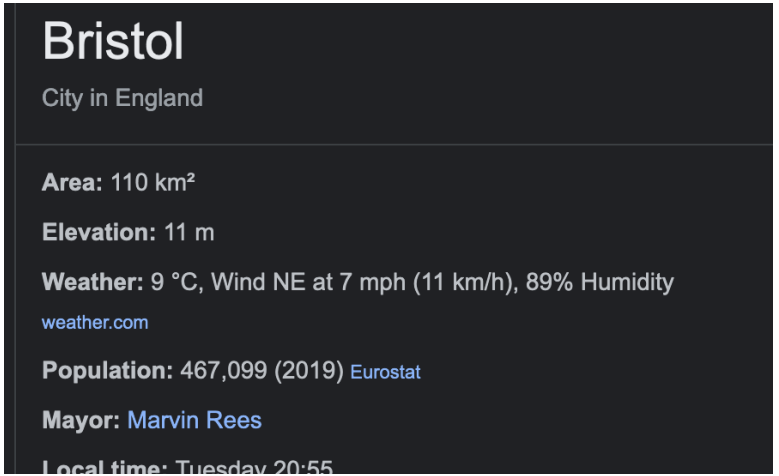
Bristol (/ˈbrɪstəl/ (listen)) is a city, ceremonial county and unitary authority in England.[7] Situated on the River Avon, it is bordered by the ceremonial counties of Gloucestershire, to the north; and Somerset, to the south. Bristol is the most populous city in South West England.[8] The wider Bristol Built-up Area is the eleventh most populous urban area in the United Kingdom.[5]

Iron Age hillforts and Roman villas were built near the confluence of the rivers Frome and Avon, and aroun the beginning of the 11th century, the settlement was known as *Brycgstow* (Old English: 'the place at the bridge'). Bristol received a royal charter in 1155 and was historically divided between Gloucestershire and Somerset until 1373 when it became a county corporate. From the 13th to the 18th century, Bristol was among the top three English cities, after London, in tax receipts.

Bristol was a starting place for early voyages of exploration to the New World. On a ship out of Bristol in 1497, John Cabot, a Venetian, became the first European to land on mainland North America. In 1499, William Weston, a Bristol merchant, was the first Englishman to lead an exploration to North America. At th height of the Bristol slave trade, from 1700 to 1807, more than 2,000 slave ships carried an estimated 500,000 people from Africa to slavery in the Americas. The Port of Bristol has since moved from Bristol Harbour in the city centre to the Severn Estuary at Avonmouth and Royal Portbury Dock.

Bristol's modern economy is built on the creative media, electronics and aerospace industries, and the city centre docks have been redeveloped as centres of heritage and culture. The city has the largest circulating community currency in the UK, the Bristol Pound, which is pegged to the pound sterling. The city has two universities, the University of Bristol and the University of the West of England, and a variety of artistic and sporting organisations and venues including the Royal West of England Academy, the Arnolfini, Spike Island, Ashton Gate and the Memorial Stadium. It is connected to London and other major UK cities by roa and rail, and to the world by sea and air: road, by the M5 and M4 (which connect to the city centre by the Portway and M32); rail, via Bristol Temple Meads and Bristol Parkway mainline rail stations; and Bristol Airport.

Bristol was named the best city in Britain in which to live in 2014 and 2017, and won the European Green Capital Award in 2015.



Bristol

City in England

Area: 110 km²

Elevation: 11 m

Weather: 9 °C, Wind NE at 7 mph (11 km/h), 89% Humidity

weather.com

Population: 467,099 (2019) Eurostat

Mayor: Marvin Rees

Local time: Tuesday 20:55

Padlet for Questions

- https://uob.padlet.org/edwinsimpson/7p23ijfjhk7j9jdr

Information Extraction (IE)

Chapter 17, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

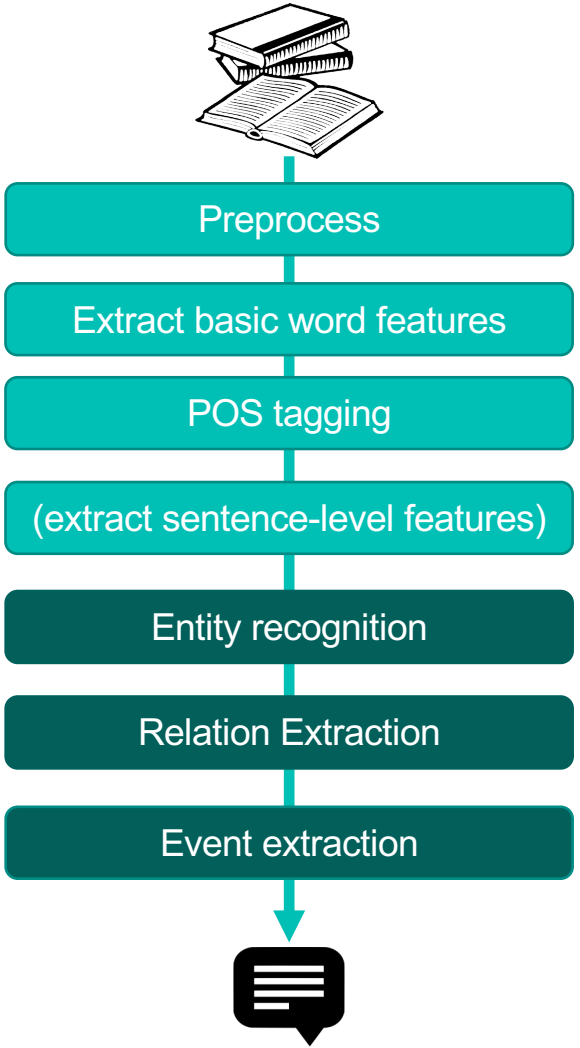
- IE involves several different steps:
  - Named entity Recognition (NER)
  - Relation Extraction (RE)
  - Event extraction

