Web and Information Retrieval Course

(Read it as: Natural Language Processing)

About me

Full Professor of Software Engineering

University of Sannio, Italy

My Research: Software Engineering

Automated tools to help developers

AI applications to software engineering

Software testing

DevOps

What about you?

- How many exams left?

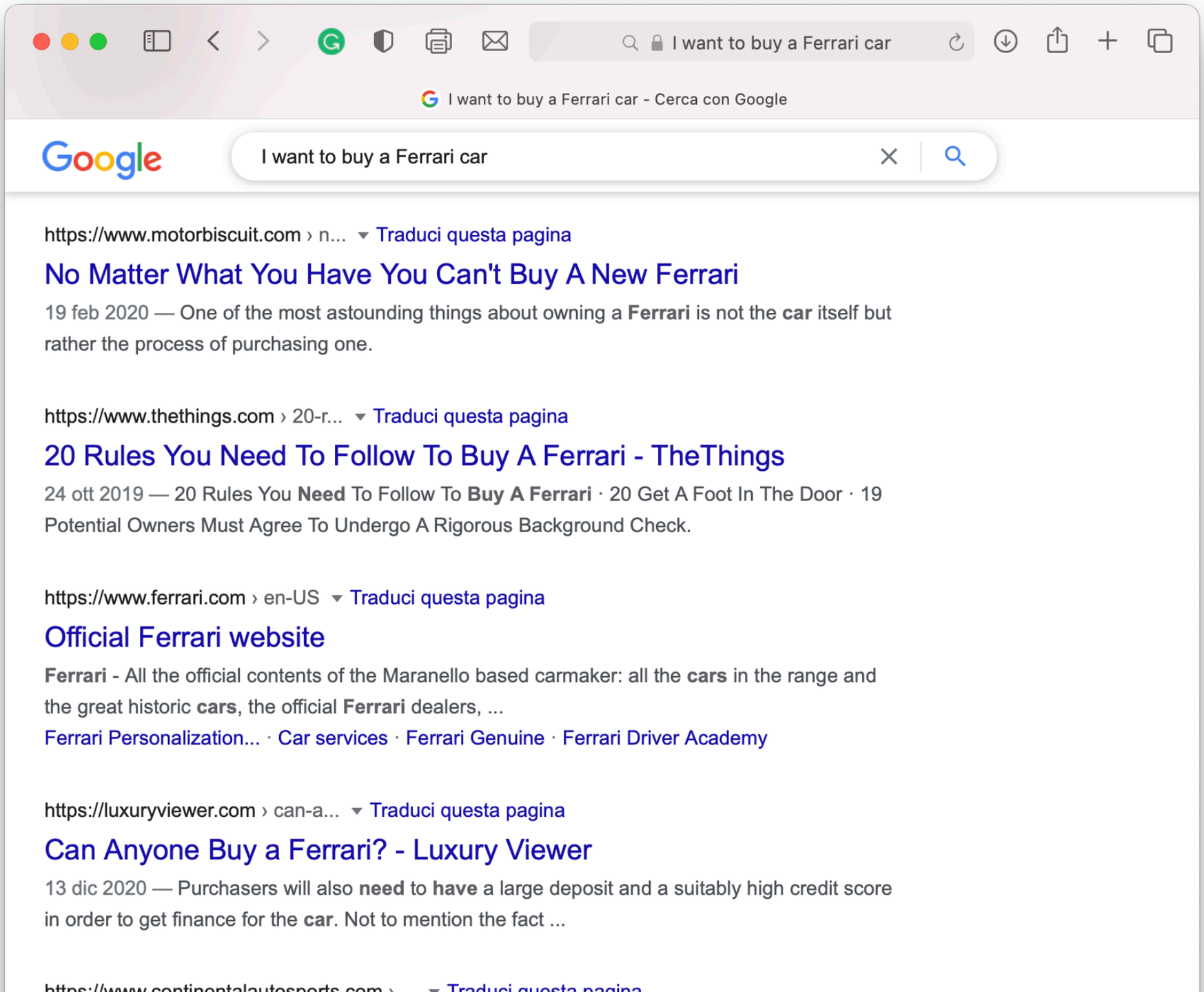
- Did you choose Data Analytics?

