Named Entity Recognition for Tweets: a Hands-on Session

Ioannis Partalas and Georgios Balikas

Grenoble Data Science Meet-up
February 2017

Short Bio

- I. Partalas: Data science researcher

  [ioannis.partalas@gmail.com]
- G. Balikas: 3rd year PhD, UGA

  [geompalik@hotmail.com]

- Meet-up 2016: e-commerce product classification

Named Entities

- Text spans from a single to a few words
- Persons, Organizations, Locations, …

Jim$_{\text{Person}}$ bought 300 shares of Acme Corp.$_{\text{Organization}}$ in 2006$_{\text{Time}}$.

Named-Entity Recognition

- Segmentation & Classification

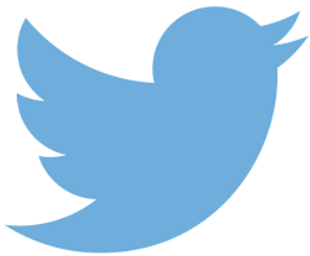
[Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.

Practical Questions

- How many entity types?
- Feature Engineering
- Model Selection
- Model Evaluation

Entity types (today)

- We will use 10 entity types
- Person, Company, Facility, Geo-loc, Movie, Music Artist, Product, Sports team, Tv show, Other



Feature Engineering

- The goal is to translate our intuition into robust features (morhpo-syntactic, contextual, ..)

Jim$_{\text{Person}}$ bought 300 shares of Acme Corp.$_{\text{Organization}}$ in 2006$_{\text{Time}}$.

Twitter Examples

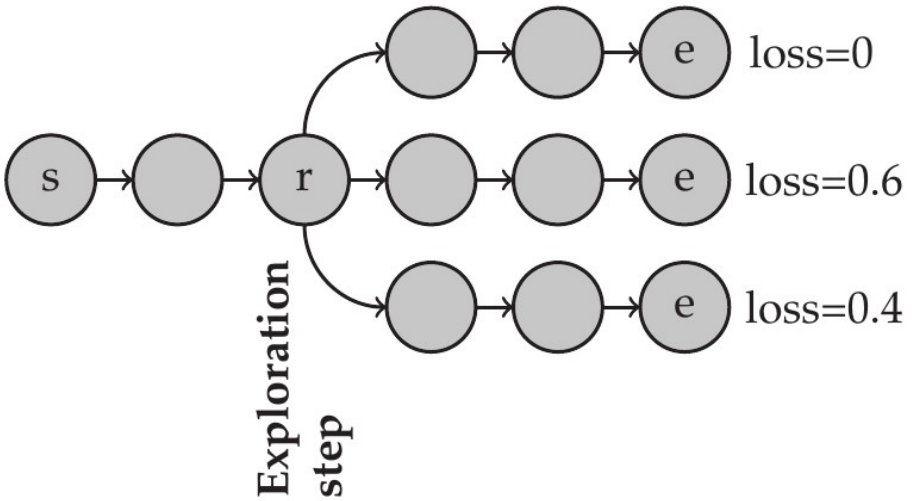
‘ Breaking$_{\text{B-movie}}$ Dawn$_{\text{I-movie}}$ ’ Returns to Vancouver$_{\text{B-geo-loc}}$

The$_{\text{B-sportsteam}}$ Wolves$_{\text{I-sportsteam}}$ to host the$_{\text{B-sportsteam}}$ Lions$_{\text{I-sportsteam}}$ for game time!

#SIUC Whats the Plan For Tonight?? Whos Goin to the$_{\text{B-facility}}$ Blast$_{\text{I-facility}}$

The model

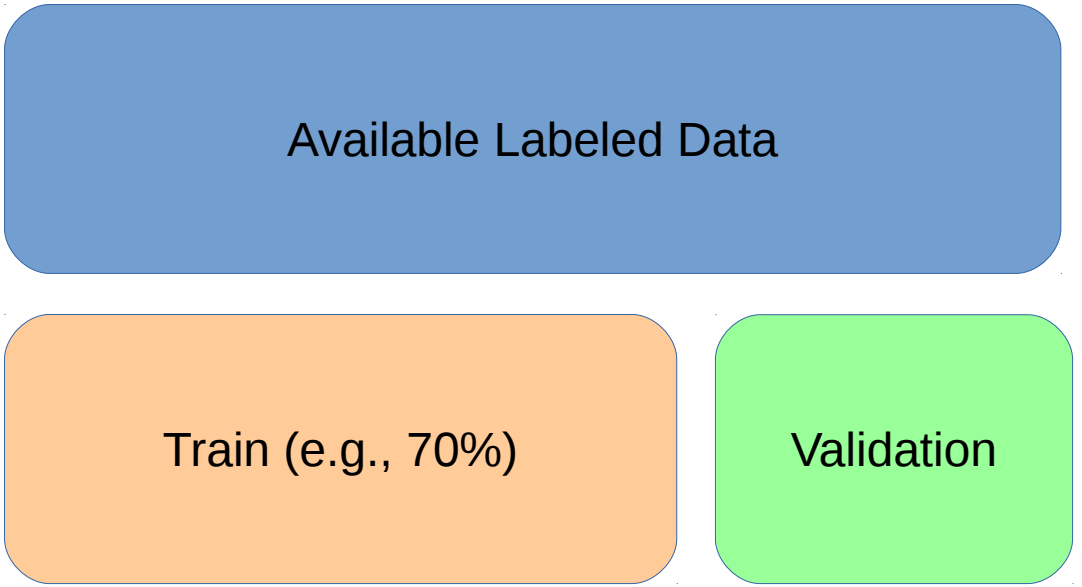
- Voppal Wabbit: memory efficient/fast/online
- Scales well, supports several models
- Today: Learning2search
- Structural learning
- Decomposes structured problems in a search space with states actions and policies



