Natural Language Processing

Introduction

2023–24
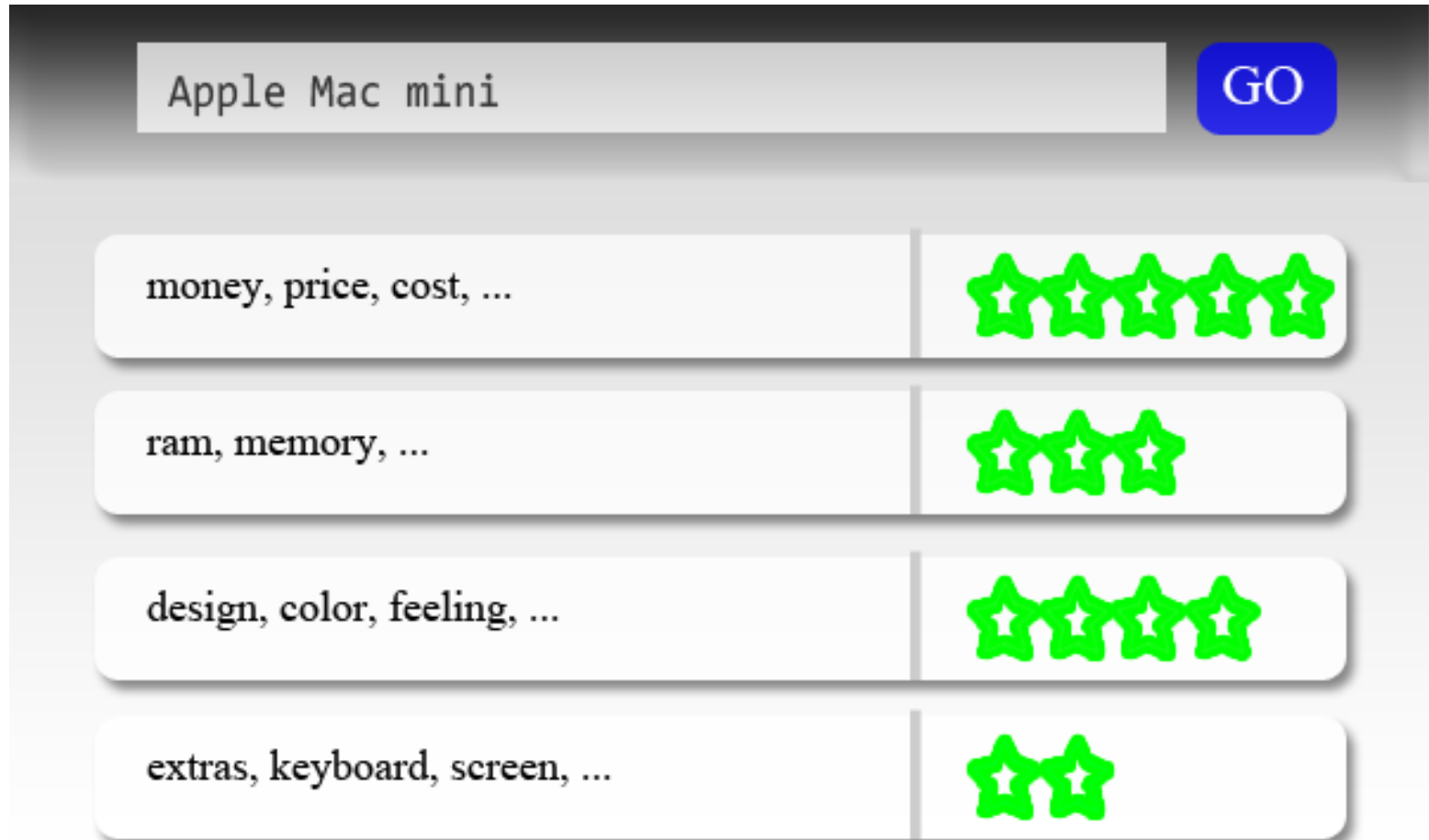
Ion Androutsopoulos

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

Contents

- **What** this course is about.

- **Examples** of applications.

- **Topics** to be covered and **prerequisites**.

