cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Distant Reading for European Literary History: Training School

Distant Reading

Milica Antić

26. novembar 2019.

JeRTeh seminar, Beograd

- Škola je održana u Budimpešti, u Centru za digitalnu humanistiku na Fakultetu društvenih nauka Univerziteta Etvoš Lorand, od 23. do 25. septembra 2019. godine.

https://www.distant-reading.net/

DAY	TIME	JOINT SESSION	TRAINING SCHOOL 1	TRAINING SCHOOL 2	TRAINING SCHOOL 3	CORE GROUP	WORKING GROUP 1	WORKING GROUP 2	WORKING GROUP 3	WORKING GROUP 4
MONDAY	13:30-14:00	Kari tanácsterem								
	14:00-15:45		A 320	A 330	Dékáni kistanácsterem	Main building 205				
	16:15-18:00		A 320	A 330	Dékáni kistanácsterem	Main building 205				
THUESDAY	9:00-10:45	A -150	A 320	A 330	Main building 206					
	11:15-13:00		A 320	A 330	Main building 206		A -150	A 329	A 429	Main building 205
	14:00-15:15		A 320	A 330	Main building 206		A -150	A 329	A 429	Main building 205
	15:45-17:00		A 320	A 330	Main building 206		A -150	A 329	A 429	Main building 205
	17:15-18:45	Kari tanácsterem								
WEDNESDAY	9:00-10:45	A 329	A 320	A 330	Dékáni kistanácsterem					
	11:15-12:45	A 329								

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

Distant Reading for European Literary History (COST Action CA16204)

TRACK 2: Natural Language Processing for Distant Reading

- Predavači - Andrew Janco (Haverford), David Lassner (Berlin), Leonard Konle (Würzburg).
- Fokus na obradi prirodnog jezika i prezentovanje biblioteke spaCy i njenih alata za programski jezik Python.

TRACK 2: Natural Language Processing for Distant Reading

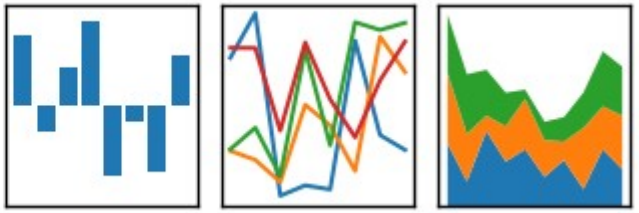
- Python programski jezik
- SpaCy biblioteka
- Sedam sesija

Sesija 1

- Uvod u programski jezik.
- Promenljive, tipovi podataka, petlje, objektno-orijentisano programiranje...

Sesija 2

- Rad sa tekstualnim datotekama
- lxml
- Xpath
- Pandas



Sesija 3 - spaCy

- Biblioteka za obradu prirodnih jezika (NLP)
- Podržava engleski, nemački, francuski, španski, portugalski, italijanski, holandski i grčki jezik, kao i multi-language.
- Prepoznavanje imenovanih entiteta i tokenizacija.

spaCy

```
nlp = spacy.load("en_core_web_sm")
text = "The rain in Spain falls mainly on the plain."
doc = nlp(text)

for token in doc:
    print(token.text, token.lemma_, token.pos_, token.is_stop)
```

	text	lemma	POS	explain	stopword
0	The	the	DET	determiner	True
1	rain	rain	NOUN	noun	False
2	in	in	ADP	adposition	True
3	Spain	Spain	PROPN	proper noun	False
4	falls	fall	VERB	verb	False
5	mainly	mainly	ADV	adverb	False
6	on	on	ADP	adposition	True
7	the	the	DET	determiner	True
8	plain	plain	NOUN	noun	False
9	.	.	PUNCT	punctuation	False

