

# News Aggregator

---

Alaa Mohamed– Marwan Ibrahim –Mayar Yasser – Mohamed Ayman

SUPERVISED BY: DR. ESLAM AMER AND ENG. MENNATULLAH GAMIL

# Introduction 1/3

---

## What is news aggregator ?

- Web application which aggregates syndicated web content such as online newspapers, blogs, podcasts, and videos in one location for easy viewing.
- A technology that helps to consolidate many websites into one page that can show only the new or updated information from many sites.

# Introduction 2/3

---

## Types of news aggregator:

- **Web-based feed readers**
  - Feed reader applications
- **Social news aggregators**
  - Frame- and media bias-aware news aggregators
- **News aggregation websites**



# Introduction 3/3

## Weekly usage of news aggregators worldwide between 2017-2018

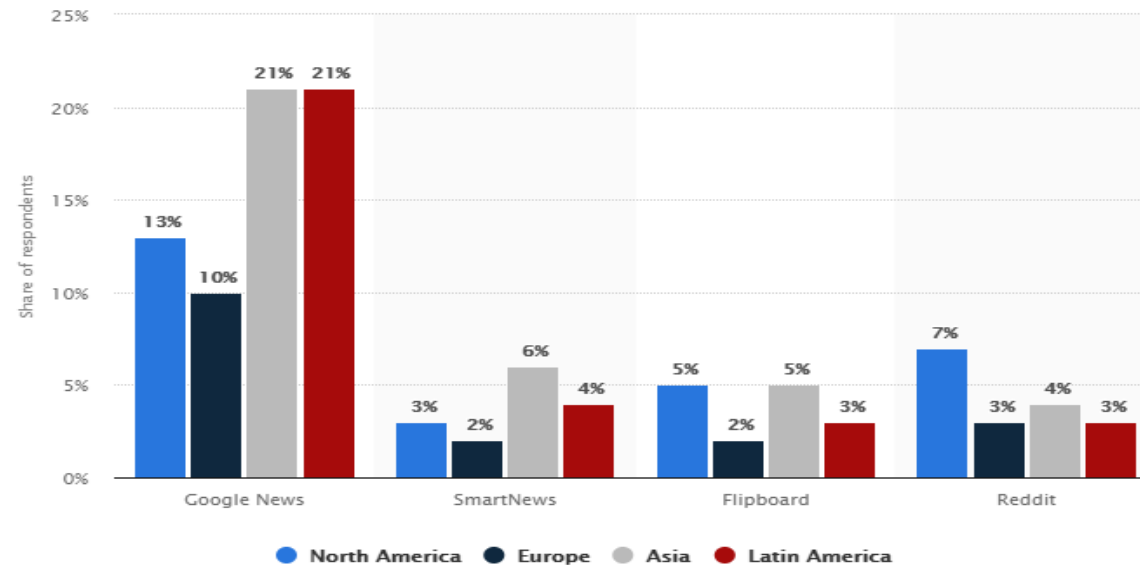


Figure1: usage of news aggregators worldwide in 2017<sup>[1]</sup>

[1]Watson, A. (2018). *Weekly usage of news aggregators worldwide in 2017, by region*. Worldwide: Statista.

# Related Work 1/3: Extracting Different News Perspective <sup>[2]</sup>.

## □ Aims to:

- Explore both common and different information in related articles using **MNA** approach.

- Writing and presenting the news could affect the reader perspective by not showing all perspectives.

## □ Methods Used:

- Article extraction using **naïve** approaches.
- Clustering using hierarchical agglomerative clustering (**HAC**).
- Summarized using (multi-document summarizer) **MEAD**.

- Their future work is to add another news categories and reducing inter-cell redundancy.

