Information extraction

1. Introduction

Simon Razniewski
Winter semester 2019/20
Outline

1. Introducing each other
2. Organization of the course
3. What&why
4. Preliminaries & Lab 1
Simon Razniewski

• Senior Researcher at MPII, Department 5
• Heading “Knowledge Base Construction and Quality” area

• Background
  • Assistant professor at FU Bozen-Bolzano, Italy, 2014-2017
  • Research stays at AT&T Labs-Research, University of Queensland, UC San Diego
  • PhD FU Bozen-Bolzano, 2014
  • Diplom at TU Dresden, 2010

• Expertise:
  • Logics, databases, Semantic Web
  • More recently IR, (applied) NLP, ML, ...

• Research focus:
  ▪ Analyzing what knowledge bases know, and what they don’t
Cuong Xuan Chu

• Doctoral researcher at D5, MPII

• Focus on information extraction for fictional domains and commonsense knowledge
Department 5

- Department 5: Database and information systems, ~25 members

- **Knowledge discovery:** extracting, organizing, searching, exploring and ranking facts from structured, semi-structured, textual and multimodal information sources

- **yago** Knowledge Base
  - Earliest prominent machine-generated knowledge base (2007)
  - Contains more than 10 million entities and more than 120 million facts

- Gerhard Weikum 259th most cited computer scientist worldwide
And you?

• Course of study
• Preknowledge
• ...
• Comments?

• https://tinyurl.com/ie-uds
Outline

1. Introducing each other
2. Organization of the course
3. What&why
4. Preliminaries & Lab 1
Learning outcomes

• Knowledge
  • What IE is about (“What”)
  • What IE is good for (“Why”)
  • What main tasks and challenges in IE are
  • What standard approaches to IE are (“How”)

• Skills
  • Analyze potentials and limitations of IE approaches
  • Learn to choose right datasource and method for right task
  • Implement simple solutions for main problems in IE
    • Scraping, typing, linking, ...

• Abilities
  • Build your own IE pipeline for an IE problem

→ Very practical focus!
Prerequisites

• Basics of ML
  • We won’t go deep

• Python programming
  • Essential
  • Still time to learn

• Helpful but not required
  • Basic notions of information retrieval (IRDM?)
  • Computational linguistics (SNLP?)
Formal organization

• Credit points: 6, hours: 180 (!)

• Registration
  ▪ Subscribe to the mailing list https://groups.google.com/d/forum/ie1920
  ▪ Register in HISPOS timely before the exam

• When?
  • Lecture (9x): Tuesday 10:00-12:00
  • Lab (9x): Tuesday 16:00-18:00

• How to pass this course?
  ▪ 8 small practical assignments
    ▪ Pass/fail
    ▪ To be admitted to exam, pass at least 6
  ▪ Oral exam
Assignments

• Published on lecture day (Tuesday)
• **Due Saturday 23:59 same week**

• Labs are there to start solving the assignments

• Discussing assignments together is allowed, but **each student must write their own solution**
  • No sharing of code!
  • Plagiarism = course failed for both
  • Avoid **triangular plagiarism** = cite sources
    • “Approach for NER adapted from stackoverflow.com/how-to-...”

• **Libraries** that solve core tasks not allowed
  • In doubt ask..

• **Weekly assignments are evil!?**
  • Psychological trick to help you learn and pass!
Assignment content

• Coding

• 3/7 are assignments in competition format
  • Crisp input/output problem specification
    • “From the first sentence of Wikipedia, extract the type of an entity”
  • Labelled training data set
  • Unseen (hidden) evaluation dataset
    • To avoid overfitting
    → Ranked list by a standard metric, e.g., precision or F1-score
      • But pass/fail does not depend on relative performance
## Schedule