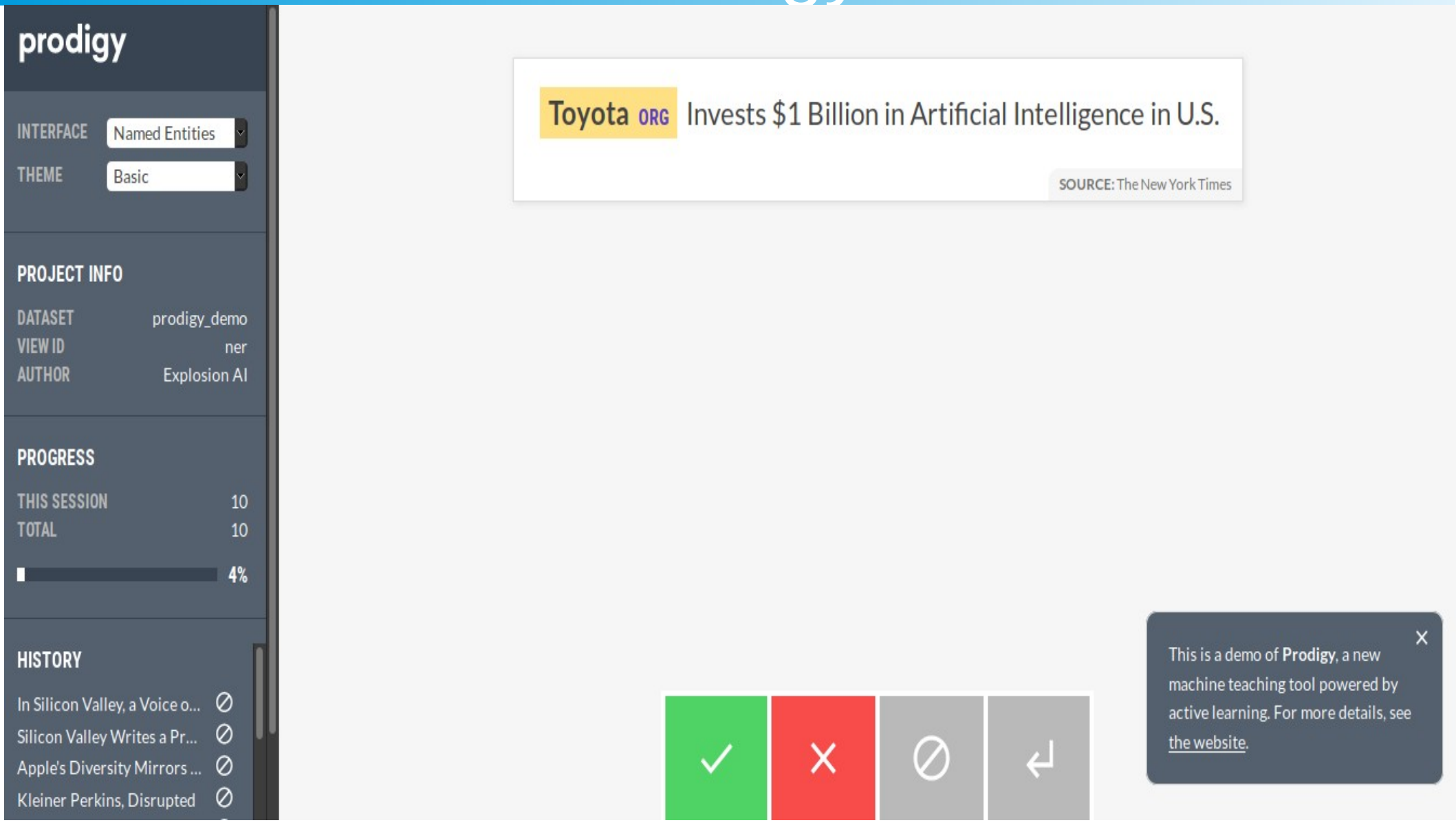
spaCy Matcher

- Pretraživanje delova teksta.
- Za razliku od regularnih izraza koji rade samo sa niskama, spaCy Matcher radi sa Doc i Token objektima.
- Moguće je ispisati pravila za pronalaženje određenih reči.

