



# Text Analytics

*Introduction*

2025–26

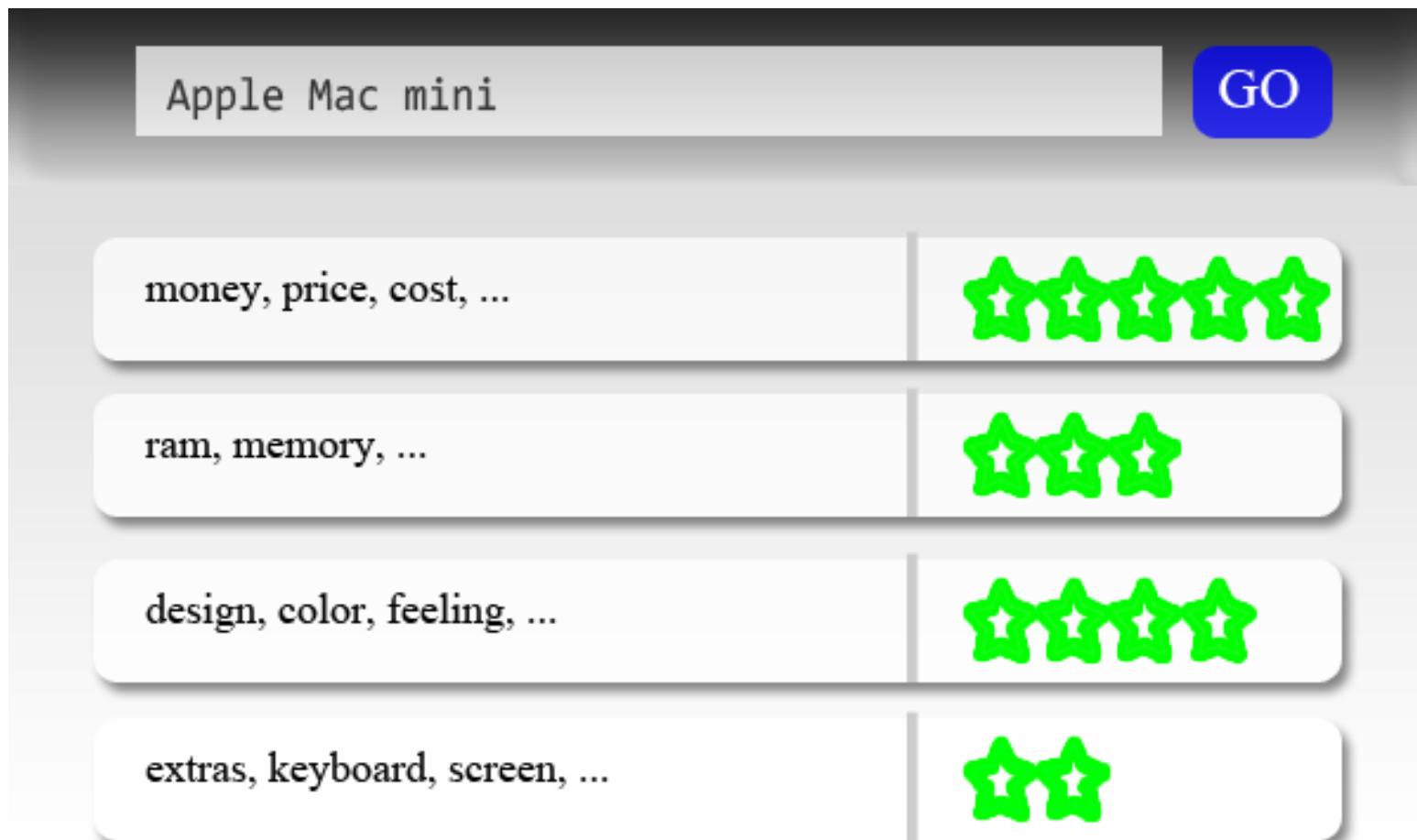
Ion Androutsopoulos

<http://www.aueb.gr/users/ion/>

# Contents

- **What** this course is about.
- **Examples** of applications.
- **Topics** to be covered.
- **Organization** of the course:
  - Lectures, demos.
  - Assignments, exams, grading policy.