- **Organization** of the course:

  o Lectures, demos.

 o 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
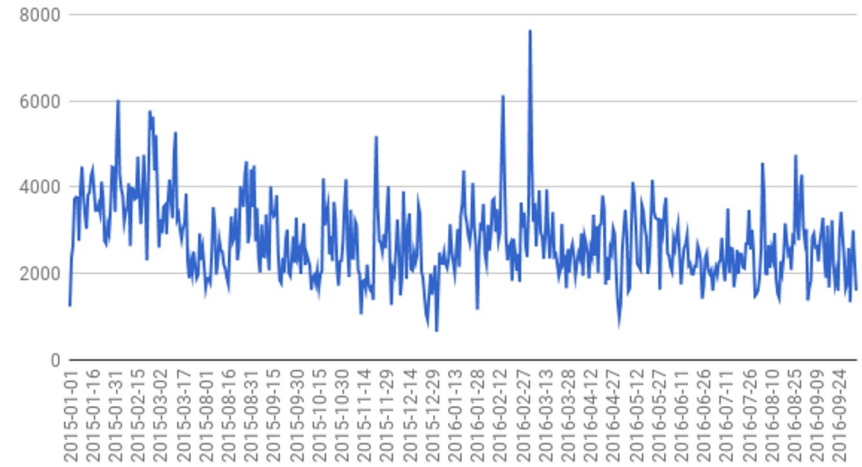
Moderation Panel

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

J. Pavlopoulos, P. Malakasiotis and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation", EMNLP 2017, http://nlp.cs.aueb.gr/pubs/emnlp2017.pdf.

4

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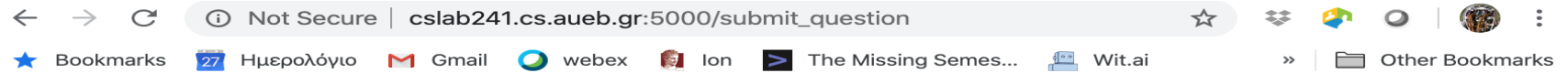
Results for the question: How is Covid-19 transmitted?

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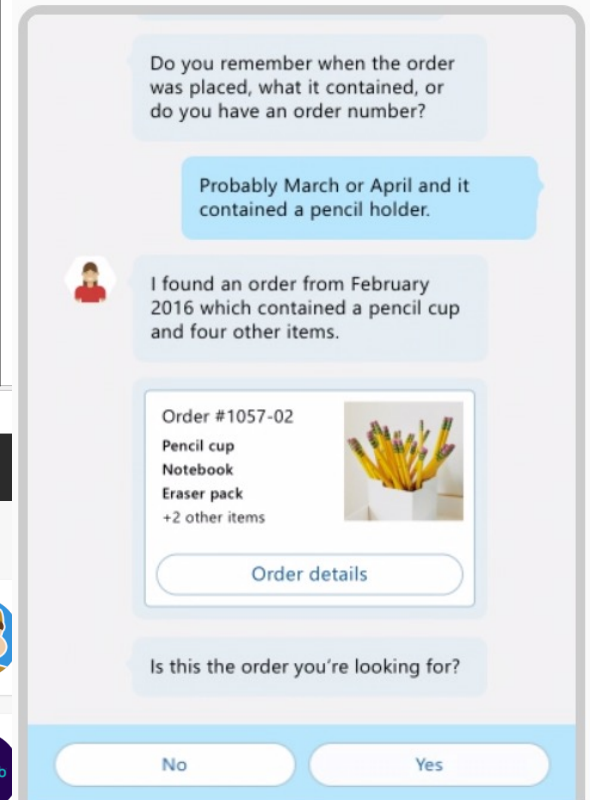
Title: Facemask shortage and the coronavirus disease (COVID-19) outbreak: Reflection on public health measures

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

D. Pappas, P. Stavropoulos and I. Androutsopoulos, "AUEB-NLP at BioASQ 8: Biomedical Document and Snippet Retrieval". Proceedings of the 8th BioASQ workshop at CLEF 2020. http://nlp.cs.aueb.gr/publications.html

Dialogue Systems

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

https://chat.openai.com/

8

Machine translation

Speech Recognition Breakthrough for the Spoken, Translated Word

Again, the results are not perfect.
They are in fact quite a few errors.
There's much work to be done in this area.
With this technology is very promising.

再次，结果并不完美。
他们实际上不少错误。
在这方面有很多工作要做。
用这个技术很有希望。

Diana Heinrichs

Hallo Diane, wie dir 's geht.
hello diane how are you doing.

3:14 / 6:36

...Y LO VA
EL OTRO DIRECCION.

D IT GOES
THE OTHER DIRECTION.

