INF 558/CSCI 563: Building Knowledge Graphs

Homework 2: Information Extraction

Released: Jan 23rd, 2020
Due: Jan 30th, 2020 @ 23:59

Ground Rules

This homework must be done individually. You can ask others for help with the tools, however, the submitted homework must be your own work.

Summary

In this homework, you will extract data from unstructured sources. Like the previous homework, the data will be acquired from the Internet Movie Database (IMDb) website (https://www.imdb.com/). To extract data from unstructured text, you will use Spacy (https://spacy.io/), an open-source software library for advanced natural language processing (NLP).

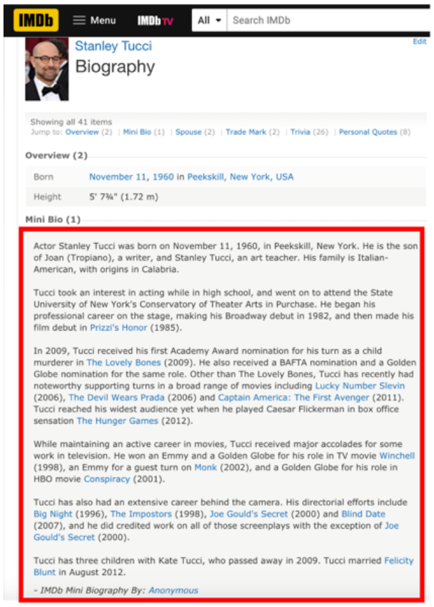
Task 1: Extraction (3 points)

Extract the biographies of the top 500 entities you scraped in the previous homework (Homework 01, Task 1.2) from IMDb.

Store your results in a tab-separated values (`tsv`) file. Each row should represent a single entity, and should contain two values: URL and biography. Figure 1 shows an example page of a cast entity with its biography marked in a red box.

Notes:

- (Hint) There is no need to re-implement a scraping bot, you should use the URLs in your `jl` file from previous homework, modify them slightly to access the biography page and use a parsing tool (e.g. BeautifulSoup) to extract the required attribute. Please note that while you're accessing pages on IMDb (parsing/scraping) you should again obey the website's politeness rules (i.e. sleep time between requests) to avoid getting banned.
- Make sure that you end up with a total of 500 entries in which each entity has a biography text (while scraping/parsing you should ignore entries with no biography text on their page).
- You can check the attached file `entities_bio_sample.tsv` to understand the format.



Stanley Tucci
Biography

Showing all 41 items
Jump to: Overview (2) | Mini Bio (1) | Spouse (2) | Trade Mark (2) | Trivia (26) | Personal Quotes (8)

Overview (2)

Born November 11, 1960 in Peekskill, New York, USA
Height 5' 7¾" (1.72 m)

Mini Bio (1)

Actor Stanley Tucci was born on November 11, 1960, in Peekskill, New York. He is the son of Joan (Tropiano), a writer, and Stanley Tucci, an art teacher. His family is Italian-American, with origins in Calabria.

Tucci took an interest in acting while in high school, and went on to attend the State University of New York's Conservatory of Theater Arts in Purchase. He began his professional career on the stage, making his Broadway debut in 1982, and then made his film debut in Prizzi's Honor (1985).

In 2009, Tucci received his first Academy Award nomination for his turn as a child murderer in The Lovely Bones (2009). He also received a BAFTA nomination and a Golden Globe nomination for the same role. Other than The Lovely Bones, Tucci has recently had noteworthy supporting turns in a broad range of movies including Lucky Number Slevin (2006), The Devil Wears Prada (2006) and Captain America: The First Avenger (2011). Tucci reached his widest audience yet when he played Caesar Flickerman in box office sensation The Hunger Games (2012).

While maintaining an active career in movies, Tucci received major accolades for some work in television. He won an Emmy and a Golden Globe for his role in TV movie Winchell (1998), an Emmy for a guest turn on Monk (2002), and a Golden Globe for his role in HBO movie Conspiracy (2001).

