**2nd Year Second Cycle Option: IASD**
**Natural Language Processing**

# Practical work N° 01 : PREPROCESSING IN NLP

## Objectif :

- ➔ Understand the basic concepts of preprocessing in NLP

## Required Libraries

Nltk

SpaCy

WordCloud

Pywaffle

## A. Data Preparation

1. Import the dataset **spooky.csv** from the URL https://github.com/GU4243-ADS/spring2018-project1-ginnyqg/raw/master/data/spooky.csv using pandas and display the first **10** samples.

## B. Text Cleaning

1. Handle repetitive characters (e.g., "cooooool" → "cool").
2. Manage homoglyphs (e.g., "$tupide" → "stupide").
3. Transform special entries such as URLs, email addresses, and HTML tags into a canonical form.
4. Convert all characters to lowercase.
5. Remove punctuation.
6. Remove stop words.

## C. Tokenization

1. Tokenize each sentence based on spaces / punctuation.
2. Tokenize each sentence using a rule-based tokenization algorithm.
3. Tokenize each sentence using a subword tokenization algorithm.

## D. Named Entity Recognition

1. Represent named entities for each sentence (using NLTK or SpaCy).

## E. Form Reduction

1. Use lemmatization and stemming with NLTK.

Optional: Perform the same tasks with SpaCy.

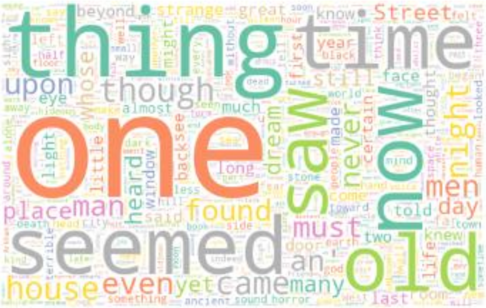
## F. Frequency Analysis

1. Count the number of sentences, for each author, where the word "**Great**" appears.

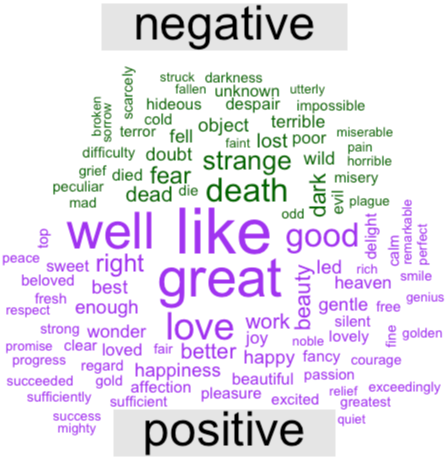
2. Use **pywaffle** to obtain a graph summarizing the number of occurrences of the word "**great**" per author.

3. Repeat the analysis with the word "**impossible**".

4. Using the **wordCloud** function, create three word clouds to represent the most used words by each author.



5. Using the **wordCloud** function, display the top 100 positive and negative words used by the authors.


negative
struck
darkness
fallen
unknown
utterly
scarcely
hideous
despair
impossible
broken
sorrow
cold
object
terrible
terror
fell
faint
lost
poor
miserable
difficulty
doubt
strange
wild
pain
horrible
grief
died
fear
misery
peculiar
dead
die
death
dark
evil
plague
mad
odd
top
well
like
good
delight
calm
remarkable
perfect
peace
sweet
right
great
beauty
led
rich
beloved
heaven
smile
best
fresh
respect
enough
gentle
free
genius
strong
wonder
love
work
silent
promise
clear
loved
fair
better
joy
noble
lovely
fine
golden
progress
regard
happy
fancy
courage
succeeded
gold
happiness
beautiful
passion
affection
sufficiently
sufficient
pleasure
excited
relief
exceedingly
greatest
success
mighty
quiet
positive