INF 558: Building Knowledge Graphs

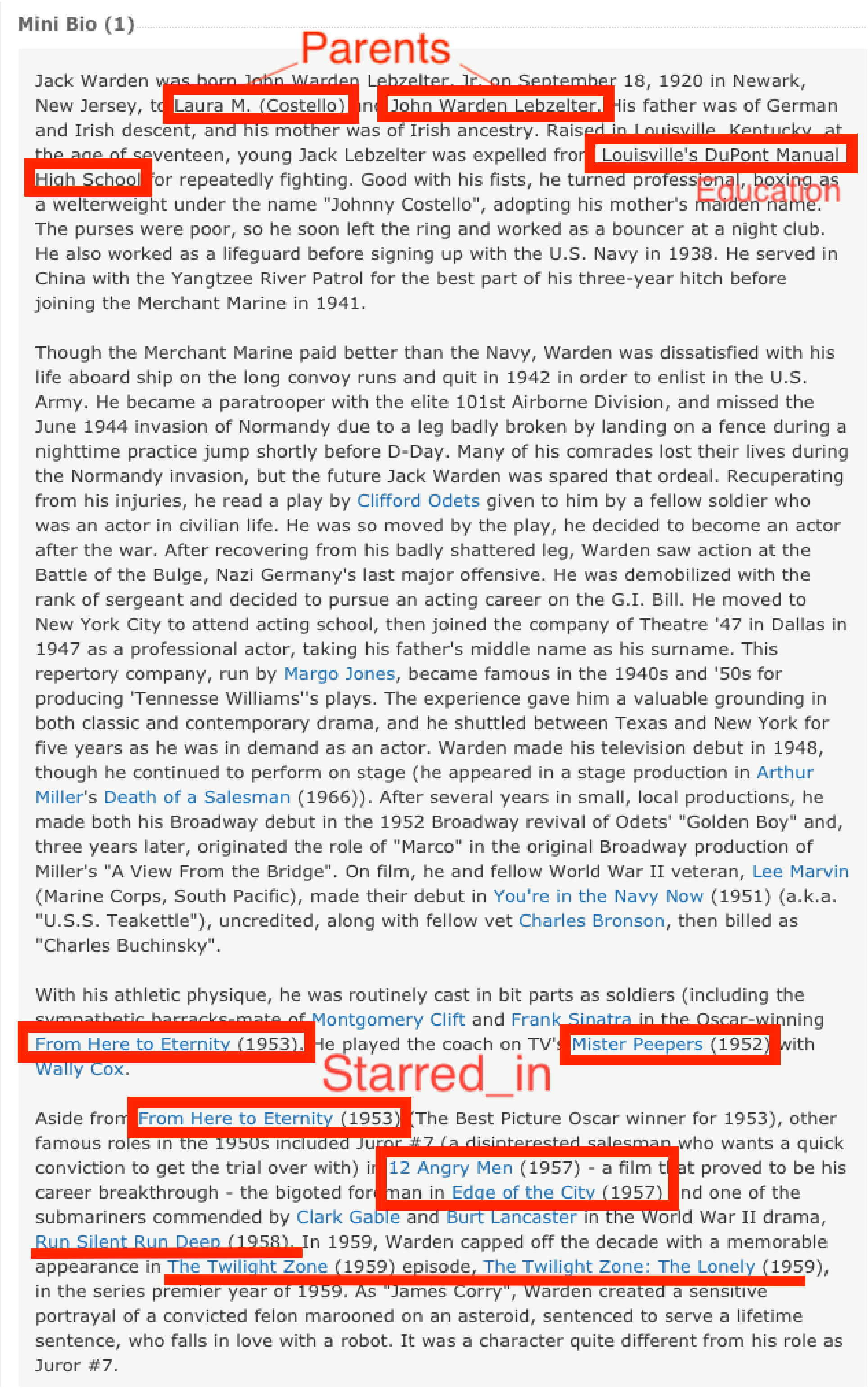
Report of Homework2: Information Extraction

Author: Zongdi Xu (USC ID 5900-5757-70)

Date: Jan 30, 2020

Task 2.1

From the page of Jack Warden, the attributes extracted from that biography text are showed as below:



