



COVID-19  
CORONAVIRUS

# Μηχανή αναζήτησης άρθρων σχετικά με τον Covid-19

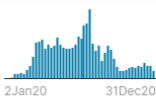
Μπαρμπαλιάς Ευάγγελος	2766
Ζέρβας Δημήτριος	2784

Ανάκτηση Πληροφορίας	28-5-21
----------------------	---------

# ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΤΜΗΜΑΤΩΝ ΤΗΣ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ

Χρησιμοποιήσαμε δεδομένα από το **kaggle** της μορφής

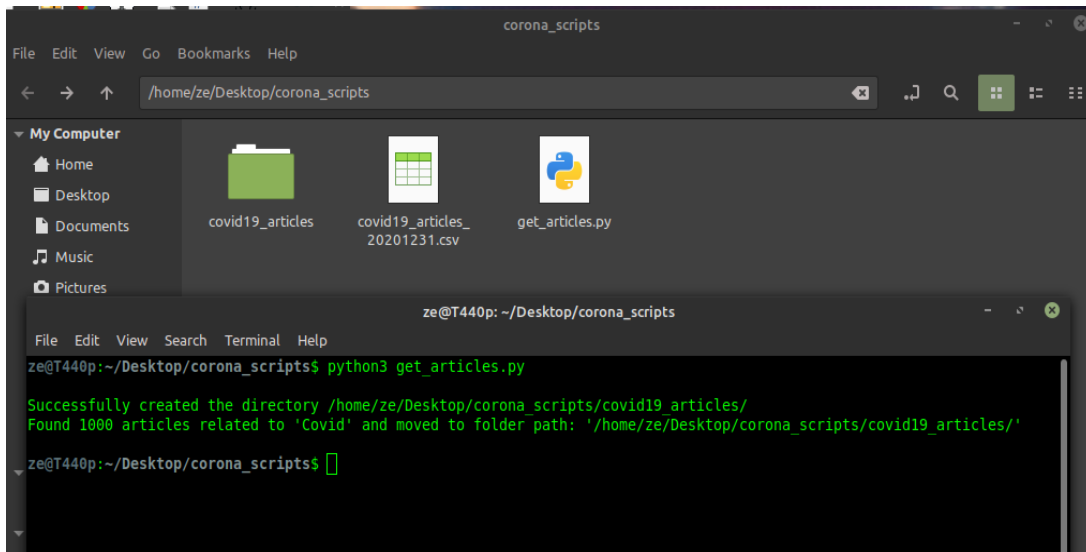
(<https://www.kaggle.com/jannalipenkova/covid19-public-media-dataset>)

< covid19_articles_20201231.csv (2.09 GB)						
Detail Compact Column						
<b>About this file</b>						
This file contains more than 380,000 online articles with full texts which were scraped from online media in the timespan January 1 - December 31, 2020 from 65 English-language websites.						
▲ author	📅 date	▲ domain	▲ title	🔗 url	▲ content	▲ topic_area
Article author	Date of initial publication	Web domain	Headline of the article	Complete article URL	Article content	The topic area is determined by the website where the article was published. Possible values: general, business,
[null] 51%		finance.yahoo 25%	354320 unique values	364827 unique values	368920 unique values	business 66%
Associated Press 0%		marketscreener 22%				general 23%
Other (180141) 49%		Other (196891) 53%				Other (39580) 11%
Thomas Hughes	2020-01-02	marketbeat	Three Industrial Giants You Should Own In 2020	<a href="https://www.marketbeat.com/originals/three-industrial-giants-you-should-own-in-2020/">https://www.marketbeat.com/originals/three-industrial-giants-you-should-own-in-2020/</a>	With the end of the year just around the corner, it's past time to think about positioning for 2020. ...	business
Thomas Hughes	2020-01-03	marketbeat	Labor Stocks Are Going To Break Out In 2020	<a href="https://www.marketbeat.com/originals/labor-stocks-are-going-to-break-out-in-2020/">https://www.marketbeat.com/originals/labor-stocks-are-going-to-break-out-in-2020/</a>	The labor markets were one of the most closely watched segment of our economy in 2019. Despite all t...	business