<table>
<thead>
<tr>
<th>Tentative date</th>
<th>Lecture</th>
<th>Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 15.10.</td>
<td>Introduction</td>
<td>Dataset familiarization (pdf)</td>
</tr>
<tr>
<td>2 22.10.</td>
<td>Knowledge representation</td>
<td>Domain modelling</td>
</tr>
<tr>
<td>3 29.10.</td>
<td>Crawling and Scraping</td>
<td>Infobox scraping</td>
</tr>
<tr>
<td>4 12.11.*</td>
<td>NER, typing and taxonomy induction</td>
<td>Entity typing from Wikipedia first sentence</td>
</tr>
<tr>
<td>5 19.11.</td>
<td>Disambiguation</td>
<td>Disambiguation</td>
</tr>
<tr>
<td>6 26.11.</td>
<td>Fact extraction</td>
<td>Pattern-fact duality exploration</td>
</tr>
<tr>
<td>7 3.12.</td>
<td>OpenIE and evaluation</td>
<td>OpenIE coding</td>
</tr>
<tr>
<td>8 10.12.</td>
<td>Rule Mining</td>
<td>Exhaustive short rule evaluation, crowdsourcing</td>
</tr>
<tr>
<td>9 17.12.</td>
<td>Applications</td>
<td>Exam preparation</td>
</tr>
<tr>
<td>(7.1.2020)</td>
<td>(Backup slot)</td>
<td></td>
</tr>
<tr>
<td>14.+15.1.2020</td>
<td>Oral exam</td>
<td></td>
</tr>
</tbody>
</table>

* Note: No lecture/lab on 5.11.
Outline

1. Introducing each other
2. Organization of the course
3. What&why
4. Preliminaries & Lab 1
3. Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Extraction techniques
V. Technologies
VI. Applications
VII. Past, present and future
I. Motivation
• [https://en.wikipedia.org/wiki/Max_Planck_Institute_for_Informatics](https://en.wikipedia.org/wiki/Max_Planck_Institute_for_Informatics)

• [https://www.wikidata.org/wiki/Q565400](https://www.wikidata.org/wiki/Q565400)
What for?

• One central hub for interlanguage interlinking of 100+ Wikipedia editions

• Your AI chatbot wants to know where MPII, MIT and KAIST are located? → structured query

• A library wants to distinguish which of the 100+ literary John Smiths wrote “A description of New England”? → Wikidata ID
Samples of advanced queries

• Who discovered the most planets: http://tinyurl.com/y7rldyqc

• Distribution of places ending with “-weiler” in Germany: https://w.wiki/67o

• Living relatives of Charlemagne: https://w.wiki/67n
The Semantic Web

• Term coined by Tim Berners-Lee for a machine-readable Web
  • Crucial for intelligent agents

• Web content originally from humans for humans

⇒ Make machines read human language, or make humans write machine-readable structured data?

*Machine reading* vs. *information extraction/knowledge base construction*
3. Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Extraction techniques
V. Technologies
VI. Applications
VII. Past, present and future
Definitions

Information extraction is the task of transforming semi/unstructured information into a machine readable format.

Collections of machine-readable information about the general world are called knowledge bases/graphs.
Common types of machine knowledge

• Lexical knowledge
  • \(<\text{shout}, \text{isA}, \text{verb}>\>
  • \(<\text{shout}, \text{subformOf}, \text{communicate}>\>

• Instance knowledge (“Encyclopedic KBs”):
  • \(<\text{Paris}, \text{capitalOf}, \text{France}>\>
  • \(<\text{MPII}, \text{foundedIn}, 1988>\>
  • \(<\text{Angela Merkel}, \text{major}, \text{Physics}>\>

• Class knowledge (“Commonsense”):
  • \(<\text{Pizza}, \text{is}, \text{tasty}>\>
  • \(<\text{Elephant}, \text{color}, \text{grey}>\>
  • \(<\text{turnOnPC}, \text{requires}, \text{power}>\>
Lexical KBs

• WordNet (1995)
• FrameNet (1998)
• (Wiktionary (2002))
• SenticNet (2010)
• ...

