University of Amsterdam
Economics & Business

Data Science Minor – 2020/2021

Text Retrieval
Text Mining

Julien ROSSI

Kick-Off

- Understand the challenges of **text representation**
- Know about the important machine learning tasks
- Know **what to expect** from this course
- Know **what is expected** from you

Who am I ?

Julien ROSSI

ABS Lecturer, Researcher

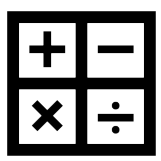
Msc Computer Sciences
MBA Big Data & Business Analytics (ABS)
PhD (in progress)

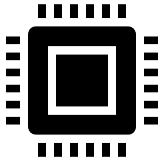
Ex-Software Developer, ICT Engineer, ICT Manager

Research in Legal Text Analytics

Co-Founder “TalkAir” – Text Analytics for Professionals

This course in a nutshell





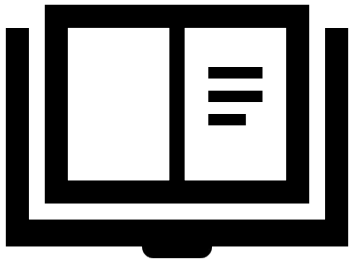
Represent texts in a way that a computer can assist us to perform useful tasks

CLASSIFICATION
What kind of book is this?

RETRIEVAL
Which books are about dinosaurs?

CLUSTERING
To what group of books is it the closest?

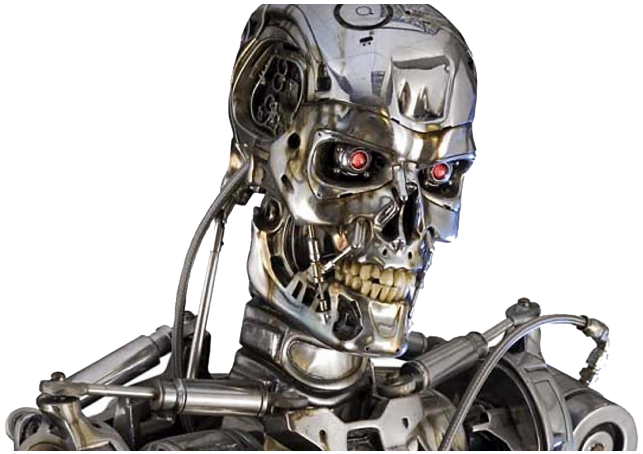
From TEXT to VECTORS

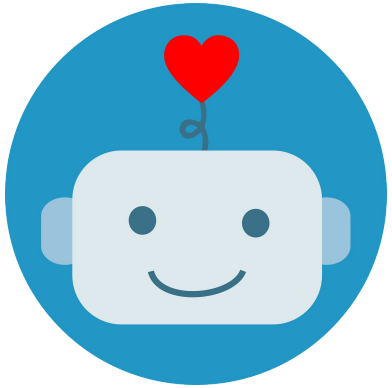


In “Small Gods”, Terry Pratchett

“Time is a drug. Too much of it kills you.”

[0.00123,-1.4567, 0.00000,..., 0.02]





What kind of text?

Tweets

- Sentiment Analysis (classification)
- Topics & Trends (clustering)
- Hate speech (classification)

Tweet illustration from TechCrunch Link
Viz llustration from a Blog Post by Chulong Li Link

What kind of text?

Parliamentary Archives

- Topics (Clustering)
- Question Answering (Retrieval)
- Argument (Retrieval)

Illustration “Handlingenkamer, Binnenhof, Den Haag, Netherlands” Link

What kind of text?

Anything, Any language!!

- News articles, Blog posts
- Mails, Company Documents
- Discussion forums, Reddit
- Contracts, Laws, Regulations, Court Cases
- Academic literature
- Meeting minutes, Video captions
- Etc…

What tasks?

Anything, Any language!!

Clustering news: "*which articles talk about the same topic?*"

Classification of Mails: "*which ones discuss a fraud scheme?*"

Retrieval of Case Law: "*which similar cases can I refer to?*"

Retrieval in Meeting minutes: "*when did we talk about this?*"

Retrieval in Documentation:

- **如何更换iPhone电池？** كيفية استبدال بطارية الايفون؟
- как заменить аккумулятор iphone? How to replace an iphone battery?

Course Schedule

- 4 weeks
  - 2h Lecture: Concepts, Theory
  - 2h Tutorial: Practice (3 groups)

- Assessment
  - 1 group assignment (40%)
  - 1 final exam (60%) (Take-home)

Course Schedule

- Group Assignment (scope: week 1 & 2)
  - Available January 11
  - Due January 21 23:59 CET

- Final Exam (Take-Home = Canvas Assignment)
  - January 28
  - Available on Canvas 09:00 CET
  - Due 11:00 CET

Assessment

- Passing grade
  - Average > 5.5
  - Exam > 5.0

- Resit
  - Average > 5.5
  - Resit > 5.0

Tools

- **ZOOM** for lectures and tutorials
  - See Canvas

- **SLACK** for interactions
  - During lectures and tutorials
  - Outside of contact hours
  - See Canvas

Tools

- **Google Colab** for notebooks
  - Your own laptop is also fine !