Το αρχείο **covid19\_articles\_20201231.csv**, έχει τα εξής πεδία

- Author
- Date
- Domain
- Title
- Url
- Content
- Topic\_area

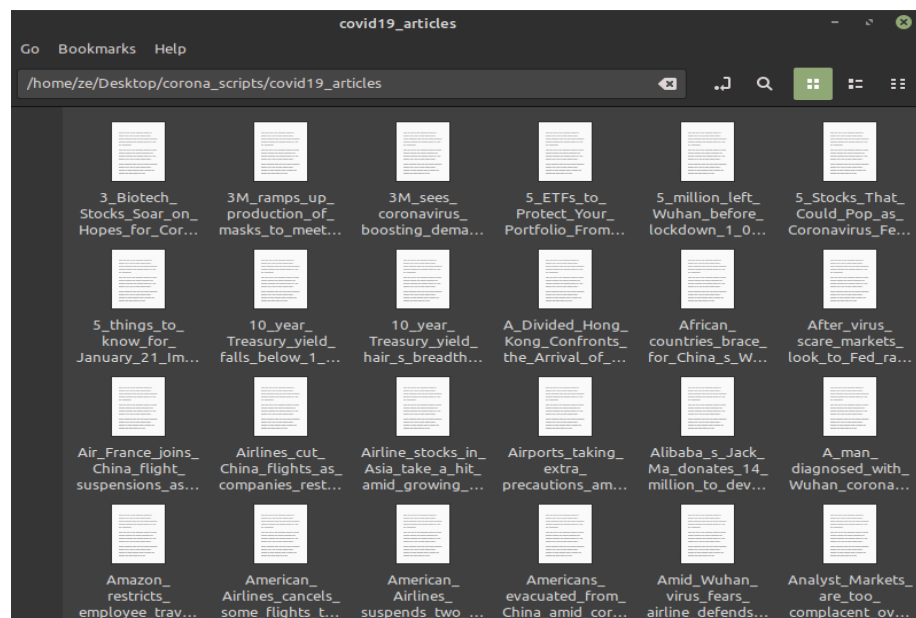
Για την συλλογή των άρθρων χρησιμοποιήσαμε το script **get articles**.



Με το αρχείο αυτό /scripts/**get\_articles.py** διαβάζουμε κάθε γραμμή του csv και ελέγχουμε αν κάποια από τις λέξεις **covid, sars, coronavirus, virus, pneumonia, flu, epidemic** εμφανίζεται στον τίτλο(title). Αν εμφανίζεται, τότε αποθηκεύει όλα τα πεδία με την παρακάτω σειρά author, date, domain, title, url, topic\_area, content με new line σε ένα νέο φάκελο που δημιουργείται.

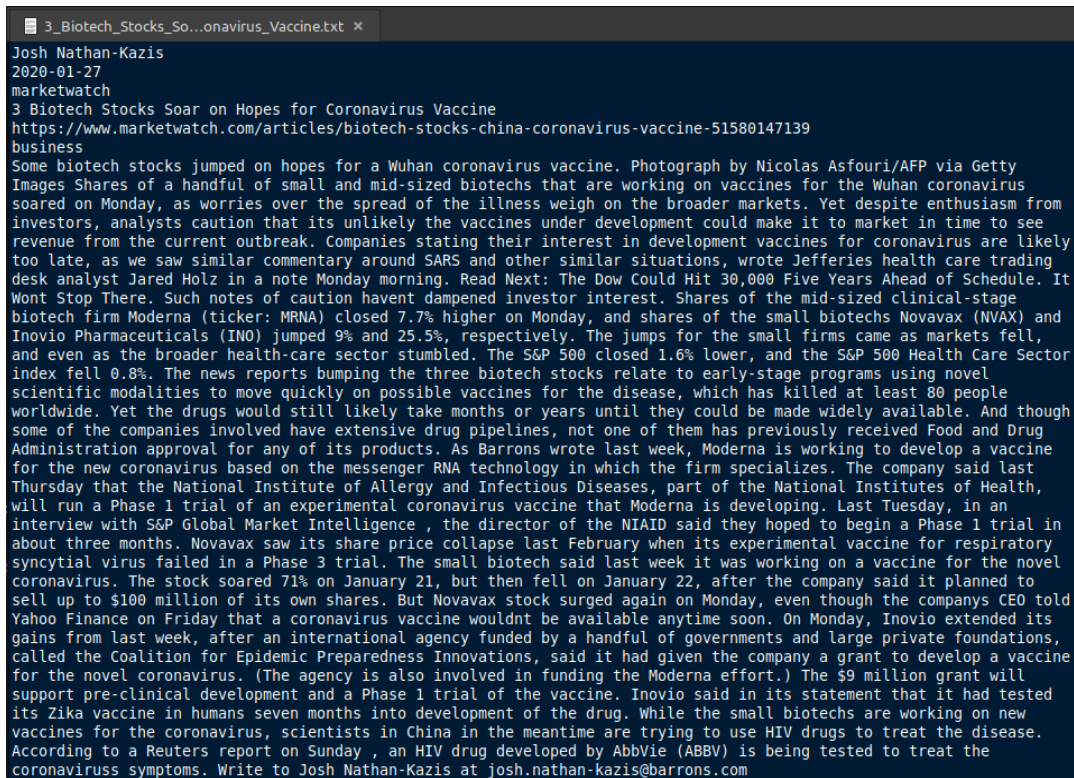
Ο φάκελος δημιουργείται αυτόματα (εάν δεν υπάρχει ήδη) και εκεί αποθηκεύονται αρχεία txt, με βάση τον τίτλο του κάθενος  
πχ 3 Biotech Stocks Soar on Hopes for Coronavirus Vaccine.txt