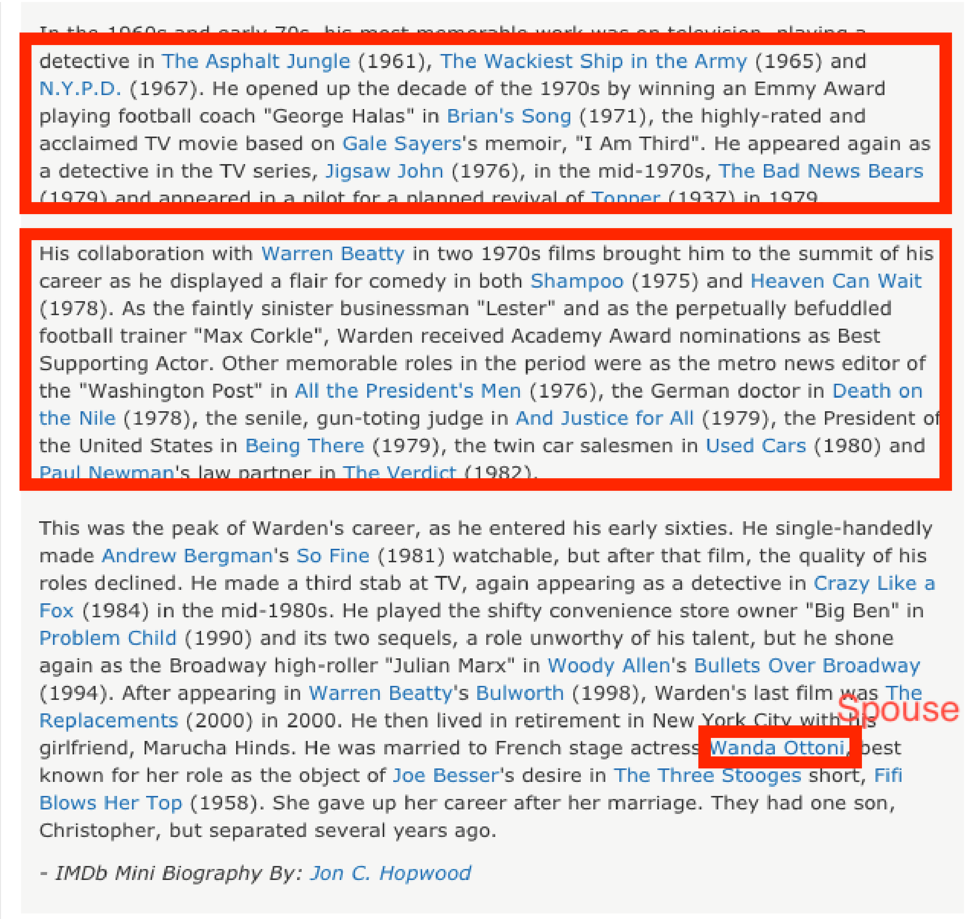
Mini Bio (1)

Jack Warden was born John Warden Lebzelter, Jr. on September 18, 1920 in Newark, New Jersey, to Laura M. (Costello) and John Warden Lebzelter. His father was of German and Irish descent, and his mother was of Irish ancestry. Raised in Louisville, Kentucky, at the age of seventeen, young Jack Lebzelter was expelled from Louisville's DuPont Manual High School for repeatedly fighting. Good with his fists, he turned professional, boxing as a welterweight under the name "Johnny Costello", adopting his mother's maiden name. The purses were poor, so he soon left the ring and worked as a bouncer at a night club. He also worked as a lifeguard before signing up with the U.S. Navy in 1938. He served in China with the Yangtzee River Patrol for the best part of his three-year hitch before joining the Merchant Marine in 1941.

Though the Merchant Marine paid better than the Navy, Warden was dissatisfied with his life aboard ship on the long convoy runs and quit in 1942 in order to enlist in the U.S. Army. He became a paratrooper with the elite 101st Airborne Division, and missed the June 1944 invasion of Normandy due to a leg badly broken by landing on a fence during a nighttime practice jump shortly before D-Day. Many of his comrades lost their lives during the Normandy invasion, but the future Jack Warden was spared that ordeal. Recuperating from his injuries, he read a play by Clifford Odets given to him by a fellow soldier who was an actor in civilian life. He was so moved by the play, he decided to become an actor after the war. After recovering from his badly shattered leg, Warden saw action at the Battle of the Bulge, Nazi Germany's last major offensive. He was demobilized with the rank of sergeant and decided to pursue an acting career on the G.I. Bill. He moved to New York City to attend acting school, then joined the company of Theatre '47 in Dallas in 1947 as a professional actor, taking his father's middle name as his surname. This repertory company, run by Margo Jones, became famous in the 1940s and '50s for producing 'Tennesse Williams''s plays. The experience gave him a valuable grounding in both classic and contemporary drama, and he shuttled between Texas and New York for five years as he was in demand as an actor. Warden made his television debut in 1948, though he continued to perform on stage (he appeared in a stage production in Arthur Miller's Death of a Salesman (1966)). After several years in small, local productions, he made both his Broadway debut in the 1952 Broadway revival of Odets' "Golden Boy" and, three years later, originated the role of "Marco" in the original Broadway production of Miller's "A View From the Bridge". On film, he and fellow World War II veteran, Lee Marvin (Marine Corps, South Pacific), made their debut in You're in the Navy Now (1951) (a.k.a. "U.S.S. Teakettle"), uncredited, along with fellow vet Charles Bronson, then billed as "Charles Buchinsky".

