Kathmandu University

Department of Computer Science and Engineering

Dhulikhel, Kavre

A Project Report

on

"Minimal News"

[Code No:COMP207]

(For partial fulfillment of 2nd Year, 2nd Semester in Computer Engineering)

Submitted by

Aashutosh Aryal (5)

Sandesh Dahal (15)

Shusan Khatiwada (25)

Sameer Jung Subedi (49)

Submitted to

Mr. Nabin Ghimire

Department of Computer Science and Engineering

28 September, 2020

Bonafide Certificate

This project work on

“Minimal News”

is the bonafide work of

Aashutosh Aryal(05)

Sandesh Dahal(15)

Shusan Khatiwada(25)

Sameer Jung Subedi(49)

who carried out the project work under my supervision.

Project Supervisor:

Mr. Nabin Ghimire

Lecturer

Acknowledgement

We are very grateful to the Department of Computer Science and Engineering, Kathmandu University for providing us a gateway to explore ideas and approaches to make an efficient web app.

Our foremost word of gratitude goes to our project supervisor Mr. Nabin Ghimire, for his continuous guidance, invaluable constructive criticism and support throughout the process.

Here again thanking those beautiful hands and great minds that have helped us to complete this project.

Abstract

For most of history, it was inconceivable that there could ever be such a thing as ‘too much news’. But since the middle of the twentieth century, the volume of information has been growing exponentially.

The large volume of information is not a problem per se. The real problems are the lack of information-quality control and our inability to control the flow of incoming information. News has been commoditized and, in this process, it has become a major – though still too little known – risk to our mental health. Always, news organizations speak of our need to know – and the need to know right now. Human minds are fragile and because we have responsibilities closer to home, we need to lead out lives rather than be torn apart by the stories of the lives of others who are remote and irrelevant to us. This is why we created MINIMAL NEWS.

MINIMAL NEWS is a web news service that takes an approach different from the mainstream media focusing on ‘the need to know’, not everything, but ‘what is relevant’. A simple registration will deliver minimal curated news to the users’ right to their inbox. No more ads, no more getting stuck in infinite scroll or a rabbit hole of hyperlinks so that the user can give their precious attention to the tasks that require it while still staying informed.

Keywords: *minimal, curated, database, client-side, server-side, development, production*

Table of Contents

Acronyms/ Abbreviation

1. API: Application Programming Interface
2. CSS: Cascading Style Sheets
3. DBMS: Database Management System
4. HTML: HyperText Markup Language
5. HTTP: HyperText Transfer Protocol
6. IDE: Integrated Development Environment
7. JS:JavaScript
8. JSON: JavaScript Object Notation
9. ORM: Object Relational Mapper
10. RDBMS: Relational Database Management System
11. SQL: Structured Query Language
12. UI: User Interface
13. UX: User Experience
14. WSGI: Web Server Gateway Interface
15. XML: eXtensible Markup Language

Chapter 1: Introduction

1.1 Background

Going online created more opportunities for newspapers, such as competing with broadcast journalism in presenting breaking news in a timelier manner. The credibility and strong brand recognition of well-established newspapers, and the close relationships they have with advertisers, are also seen by many in the newspaper industry as strengthening their chances of survival. The movement away from the printing process can also help decrease costs.

News reporters are being taught to shoot video and to write in the succinct manner necessary for internet news pages. Some newspapers have attempted to integrate the internet into every aspect of their operations, e.g., the writing of stories for both print and online, and classified advertisements appearing in both media, while other newspaper websites may be quite different from the corresponding printed newspaper.

An early example of an "online-only" newspaper or magazine was (PLATO) News Report, an online newspaper created by Bruce Parrello in 1974 on the PLATO system at the University of Illinois. Beginning in 1987, the Brazilian newspaper Jornaldodia ran on the state-owned Embratel network, moving to the internet in the 1990s. By the late 1990s, hundreds of U.S. newspapers were publishing online versions, but did not yet offer much interactivity. One example is Britain's Weekend City Press Review, which provided a weekly news summary online beginning in 1995. Today, online news has become a huge part of society which leads people to argue whether or not it is good for society. Some people argue that online news does not provide the detail needed to fully understand what actually happened. It is more just a fast summary to inform people what happened, but does not give a solution or fixation to the problem.