VOWPAL WABBIT

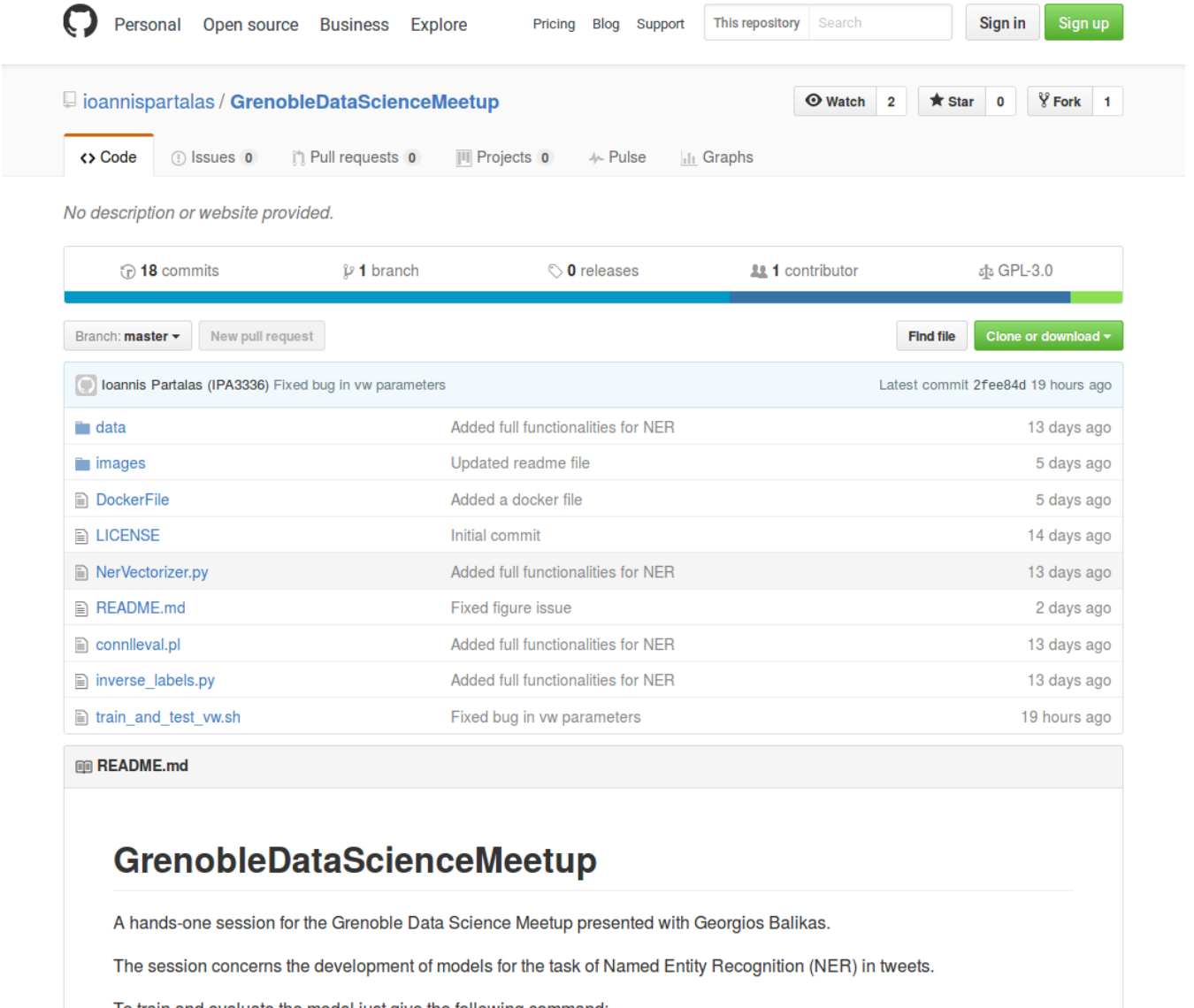
Model Validation

- Today: Train-Validation split



Hands-on

- https://github.com/ioannispartalas/GrenobleDataScienceMeetup



Steps

- Assuming you have installed vw, python,...
- Download the code and data
- git clone https://github.com/ioannispartalas/GrenobleDataScienceMeetup.git

“NerVectorizer.py”: vectorization

./train_and_test_vw.sh train_dev2015 dev

How to proceed

- Discuss the contents of NerVectorizer.py
- Imagine more/better features
- Implement them
- Improve the F_1-measure !!

What's next

- http://cap2017.imag.fr/competition.html
- NER task with French tweets
- 600e prize for the first