With his athletic physique, he was routinely cast in bit parts as soldiers (including the sympathetic barracks-mate of Montgomery Clift and Frank Sinatra in the Oscar-winning From Here to Eternity (1953). He played the coach on TV's Mister Peepers (1952) with Wally Cox.

Aside from From Here to Eternity (1953) (The Best Picture Oscar winner for 1953), other famous roles in the 1950s included Juror #7 (a disinterested salesman who wants a quick conviction to get the trial over with) in 12 Angry Men (1957) - a film that proved to be his career breakthrough - the bigoted foreman in Edge of the City (1957) and one of the submariners commended by Clark Gable and Burt Lancaster in the World War II drama, Run Silent Run Deep (1958). In 1959, Warden capped off the decade with a memorable appearance in The Twilight Zone (1959) episode, The Twilight Zone: The Lonely (1959), in the series premier year of 1959. As "James Corry", Warden created a sensitive portrayal of a convicted felon marooned on an asteroid, sentenced to serve a lifetime sentence, who falls in love with a robot. It was a character quite different from his role as Juror #7.



In the 1960s and early 70s, his most memorable work was on television, playing a detective in The Asphalt Jungle (1961), The Wackiest Ship in the Army (1965) and N.Y.P.D. (1967). He opened up the decade of the 1970s by winning an Emmy Award playing football coach "George Halas" in Brian's Song (1971), the highly-rated and acclaimed TV movie based on Gale Sayers's memoir, "I Am Third". He appeared again as a detective in the TV series, Jigsaw John (1976), in the mid-1970s, The Bad News Bears (1979) and appeared in a pilot for a planned revival of Topper (1937) in 1979.

His collaboration with Warren Beatty in two 1970s films brought him to the summit of his career as he displayed a flair for comedy in both Shampoo (1975) and Heaven Can Wait (1978). As the faintly sinister businessman "Lester" and as the perpetually befuddled football trainer "Max Corkle", Warden received Academy Award nominations as Best Supporting Actor. Other memorable roles in the period were as the metro news editor of the "Washington Post" in All the President's Men (1976), the German doctor in Death on the Nile (1978), the senile, gun-toting judge in And Justice for All (1979), the President of the United States in Being There (1979), the twin car salesmen in Used Cars (1980) and Paul Newman's law partner in The Verdict (1982).

This was the peak of Warden's career, as he entered his early sixties. He single-handedly made Andrew Bergman's So Fine (1981) watchable, but after that film, the quality of his roles declined. He made a third stab at TV, again appearing as a detective in Crazy Like a Fox (1984) in the mid-1980s. He played the shifty convenience store owner "Big Ben" in Problem Child (1990) and its two sequels, a role unworthy of his talent, but he shone again as the Broadway high-roller "Julian Marx" in Woody Allen's Bullets Over Broadway (1994). After appearing in Warren Beatty's Bulworth (1998), Warden's last film was The Replacements (2000) in 2000. He then lived in retirement in New York City with his girlfriend, Marucha Hinds. He was married to French stage actress Wanda Ottoni, best known for her role as the object of Joe Besser's desire in The Three Stooges short, Fifi Blows Her Top (1958). She gave up her career after her marriage. They had one son, Christopher, but separated several years ago.

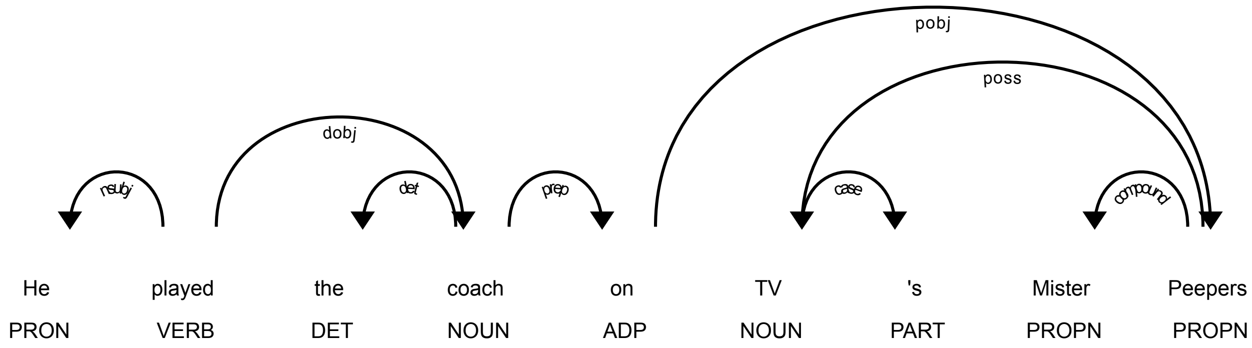
- IMDb Mini Biography By: Jon C. Hopwood