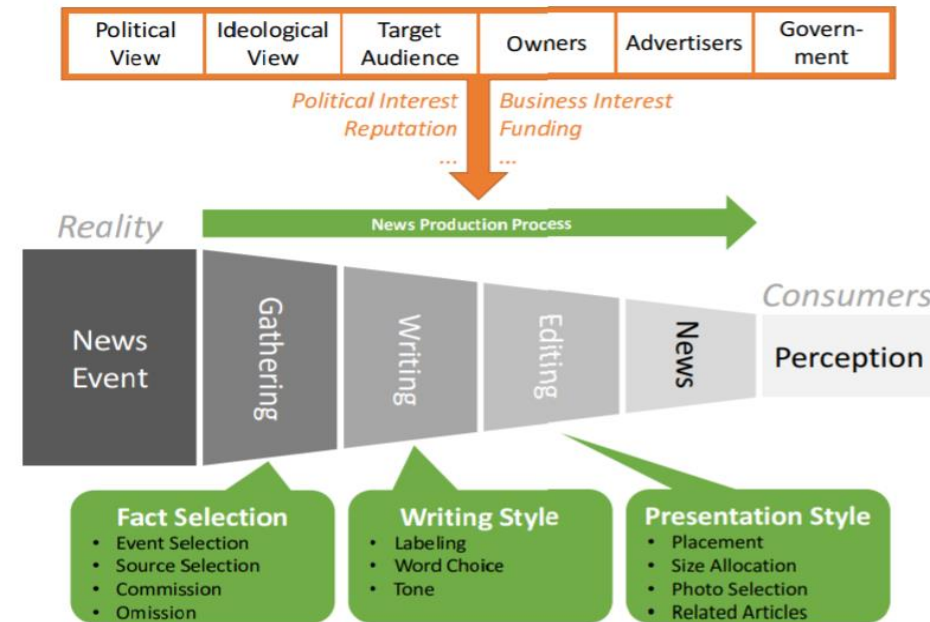


Figure2: Reasons and forms of media bias<sup>[2]</sup>

# Related Work 2/3: Merging News Aggregator with social media<sup>[3]</sup>.

## ❑ Aims to:

- Improve the news aggregator platform by **merging** the RSS news with both **social media** as Reddit , Twitter.
- Expose in **Multi language**.
- Create **account** for each user to have his own history.

❑ To drag together data from numerous channels but without getting what is written in Social media

## ❑ Methods Used:

- REST APIs :Accounts
- Web framework :web page
- HTTP requests : API

❑ The user has created an account then greeted by the main page which contains a list of summarized articles with all information.

❑ Their future work is to integrate with **YouTube** and Integration with other services such as **APIs** providing currency exchange information and weather conditions.

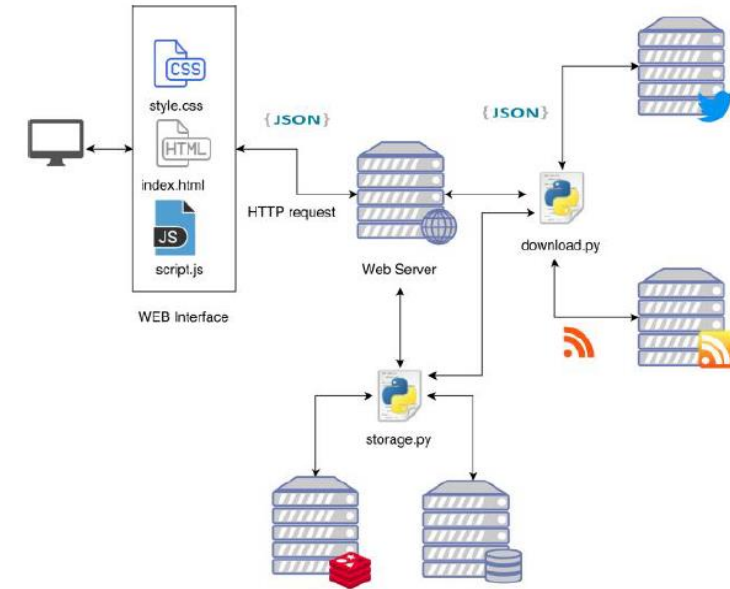


Figure3: Overall functional architecture<sup>[3]</sup>

# Related Work 3/3: The usage of FastText Classification<sup>[4]</sup>

## □ Aims to:

- **Clustering** data from Social Media into important and insignificant things with respect to landslide as a normal disaster.
- Make a large external repository of knowledge **tractable** first.

