Natural language processing

Introduction and course overview

X. Hilaire

ESIEE Paris, IT department
x.hilaire@esiee.fr

August 27, 2023

Outline

1. Introduction

2. Course overview

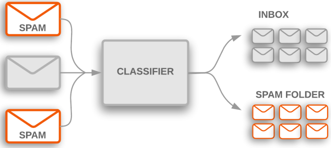
What natural language processing (NLP) ?

- In a nutshell: NLP is a field of computer science which aims to reproduce everything that humans know to do with language by a computer.
- Includes (but not limited to) various tasks such as :
  - Text classification [1]
  - Text clustering and topic detection [1]
  - Sentiment and opinion analysis [1]
  - Text summarization
  - Machine translation [1]
  - Text generation
  - Question answering (ChatGPT)

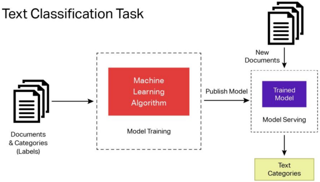
[1] Problem tackled in the course, at least once in a lab

Text classification

Text classification is to label a given text using predefined labels. Examples: news, finance, sports, cars, internet, etc. A simple, yet useful example is SPAM/NON SPAM classification.



[2]



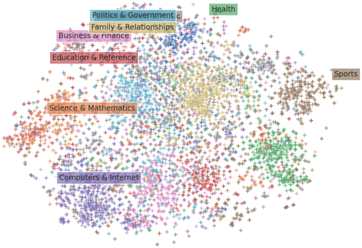
[3]

[2] https://towardsdatascience.com

[3] https://imerit.net

Text clustering

- In essence, text clustering is the unsupervised version of text classification
- Goal = to segment a text or corpus in different parts, but those don't have any predefined label



[4]

A reference library for text clustering: *Carrot*[2]
`https://github.com/carrot2/carrot2`
`https://search.carrot2.org/`

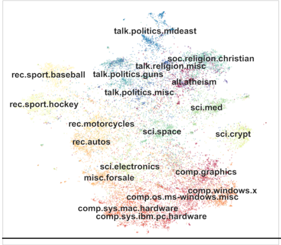
[4] ZSL Blog Post

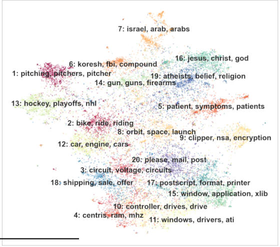
Topic detection and modeling

- Closely related to text clustering
- A **topic** is a set of words which are highly correlated one to each other (a "strong" cluster)
- Assuming a document is related to several topics:

$$p(word_i|doc_j) = \sum_k p(word_i|topic_k)p(topic_k|doc_j)$$

This bayesian expansion turns out to work better than working on words directly.





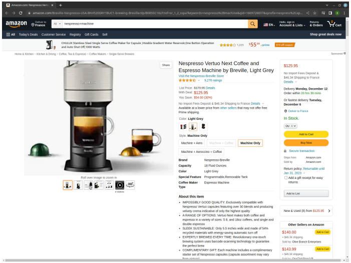
[5]

[5] sbert.net

Sentiment and opinion analysis

Extracting subjective information from text :

- Has a given text a positive / negative connotation ?
- Is it enthusiastic / critical, joyful / sad, interesting / not interesting ?
- Has a given product received a good / bad review ?



Amazon.com: Nespresso Vertu ×

amazon.com/Breville-Nespresso-USA-BNV520GRY1BUC1-brewing-Breville/dp/B085SC16LF/ref=sr_1_2_sspa?keywords=nespresso%2Bmachine&qid=1669728837&sprefix=nespress%2Cap...

CHULUX Stainless Steel Single Serve Coffee Maker for Capsule ,Visiable Gradient Water Reservoir,One Button Operation and Auto Shut Off,1000 Watts ★★★★☆ 1,225 $55⁹⁹ prime $13 off coupon

Home & Kitchen › Kitchen & Dining › Coffee, Tea & Espresso › Coffee Makers › Single-Serve Brewers

Sponsored

Share

Roll over image to zoom in

Nespresso Vertuo Next Coffee and Espresso Machine by Breville, Light Grey

Visit the Nespresso-Breville Store

★★★★½ 9,276 ratings

List Price: ~~$179.95~~ Details

With Deal: $125.95

You Save: $54.00 (30%)

No Import Fees Deposit & $46.34 Shipping to France Details

Available at a lower price from other sellers that may not offer free Prime shipping.

