Natural Language

Practical Classes

Luísa Coheur

2025

P4

Preprocessing, **Regular Expressions** and **N-grams**

Image generated by ChatGPT

- **Summary**:
  - Pre-processing
  - Regular Expressions
  - N-grams
- **Operational objectives**:
  - Practice pre-processing with a widely used NLP tool
  - Practice regular expressions
  - Practice N-grams
- **This class needs**: paper, a pen/pencil and computer.
- **Class materials**: this guideline and a notebook.

Another client

Again, it is hard to believe who is your next client:

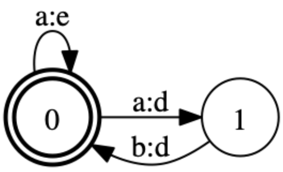
> *Dear Detective,*
> *My name is Anna Early, and I am the mayor of a small town in US. There is a group of fans of the awful series Friends who want to erect a statue of one of the characters. I complained, but I used incorrect arguments, and they won the first round. However, I am almost sure we can prove that this is a very poor series with a very limited vocabulary of words with less than four characters (including the ridiculous word "yeah"). I know that you are an NLP detective, and maybe you can help me prove this. I will pay you substantially. Sincerely yours,*
> *Anna Early*

You start thinking about whether you should accept the job or not. Damn it! You decided to enter computer science because you thought it would keep you away from ethical problems. You figured that focusing on algorithms and code would mean fewer moral issues to deal with. But now, you see that ethical challenges are everywhere, not just in AI. Again: damn it! While pondering about whether you should accept the job, you decide to do some NLP exercises.

1 Towards expertise

Exercises on Regular Expressions

1. Consider the transducer (don't panic if you do not know what a transducer is[1]). Assume that 0 is the initial and final state.



   (a) Considering that "abab" is part of the language of the given transducer, given "abab", the output of the transdutor is:

   A) eded
   B) edad
   C) dddd
   D) dede
   E) dedd

   (b) A sequence that will NEVER be the output of the transducer is:

   A) dd
   B) edd
   C) eed
   D) e
   E) edde

   (c) Choose the only correct option. ALL the possible output sequences of this transducer can be represented with the following regular expression:

   A) (e*(dd)?e*)*
   B) (e*[dd]e*)*
   C) (e+(dd)?e+)*
   D) (e+[dd]e*)*

[1] In NLP, a transducer is a computational model that converts input sequences (like text) into output sequences (like tagged entities or translated text). It was often used in NLP some years ago.

E) (e+(dd)e*)*

2. Write in the empty cell, the sequence that respects (both) regular expressions.

	[^AB]
[ABC]	

3. Consider a regular expression composed of the words hello and bye. Define a regular expression that (only) recognizes:

   (a) The set of sequences that end in byebye.

   (b) The set of sequences where each pair of ‘hello’s is followed by a pair of ‘bye’s (not easy!).

An exercise with N-grams

4 Consider the following pre-processed text:

in the summer it is hot
in the summer it is very hot
the butterflies are very hot
the butterflies fly a lot in the summer
in the summer the butterflies fly
the butterflies fly when it is hot

(a) Indicate two types of preprocessing that have been applied to the previous text and that are part of the usual (somewhat old-fashioned) text pre-processing in NLP.

(b) Fill in with bigrams counts the empty cells of the following table (for example, the bigram “the butterflies” occurs 4 times).

	<s>	the	butterflies	fly	</s>
<s>					
the			4		
butterflies					
fly					
</s>					

(c) Taking into account the previous table and that there are 6 sentences in the corpus, what is the probability of the sentence “the butterflies fly”? Consider the beginning and end of the sentence. Indicate all calculus.

2 Let’s go!

You create a new notebook, P4_friends_dataset_analysis (as usual, use Google Colab[2]). You use your expertise in UNIX to create a new dataset with the lines of the 6 main characters only. You play with the notebook and, in particular, with the stop words until you decide that they will not make a big difference: the most used vocabulary in the series is indeed limited and short on characters. Should you tell this to Anna Early? Should you accept the job? You are not in the mood of starting a statue war. Besides, you managed to sell the Friends series in VHS in OLX, although for almost no money. Hum...

While you torment yourself with the choice you have to make, you receive an email and it looks so much more interesting that you totally forget about Anna Early. OMG! You can’t believe it! An email from HIM, the one!

[2] https://colab.research.google.com.

3 A new challenge

> *Youngster,*
> *Here Morcela. Need your services. Attached there is a file with 300 emails. Three are fake. Please, find which ones. Important. Lives in danger. Fast.*

Youngster? This is definitely Inspector Morcela. Well, let us see how to help him. You start by asking ChatGPT about syntax rules for email addresses. Then you remember that chatGPT is not the most reliable source of information and you search for another sources. Although the correct syntax is a little bit more complicated, you assume that a valid email address has:

- A username constituted of:
  - letters (a-z, A-Z)
  - numbers (0-9)
  - dots (.), hyphens (-), underscores (_) (and other characters that we will ignore)
- The @ symbol
- A domain part that includes a name followed by a top-level domain split by a dot (.). The name part includes:
  - letters (a-z, A-Z)
  - numbers (0-9)
  - hyphens (-)
- In addition, there are some constraints:
  - the dot (.) is not allowed at the start or end of the username and cannot appear consecutively (e.g., john..doe@example.com is invalid)
  - the domain part must not start or end with a hyphen
  - the top-level domain must be at least two characters long and have primarily letters (e.g., .com, .org)

You go to the site `https://regex101.com` to test your colossal Regular Expression. Find it!

After some tries you reply to Morcela:

> *Dear Inspector Morcela,*
>
> *The fake emails are: ...*

The inspector responds in the early hours of the morning:

> *Youngster,*
>
> *People saved. We couldn't catch this AI freak, specialized in NLP, who only causes problems. You will be useful soon.*

Well, no thanks and no money, but you saved lives. You go back to sleep feeling like a hero.