Έχουμε  
συλλέξει 1000  
τέτοια άρθρα,  
οπότε ο  
φάκελος  
covid19\_article  
s περιέχει:



Το κάθε αρχείο είναι χωρισμένο με βάση τα πεδία του και new line χαρακτήρες, δηλαδή

- όπου η 1η γραμμή είναι -> author
- η 2η είναι -> date
- η 3η είναι -> domain
- η 4η είναι -> title
- η 5η είναι -> topic\_area
- η 6η είναι -> content



3\_Biotech\_Stocks\_So...onavirus\_Vaccine.txt x

Josh Nathan-Kazis  
2020-01-27  
marketwatch  
3 Biotech Stocks Soar on Hopes for Coronavirus Vaccine  
<https://www.marketwatch.com/articles/biotech-stocks-china-coronavirus-vaccine-51580147139>  
business

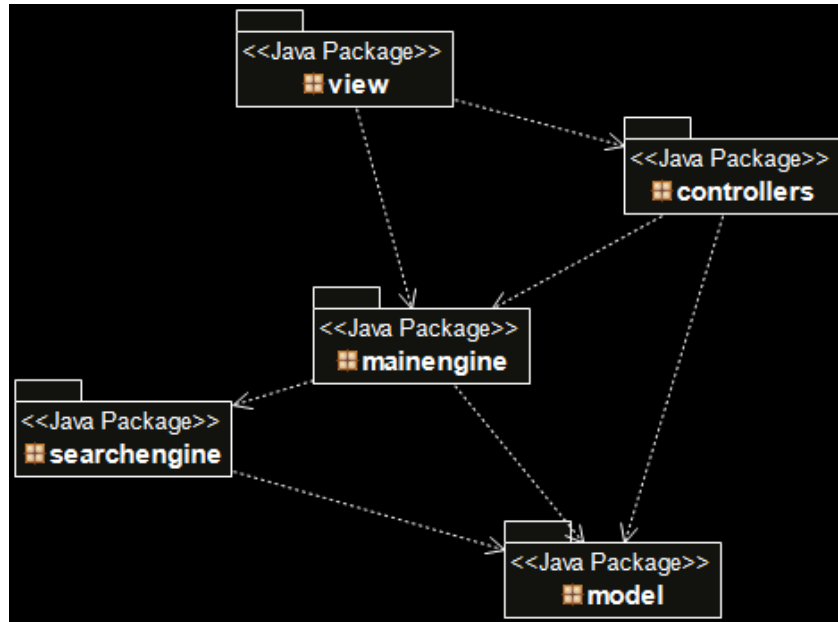
Some biotech stocks jumped on hopes for a Wuhan coronavirus vaccine. Photograph by Nicolas Asfour/AFP via Getty Images Shares of a handful of small and mid-sized biotechs that are working on vaccines for the Wuhan coronavirus soared on Monday, as worries over the spread of the illness weigh on the broader markets. Yet despite enthusiasm from investors, analysts caution that its unlikely the vaccines under development could make it to market in time to see revenue from the current outbreak. Companies stating their interest in development vaccines for coronavirus are likely too late, as we saw similar commentary around SARS and other similar situations, wrote Jefferies health care trading desk analyst Jared Holz in a note Monday morning. Read Next: The Dow Could Hit 30,000 Five Years Ahead of Schedule. It Wont Stop There. Such notes of caution havent dampened investor interest. Shares of the mid-sized clinical-stage biotech firm Moderna (ticker: MRNA) closed 7.7% higher on Monday, and shares of the small biotechs Novavax (NVAX) and Inovio Pharmaceuticals (INO) jumped 9% and 25.5%, respectively. The jumps for the small firms came as markets fell, and even as the broader health-care sector stumbled. The S&P 500 closed 1.6% lower, and the S&P 500 Health Care Sector index fell 0.8%. The news reports bumping the three biotech stocks relate to early-stage programs using novel scientific modalities to move quickly on possible vaccines for the disease, which has killed at least 80 people worldwide. Yet the drugs would still likely take months or years until they could be made widely available. And though some of the companies involved have extensive drug pipelines, not one of