Pick the sentence `He played the coach on TV's Mister Peepers` for visualization:

```
# !python3 -m spacy download en_core_web_sm
import spacy
import en_core_web_sm
import csv
nlp = en_core_web_sm.load()


sentence = nlp("He played the coach on TV's Mister Peepers")

from spacy import displacy
options = {"distance": 96}
displacy.render(sentence, style="dep", options=options)
```



Task 2.2

The template to apply extrators to the original entries:

```
import csv
from spacy.matcher import Matcher
tsv_reader = csv.reader(open('entities_bio.tsv'), delimiter='\t')

limit = 500
count = 0

def matching(doc, pattern):
    result = []
    for sent in doc.sents:
        matcher = Matcher(nlp.vocab)
        matcher.add("matching", None, pattern)

        matches = matcher(nlp(str(sent)))
        if len(matches)>0:
            match = matches[-1]
            span = sent[match[1]:match[2]]
            result.append(span.text)

    return result

def max_length(list1, list2):
    if len(list1)>len(list2):
        return list1
    else:
        return list2

with open('cast.jl', 'w') as fout:
    for (idx, (url, bio)) in enumerate(tsv_reader):
        count += 1
        result = {}
        result['url'] = url
        result['spouse'] = max_length(matching(nlp(bio), pattern_spouse_lexical), matching(nlp(bio),
        result['education'] = max_length(matching(nlp(bio), pattern_education_lexical), matching(nlp
        result['parent'] = max_length(matching(nlp(bio), pattern_parent_lexical), matching(nlp(bio),
        result['starred_in'] = max_length(matching(nlp(bio), pattern_starred_in_lexical), matching(
#         for idx, sent in enumerate(nlp(bio).sents):
#             pass
        fout.write(str(result)+'\n')
        if count>=limit:
            break
        pass
    fout.close()
```

Lexical Extractors

```
pattern_spouse_lexical = [
            {'LOWER': 'married'},
            {'OP': '*'},
            {'LOWER': 'to'},
            {'TEXT': {'REGEX': '\s*'}, 'OP': '*'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'TEXT': {'REGEX': '\s*'}, 'OP': '+'},
            ]



pattern_parent_lexical = [
            {'LOWER': 'born'},
            {'OP': '*'},
            {'LOWER': 'to'},
            {'TEXT': {'REGEX': '\s*'}, 'OP': '*'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'TEXT': {'REGEX': '\s*'}, 'OP': '+'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'LOWER': 'and','OP': '?'},
            {'TEXT': {'REGEX': '\s*'}, 'OP': '*'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'TEXT': {'REGEX': '\s*'}, 'OP': '+'},
            {'IS_PUNCT': True, 'OP': '*'},
            ]
```

```
pattern_education_lexical = [
            {'TEXT': {'REGEX': '^(attend|attended)$'}},
            {'OP': '+'},
            ]
```

```
pattern_starred_in_lexical = [
            {'TEXT': {'REGEX': '^(star|starred)$'}},
            {'LOWER': 'in'},
            {'OP': '+'},
            ]
```

Syntactic Extractors

```
# define the pattern
pattern_spouse_syntactic = [
            {'POS': 'ADJ', 'LOWER': 'married'},
            {'OP': '*'},
            {'LOWER': 'to', 'POS': 'ADP'},
            {'POS': 'ADJ', 'OP': '*'},
            {'POS': 'NOUN', 'OP': '*'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'ENT_TYPE': 'PERSON', 'OP': '+'},
            ]
```

```
pattern_parent_syntactic = [
            {'POS': 'VERB', 'ORTH': 'born'},
            {'OP': '*'},
            {'LOWER': 'to', 'POS': 'ADP'},
            {'POS': 'ADJ', 'OP': '*'},
            {'POS': 'NOUN', 'OP': '*'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'ENT_TYPE': 'PERSON', 'OP': '+'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'LOWER': 'and', 'POS': 'CCONJ', 'OP': '?'},
            {'POS': 'ADJ', 'OP': '*'},
            {'POS': 'NOUN', 'OP': '*'},
            {'IS_PUNCT': True, 'OP': '*'},
            {'ENT_TYPE': 'PERSON', 'OP': '+'},
            {'IS_PUNCT': True, 'OP': '*'},
            ]
```

```
pattern_education_syntactic = [
            {'POS': 'VERB', 'LEMMA': 'attend'},
            {'OP': '+'},
            ]
```

```
pattern_starred_in_syntactic = [
            {'POS': 'VERB', 'LEMMA': 'star'},
            {'POS': 'ADP', 'LOWER': 'in'},
            {'OP': '+'},
            ]
```

Task 2.3

Recall rates for every extractor:

Attribute	spouse	education	parent	starred_in
Lexical	90%	80%	95%	70%
Syntactic	85%	85%	80%	70%