

# Ioannis Daras

## **Adding Greek language to spaCy**

Mentors:

Gogoulos Markos

Louridas Panos

# About me

---

Third year student at National Technical University of Athens, studying Electrical and Computer Engineering.

Academic interests: Algorithms, Machine Learning, Automatic Control.

Competition-holic: from Robotics to Machine Learning and back.

Founder of Rattle, a start-up that makes shopping easy.

# Why NLP?

and even worse;  
WHY GREEK NLP?

- 3.8 billion internet users,
- 120 million emails
- 500.000 Facebook comments
- 3 million Google searches

**EVERY MINUTE**

**Not enough tools in Greek  
NLP :(**

---

# Why spaCy?

- ❑ Pythonic
- ❑ State of the art algorithms
- ❑ Industrial length to academic research

# The project

- Lemmatizer ✓
- Tokenizer
- Stop-words ✓
- Norm-exceptions ✓
- Lexical Attributes ✓
- POS tagger ✓
- DEP tagger ✓
- NER extraction
- Text classification
- Sentiment Analysis



# Challenges & Linguistic Approach

---

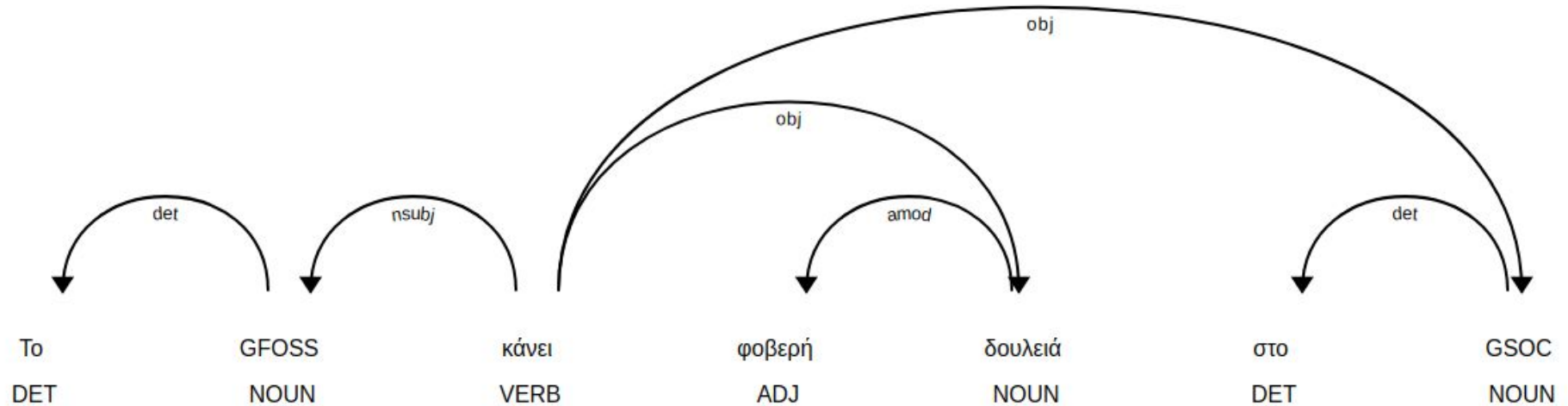
## Lemmatization Problem

```
NOUN_RULES = [  
    ["ες","α"], # κεραμίδες  
    ["ές","ά"],  
    ["ές","ά"],  
    ["ες","α"], # εσπερινές  
    ["ες","ης"], # φυλακές ]
```

## Stop-words Problem

**Gensim & Wikipedia!**

# POS and DEP tagger



PERSON 1

ORG 2

NORP 3

FAC 4

PRODUCT 5

EVENT 6

WORK\_OF\_ART 7

LOC 8

Ο **Β. Γκόρντον Τσάιλντ** PERSON εντόπισε έναν αριθμό παρόμοιων χαντρών στα μουσεία της κεντρικής και της ανατολικής **Ευρώπης.**

LOC

---

## NER Extraction

Learn history, organizations, art.  
**Understand the world.**





---

## Text Classification

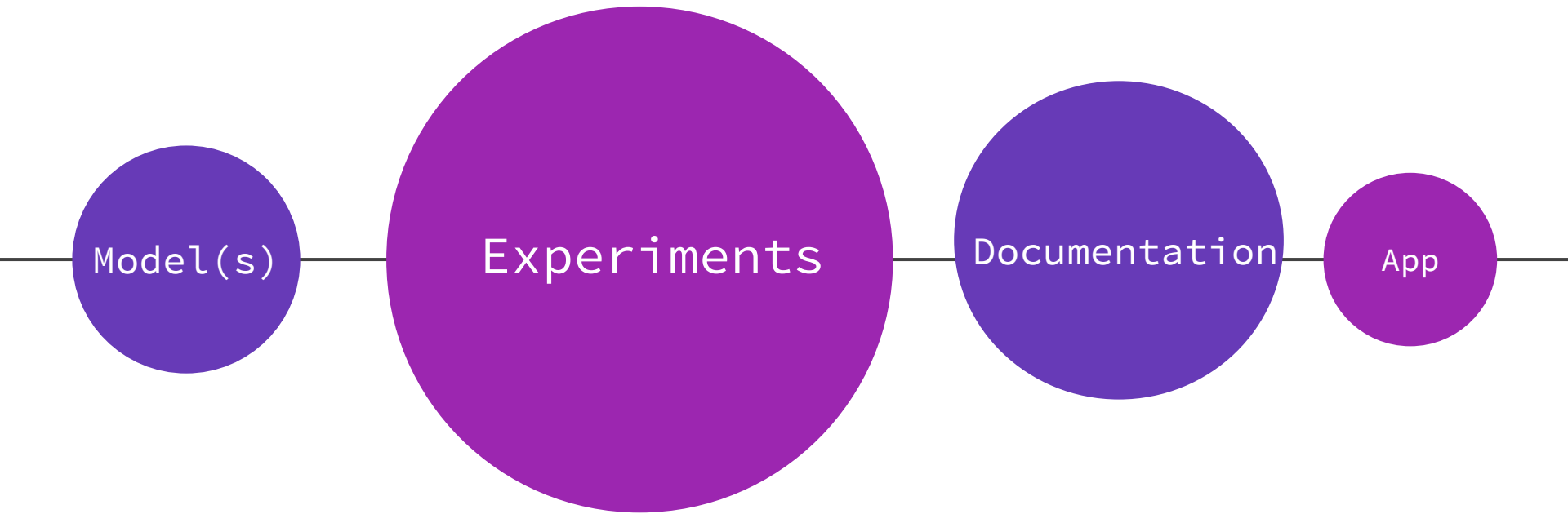
What the article is about? Racism?  
Politics? Art? **Learn topics.**

# TODOs

- ★ Tokenizer
- ★ NER
- ★ Text Classification
- ★ Sentiment



# Project Milestones



# Contact



Ioannis Daras

[daras.giannhs@gmail.com](mailto:daras.giannhs@gmail.com)

[www.linkedin.com/in/ioannis-daras](http://www.linkedin.com/in/ioannis-daras)

<https://github.com/giannisdaras>