them has previously received Food and Drug Administration approval for any of its products. As Barrons wrote last week, Moderna is working to develop a vaccine for the new coronavirus based on the messenger RNA technology in which the firm specializes. The company said last Thursday that the National Institute of Allergy and Infectious Diseases, part of the National Institutes of Health, will run a Phase 1 trial of an experimental coronavirus vaccine that Moderna is developing. Last Tuesday, in an interview with S&P Global Market Intelligence , the director of the NIAID said they hoped to begin a Phase 1 trial in about three months. Novavax saw its share price collapse last February when its experimental vaccine for respiratory syncytial virus failed in a Phase 3 trial. The small biotech said last week it was working on a vaccine for the novel coronavirus. The stock soared 71% on January 21, but then fell on January 22, after the company said it planned to sell up to \$100 million of its own shares. But Novavax stock surged again on Monday, even though the companys CEO told Yahoo Finance on Friday that a coronavirus vaccine wouldnt be available anytime soon. On Monday, Inovio extended its gains from last week, after an international agency funded by a handful of governments and large private foundations, called the Coalition for Epidemic Preparedness Innovations, said it had given the company a grant to develop a vaccine for the novel coronavirus. (The agency is also involved in funding the Moderna effort.) The \$9 million grant will support pre-clinical development and a Phase 1 trial of the vaccine. Inovio said in its statement that it had tested its Zika vaccine in humans seven months into development of the drug. While the small biotechs are working on new vaccines for the coronavirus, scientists in China in the meantime are trying to use HIV drugs to treat the disease. According to a Reuters report on Sunday , an HIV drug developed by AbbVie (ABBV) is being tested to treat the coronavirus symptoms. Write to Josh Nathan-Kazis at [josh.nathan-kazis@barrons.com](mailto:josh.nathan-kazis@barrons.com)

Κατά το pre-processing αφαιρούμε από το κυρίως κείμενο (content) special characters, new lines, javascript code ( document.write(...); )  
...

Τα υπόλοιπα πεδία δεν χρειάζονται pre-processing.