In this era of ‘too much news’, MINIMAL NEWS focuses only on the necessary and specific news that a reader wants. Instead of falling on a rabbit hole and wasting time on the unnecessary, irrelevant news, a reader has the privilege of reading what he wants to read. This will not only save time but also prevent the reader to get mentally affected by the not needed news.

1.2 Objectives

There are some specific goals and objectives for our project. Some of them are listed below:

- To allow the reader to get specific categories of news.
- To eliminate the risk of possible mental disturbance caused by the irrelevant online news.
- To eliminate the irritation caused by the ads and pop-ups while reading news online.
- To provide minimal curated news to the users as per their preference.
- To relieve users of the inconvenience of having to install separate apps for news.

1.3 Motivation and Significance

Today's era is truly an era of information overload with News not being an exception. It is an important part of people's lives. There are so many sites and applications regulating news online. So, it is an equally important and difficult task to select the most promising and reliable news source on the Internet. Even with a good news source, we have to bear with all the advertisements and pop-ups which can be annoying and can ruin our reading experience. This is where MINIMAL NEWS steps in.

MINIMAL NEWS is a web news service that takes an approach different from the mainstream media focusing on ‘the need to know’, not everything, but ‘what is relevant’. A simple registration will deliver minimal curated news to the users’ right to their inbox. No more ads, no more getting stuck in an infinite scroll or a rabbit hole of hyperlinks so that the user can give their precious attention to the tasks that require it while still staying informed.

Chapter 2: Related Works

The idea used in this project, in recent days, have been used in various aspects of everyday world ranging from news portals' email services, news websites/web-apps to newsletters. Researching and observing services like "Minimal News" over the Internet, we intend to use the features matching with our project as a reference to enhance our project, using our original ideas and things we learn from the related works to create a unique project. Some of the most relevant works similar to our project are:

1. BBC Newsletter

BBC News is an operational business division of the British Broadcasting Corporation (BBC) responsible for the gathering and broadcasting of news and current affairs. The service maintains 50 foreign news bureau with more than 250 correspondents around the world. The newsletter service of BBC has a feature of direct inbox of news through email. The requirements include a BBC account and the user must be 13 years or over to apply for this service. The user can unsubscribe this service anytime.

Link: http://pages.email.bbc.com/subscribe

2. CNN Newsletter

CNN (Cable News Network) is an American news-based pay television channel owned by CNN Worldwide, a unit of the WarnerMedia News & Sports division of AT&T's WarnerMedia. Upon its launch in 1980, CNN was the first television channel to provide 24-hour news coverage, and was the first all-news television channel in the United States. The newsletter service of CNN has a feature of direct inbox of news through email with a wide range of content filtering. The requirements include a CNN account to apply for this service. The user can unsubscribe this service anytime.

Link: https://edition.cnn.com/email/subscription

Chapter 3: Design and Implementation

We built our web application using the very popular PostgreSQL for database, Flask and Gunicorn for backend and HTML/CSS for frontend.

3.1 Implementation Specifications

Grabber was built upon the strong yet flexible foundation of Python and Flask with the help of the following tools.

3.1.1 Client-Side (Frontend)

Front-end involves the codebase that runs on the client computer. It is the part that users interact with. Everything that the user sees when they're navigating around the program, from fonts and colors to menus, is built using HTML, CSS, and JavaScript that is being controlled by the computer's browser.

HTML

Hypertext Markup Language is the standard markup language for documents designed to be displayed in a web browser. HTML5 is to be used for structuring and presenting content on the World Wide Web. It is the fifth and latest major version of HTML that is a World Wide Web Consortium recommendation.