# Opinion mining



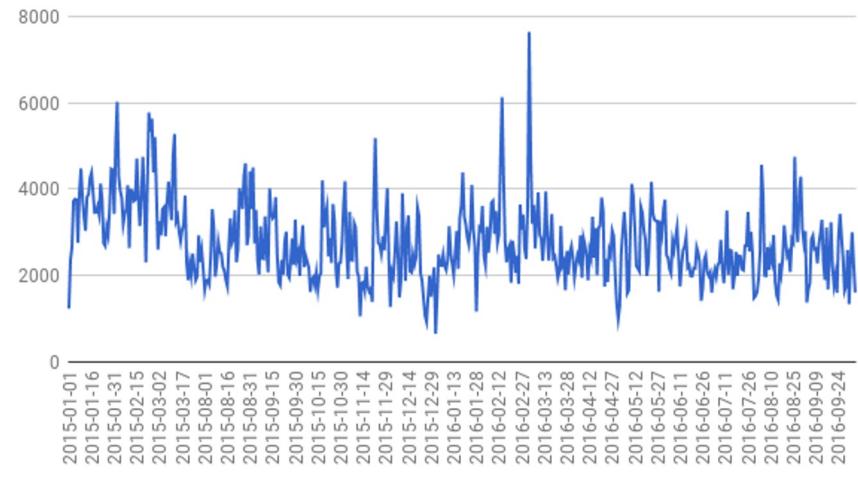
- From **customer reviews** (<http://alt.qcri.org/semeval2016/task5/>).
- From **tweets** (<http://alt.qcri.org/semeval2016/task4/>), **FB updates**, **blogs**, **newspapers**, **radio/TV broadcasts**, ...

# User comment moderation

A moderation panel assists the moderators to detect abusive comments, and leads to quicker publication of non-abusive comments.

Highlighting suspicious words using an RNN with self-attention.

Number of comments per day



## Moderation Panel

Go	and	hang	yourself	!	85%						
You	are	ignorant	and	vandal	!	88%					
Hello	there	try	to	relax	0%						
Thanks	.	Please	go	f#\$@	yourself	.	Ty	!	85%		

# Extracting Contract Elements

THIS AGREEMENT is made the **15th day of October 2009** (The “**Effective Date**”) BETWEEN:

- (1) **Sugar 13 Inc.**, a corporation whose office is at James House, 42-50 Bond Street, London, EW2H TL (“**Sugar**”);  
(2) **E2 UK Limited**, a limited company whose registered office is at 260 Bathurst Road, Yorkshire, SL3 4SA (“**Provider**”).

## RECITALS:

- A. The Parties wish to enter into a framework agreement which will enable Sugar, from time to time, to [...]  
B. [...]

## NO THEREFORE IT IS AGREED AS FOLLOWS:

### ARTICLE I - DEFINITIONS

- “**Sugar**” shall mean: Sugar 13 Inc.  
“**Provider**” shall mean: E2 UK Limited  
“**1933 Act**” shall mean: **Securities Act of 1933**

### ARTICLE II - TERMINATION

The Service Period will be for **five (5) years** from the Effective Date (The “**Initial Term**”). The agreement is considered to be terminated in **October 16, 2014**.

### ARTICLE III - PAYMENT - FEES

During the service period monthly payments should occur. The estimated fees for the Initial Term are **£154,800**.

### ARTICLE IV - GOVERNING LAW

This agreement shall be governed and construed in accordance with the **Laws of England & Wales**. Each party hereby irrevocably submits to the exclusive jurisdiction of the courts sitting in **Northern London**.

**IN WITNESS WHEREOF**, the parties have caused their respective duly authorized officers to execute this Agreement.

**BY:** George Fake  
Authorized Officer  
Sugar 13 Inc.

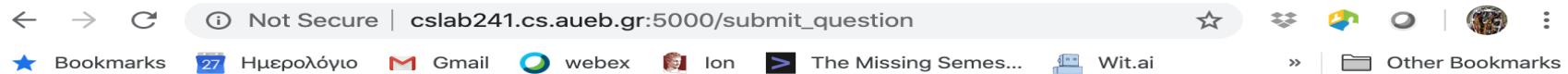
**BY:** Olivier Giroux  
CEO  
E2 UK LIMITED

Extract start/end dates, durations, contractors, amount, legal references, jurisdiction etc.

I. Chalkidis, I. Androutsopoulos and A. Michos, “Extracting Contract Elements”, ICAIL 2017, <http://nlp.cs.aueb.gr/pubs/icail2017.pdf>.

I. Chalkidis and I. Androutsopoulos, “A Deep Learning Approach to Contract Element Extraction”, JURIX 2017, <http://nlp.cs.aueb.gr/pubs/jurix2017.pdf>.