□ Problem is to assign a document using its text to one or more categories from a predefined set.

## □ Methods Used:

- **FastText**(randomized explicit semantic analysis)

□ Text categorization is a important problem. RS-ESA approach separating real texts from imaginary ones.

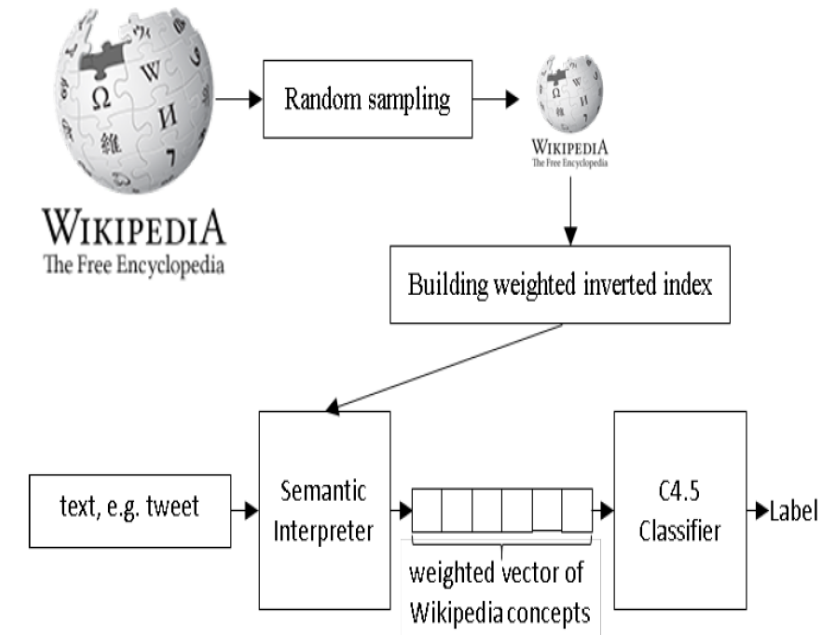