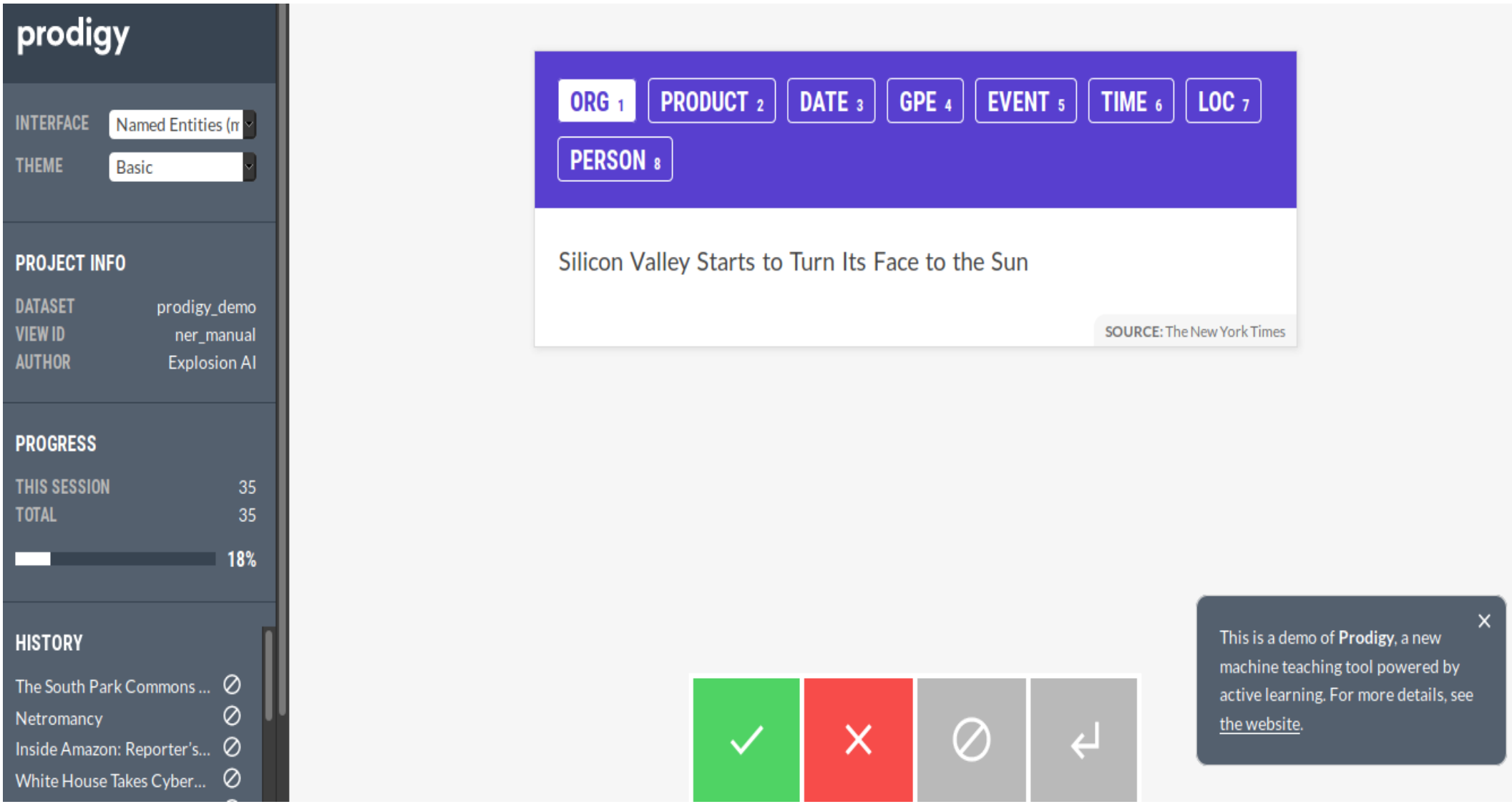
Prodigy

- Alat za anotaciju koji podržava aktivno mašinsko učenje.
- Koristi se veb interfejs kako bi proces obučavanja bio što brži i u realnom vremenu.
- Moguće je prepoznavanje imenovanih entiteta, part-of-speech, slika, objekata i slično.