# IR/QA for Document Collections



## Results for the question: How is Covid-19 transmitted?

Title: Coronavirus (COVID-19) Outbreak: What the Department of Radiology Should Know

Date: 2020-02-19 || Section: PRECAUTIONS FOR RADIOLOGY DEPARTMENT PERSONNEL

Available on: [PMID: 32092296](#)

Available on: [Doi : 10.1016/j.jacr.2020.02.008](#)

**Coronavirus (COVID-19) Outbreak: What the Department of Radiology Should Know**

Radiographers are among the first-line health care workers who might be exposed to 2019 novel COVID-19.

Diagnostic imaging facilities should have guidelines in place to manage individuals with known or suspected COVID-19 infection.

The novel COVID-19 is highly contagious and is believed to transmit mostly through respiratory droplets, but there is uncertainty as to whether the virus can be transmitted by touching a surface or an item that is contaminated (ie, a fomite).

A thorough understanding of the routes of virus transmission will be essential for patients' and health care professionals' safety.

Droplets have the greatest risk of transmission within 3 ft (91.44 cm), but they may travel up to 6 ft (183 cm) from their source [19].

For the purpose of diagnostic imaging in individuals with NCIP, whenever possible, portable radiographic equipment should be used to limit transportation of patients.

On the basis of experience with SARS, the use of a satellite radiography center and dedicated radiographic equipment can decrease the risk for transmission from known infected individuals.

If a patient needs to be transported to the radiology department, he or she should wear a surgical mask during transport to and from the department.

Title: Facemask shortage and the coronavirus disease (COVID-19) outbreak: Reflection on public health measures

Title: (Not Available)

Title: Epidemiological and clinical features of COVID-19 patients with and without pneumonia in Beijing, China

Title: Journal Pre-proof COVID-19 Spike-host cell receptor GRP78 binding site prediction COVID-19 Spike-host cell receptor GRP78 binding site prediction COVID-19 Spike -GRP78 binding site prediction

# Dialogue Systems



Microsoft

Bot Framework PREVIEW My bots Documentation Bot Directory Blog

## Featured bots

 **Skyscanner**  
Skyscanner  
The world's travel search engine

 **AzureBot**  
Microsoft  
Be more productive with your own Microsoft Azure subscriptions using

 **Jessie Humani**  
PullString, Inc.  
Jessie Humani's life is in your hands. Will you help her make some

 **SI x GameOn ...**  
GameOn Technology  
Love the NFL? Add the SI x GameOn NFL bot to get unparalleled

 **Baymax**  
TrueCare24  
Your personal healthcare companion. The service is available

 **Cardea**  
Powered by RingMD  
Meet Cardea, your personal medical aid. She will answer your health

Do you remember when the order was placed, what it contained, or do you have an order number?

Probably March or April and it contained a pencil holder.

 I found an order from February 2016 which contained a pencil cup and four other items.

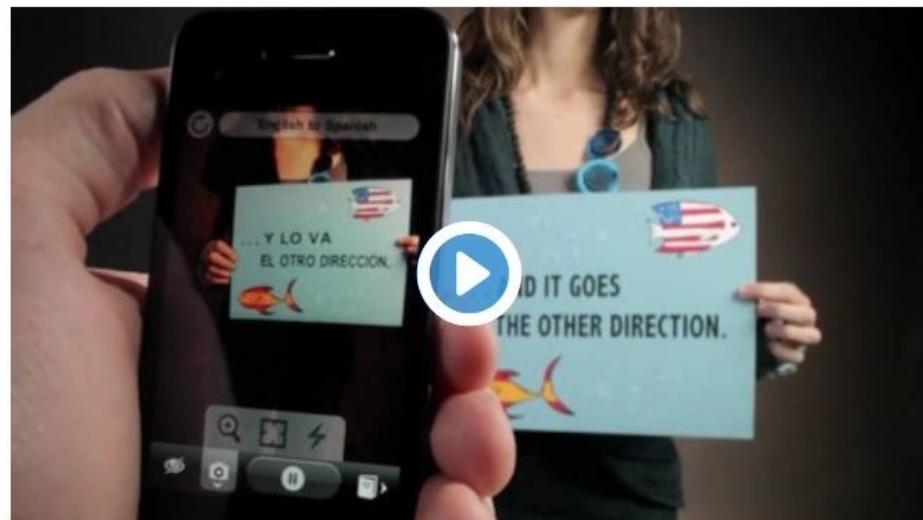
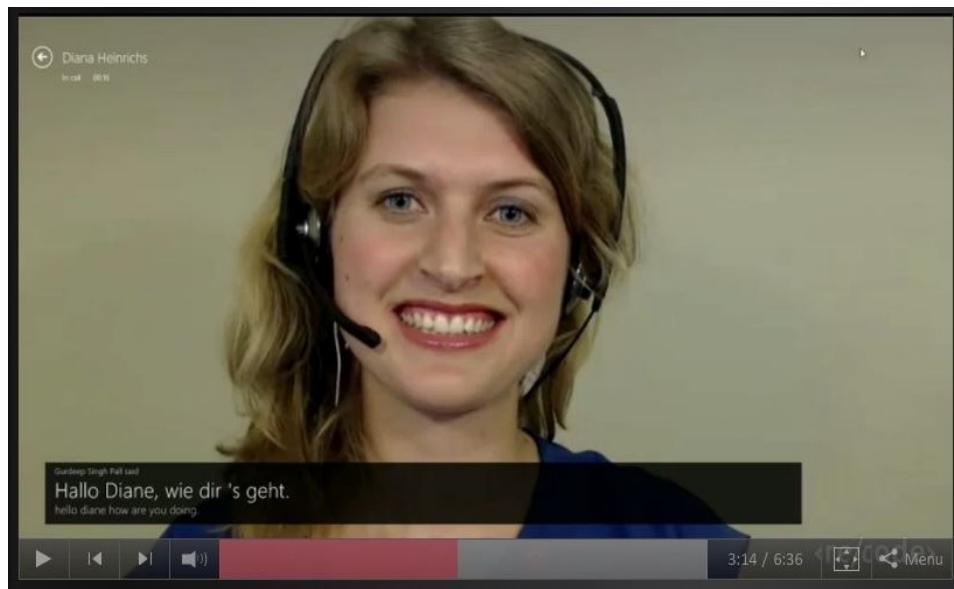
Order #1057-02  
Pencil cup  
Notebook  
Eraser pack  
+2 other items 