Color: Light Grey

Style: Machine Only

Machine + Aero | Machine + Coffee | Machine Only | Machine + Aeroccino + Coffee

Brand	Nespresso-Breville
Capacity	18 Fluid Ounces
Color	Light Grey
Special Feature	Programmable,Removable Tank
Coffee Maker Type	Espresso Machine

About this item

- IMPOSSIBLY GOOD QUALITY: Exclusively compatible with Nespresso Vertuo capsules featuring over 30 blends and producing velvety crema indicative of only the highest quality
- A RANGE OF OPTIONS: Vertuo Next makes both coffee and espresso in a variety of sizes: 5 8, and 18oz coffees, and single and double espresso
- SLEEK SUSTAINABLE: Only 5.5 inches wide and made of 54% recycled materials with energy-saving automatic turn off
- EXPERTLY BREWED EVERY TIME: Revolutionary one-touch brewing system uses barcode-scanning technology to guarantee the perfect brew
- COMPLIMENTARY GIFT: Each machine includes a complimentary starter set of Nespresso capsules (capsule assortment may vary from picture)

$125.95

No Import Fees Deposit & $46.34 Shipping to France Details

Delivery **Monday, December 12**. Order within 20 hrs 39 mins

Or fastest delivery **Tuesday, December 6**

Deliver to France

In Stock.

Qty: 1

Secure transaction

Ships from	Amazon.com
Sold by	Amazon.com

Return policy: Returnable until Jan 31, 2023

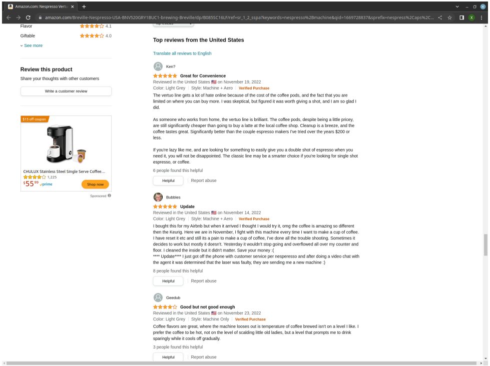
□ Add a gift receipt for easy returns

New & Used (8) from $125.95

Other Sellers on Amazon

$140.00 + $46.34 shipping Sold by: Olive Branch Enterprises — Add to Cart

$143.99 + $46.34 shipping Sold by: OneClickShop148 — Add to Cart



Amazon.com: Nespresso Vertu ×

amazon.com/Breville-Nespresso-USA-BNV520GRY1BUC1-brewing-Breville/dp/B085SC16LF/ref=sr_1_2_sspa?keywords=nespresso%2Bmachine&qid=1669728837&sprefix=nespress%2Caps%2C...

Flavor	★★★★☆	4.1
Giftable	★★★★☆	4.0

See more

Review this product

Share your thoughts with other customers

$13 off coupon

CHULUX Stainless Steel Single Serve Coffee...

★★★★☆ 1,225

$55⁹⁹ prime

Sponsored

Top reviews from the United States

Translate all reviews to English

Ken?

★★★★★ **Great for Convenience**

Reviewed in the United States on November 19, 2022

Color: Light Grey | Style: Machine + Aero | Verified Purchase

The vertuo line gets a lot of hate online because of the cost of the coffee pods, and the fact that you are limited on where you can buy more. I was skeptical, but figured it was worth giving a shot, and I am so glad I did.

As someone who works from home, the vertuo line is brilliant. The coffee pods, despite being a little pricey, are still significantly cheaper than going to buy a latte at the local coffee shop. Cleanup is a breeze, and the coffee tastes great. Significantly better than the couple espresso makers I've tried over the years $200 or less.

If you're lazy like me, and are looking for something to easily give you a double shot of espresso when you need it, you will not be disappointed. The classic line may be a smarter choice if you're looking for single shot espresso, or coffee.