Course Goal

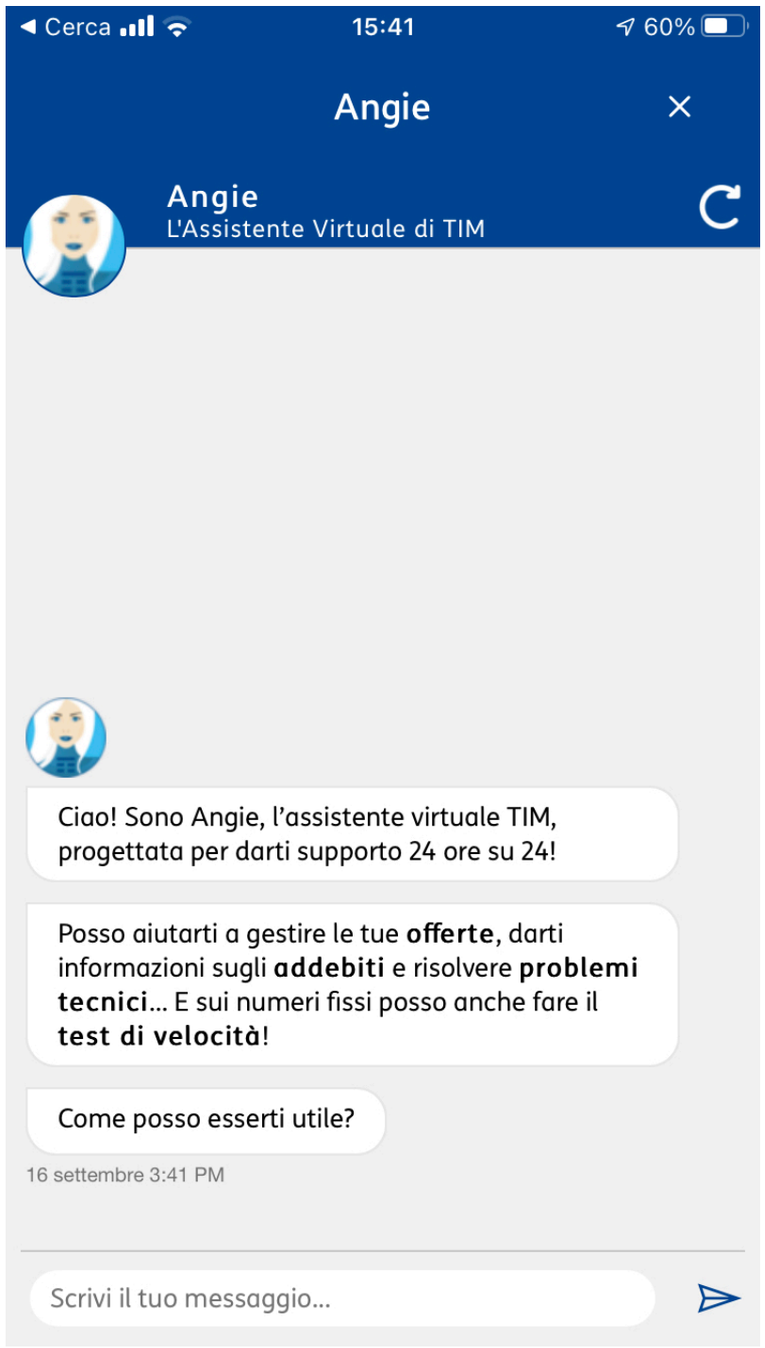
We will study the application of techniques from Artificial Intelligence, computational linguistics, and Information Retrieval to process natural language

So what?