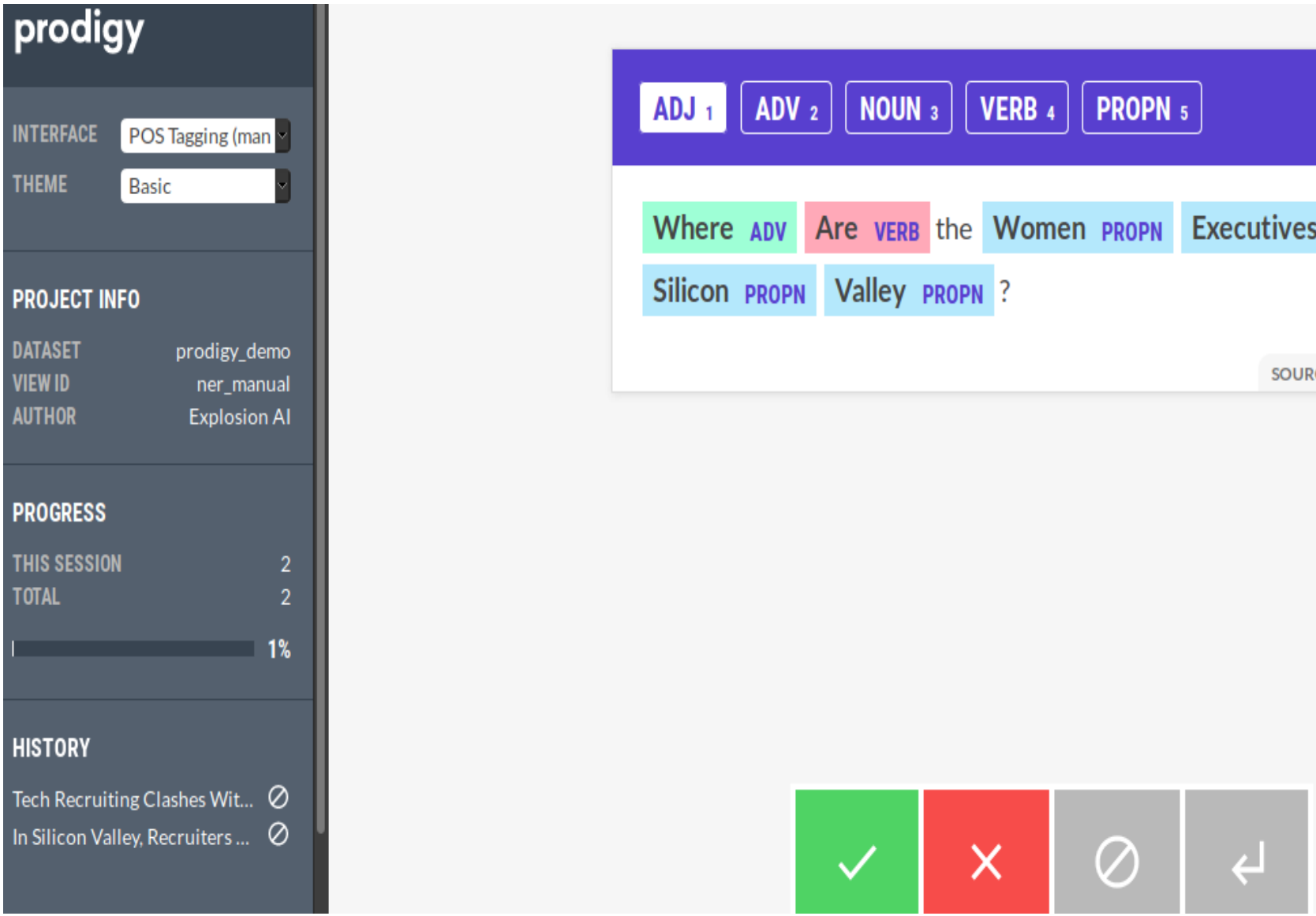
Prodigy


prodigy
INTERFACE
Named Entities
THEME
Basic
PROJECT INFO
DATASET
prodigy_demo
VIEW ID
ner
AUTHOR
Explosion AI
PROGRESS
THIS SESSION
10
TOTAL
10
4%
HISTORY
In Silicon Valley, a Voice o...
Silicon Valley Writes a Pr...
Apple's Diversity Mirrors ...
Kleiner Perkins, Disrupted
Toyota ORG Invests $1 Billion in Artificial Intelligence in U.S.
SOURCE: The New York Times
This is a demo of Prodigy, a new machine teaching tool powered by active learning. For more details, see the website.

Prodigy


prodigy
INTERFACE
Named Entities (n
THEME
Basic
PROJECT INFO
DATASET
prodigy_demo
VIEW ID
ner_manual
AUTHOR
Explosion AI
PROGRESS
THIS SESSION
35
TOTAL
35
18%
HISTORY
The South Park Commons ...
Netromancy
Inside Amazon: Reporter's...
White House Takes Cyber...
ORG 1
PRODUCT 2
DATE 3
GPE 4
EVENT 5
TIME 6
LOC 7
PERSON 8
Silicon Valley Starts to Turn Its Face to the Sun
SOURCE: The New York Times
This is a demo of Prodigy, a new machine teaching tool powered by active learning. For more details, see the website.

Prodigy


prodigy
INTERFACE
POS Tagging (man
THEME
Basic
PROJECT INFO
DATASET
prodigy_demo
VIEW ID
ner_manual
AUTHOR
Explosion AI
PROGRESS
THIS SESSION
2
TOTAL
2
1%
HISTORY
Tech Recruiting Clashes Wit...
In Silicon Valley, Recruiters ...
ADJ 1
ADV 2
NOUN 3
VERB 4
PROPN 5
Where ADV
Are VERB
the
Women PROPN
Executives PROPN
in
Silicon PROPN
Valley PROPN
?
SOURCE: The New York Times

Linkovi

Python - https://www.python.org/

lxml - https://lxml.de/

pandas - https://pandas.pydata.org/

spaCy - https://spacy.io/

Prodigy - https://prodi.gy/

Resursi - https://github.com/apjanco/spacY_workshops

Hvala na pažnji.