Word to search for: shout

Display Options: (Select option to change) ▼ Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) cry, outcry, call, yell, shout, vociferation (a loud utterance; often in protest or opposition) "the speaker was interrupted by loud cries from the rear of the audience"

Verb

- S: (v) shout (utter in a loud voice; talk in a loud voice (usually denoting characteristic manner of speaking)) "My grandmother is hard of hearing--you'll have to shout"
- S: (v) shout, shout out, cry, call, yell, scream, holler, hollo, squall (utter a sudden loud cry) "she cried with pain when the doctor inserted the needle"; "I yelled to her from the window but she couldn't hear me"
  ○ direct troponym / full troponym
  ○ verb group
  ○ direct hypernym / inherited hypernym / sister term
  ○ derivationally related form
  ○ phrasal verb
  ○ sentence frame
- S: (v) exclaim, cry, cry out, outcry, call out, shout (utter aloud; often with surprise, horror, or joy) "I won!" he exclaimed"; "Help!" she cried"; "I'm here, the mother shouted when she saw her child looking lost"
- S: (v) abuse, clapperclaw, blackguard, shout (use foul or abusive language towards) "The actress abused the policeman who gave her a parking ticket"; "The angry mother shouted at the teacher"
FrameNet

• Example Frame – “Revenge”: Because of some injury to something-or-someone important to an avenger (maybe himself), the avenger inflicts a punishment on the offender. The offender is the person responsible for the injury.

• Frame elements:
  • avenger, offender, injury, injured_party, punishment.

• Invoking terms:
  • Nouns: revenge, vengeance, reprisal, retaliation
  • Verbs: avenge, revenge, retaliate (against), get back (at), get even (with), pay back
  • Adjectives: vengeful, vindictive
Encyclopedic KBs ("Instance-oriented KBs")

• Cyc (1984)
• YAGO (2007)*
• DBpedia (2007)
• Wikidata (2012)

* developed at MPII
dbo:activeYearsEndDate
- 2004-11-04 (xsd:date)
- 2008-11-16 (xsd:date)

dbo:activeYearsStartDate
- 1997-01-08 (xsd:date)
- 2005-01-03 (xsd:date)
- 2009-01-20 (xsd:date)

dbo:almaMater
- dbo:Occidental_College
- dbo:Columbia_College._Columbia_University
- dbo:Harvard_Law_School

dbo:award
- dbo:Nobel_Peace_Prize

dbo:birthday
- 1961-08-04 (xsd:date)
- 1961-8-4

dbo:birthPlace
- dbo:Hawaii
- dbo:Honolulu
- dbo:Kapiolani_Medical_Center_for_Women_and_Children

dbo:office
- 44th President of the United States

dbo:party
- dbo:Democratic_Party_(United_States)

dbo:region
- dbo:Illinois
Commonsense KBs (class-oriented)

- Cyc (1984)
- ConceptNet (1999)
- WebChild (2014)*
- TupleKB (2017)
- Quasimodo (2019)*

* Developed at MPII
ConceptNet
A wheeled vehicle that has two wheels and is moved by foot pedals.

<table>
<thead>
<tr>
<th>TYPE OF</th>
<th>wheeled_vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related to</td>
<td>artifact, under the category of cycling</td>
</tr>
<tr>
<td>COMPARABLES</td>
<td>bicycle, bike, bicycle,motorcycle, unicycle,bicycle, bicycle,wheel, bicycle,mountain,bike</td>
</tr>
<tr>
<td>ACTIVITIES</td>
<td>ride bicycle, buy bicycle, use bicycle, sell bicycle, steal bicycle</td>
</tr>
<tr>
<td>HAS PHYSICAL PARTS</td>
<td>axle, bicycle seat, bicycle wheel, brake, casing</td>
</tr>
<tr>
<td>HAS SUBSTANCE</td>
<td>suspension, hydrogen, oxygen, air, water</td>
</tr>
<tr>
<td>IN SPATIAL PROXIMITY WITH</td>
<td>street, chain, park, city, rack</td>
</tr>
<tr>
<td>PHYSICAL PROPERTIES</td>
<td>sensitive, fast, cool, light, small</td>
</tr>
<tr>
<td>ABSTRACT PROPERTIES</td>
<td>welcome, old, safe, good, important</td>
</tr>
<tr>
<td>OTHER PROPERTIES</td>
<td>cheap, dangerous, lucky, wobbly, hard</td>
</tr>
<tr>
<td>ASSOCIATED WITH COUNTRY</td>
<td>united_states, denmark, europe, vietnam, germany</td>
</tr>
</tbody>
</table>
3. Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Extraction techniques
V. Technologies
VI. Applications
VII. Past, present and future
Facts (triples) and their constituents

