Natural Language Processing with Python

github.com/bonzanini/nlp-tutorial

@MarcoBonzanini and @MiguelMAlvarez

Nice to Meet You

Marco Bonzanini
Freelance Data Scientist

github.com/bonzanini/nlp-tutorial

Schedule

- Intro & Logistics (10m)

- Environment Set Up (10m)

- Exploring Text Data (1h + 15m QA)

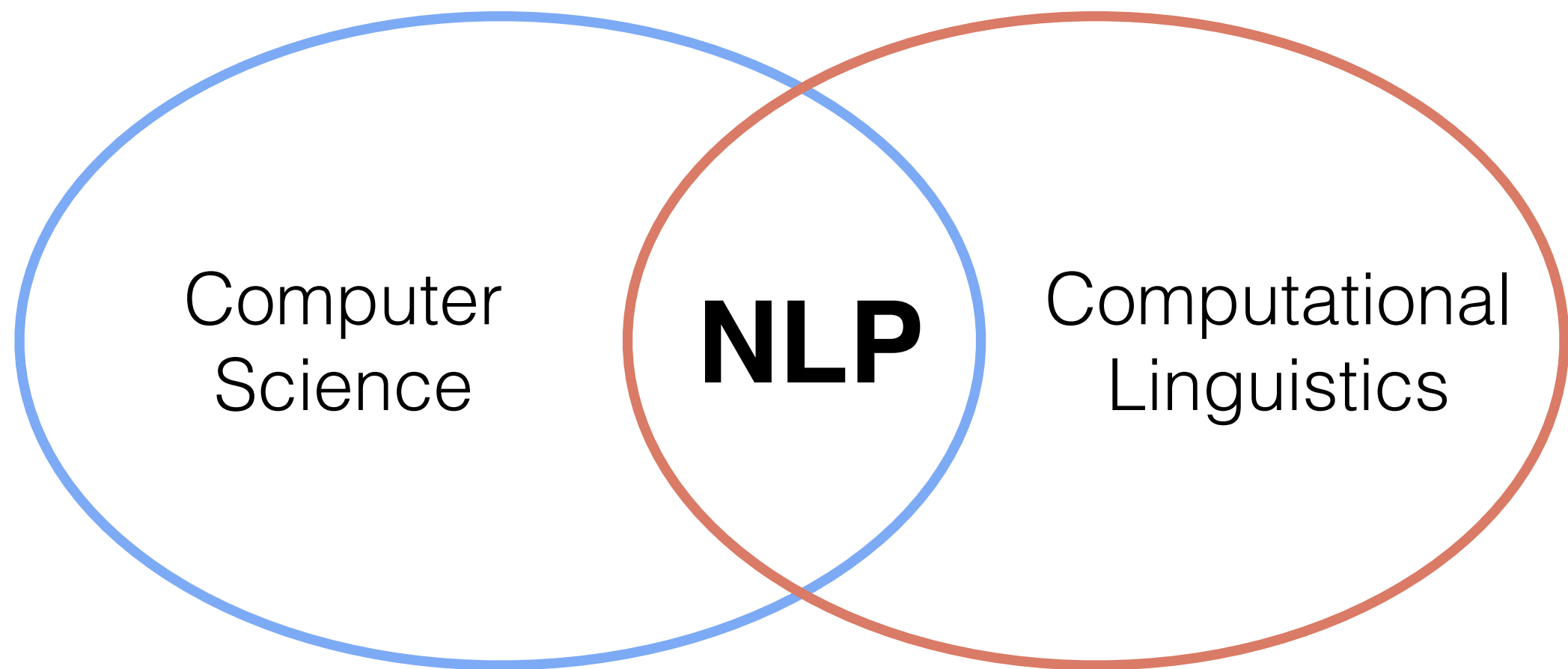- (+ a lot of bonus content in the repo)

`github.com/bonzanini/nlp-tutorial`

The Audience (You!)

- Know some Python already?

- Know some NLP already?

- Both / None of the above?

github.com/bonzanini/nlp-tutorial

Natural Language Processing

Computer Science

NLP

Computational Linguistics

github.com/bonzanini/nlp-tutorial

NLP Goals

Text Data $\Longrightarrow$ Useful Information
Actionable Insights

github.com/bonzanini/nlp-tutorial

Formal vs Natural

```
SELECT name, address
FROM businesses
WHERE business_type = 'pub'
AND postcode_area = 'CF10'
```

vs

Where is the nearest pub?

github.com/bonzanini/nlp-tutorial

NLP Applications

- Text Classification

- Text Clustering

- Text Summarisation

- Machine Translation

- Semantic Search

- Sentiment Analysis

- Question Answering

- Information Extraction

`github.com/bonzanini/nlp-tutorial`

Environment Set Up

- Tested with Python 3.4 and 3.5

- Clone the repository:

```
git clone https://github.com/bonzanini/nlp-tutorial
cd nlp-tutorial
```

Environment Set Up (cont'd)

- Set up virtual environment:

```
virtualenv nlp-venv
source nlp-venv/bin/activate
pip install -r requirements.txt
```

Environment Set Up (cont'd)

- Set up virtual environment (alternative):

```
conda create --name nlp-venv python=3.5
source activate nlp-venv
pip install -r requirements.txt
```

Environment Set Up (cont'd)

- Download NLTK data:

```
python -m nltk.downloader \
    punkt stopwords reuters
```

Environment Set Up (cont'd)

- Start up Jupyter notebook:

```
jupyter notebook
```

Exploring Text Data

Exercise

What are the most important ingredients in Italian cuisine?

recipes_exploratory_analysis.ipynb

Recipe Analysis: Summary

- Tokenisation

- Counting words

- Stop-words

- Normalisation

- Stemming

- n-grams

pyconuk_exporatory_analysis.ipynb

PyConUK Analysis Summary

- Data Visualisation

- "This talk will …"

- TF-IDF

Questions?

MSMMPEB50 (ebook 50%)
MSMMPPB15 (paperback 15%)

on packtpub.com

until 30th Nov

Community Experience Distilled

Mastering Social Media Mining with Python

Acquire and analyze data from all corners of the social web with Python

Marco Bonzanini

[PACKT]
PUBLISHING