6 people found this helpful

Helpful | Report abuse

Bubbles

★★★★★ **Update**

Reviewed in the United States on November 14, 2022

Color: Light Grey | Style: Machine + Aero | Verified Purchase

I bought this for my Airbnb but when it arrived I thought I would try it, omg the coffee is amazing so different then the Keurig. Here we are in November, I fight with this machine every time I want to make a cup of coffee. I have reset it etc and still its a pain to make a cup of coffee, I've done all the trouble shooting. Sometimes it decides to work but mostly it doesn't. Yesterday it wouldn't stop going and overflowed all over my counter and floor. I cleaned the inside but it didn't matter. Save your money :(
**** Update**** I just got off the phone with customer service per nesperesso and after doing a video chat with the agent it was determined that the laser was faulty, they are sending me a new machine :)

8 people found this helpful

Helpful | Report abuse

Geedub

★★★★☆ **Good but not good enough**

Reviewed in the United States on November 23, 2022

Color: Light Grey | Style: Machine Only | Verified Purchase

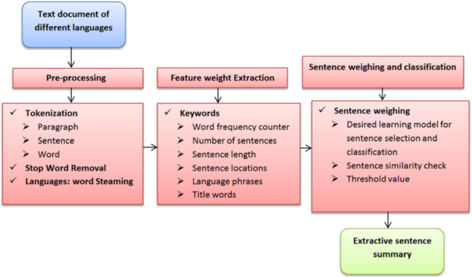
Coffee flavors are great, where the machine looses out is temperature of coffee brewed isn't on a level I like. I prefer the coffee to be hot, not on the level of scalding little old ladies, but a level that prompts me to drink sparingly while it cools off gradually.

3 people found this helpful

Helpful | Report abuse

Text summarization

- Aim = to compress a text while minimizing the loss of its meaning
- Very challenging task, as it requires a deep understanding of the text
- Two broad families of methods : extractive and abstractive
- Extractive methods
    - Extract important sentences from the text, and drop others.
    - A scoring function is needed to rank the importance of sentences. LSA (Latent semantic analysis), which we will see, is often used to do so.



Text document of different languages

Pre-processing
- ✓ Tokenization
    - ➢ Paragraph
    - ➢ Sentence
    - ➢ Word
- ✓ Stop Word Removal
- ✓ Languages: word Steaming

Feature weight Extraction
- ✓ Keywords
    - ➢ Word frequency counter
    - ➢ Number of sentences
    - ➢ Sentence length
    - ➢ Sentence locations
    - ➢ Language phrases
    - ➢ Title words

Sentence weighing and classification
- ✓ Sentence weighing
    - ➢ Desired learning model for sentence selection and classification
    - ➢ Sentence similarity check
    - ➢ Threshold value

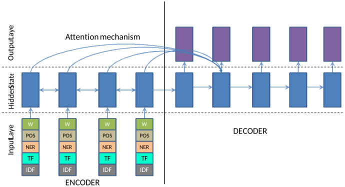
Extractive sentence summary

"Study of automatic text summarization approaches in different languages" [2]

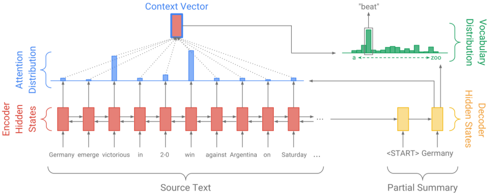
Text summarization

- Abstractive methods
    - Analyze the semantic information of every sentence of the text
    - Produce shorter, completely new sentences having the same meaning
- Two representative methods of this family are that of:
    - Nallapati et al. (IBM) "*Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond*" [3]
    - See et al. (Google) "*Get To The Point: Summarization with Pointer-Generator Networks*" [5] : both use sequence-to-sequence alignments of words.
- For a survey, see Li Jinpeng et al. "*Survey on Automatic Text Summarization*" [1]

Text summarization



Feature-rich-encoder. From Nallapati et al. [3]



Pointer-generator model. From See et al. [5]

Machine translation

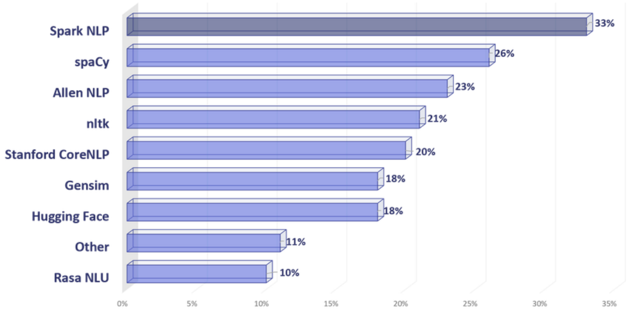
- Unmissable task, as very useful for centuries.
- Difficult task, as even human translators can disagree (*traduttore, traditore*)
- Very old problem, plethora of methods and products available (DeepL, Systran, Google translate, Microsoft Translator, Reverso, ...)
- For a recent survey, see, e.g. I. Rivera-Trigueros "*Machine translation systems and quality assessment: a systematic review*" [4]

Some NLP libraries

From `https://nlp.johnsnowlabs.com`

The most widely used NLP library in the enterprise

Which NLP libraries does your organization use [check all that apply]?