CSS

Cascading Style Sheets is a stylesheet language used to describe the presentation of a document written in HTML or XML. CSS describes how elements should be rendered on screen, on paper, in speech, or on other media.

JavaScript

JavaScript is a lightweight, interpreted, or just-in-time compiled programming language with first-class functions. While it is most well-known as the scripting language for Web pages, many non-browser environments also use it. JavaScript is a prototype-based, multi-paradigm, single-threaded, dynamic language, supporting object-oriented, imperative, and declarative (e.g. functional programming) styles.

3.1.2 Server-Side (Backend)

The server side for Minimal News is built using Flask. Since the default web-server provided by flask does not scale well and is not recommended for production, we used Gunicorn for our web server.

Flask

Flask is an open source Python web framework built with a small core and easy-to-extend philosophy. The flask app follows the Application Factory design pattern. This allows configuration changes to be applied dynamically. This also makes it easier for unit tests since it is necessary to run the application under different configuration settings for better test coverage. Not only this, but we also get the ability to create multiple app instances at runtime which helps maintain scalability.

Gunicorn

Gunicorn is a Python WSGI HTTP server for Unix. The Gunicorn server is broadly compatible with a number of web frameworks, simply implemented, light on server resources and fairly fast.

3.1.3 Database (Storage)

A relational Database system was used for our storage needs. We used two relations to information related to users and news respectively. We used SQLite as our database during production and for testing due to its portability and ease of use and switched to PostgreSQL for production because it provides better reliability and scalability compared to SQLITE.

SQLAlchemy

SQLAlchemy is the Python SQL toolkit and Object Relational Mapper that gives application developers the full power and flexibility of SQL. It provides a full suite of well-known enterprise-level persistence patterns, designed for efficient and high-performing database access, adapted into a simple and Pythonic domain language.

SQLite

SQLite is a cross platform relational database management system (RDBMS) which implements a small, fast, self-contained, high-reliability and full-featured, SQL engine. SQLite is ACID-compliant and implements most of the SQLstandard, generally following PostgreSQL syntax.

PostgreSQL

PostgreSQL is a free and open source relational database management system (RDBMS) that emphasizes extensibility and SQL compliance. Its free, open source, highly customizable nature has made it the most popular option for an open source database on cloud hosting platforms.

3.1.4 Scraper

When scraping the web, a web scraper sends a GET request using HTTP protocol to the target URL you mention. The web server if considers the request legitimate it allows you to read html of the web page and you store it in your program environment. We used scrapy for our web scraper. The scraping script takes a seed url and dynamically generates further urls to crawl based on the news categories provided and asynchronously crawls them scraping relevant information.

Scrapy

Scrapy is a fast-high-level python web crawling and web scraping framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.

3.1.5 Summarizer

Text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document. Automatic text summarization is a common problem in machine learning and natural language processing (NLP).

There are two main approaches used in summarizing a text using machine learning:

I. **Abstractive Summarization**
II. **Extractive Summarization**

We've decided to use the extractive text summarization approach in our program because although abstraction performs better at text summarization, developing its algorithms requires complicated deep learning techniques and sophisticated language modeling.
We followed these steps while using the extractive approach:

Step 1: Text Processing

This was done using the python's natural language processing module, nltk. The whole paragraph is divided into sentences. It extracts a sentence whenever a period appears but also considers periods in a part of a sentence like "Mr.".

Step 2: Text Cleaning

After the text has been divided into sentences we remove the extremely common words with little meaning such as "and" and "the". These words are called stopwords. Performing the filtering, assists in removing redundant and insignificant information which may not provide any added value to the text's meaning.

Step 3: Tokenization

Tokenizing the sentences is done to get all the words present in the sentences. After tokenization, each word is evaluated as a separate entity.

Step 4: Word frequency table creation

To calculate the weighted occurrence frequency of all the words, we divide the occurrence frequency of each of the words by the frequency of the most recurrent word in the paragraph along with its weighted frequency. The weighted frequency is the frequency relative to the word with maximum frequency. The weighted frequency is always less than or equal to one since its relative to the maximum frequency.