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

9

Natural language generation

Touring Machine (G. Karakatsiotis and V. Pterneas, Dept. of Informatics, AUEB). Won the Greek Imagine Cup of Microsoft in 2010 and 1st prize in interoperability at the International Imagine Cup of 2011. The texts are generated from an OWL ontology of the Ancient Agora of Athens, which was created during IST INDIGO (www.ics.forth.gr/indigo/) from information provided by the Foundation of the Hellenic World (www.ime.gr).

Language technology examples

- Algorithms, models, systems to **analyze** and **generate written** or **spoken natural language**.

  o **Spelling/syntax checking**, smart **keyboards**.

  o **Machine translation**, of texts or speech.

  o **Filtering/routing** (e.g., spam filters, call centers).

  o **Information extraction** and **opinion mining**.

  o **Question answering** (e.g., biomedical QA).

  o **Spoken dialogue systems** and **chatbots** (e.g., booking tickets, requesting information, in-car dialogues).

  o **Text generation** (e.g., describing products in multiple languages, generating reports from sensor data).

  o **Text summarization** (e.g., of multiple retrieved docs).

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.

Topics to be covered (hopefully)

- **W1:** Introduction, n-gram language models, spelling correction, beam search decoding.

- **W2–3:** Text classification/regression with (mostly) linear models, including logistic regression, SGD, evaluation measures.

- **W4–5:** Intro to deep learning, text/token classification with MLPs.

- **W6–8 :** NLP with RNNs, including text/token classification, attention, encoders-decoders, machine translation.

- **W9–11:** NLP with CNNs and Transformers, incl. (some) image-to-text, BERT, GPT-x and friends, fine-tuning, prompting, …

- **W12–13**: Speech processing with Transformers, dialogs (?).

Prerequisites, recommended courses

- **Prerequisites**:
  - **Calculus** (e.g., derivatives), **linear algebra** (vector and matrix operations), **probability theory** (e.g., conditional probabilities).
  - **Python** for the programming assignments.

- Other **recommended (optional)** companion **courses**:
 - **Deep Learning** (next term). The NLP course also serves as an introduction to DL.

Lectures, demos, assignments

- **13 lectures/demos** (3 hours/week):

  o **Lectures** (key ideas, theory). Slides, study exercises (most solved), references provided.

  o **Demos**: in ~5 weeks, the **last hour** will be a mini-lab of related tools (e.g., NLTK, spaCy, scikit-learn, Tensorflow/Keras, PyTorch). Code examples provided.

- **Group assignments** (50% if exam ≥ 5, else 0%):

  o Some of the unsolved **exercises** (probably 5, most requiring **programming**) to be handed in (**50%**).

- **Final exam** (50% if exam ≥ 5, else 100%).

Recommended books

- **NLP textbooks** (optional):
  - *Speech & Language Processing* by D. Jurafsky and J.H. Martin, 2^η edition, Pearson, 2009. (Available at AUEB's library. Free draft 3rd edition: http://web.stanford.edu/~jurafsky/slp3)
  - *Neural Network Models for Natural Language Processing*, by Y. Goldberg, Morgan & Claypool, 2017. (Available at AUEB's library.)
  - *Introduction to Natural Language Processing* by J. Eisenstein, MIT Press, 2019. (Free draft: https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf)
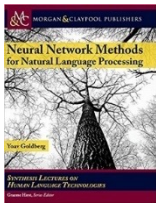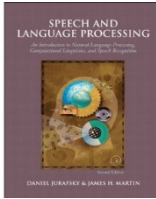
Recommended books

- **Older but still useful NLP/IR textbooks** (optional):

  - *Foundations of Statistical Natural Language Processing* by C.D. Manning and H. Schutze, MIT Press, 1999. (Available at AUEB's library.)

  - *An Introduction to Information Retrieval* by C.D. Manning, P. Raghavan and H. Schütze, Cambridge University Press, 2008. (Freely available from: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.)
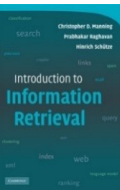
Recommended books

- **DL textbooks** (optional):
  - *Deep Learning in Python,* by F. Chollet, Manning Publications, 2nd edition, 2021. (Free 1st edition: https://www.manning.com/books/deep-learning-with-python. But 2nd edition highly recommended.)
  - *Understanding Deep Learning*, by S.J.D. Prince, MIT Press (in press). (Free pre-print: https://udlbook.github.io/udlbook/)
  - *Introduction to Deep Learning*, by E. Charniak, MIT Press, 2019. (Available at AUEB's library.)
  - *Dive into Deep Learning*, by Zhang et al. (Freely available: https://d2l.ai/)