Source:2020 NLP Industry Survey, by **Gradient Flow**.

Outline

1 Introduction

2 Course overview

Objectives

To know the most important algorithms used in NLP :

1. Text preprocessing (tokenization) ☞ painful but mandatory step, needed to :
   - extract the textual parts of documents, often much richer (LaTeX, HTML, XML, etc.)
   - normalize words, possibly correcting typos on-the-fly
2. Word embeddings ☞ maps words from a sparse space to a dense space. Actually very important :
   - moves words which are semantically close to close representations in the dense space
   - allows to substitute a word by a synonym with minimal distortion
3. Topic detection, text classification, sentiment analysis ☞ old technique, borrowed from dimensionality reduction.
4. Machine translation ☞ automatically translates a text from one language to another with minimal semantic distortion
5. Transformers are not covered in this course due to lack of time, but in AIC-5201C "Deep-Learning and its Applications"

Course overview

The course heavily relies on some fundamental tools :

- Singular value decomposition (SVD) for dimensionality reduction and topic detection
- Fisher kernel linear discriminant analysis (k-LDA) and SVM for text classification and sentiment analysis
- Ordinary (non-recurrent) neural network and **backpropagation** ← very important
- Elman networks, recurrent neural networks (RNN)
- Long short term memory (LSTM) and gated recurrent units (GRU) networks

Don't underestimate them !

Course overview

Schedule:

- Lectures and exercise sessions / labs are alternated
  - about 20–40mn of lecture
  - then one exercise on the topic seen
- Do not pay strict attention to your schedule on `https://planif.esiee.fr` : the C/TD/TP breakdown is only meaningful in terms of accounting

About labs:

- Programming language is Python (v3)
- The NLP part involves NLTK, Gensim, and basic Python libraries (numpy, scikit-learn, matplotlib, ...)
- Neural networks can be implemented with Keras (simple, but slow) or PyTorch (more complex, but faster). TensorFlow allowed (for efficiency reasons), but beware of complex API, and expect limited support from me.

Overview

- Labs have been validated on ESIEE's machines, Linux OS (Debian 11.5)
- You can work on your own laptop or environment, but the final lab you will submit must work on ESIEE's machines, where it will be assessed
- RAM consumption is limited, but training is CPU intensive

Grading policy

- Final grade = 10% lab2 (sentiment analysis) + 40% lab4 (machine translation) + 50% written exam
- Labs must be done by pairs. Instructions for declaring pairs and submitting will be provided by mail, and involve my server `https://mvproxy.esiee.fr`
- The written exam has two parts:
  - First is 20–30mn, close book (no documents), rewarded 4–5 points. It typically consists of short questions, testing your knowledge about the lectures.
  - Second is 1h30–1h40mn, open book (all documents are allowed), rewarded 15–16 points. It typically consists of case studies.
- Electronic devices of all kinds, are forbidden during the written exam.
- Should you be granted a reset / catch up exam, the conditions remain exactly the same.

Documents, post-assistance

Course web page:

- The homepage of the course is `https://perso.esiee.fr/~hilairex/AIC-5102B`
- All the information and documents you need will be added to it on the fly

Post-assistance:

- I am usually available on Thursday afternoons (1pm-7pm) at my office 5352 for post-assistance
- If you have an urgent problem, you can always reach me by email: `x.hilaire@esiee.fr`

Li Jinpeng, Zhang Chuang, Xiaojun Chen, Yue Hu, and Pengcheng Liao.
Survey on automatic text summarization.
Journal of Computer Research and Development, 58(1):1, 2021.

Yogesh Kumar, Komalpreet Kaur, and Sukhpreet Kaur.
Study of automatic text summarization approaches in different languages.
Artif. Intell. Rev., 54(8):5897–5929, 2021.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang.
Abstractive text summarization using sequence-to-sequence rnns and beyond.
2016.

Irene Rivera Trigueros.
Machine translation systems and quality assessment: a systematic review, 4 2021.

Abigail See, Peter J. Liu, and Christopher D. Manning.

Get to the point: Summarization with pointer-generator networks, 2017.