Text-based search and comparison

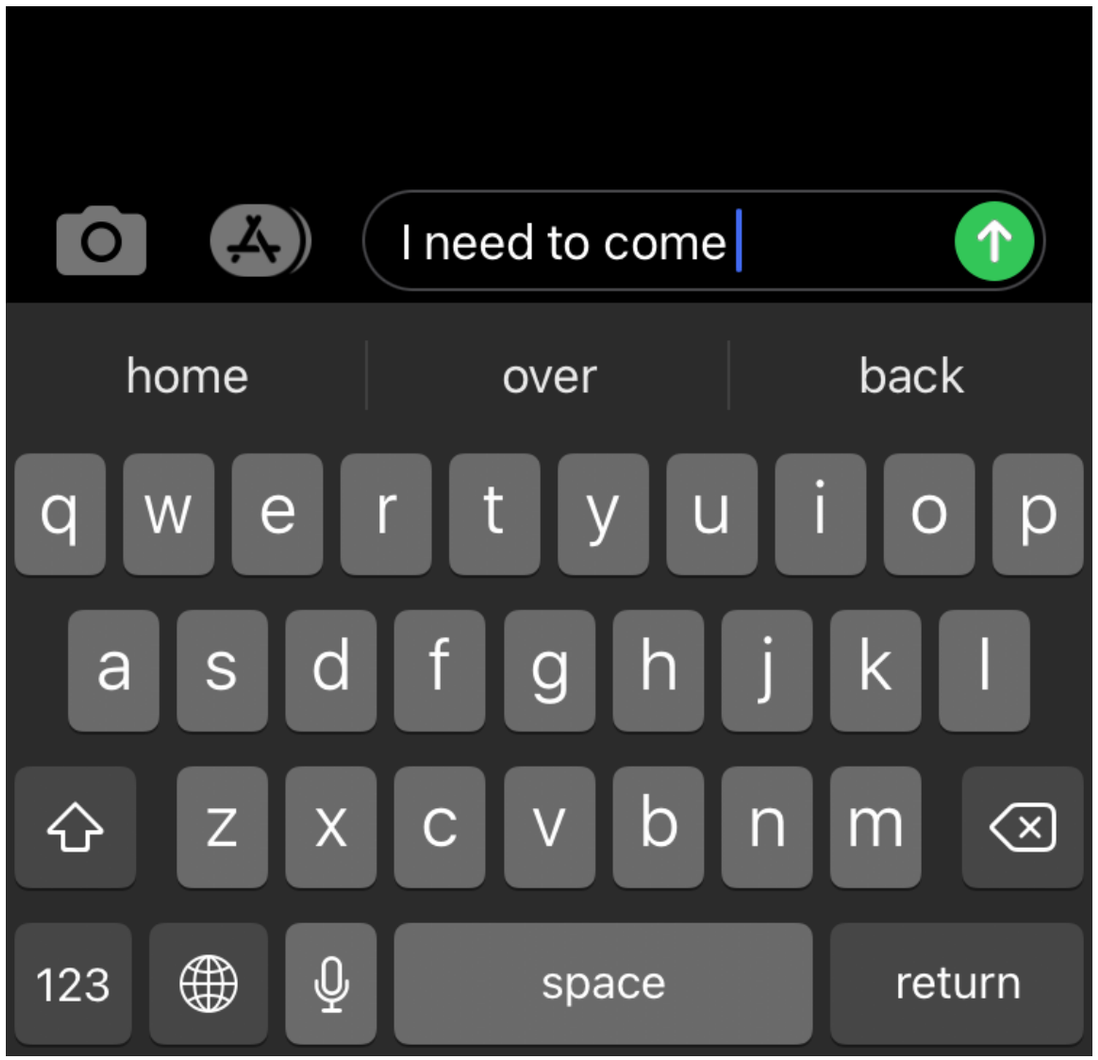


Chatbots


Cerca
15:41
60%
Angie
Angie
L'Assistente Virtuale di TIM
Ciao! Sono Angie, l'assistente virtuale TIM, progettata per darti supporto 24 ore su 24!
Posso aiutarti a gestire le tue offerte, darti informazioni sugli addebiti e risolvere problemi tecnici... E sui numeri fissi posso anche fare il test di velocità!
Come posso esserti utile?
16 settembre 3:41 PM
Scrivi il tuo messaggio...

Personal Assistants

More Assistants...



Large Language Models

DI Write a Python regular expression matching a date and time

Here's a regular expression in Python that matches a date and time in the format of "YYYY-MM-DD HH:MM:SS":

python Copy code

```python
import re

datetime_regex = r"\d{4}-\d{2}-\d{2} \d{2}:\d{2}:\d{2}"
```

This regex matches a date in the format of "YYYY-MM-DD" followed by a space, then the time in the format of "HH:MM:SS". The `\d` character class matches any digit, and `{4}`, `{2}` and `{2}` are quantifiers that specify how many digits to match for the year, month and day respectively, as well as hours, minutes and seconds. The `\s` matches any whitespace character and the `:` matches the colon character.

Analytics

facebook

LinkedIn

Google

twitter

Pre-requirements

- Courses that must be already taken: Math, Math for information engineering, Programming 2
- In essence we need knowledge about:
  - Programming (though we will use a different programming language)
  - Statistics, probability
  - Vector algebra

Course Syllabus

Overview

- Introduction
- Python crash course
- Text processing
- Vector Space models
- Statistical language models
- Basics of text classification
- Text processing with neural networks
- Advanced deep learning models

Introduction

- Basics of machine learning
- Applications in the area of natural language processing
- NLP: the genesis and trends

Python crash course

- Introduction to the language - differences with Java
- Main language constructs
- Object-oriented programming
- Using and installing libraries
- Popular data science libraries

Text processing

- Textual analysis through regular expressions
- Text editing distance
- Typical information retrieval process
- Text normalization
- Stemming and lemmatization