Step 5: Summarization

After substituting each of the words found in the original sentences with their weighted frequencies, we'll compute their sum. This will give us a score for each sentence. Since the weighted frequencies of the insignificant words, such as stop words and special characters, were removed during the processing stage, our score is more accurate.

From the sum of the weighted frequencies of the words, we can add the sentences with the most weighty scores and get a summary based on our needs.

3.2 System Requirement Specification

Since the project is web based, any web browser with support for HTML5, CSS and JavaScript will be capable of running our project. We recommend Mozilla Firefox and Google Chrome.

Chapter 4: Discussion on Achievements

When we came up with the idea for ‘minimal news’ we simply wanted to send news to the user according to their preferences. But soon we realized that some contents were too much and it would be against the concept of ‘minimal news’. Hence we decided to summarize the text using machine learning and natural language processing. While it was a small part in our program it let us learn about the basics of machine learning and natural language processing.

In our last semester’s project we focused more on desktop application rather than a web based program. During the course of this project we became familiar with the workings of a web based application as well.

Initially we intended to use ‘newsapi’ to get the contents of the news but later on after consulting with our supervisor we decided against it and scraped ‘thehimalayantimes.com’ for news ourselves. This way we had the content from a reputable news source in Nepal instead of relying on API which didn’t support a Nepali news site.

One of the slight changes to our project was the decision to add a corona statistics counter for Nepal. We decided that it was a very practical feature and would be used regularly in these difficult times.

Features:

- **Tailored news feed:** Users will get news according to their preferences directly to their email which reduces the hassle for the users.
- **Time of the day:** Users can choose the time of the day on which they want to receive the news.
- **Corona statistics:** Users can get the latest information about the covid-19 in Nepal which includes total cases, new cases today, number of recovered patients, etc. among other things.
- **Summarized content:** Users get a summarized version of the article.

Chapter 5: Conclusion

News informs our view of the world, and in response we take action and make choices based on how we perceive the world to be. MINIMAL NEWS is a web news service that takes an approach different from the mainstream media. Instead of bombarding the user with every piece of information available in every subject matter we aim to deliver news to the user tailored to their needs. A simple registration will deliver minimal curated news to the users' right to their inbox as per their subscription. No more ads, no more getting stuck in infinite scroll or a rabbit hole of hyperlinks so that the user can give their precious attention to the tasks that require it while still staying informed.

5.1. Limitations

- Application has not been deployed on the web.
- Only works with one news source.
- Summary is extractive, hence it's not always accurate.
- Sent email is in plain text.

5.2. Future enhancements

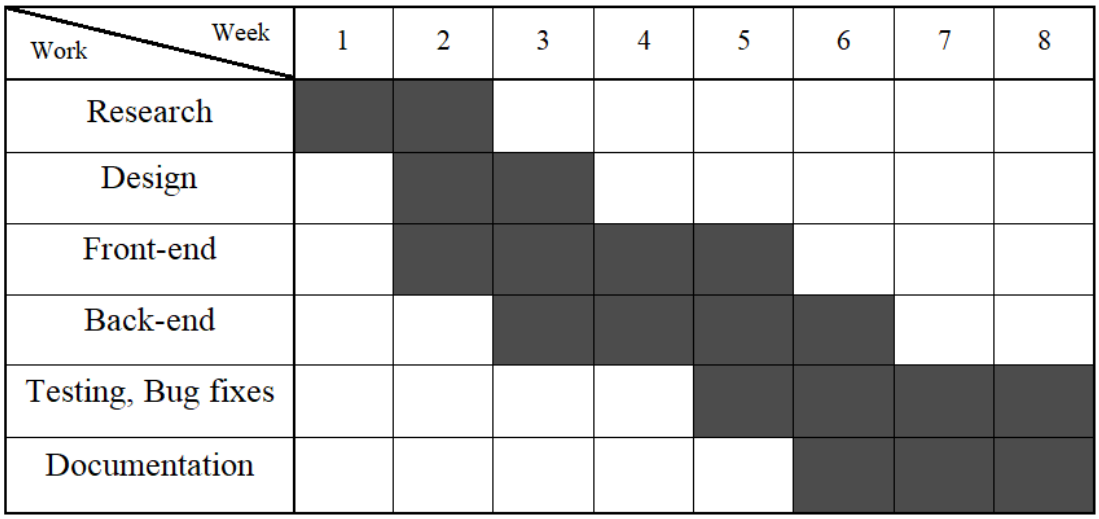
- Deploy the application on a server.
- Using other news sources for the news content.
- Use supervised learning algorithms and an abstractive approach for better summary.
- Format the news to make it more appealing.

