

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
[\[ioannis.partalas@gmail.com\]](mailto:ioannis.partalas@gmail.com)
- G. Balikas: 3rd year PhD, UGA
[\[geompalik@hotmail.com\]](mailto: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_{Person} bought 300 shares of Acme
Corp._{Organization} in 2006_{Time}.

Named-Entity Recognition

- Segmentation & Classification

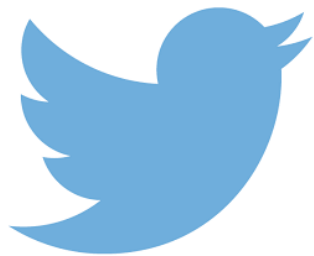
[Jim]_{Person} bought 300 shares of [Acme Corp.]_{Organization} in [2006]_{Time}.

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 (morpho-syntactic, contextual, ..)

Jim_{Person} bought 300 shares of Acme
Corp._{Organization} in 2006_{Time}.

Twitter Examples

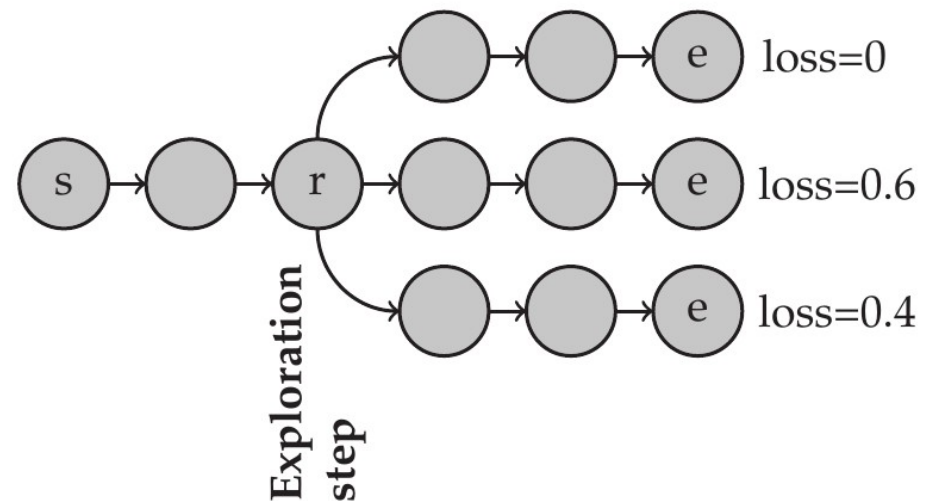
‘ **Breaking**_{B-movie} **Dawn**_{I-movie} ’ Returns to **Vancouver**_{B-geo-loc}

The_{B-sportsteam} **Wolves**_{I-sportsteam} to host **the**_{B-sportsteam} **Lions**_{I-}
_{sportsteam} for game time!

#SIUC Whats the Plan For Tonight?? Whos Goin to
the_{B-facility} **Blast**_{I-facility}

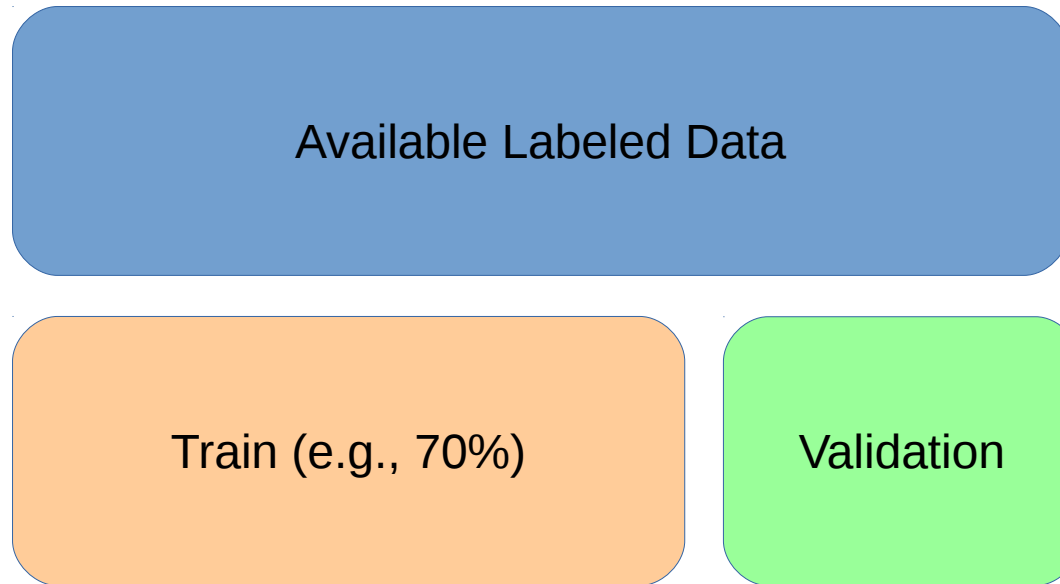
The model

- 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



Model Validation

- Today: Train-Validation split



Hands-on

- <https://github.com/ioannispatalas/GrenobleDataScienceMeetup>

The screenshot shows the GitHub repository page for `ioannispatalas / GrenobleDataScienceMeetup`. The repository has 2 watchers, 0 stars, and 1 fork. It is currently on the `master` branch. The repository statistics show 18 commits, 1 branch, 0 releases, 1 contributor, and a GPL-3.0 license. The commit history is as follows:

Commit	Description	Time
Ioannis Patalas (IPA3336)	Fixed bug in vw parameters	Latest commit 2fee84d 19 hours ago
data	Added full functionalities for NER	13 days ago
images	Updated readme file	5 days ago
DockerFile	Added a docker file	5 days ago
LICENSE	Initial commit	14 days ago
NerVectorizer.py	Added full functionalities for NER	13 days ago
README.md	Fixed figure issue	2 days ago
connlleval.pl	Added full functionalities for NER	13 days ago
inverse_labels.py	Added full functionalities for NER	13 days ago
train_and_test_vw.sh	Fixed bug in vw parameters	19 hours ago

The `README.md` file content is as follows:

GrenobleDataScienceMeetup

A hands-one session for the Grenoble Data Science Meetup presented with Georgios Balikas.

The session concerns the development of models for the task of Named Entity Recognition (NER) in tweets.

To train and evaluate the model just give the following commands:

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

CAp 2017

Conférence sur l'Apprentissage Automatique



28 – 30 juin 2017 *Grenoble*