- **Entities**: Objects about which statements can be made  
  *Paris; Trump; Irony*

- **Property/predicate/relation/attribute**: What can be said  
  *locatedIn(entity, location), worksAt(person, organization), antonymOf(term, term)*

- **Fact/statement/claim/triple**: Core building block of KBs  
  *<Paris, locatedIn, France>*

→ General form:  

  `<subject, predicate, object>`

  `<s, p, o>`
Subjects and objects

- Machine-generated identifiers
  - Wikidata: Q4262, Q67245

- Canonical name strings
  - DBpedia, YAGO: “John_Smith_(politician)”

- Internationalized resource identifier (IRI)
  - Semantic web: http://dbpedia.org/resource/Max_Planck

- General phrases
  - TupleKB: <industry, grow over, past few decade>

- Literals: Attribute values that are no entities
  - www.mpi-inf.mpg.de
  - Often with units: 1.63m; 54.85° N

- Same for predicates, sometimes canonicalized, sometimes just text
Classes and class hierarchies

• **Classes/types**: Allow to group similar entities
  *Presidents, nouns, Greek gods*

• **Type/property hierarchy**: Tree-like hierarchy among types/properties (cf. inheritance in object-oriented programming)
  `<Town, subclassOf, Administrative_unit>`
Classes

Saarbrücken (Q1724)
capital of the German state of Saarland
Saarbrücken

Most relevant properties which are absent
In more languages

Statements

instance of

big city
  0 references

college town
  0 references

urban municipality of Germany
  0 references

state capital in Germany
  0 references

municipality of Germany
  0 references
Taxonomies

https://angryloki.github.io/wikidata-graph-builder/?property=P279&item=Q5
Embedding-based knowledge

• Apple (0.72 0.35 0.91)
• Pear (0.80 0.33 0.55)
• Penguin (0.12 0.58 0.27)

→ Not human-readable
→ Limited machine-readable (meaning of dim. 2?)
• Often impressive performance (e.g., analogies)
3. Introduction to Information Extraction

I. Motivation

II. Definition and topics

III. Formal foundations

IV. Extraction techniques

V. Technologies

VI. Applications

VII. Past, present and future
How to extract information?
Possible approaches

A. Humans (CYC, ConceptNet, Wikidata)

B. Structured extraction (YAGO, DBpedia)

C. Text extraction (NELL, Textrunner)

D. Constraints and pattern mining
A. Humans: Experts

• Potentially best quality

• Difficult to scale
  • CYC: “In 1986, Doug Lenat estimated the effort to complete the KB to be 250,000 rules and 350 man-years of effort.”
Humans: Crowdsourcing/Gamification

• Make work fun (?)
Humans: Volunteers

- Wikidata: 18k active users
- Intrinsic motivation achieves great things
- Broad expertise, compared with selected experts or paid crowdsourcing

- [https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all](https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all)
Humans: Challenges

• ConceptNet:
  • Common knowledge, normalization

• Crowdsourcing: Quality assurance

• Wikidata: Modelling and agreement
  • E.g., ethnicity, notable_work, ...
  • Multilingual concept alignment

<table>
<thead>
<tr>
<th>elephant is capable of...</th>
</tr>
</thead>
<tbody>
<tr>
<td>carry a trunk</td>
</tr>
<tr>
<td>forget to go on the paper</td>
</tr>
<tr>
<td>lift logs from the ground</td>
</tr>
<tr>
<td>to lift the tree</td>
</tr>
<tr>
<td>remember water sources</td>
</tr>
<tr>
<td>visit the grocery store</td>
</tr>
<tr>
<td>weigh up to 14000 pounds</td>
</tr>
<tr>
<td>weight 1000 kilos</td>
</tr>
</tbody>
</table>
B. Structured extraction

- Wikipedia already provides structured data

- All we need to do is harvest...
Work done?

- Noise
- Canonicalization of entities and predicates
- Usage of category system

Examples: YAGO, DBpedia
C. Text extraction

• In principle **most powerful**
  • No need for humans
  • No restriction to Wikipedia existence

• In practice **very noisy**
  • Canonicalization
  • Consistency
  • ...