Vector space models

- Term weighting
- Introduction to vector space models
- Similarity measures
- Representations - Inverted index
- Feedback models - Rocchio

Statistical language models

- Language models - n-grams
- Recommenders based on n-grams
- Probabilistic models for text classification - Naive Bayes Models

Advanced Models

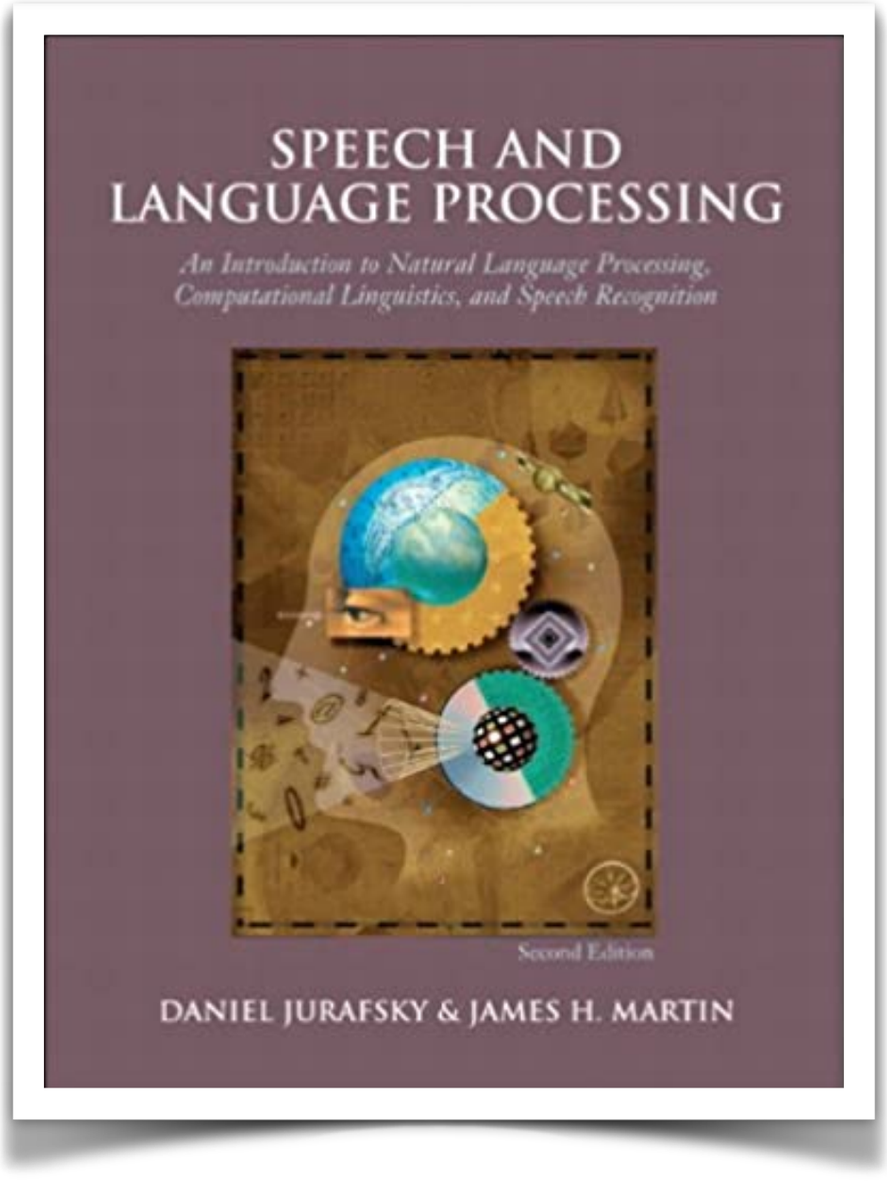
- Latent Semantic Indexing - Latent Dirichlet Allocation
- Word Embedding Models
- Neural Networks, Deep Learning
- Pre-trained transformers for natural language processing
    - Classification, summarization, translation tasks
- Large language models, hybrid models
(e.g., Retrieval Augmented Generation)

Applications

- Text classification
- Document retrieval systems
- Recommender systems
- Chatbots