Figure4:RS-ESA Overview<sup>[4]</sup>

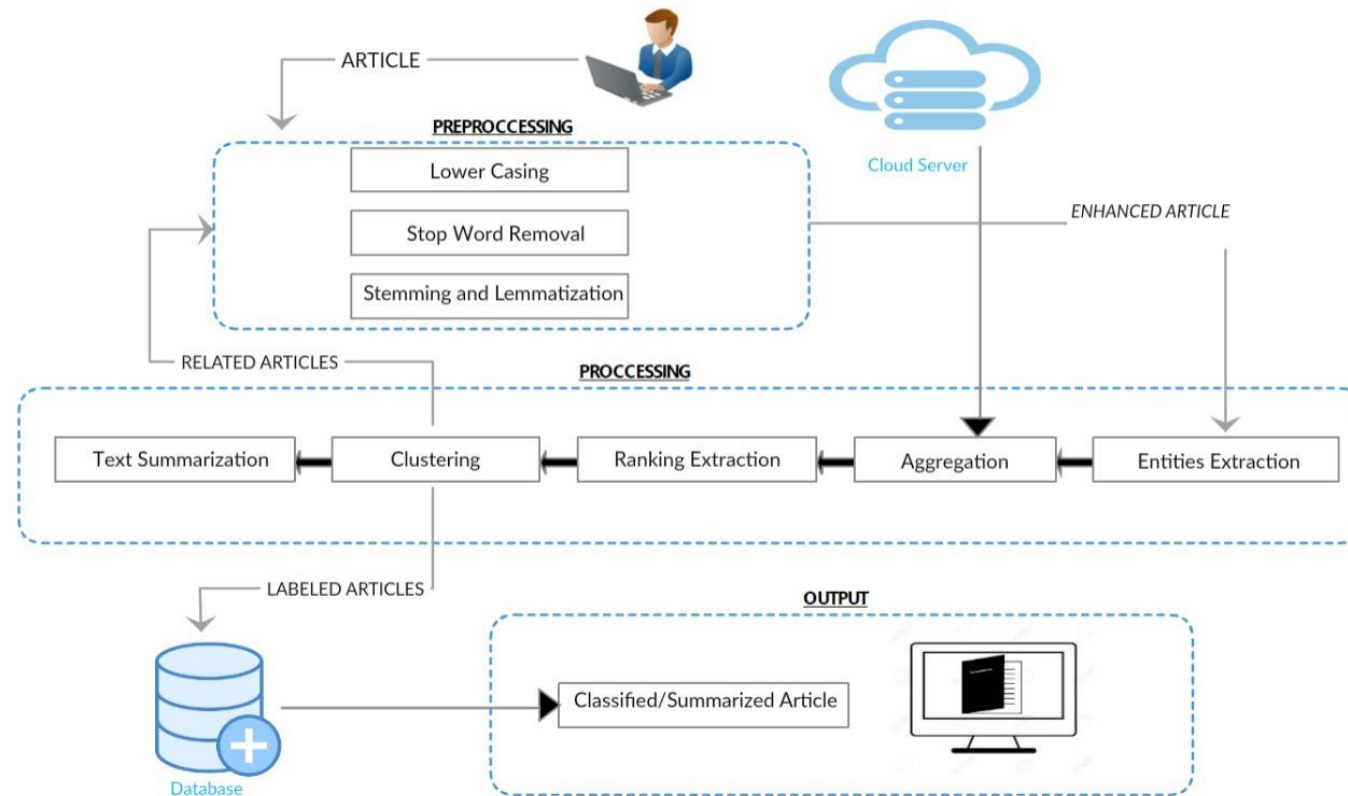
# Problem Statement

---

- Aggregate **highly relevant** news which are related to a certain event from **heterogeneous** sources according to the **trustworthy** of those news agencies, **summarizing** them into single page with all perspectives.