Tucci has also had an extensive career behind the camera. His directorial efforts include Big Night (1996), The Impostors (1998), Joe Gould's Secret (2000) and Blind Date (2007), and he did credited work on all of those screenplays with the exception of Joe Gould's Secret (2000).

Tucci has three children with Kate Tucci, who passed away in 2009. Tucci married Felicity Blunt in August 2012.

- IMDb Mini Biography By: Anonymous

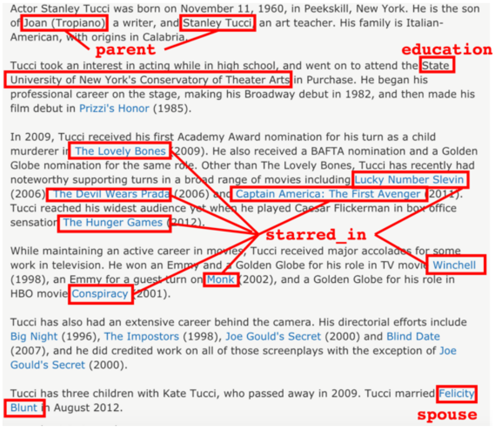
Figure 1: An example entity webpage with his biography marked

Task 2: SpaCy (7 points)

In this task, you will extract structured data, for each entity, from the unstructured biography text you have scraped in the previous task. You will perform Rule-Based Extraction. We are interested in the following attributes:

`spouse`	Name(s) of the husband(s) or wife(s) of the entity
`education`	Educational institution(s) attended by the entity
`parent`	Name(s) of the entity's parents
`starred_in`	Name(s) of the movies in which the entity had some role

Figure 2 shows these attributes over the biography text from the page shown in Figure 1.



Actor Stanley Tucci was born on November 11, 1960, in Peekskill, New York. He is the son of Joan (Tropiano), a writer, and Stanley Tucci, an art teacher. His family is Italian-American, with origins in Calabria.

Tucci took an interest in acting while in high school, and went on to attend the State University of New York's Conservatory of Theater Arts in Purchase. He began his professional career on the stage, making his Broadway debut in 1982, and then made his film debut in Prizzi's Honor (1985).

In 2009, Tucci received his first Academy Award nomination for his turn as a child murderer in The Lovely Bones (2009). He also received a BAFTA nomination and a Golden Globe nomination for the same role. Other than The Lovely Bones, Tucci has recently had noteworthy supporting turns in a broad range of movies including Lucky Number Slevin (2006), The Devil Wears Prada (2006) and Captain America: The First Avenger (2011). Tucci reached his widest audience yet when he played Caesar Flickerman in box office sensation The Hunger Games (2012).

While maintaining an active career in movies, Tucci received major accolades for some work in television. He won an Emmy and a Golden Globe for his role in TV movie Winchell (1998), an Emmy for a guest turn on Monk (2002), and a Golden Globe for his role in HBO movie Conspiracy (2001).

Tucci has also had an extensive career behind the camera. His directorial efforts include Big Night (1996), The Impostors (1998), Joe Gould's Secret (2000) and Blind Date (2007), and he did credited work on all of those screenplays with the exception of Joe Gould's Secret (2000).

Tucci has three children with Kate Tucci, who passed away in 2009. Tucci married Felicity Blunt in August 2012.

Figure 2: An example biography text with the required attributes marked

Before moving to this section's subtasks, familiarize yourself with SpaCy. We provide a python notebook (`Task2.ipynb`), which contains instructions, code and descriptions on how to perform information extraction using SpaCy and how to implement Rule-Based Matching patterns and functionalities.

Task 2.1 (1 point)

Pick one webpage of a single entity from what you have scraped (out of the 500). Screenshot the page and highlight the required attributes over the screenshot (similar to what is shown in Figure 2).
Then, pick one sentence from the bio which includes a reference to a movie/show (such as the sentence shown in Figure 2 that is marked with a `starred_in` relation). Use the provided notebook (or use the code provided in there) to visualize its dependency parse-tree (using displaCy). An example of such tree is shown in Figure 3.