Course Material

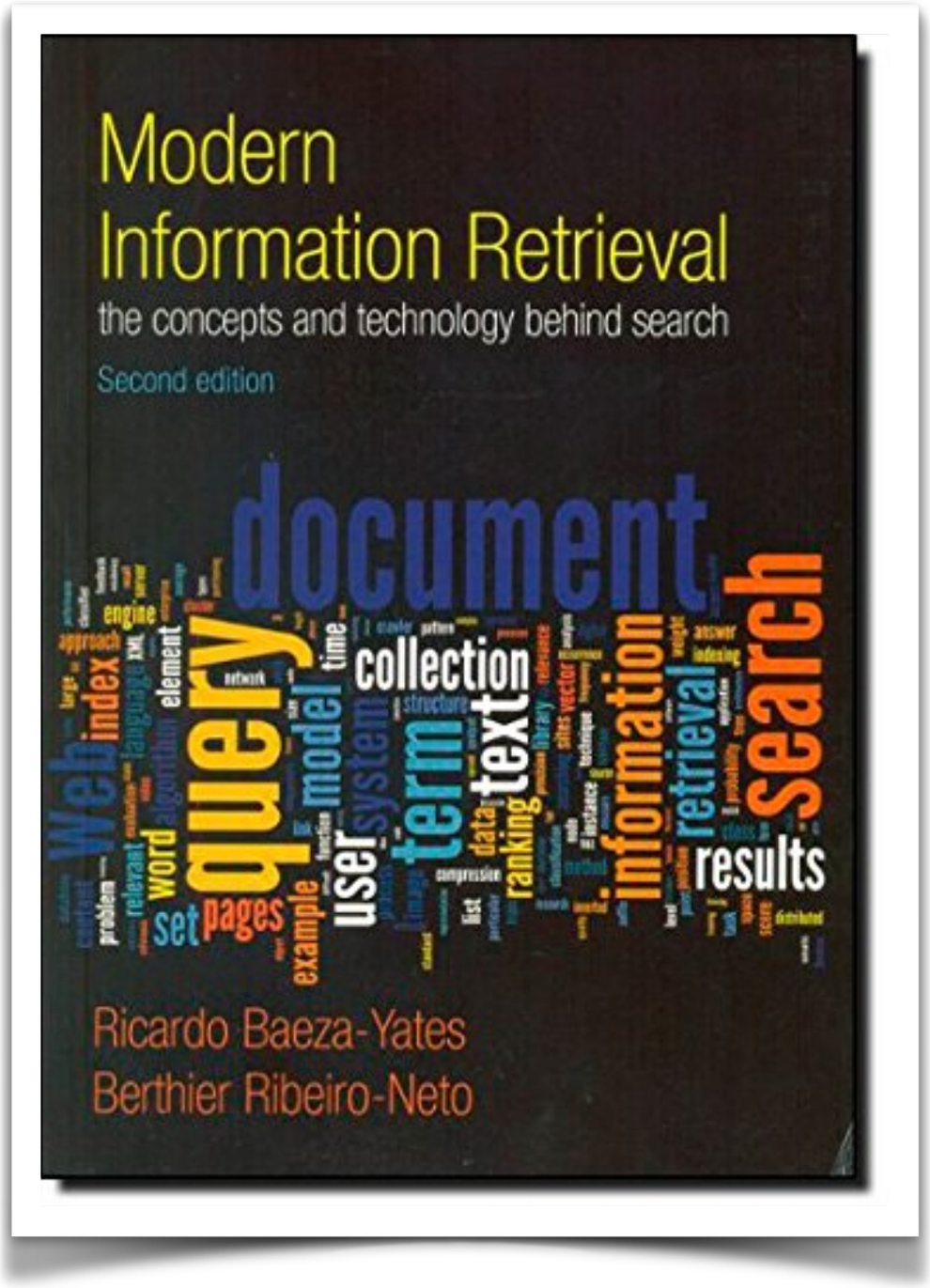
NLP Book



Daniel Jurafsky, James H. Martin - Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

https://web.stanford.edu/~jurafsky/slp3/

(Optional) IR Book



Ricardo Baeza-Yates, Berthier Ribeiro-Neto - Modern Information Retrieval: The Concepts and Technology Behind Search. Addison-Wesley Professional; 2° ed. (23 Dec 2010)

Course Handy

- https://handy.unisannio.it/

- Course slides, scripts, datasets

- Working group & communication

Exam

The Exam

- Project (50%)

- Oral (50%)

Project types

- Individual work, or small team (2 students)
- Realize a small natural language processing tool, and analyze a dataset
- Information retrieval, document classification, simple chatbot/personal assistant

Projects - Notes

- **It cannot be one of the projects we have seen during lectures**
- **It must be a machine learning application to text** (natural language, but also source code)
    - Applications of machine learnings to other datasets are not considered as valid projects

Contact

- Email: dipenta@unisannio.it
- Office: Via Traiano 9 (RCOST) 2nd floor
- Web: https://mdipenta.github.io