[Order details](#)

Is this the order you're looking for?

[No](#) [Yes](#)

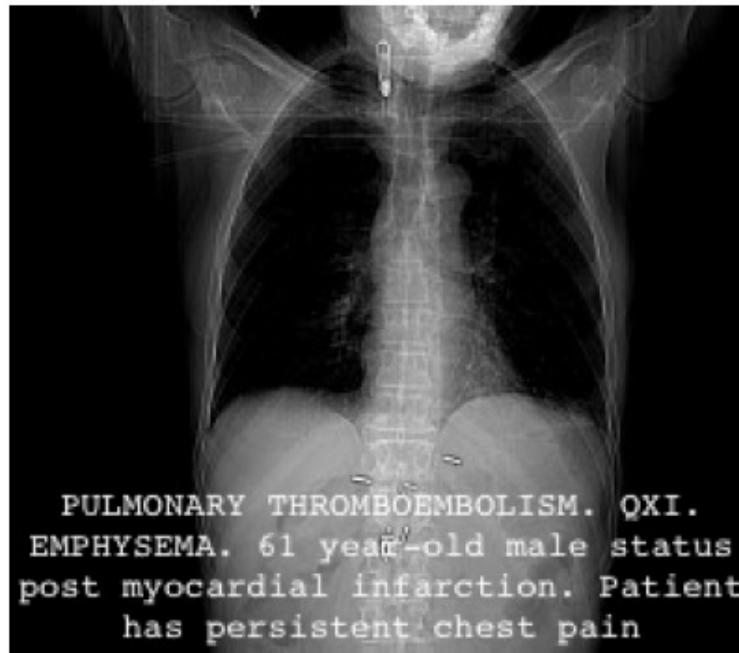
# Machine translation



<https://www.microsoft.com/en-us/research/video/speech-recognition-breakthrough-for-the-spoken-translated-word-short/>  
<https://www.youtube.com/watch?v=RuAp92wW9bg>  
<https://www.youtube.com/watch?v=h2OfQdYrHRs>



(a) General



(b) Biomedical

# Image captioning

- Image retrieval via captions.
- Eyesight problems.
- Drafting medical reports.

ΠΡΑ  
ΚΤΟ  
ΡΕΙΟ ΑΘΗΝΑΪΚΟ - ΜΑΚΕΔΟΝΙΚΟ ΠΡΑΚΤΟΡΕΙΟ ΕΙΔΗΣΕΩΝ

ΜΑΚΕΔΟΝΙΑ SPORT BUSINESS ΥΓΕΙΑ ΤΟΥΡΙΣΜΟΣ LIFE MICROSITES ΠΕΡΙΣΣΟΤΕΡΑ ▾

READ WITH  
3sayR 104.9FM

Ερευνητική ομάδα του Οικονομικού Πανεπιστημίου Αθηνών κατακτά συνεχείς διακρίσεις σε διεθνή διαγωνισμό Τεχνητής Νοημοσύνης