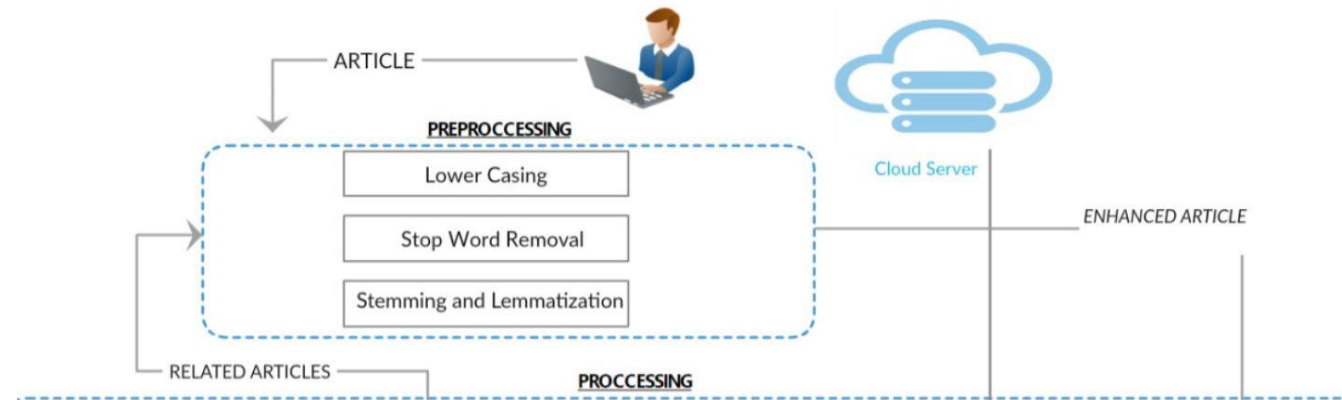


# System Overview



# Pre Processing

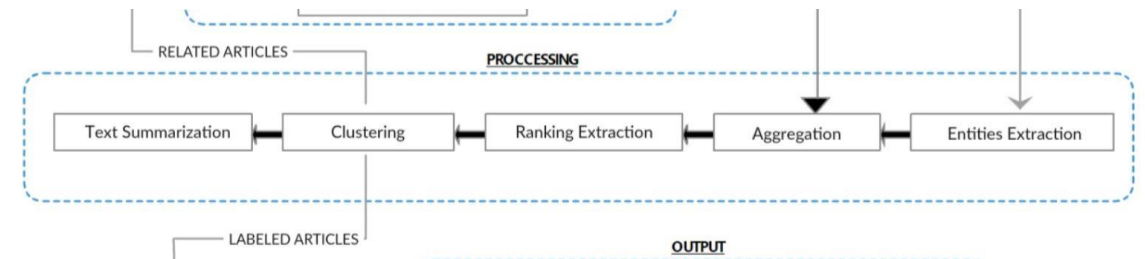
- Lowercasing.
- Stop word removal.
- Stemming.
- Text Classification using **FastText** Library.



# Processing

---

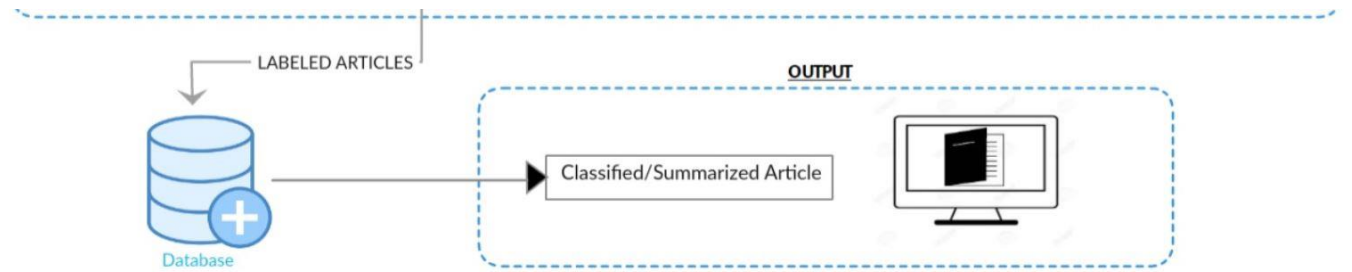
- Extracting relevant articles using **Naïve** approach.
- Clustering all of the related news using **hierarchical agglomerative clustering (HAC)**.
- Summarization in one page using **MEAD** toolkit.



# Output

---

- Getting Classified/Summarized article.
- Upload to database.
- Output to the user interface.