References

- Scrapy, (2020, August). Retrieved fromhttps://docs.scrapy.org/en/latest/
- PostgreSQL 13.0, (2020, September 24). Retrieved from PostgreSQL: https://www.postgresql.org/docs/current/index.html
- Mihalcea and Tarau, (2004). TextRank:BringingOrderintoTexts. Retrieved from https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf
- Rocha, B. [Bruno Rocha]. (2020, July7*). The Application Factory Architecture - Bruno Rocha*[Video file]. Retrieved from YouTube: https://www.youtube.com/watch?v=xNo-eOfZH5Q
- Flask, (2020, April). *Flask Documentation*. Retrieved from Flask: https://flask.palletsprojects.com/en/1.1.x/
- SQLAlchemy 1.3.19, (2020, August 17). Retrieved from SQLAlchemy https://docs.sqlalchemy.org/en/13/
- Javascript, (2015). Retrieved from javascript https://devdocs.io/javascript/
- Luís Gonçalves. (2020, April 12). *Automatic Text Summarization with Machine Learning — An overview.* Retrieved from Medium: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25

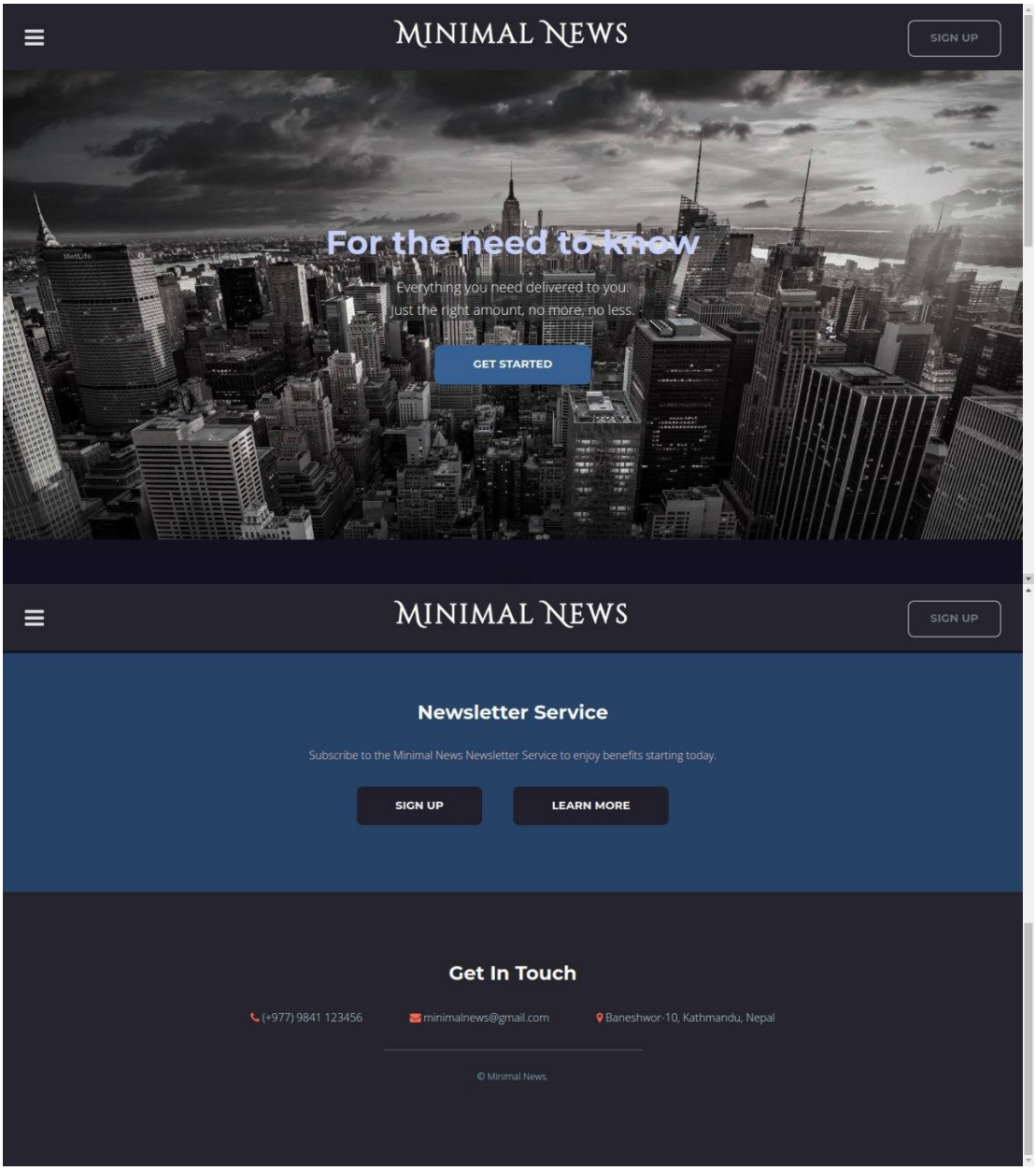
Appendices

Appendix A

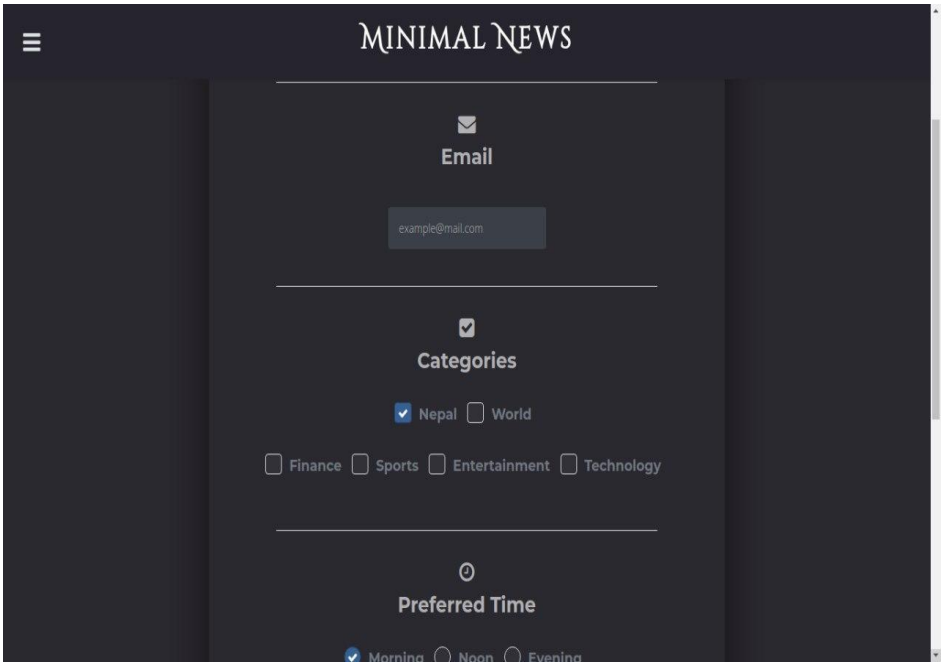


Gantt Chart

Appendix B

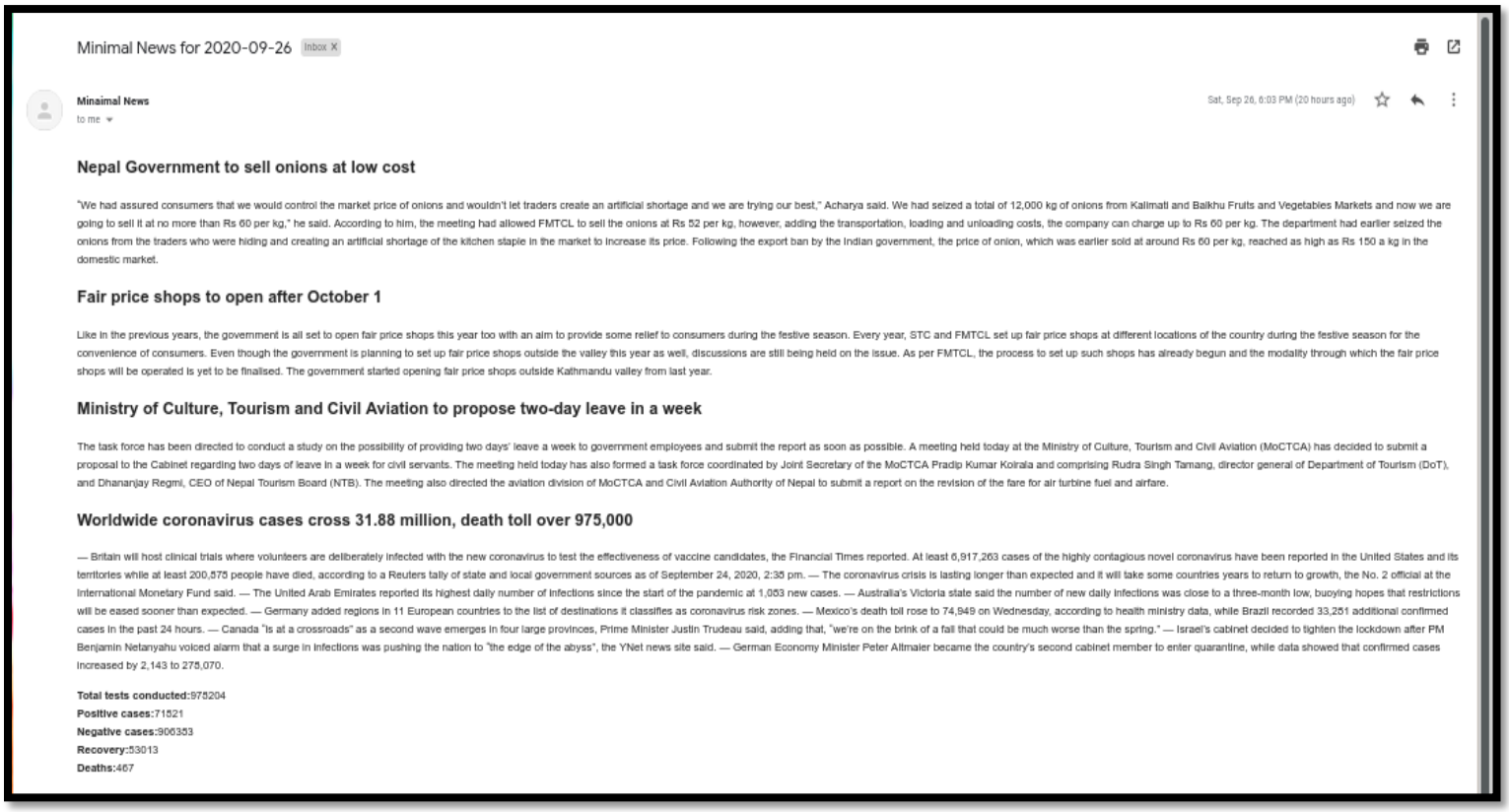


Front page of the website



Registration Form

Success Message



Minimal News for 2020-09-26 Inbox ×

Minaimal News
to me

Sat, Sep 26, 6:03 PM (20 hours ago)