• Examples: NELL, Textrunner

William Henry Gates III (born October 28, 1955),[2] commonly known as Bill Gates, is an American businessman, co-founder and chairman of Microsoft. He is the second richest person in the world just behind Jeff Bezos as of October 2017.[3]
IE demo

- [https://www.rosette.com/capability/relationship-extraction/#try-the-demo](https://www.rosette.com/capability/relationship-extraction/#try-the-demo)

- Merkel is of German and Polish descent. Her paternal grandfather, Ludwik Kasner, was a German policeman of Polish ethnicity, who had taken part in Poland’s struggle for independence in the early 20th century.[22] He married Merkel's grandmother Margarethe, a German from Berlin, and relocated to her hometown where he worked in the police. In 1930, they Germanized the Polish name Kaźmierczak to Kasner.[23][24][25][26] Merkel's maternal grandparents were the Danzig politician Willi Jentzsch, and Gertrud Alma née Drange, a daughter of the city clerk of Elbing (now Elbląg, Poland) Emil Drange. Since the mid 1990s, Merkel has publicly mentioned her Polish heritage on several occasions and described herself as a quarter Polish, but her Polish roots became better known as a result of a 2013 biography.

- In 1968, Merkel joined the Free German Youth (FDJ), the official communist youth movement sponsored by the ruling Marxist–Leninist Socialist Unity Party of Germany.[30][31][32] Membership was nominally voluntary, but those who did not join found it difficult to gain admission to higher education.[33] She did not participate in the secular coming of age ceremony Jugendweihe, however, which was common in East Germany. Instead, she was confirmed.[34] During this time, she participated in several compulsory courses on Marxism-Leninism with her grades only being regarded as "sufficient".
Challenges

- Entity identification
- Entity disambiguation
- Relation identification
- Relation normalization
- ...

- End-to-end models can alleviate these to some extent, but are specific to their training data
  - E.g., DeepDive
D. Constraints

Databases
- Key, foreign key, range, ...

Knowledge bases:
- Events start earlier than they end
- Every human must have two parents
- Mayors of cities must be humans
- The parent of a person’s sibling is the person’s parent

- Can be used to...
  - ... reject KB modifications
  - ... indicate missing information
  - ... infer new facts

- But reality is messy..
Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Construction techniques
V. Technologies
VI. Applications
VII. Past, present and future
Which technologies every information extraction engineer should know about?
Technologies (1): Scraping

- **BeautifulSoup** for Python web scraping
Technologies (2): Storing

- **RDF** for representing data
  - Resource description framework
  - Turtle syntax for triples and data types:

  `<Mark_Twain> <author> <Huckleberry_Finn>.
  <Huckleberry_Finn> <description> “A 19th century classic novel”.

IRIs for unique identification of entities:

  <http://yago-knowledge.org/resource/Mark_Twain>

Prefixes for shorthand notation:

  @prefix yago: <http://yago-knowledge.org/resource>
yago:Mark_Twain  yago:dateOfBirth 30.11.1835
Technologies (3): Querying

- **SPARQL** for posing queries
  - Query language inspired by SQL

Wikidata cats: [https://w.wiki/33a](https://w.wiki/33a)
Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Construction techniques
V. Technologies
VI. Applications
VII. History and future
What KBs are good for

• Master data
• Data mining
• Search enhancements
• Question answering
• Language generation
• Entity linking
• Learning more knowledge
• ....
### Master data (1)

| wd:Q6258248 | John Smith |
| wd:Q6258251 | John Smith |
| wd:Q6258255 | John Smith |
| wd:Q6258259 | John Smith |
| wd:Q6258261 | John Smith |
| wd:Q6258263 | John Smith |
| wd:Q6258265 | John Smith |
| wd:Q6258267 | John Smith |
| wd:Q6258270 | John Smith |
| wd:Q6258271 | John Smith |
| wd:Q6258276 | John Smith |
| wd:Q6258278 | John Smith |
| wd:Q6258281 | John Smith |
| wd:Q6258284 | John Smith |
| wd:Q6258286 | John Smith |
| wd:Q6258288 | John Smith |
| wd:Q6258290 | John Smith |
| wd:Q6258293 | John Smith |
| wd:Q6258294 | John Smith |
| wd:Q6258296 | John Smith |