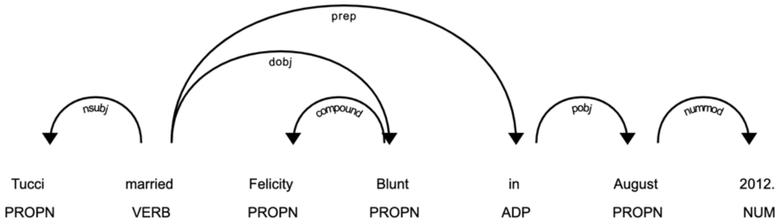


Figure 3: A dependency parse-tree of a sentence

Task 2.2 (4 points)

Implement two extractors for each attribute (`spouse`, `education`, `parent`, `starred_in`):

- Lexical: One that uses only relation phrases.
- Syntactic: One that uses the POS tags, the dependency parse-tree and lexical phrases.

The code should be stored in python script file; the script should accept a single argument (your `tsv` file from previous task) and produce a `jl` file which includes all the extracted values of each attribute and the original `url` (for each entry). This file will be submitted as well. Each value should be a list (except for the `url`). An example file is provided (`sample__task_2_2.jl`).

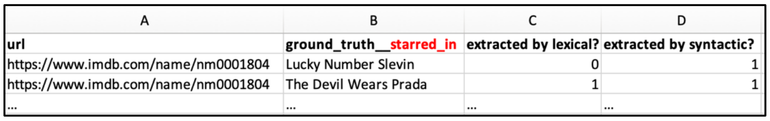
Notes:

- Your script file (`.py`) should include 8 extractors (4 attributes x 2 types of extractors)
- Your output file (`.jl`) should include 500 entries.
- In the provided notebook, we present a sample code that can be used as a starting point for extracting the `spouse` attribute.
- You can choose any of the extractors you have implemented to generate the output `jl` file. Preferably, you should use the one that has higher recall (from your validation results in task 2.3).

Task 2.3 (2 points)

In this task you will validate your extractors. Start by building a ground-truth by picking 20 entries out of the 500 entries, then label them manually. Then, indicate `0` if your extractor failed to acquire the value you labeled or `1` if it succeeded. Store your labeled data in 4 excel files (one for each attribute). Finally, report the recall of each one of your extractors on the labeled data (success divided by total) in the report file.

For example, for the `starred_in` attribute the file would look as shown in Figure 4. As seen in the figure, the first entity has two values (ground truth, which we manually inserted/labeled). The first extractor failed to extract the first movie but succeeded in the second one. The second extractor succeeded in both. So, the recall of the lexical extractor of this attribute would be 50% and the recall for the syntactic one would be 100%.



A	B	C	D
url	**ground_truth__starred_in**	**extracted by lexical?**	**extracted by syntactic?**
https://www.imdb.com/name/nm0001804	Lucky Number Slevin	0	1
https://www.imdb.com/name/nm0001804	The Devil Wears Prada	1	1
...	...	...	...

Figure 4: An example of an evaluation file

Submission Instructions

You must submit (via Blackboard) the following files/folders in a single `.zip` archive named `Firstname_Lastname_hw02.zip`:

- `Firstname_Lastname_hw02_report.pdf`: pdf file with your answers to Task 2.1 and 2.3
- `Firstname_Lastname_hw02_bios.tsv`: as described in Task 1
- `Firstname_Lastname_hw02_cast.jl`: as described in Task 2.2
- Excel files containing the evaluation results as described in Task 2.3:
  - `Firstname_Lastname_hw02_task_2_3__spouse.xlsx`
  - `Firstname_Lastname_hw02_task_2_3__education.xlsx`
  - `Firstname_Lastname_hw02_task_2_3__parent.xlsx`
  - `Firstname_Lastname_hw02_task_2_3__starred_in.xlsx`
- `source`: This folder includes all the code you wrote to accomplish Tasks 1 and 2 (i.e. your crawler/parser, your extractors code, etc…)