Nepal Government to sell onions at low cost

"We had assured consumers that we would control the market price of onions and wouldn't let traders create an artificial shortage and we are trying our best," Acharya said. We had seized a total of 12,000 kg of onions from Kalimati and Balkhu Fruits and Vegetables Markets and now we are going to sell it at no more than Rs 60 per kg," he said. According to him, the meeting had allowed FMTCL to sell the onions at Rs 52 per kg, however, adding the transportation, loading and unloading costs, the company can charge up to Rs 60 per kg. The department had earlier seized the onions from the traders who were hiding and creating an artificial shortage of the kitchen staple in the market to increase its price. Following the export ban by the Indian government, the price of onion, which was earlier sold at around Rs 60 per kg, reached as high as Rs 150 a kg in the domestic market.

Fair price shops to open after October 1

Like in the previous years, the government is all set to open fair price shops this year too with an aim to provide some relief to consumers during the festive season. Every year, STC and FMTCL set up fair price shops at different locations of the country during the festive season for the convenience of consumers. Even though the government is planning to set up fair price shops outside the valley this year as well, discussions are still being held on the issue. As per FMTCL, the process to set up such shops has already begun and the modality through which the fair price shops will be operated is yet to be finalised. The government started opening fair price shops outside Kathmandu valley from last year.

Ministry of Culture, Tourism and Civil Aviation to propose two-day leave in a week

The task force has been directed to conduct a study on the possibility of providing two days' leave a week to government employees and submit the report as soon as possible. A meeting held today at the Ministry of Culture, Tourism and Civil Aviation (MoCTCA) has decided to submit a proposal to the Cabinet regarding two days of leave in a week for civil servants. The meeting held today has also formed a task force coordinated by Joint Secretary of the MoCTCA Pradip Kumar Koirala and comprising Rudra Singh Tamang, director general of Department of Tourism (DoT), and Dhananjay Regmi, CEO of Nepal Tourism Board (NTB). The meeting also directed the aviation division of MoCTCA and Civil Aviation Authority of Nepal to submit a report on the revision of the fare for air turbine fuel and airfare.

Worldwide coronavirus cases cross 31.88 million, death toll over 975,000

— Britain will host clinical trials where volunteers are deliberately infected with the new coronavirus to test the effectiveness of vaccine candidates, the Financial Times reported. At least 6,917,263 cases of the highly contagious novel coronavirus have been reported in the United States and its territories while at least 200,575 people have died, according to a Reuters tally of state and local government sources as of September 24, 2020, 2:35 pm. — The coronavirus crisis is lasting longer than expected and it will take some countries years to return to growth, the No. 2 official at the International Monetary Fund said. — The United Arab Emirates reported its highest daily number of infections since the start of the pandemic at 1,053 new cases. — Australia's Victoria state said the number of new daily infections was close to a three-month low, buoying hopes that restrictions will be eased sooner than expected. — Germany added regions in 11 European countries to the list of destinations it classifies as coronavirus risk zones. — Mexico's death toll rose to 74,949 on Wednesday, according to health ministry data, while Brazil recorded 33,251 additional confirmed cases in the past 24 hours. — Canada "is at a crossroads" as a second wave emerges in four large provinces, Prime Minister Justin Trudeau said, adding that, "we're on the brink of a fall that could be much worse than the spring." — Israel's cabinet decided to tighten the lockdown after PM Benjamin Netanyahu voiced alarm that a surge in infections was pushing the nation to "the edge of the abyss", the YNet news site said. — German Economy Minister Peter Altmaier became the country's second cabinet member to enter quarantine, while data showed that confirmed cases increased by 2,143 to 275,070.

Total tests conducted:975204
Positive cases:71521
Negative cases:906353
Recovery:53013
Deaths:467

Mail delivered to user's inbox

Mail delivered to user's inbox (phone)