ΜΕ ΤΗ ΜΑΤΙΑ ΤΟΥ ΑΠΕ-ΜΠΕ /Κυριακή 24 Σεπτεμβρίου 2023, 09:56:32 / Τελευταία Ενημέρωση: 09:59 / Πηγή: ΑΠΕ-ΜΠΕ

ΠΕΡΙΣΣΟΤΕΡΕΣ ΕΙΔΗΣΕΙΣ

Με τη ματιά του ΑΠΕ-ΜΠΕ / μέρα πριν

Μια νέα εποχή ξεκινά για το τετράμωνο που περιλαμβάνει τους ιστορικούς κινηματογράφους Αττική Απόλλων στο κέντρο τη Αθήνας

Figure 1: Example of a caption produced by the model of [Vinyals et al. \(2017\)](#) for a non-biomedical image (1a) and an example of a PEIR Radiology image with its associated caption (1b).

From I. Pavlopoulos, V. Kougia, I. Androutsopoulos, “A Survey on Biomedical Image Captioning”. <https://www.aclweb.org/anthology/W19-1803/>

# Large Language Models

Default (GPT-3.5)



I teach a 12 week long AI course. Can you suggest what I should cover per week?



Designing a 12-week AI course is a great endeavor. Below is a week-by-week breakdown of topics and concepts you could cover. Keep in mind that the course can be adjusted based on the prior knowledge of your students and the depth you want to go into each topic. Here's a general outline:



## **Week 1: Introduction to AI**

- What is AI? History and milestones.
- Types of AI: Narrow (Weak) vs. General (Strong) AI.
- AI in popular culture and misconceptions.

## **Week 2: Machine Learning Fundamentals**

- Introduction to Machine Learning (ML).
- Supervised, Unsupervised, and Reinforcement Learning.
- Data preprocessing and feature engineering.

## **Week 3: Supervised Learning**

- Linear Regression.
- Logistic Regression.
- Evaluation metrics (MSE, MAE, accuracy, precision, recall, F1-score).

 Regenerate

Send a message 

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT September 25 Version](#)

# Many related terms

- **Natural Language Processing (NLP):**
  - Usually considered a sub-field of **Artificial Intelligence**.
  - **Algorithms, models, systems** to “understand” or generate mostly **written** natural language.
- **Computational Linguistics:**
  - Also mostly about **written** NL, originally more emphasis on computational *linguistic* theories, but now in effect a **synonym** of NLP, and (like NLP) probably closer to CS than linguistics.
  - The ***Association for Computational Linguistics*** (ACL) organizes the top NLP conferences and publishes the top journals.
- **(Human) Language Technology/Engineering:**
  - Less established term, often **includes speech processing**, more emphasis on **systems/engineering**.
- **Text (and Speech) Analytics:**
  - Business term, usually meaning information extraction, sentiment analysis, opinion mining.

# Course units ( $\sim 1$ unit/week)

- **U1:** Introduction, n-gram language models, estimating probabilities from corpora, text generation with beam search.
- **U2:** Text classification/regression with (mostly) linear models.
- **U3&4:** Text classification/regression with MLPs.
- **U5&6:** NLP with RNNs (incl. machine translation).
- **U7:** NLP with CNNs, image-to-text with CNNs-RNNs.
- **U8&9:** NLP with Transformers, Large Language Models.
- **U10:** Speech recognition and dialogue systems (?).

# Lectures, demos, assignments

- **10 lectures/demos** (3 hours/week):
  - **Lectures** (key ideas, theory). Slides, study exercises (most solved), references provided.
  - **Demos**: in ~5 weeks, the **last hour** will be a mini-lab of related tools (e.g., NLTK, spaCy, scikit-learn, PyTorch). Code examples.
- **Assignments** (50% if exam  $\geq 5$ , else 0%):
  - Some of the unsolved **exercises** (probably 5, most requiring **programming**) to be handed in.
- **Final exam** (50% if exam  $\geq 5$ , else 100%).

# Recommended books

- **General NLP textbooks:**

- *Speech & Language Processing* by D. Jurafsky and J.H. Martin, 2<sup>nd</sup> ed., Pearson, 2009. But continuously updated. (Free updated draft: <http://web.stanford.edu/~jurafsky/slp3.>)
- *Deep Learning for Natural Language Processing: A Gentle Introduction*, by M. Surdeanu and M.A. Valenzuela-Escarcega, Cambridge Univ. Press, 2024. (Free draft: <https://clulab.org/gentlenlp/text.html.>)
- *Neural Network Models for Natural Language Processing*, by Y. Goldberg, Morgan & Claypool, 2017. (Available at AUEB's library.)