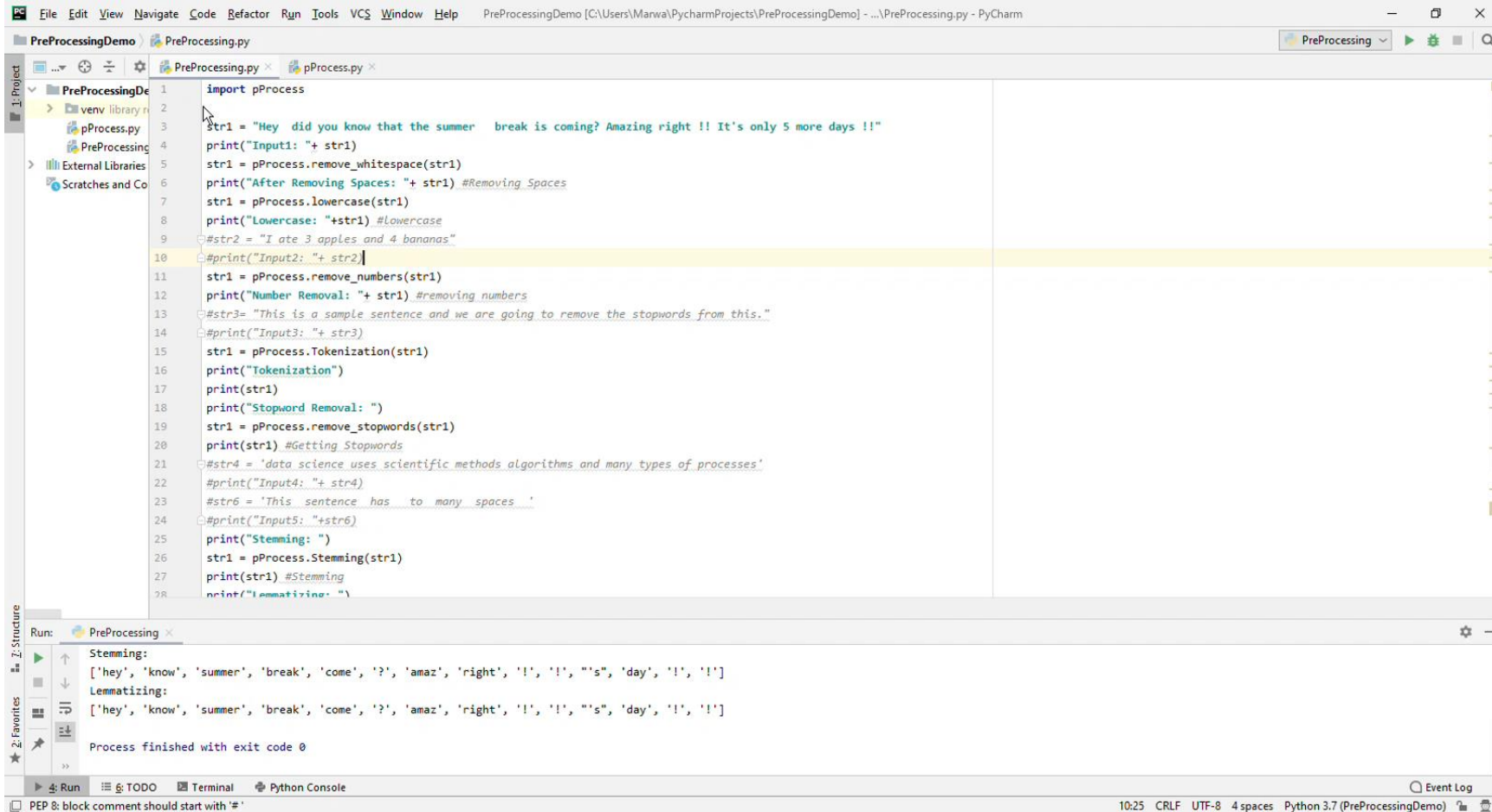


# Expected results

---

- Having related news with high **relevancy**.
- Articles Summarization to reduce **redundancy**  
and time consumed to visit several websites.
- Collecting different **perspectives** of articles and giving the user variety of options (Specialists , Tweets , Blogs or Feed).

# Demo 1/2



The image shows a PyCharm IDE window titled 'PreProcessingDemo'. The main editor displays a Python script named 'PreProcessing.py' with the following code:

```
1 import pProcess
2
3 str1 = "Hey did you know that the summer break is coming? Amazing right !! It's only 5 more days !!"
4 print("Input1: "+ str1)
5 str1 = pProcess.remove_whitespace(str1)
6 print("After Removing Spaces: "+ str1) #Removing Spaces
7 str1 = pProcess.lowercase(str1)
8 print("Lowercase: "+str1) #Lowercase
9 #str2 = "I ate 3 apples and 4 bananas"
10 #print("Input2: "+ str2)
11 str1 = pProcess.remove_numbers(str1)
12 print("Number Removal: "+ str1) #removing numbers
13 #str3= "This is a sample sentence and we are going to remove the stopwords from this.."
14 #print("Input3: "+ str3)
15 str1 = pProcess.Tokenization(str1)
16 print("Tokenization")
17 print(str1)
18 print("Stopword Removal: ")
19 str1 = pProcess.remove_stopwords(str1)
20 print(str1) #Getting Stopwords
21 #str4 = "data science uses scientific methods algorithms and many types of processes"
22 #print("Input4: "+ str4)
23 #str6 = "This sentence has too many spaces"
24 #print("Input5: "+str6)
25 print("Stemming: ")
26 str1 = pProcess.Stemming(str1)
27 print(str1) #Stemming
28 print("Lemmatizing: ")
```