UNITED AIRLINES:	SPOKESPERSON	TIM WAGNER

FAIR RAISE ATTEMPT:	LEAD AIRLINE	UNITED AIRLINES
	AMOUNT	$6
	EFFECTIVE DATE	2006-10-26

Information Extraction (IE)

- IE processes the features extracted at lower levels, such as word and syntax features.
- IE processes text at the semantic level to extract meaning.
- Its results are used in downstream tasks



NER as Sequence Labelling:

- Tag the individual tokens that make up a span:

“The bus service to Old Sodbury runs on Weekdays.”

NER as Sequence Labelling:

- Tag the individual tokens that make up a span:
- It depends what kind of entities we want to extract!
- Here, assume that the ‘bus service’ is something we want to extract information about.

O	B-Misc	I-Misc	O	B-Loc	I-Loc	O	O	B-Time	
“The	*bus*	*service*	*to*	*Old*	*Sodbury*	*runs*	*on*	*Weekdays.”*	

NER as Sequence Labelling:

- Features: the input variables to a sequence tagger or classifier that represent the characteristics of the object we want to label.
- How would the features on the right help the NER sequence tagger?

Feature	Token 1	Token 2
Unigram	”Old”	“Sodbury”
Bigram	[“to”, “Old”]	[“Old”, “Sodbury”]
Bigram	[“Old”, “Sodbury”]	[“Sodbury”, “on”]
Prefix	None	None
Suffix	None	“bury”
InPlaceList	No	Yes
POS	PROPN	PROPN
Chunk	NP	NP

Relation Extraction

The [bus service] to [Old Sodbury] runs on [weekdays]

Extract Feature Vector

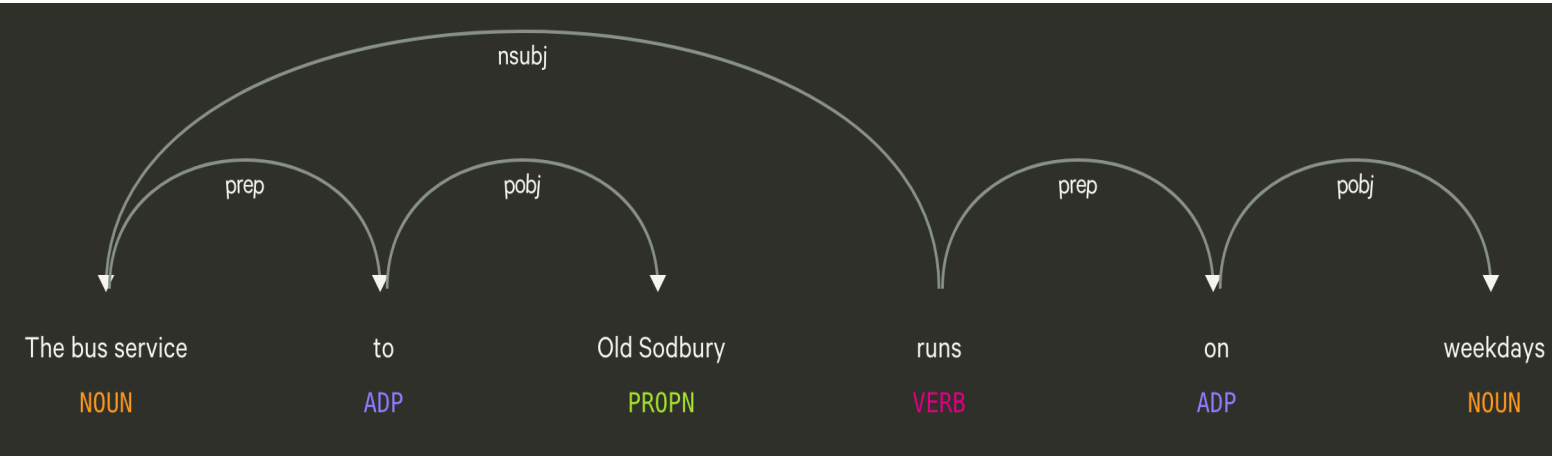
Feature	Entity 1	Entity 2
Unigram	“bus”, “service”	“old”, “sodbury”
UnigramNextToken	“to”	“runs”
UnigramPrevToken	“the”	“to”
EntityType	MISC	LOC
	Relation Features	
ConcatenatedTypes	MISC-LOC	
DependencyPath	NOUN→prep→ADP→ pobj→PROPN	

Binary Classifier for each relation type, e.g., logistic regression

[Bus service] [**TransportTo**] [Old Sodbury]

Dependency Parsing

- *runs* → nsubj → *the bus service*
- *the bus service* → prep → *to*
- *to* → pobj → *Old Sodbury*



Preprocess

Basic word features

POS tagging

Parsing

Entity recognition

Relation Extraction

Event extraction

University of BRISTOL

Quiz

bristol.ac.uk