(300 more)
Master data (2)

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>Value</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase ID</td>
<td>/m/03mb4s</td>
<td>1 reference</td>
</tr>
<tr>
<td>GND ID</td>
<td>5086841-9</td>
<td>1 reference</td>
</tr>
<tr>
<td>VIAF ID</td>
<td>157458492</td>
<td>1 reference</td>
</tr>
<tr>
<td>ISNI</td>
<td>0000 0004 0491 9823</td>
<td>1 reference</td>
</tr>
<tr>
<td>GRID ID</td>
<td>grid.419528.3</td>
<td>2 references</td>
</tr>
</tbody>
</table>

Relevant for:
- Museums
- Libraries
- Scientific publications
...
Data mining

• Use input facts to extract patterns that allow to predict new facts

\[
\begin{align*}
\text{isCitizenOf}(x, y) &\Rightarrow \text{livesIn}(x, y) \\
\text{hasAdvisor}(x, y) \land \text{graduatedFrom}(x, z) &\Rightarrow \text{worksAt}(y, z) \\
\text{wasBornIn}(x, y) \land \text{isLocatedIn}(y, z) &\Rightarrow \text{isCitizenOf}(x, z) \\
\text{hasWonPrize}(x, \text{G. W. Leibniz}) &\Rightarrow \text{livesIn}(x, \text{Germany})
\end{align*}
\]

\[\text{isCitizenOf}(\text{John, France}) \Rightarrow \text{livesIn}(\text{John, France})\]

• Various approaches based on association rule mining and latent models
Entity linking

https://gate.d5.mpi-inf.mpg.de/webaida/
Search enhancements
Question answering

Try yourself:
- When was Trump born?
- What is the nickname of Ronaldo?
- Who invented the light bulb?
Question answering (2)

• Knowledge bases **key component in question answering systems**
  • E.g., IBM Watson

• **AllenAI science challenge:** Computers currently in 8\(^{th}\) grade
  • Knowledge acquisition still major bottleneck
Language generation

- Wikipedia in world’s most spoken language: \( \frac{1}{10} \) as many articles as English Wikipedia
- World’s fourth most spoken language: \( \frac{1}{100} \)

\( \Rightarrow \) Wikidata intended to help resource-poor languages

Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Construction and maintenance
V. Technologies
VI. Applications
VII. Past, present and future
Past

1984

2001

2007

2012

2018

Cyc

Wikipedia

WordNet

Memex

(1945)

Wolfram Alpha

Yago

DBpedia

FreeBase

Knowledge Graph

WIKIDATA

Past

($\text{relationAllExists}
\text{biologicalMother}
\text{ChordataPhylum}
\text{FemaleAnimal})
Present

• **IE and KBs at most major tech companies** and beyond
  • Google, Microsoft, Alibaba, Bloomberg, ...

• **Feb 2018: $125 million investment** by Microsoft cofounder Paul Allen into non-profit research on common sense knowledge extraction and reasoning

• Research: Major part of NLP conferences taken up by IE research
Future

• ?
Outline

1. Introducing each other
2. Organization of the course
3. What&why
4. Lab 1
Lab 1

• Information extraction where from?
  • Actual web crawling nontrivial
  • Wikipedia a popular high-quality resource

• For a change, we work on a Wiki about Game of Thrones (data dump)

• **Task 1:** Find pages of certain types

• **Task 2:** Find the different surface forms of links to a page

• **Task 3:** Formulate and run some SPARQL queries over Wikidata
Regular expressions

• Search patterns for String data

```python
import re
str = "No pain no gain"
x = re.findall("\Sain", str)
print(x)
['pain', 'gain']
```

https://www.w3schools.com/python/python_regex.asp
Take home

• Information extraction translates unstructured/semistructured content into machine-readable structured formats

• Structured data is relevant for a range of knowledge-intensive and AI tasks

• More about how to do IE follows..