- **Python**
  - 3.7+
  - Modules: NumPy, Pandas, Matplotlib, NLTK, SpaCy, Gensim, Sci-Kit Learn, Transformers

Class Material

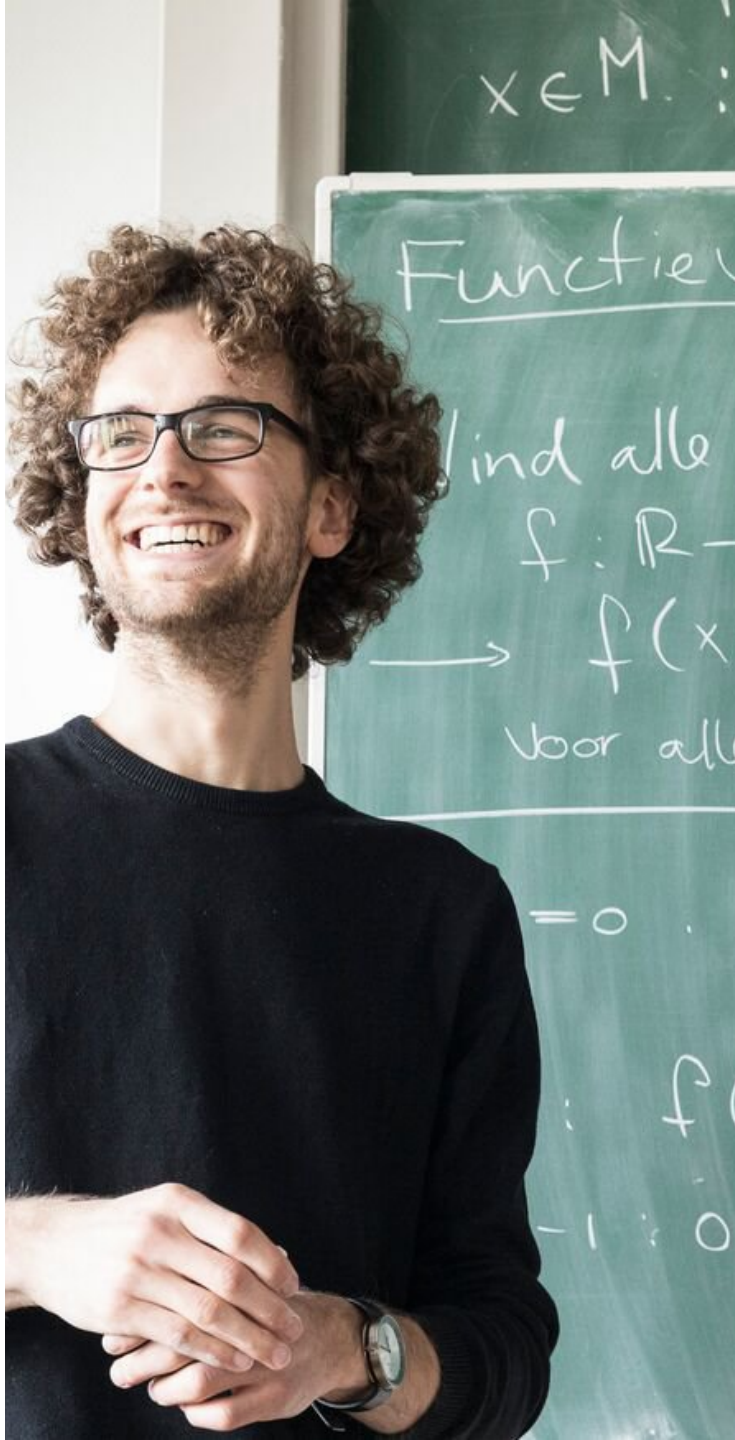
- **Available on Canvas**

- Lecture Slides
- Lecture Notebooks (links to Google Colab notebooks)
- Lecture recordings
- Tutorial Notebooks with Exercises
- Solutions for Tutorial Notebooks

Lectures

- Slides + Notebooks
- Interactive
  - Ask questions through the **Slack** chat at any time !

- Illustrated
  - Notebook will illustrate concepts
  - Whiteboard is available



Tutorials

- Notebooks with Exercises
- Interactive
  - Ask questions through the **Slack** chat at any time !

- Practical
  - Solve the exercises during the tutorial
  - Apply knowledge to a typical task
  - Use real text corpus