The Run console at the bottom shows the output of the script:

```
Stemming:
['hey', 'know', 'summer', 'break', 'come', '?', 'amaz', 'right', '!', '!', '!', '!', 's', 'day', '!', '!']
Lemmatizing:
['hey', 'know', 'summer', 'break', 'come', '?', 'amaz', 'right', '!', '!', '!', '!', 's', 'day', '!', '!']
Process finished with exit code 0
```

The status bar at the bottom indicates the current settings: 10:25 CRLF UTF-8 4 spaces Python 3.7 (PreProcessingDemo).

# Demo 2/2

---

The screenshot shows a Windows File Explorer window with a sidebar on the left and a main pane on the right. The sidebar contains 'Quick access' and 'This PC' sections. The main pane displays a table of files and folders.

Name	Date modified	Type	Size
.idea	10/8/2019 7:48 PM	File folder	
tests	10/4/2019 11:53 AM	File folder	
venv	10/1/2019 8:15 PM	File folder	
cooking.stackexchange.id	4/28/2017 8:37 PM	ID File	88 KB
cooking.stackexchange	4/28/2017 8:37 PM	Text Document	1,370 KB
cooking.train	10/4/2019 11:57 AM	Text Document	684 KB
cooking.valid	10/4/2019 11:57 AM	Text Document	686 KB
FastText	10/7/2019 11:24 AM	Python File	1 KB
readme	5/2/2017 1:27 AM	Text Document	1 KB

# Preliminary visualization

The screenshot shows the homepage of 'News Aggregator EGY'. The header is dark red with the site name in white. Below the header is a navigation menu with links for Home, Urgent News, Search By Article, Reports, Technology, Policy, Economic, Accidents, and More. A search bar is located on the right side of the header. The main content area features a 'Welcome to News Aggregator' message with the date and time: 'Tue Oct 08 2019 21:21:49 GMT+0200 (Eastern European Standard Time)'. Below this, there are three article thumbnails. The first thumbnail is titled 'Barack Obama and Donald Trump Cannot Stand Each Other' and features a photo of Barack Obama. The second thumbnail is titled 'US migrant caravan: Trump's asylum ban halted by judge' and features a photo of Donald Trump. The third thumbnail is titled 'Bale can handle Real pressure - Giggs' and features a photo of Gareth Bale. At the bottom of the page, there is a footer with links for 'Terms Of Use', 'About Us', and 'Contact Us', and a copyright notice: 'Copyright © 2019 EgyptToday'.

This screenshot shows the same homepage as the previous one, but with a 'Search by Article' modal window open. The modal has a title bar with a close button (X). Below the title is a search input field with the placeholder text 'Keyword to the event'. To the left of the search field is a dropdown menu labeled 'Article Type -' with a list of categories: Urgent News, Articles, Reports, Politics, Accidents, Economics & Currency, Sport, Weather, Technology, Arts & Entertainment, Education, and Researches. At the bottom of the modal are two buttons: 'Search' and 'Close'. The background of the page is dimmed, showing the same article thumbnails as in the previous screenshot. At the bottom of the browser window, several open tabs are visible, including 'announcements-online', 'custom-quantum', 'BraDnascinnDem', 'Graduation Dinner', and 'News Aggregator'.



*Any Questions ?*

*Thank  
you!*