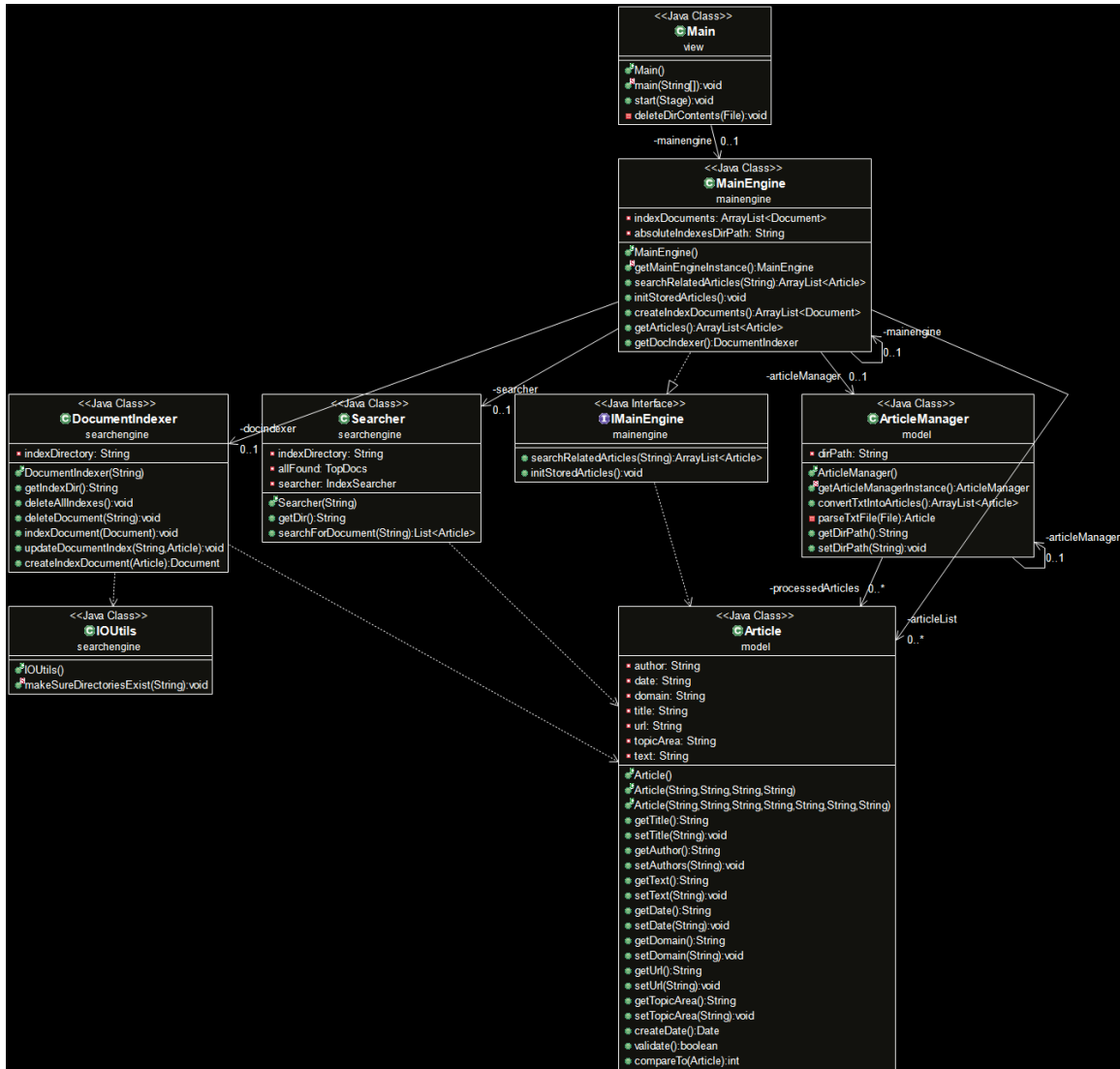
## ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΛΟΓΙΣΜΙΚΟΥ

Το μοντέλο αρχιτεκτονικής που χρησιμοποιήθηκε για τη δημιουργία της εφαρμογής είναι το Model-View-Controller (MVC). Στο μοντέλο αυτό η εφαρμογή διαιρείται σε τρία διασυνδεδεμένα μέρη ώστε να διαχωριστεί η παρουσίαση της πληροφορίας στον χρήστη από την μορφή που έχει αποθηκευτεί στο σύστημα.



- **view package** : Περιέχει τις boundary classes, υπεύθυνες για την αναπαράσταση της πληροφορίας που περιέχει το model, δημιουργώντας γραφική παρουσίαση στον χρήστη.
- **controllers package** : Το controllers πακέτο περιέχει κλάσεις-controllers για τον χειρισμό των διάφορων components που απαρτίζουν το UI, καθώς και των διάφορων ενεργειών που υλοποιούνται κατά την αλληλεπίδραση του χρήστη με το UI.
- **mainengine package** : Κεντρική business logic engine, με interface προς υλοποίηση.
- **searchengine package** : Υποσύστημα για τη λειτουργία της μηχανής αναζήτησης.
- **model package** : Domain classes του συστήματος.

## ΔΙΑΓΡΑΜΜΑ UML



## ΛΕΙΤΟΥΡΓΙΕΣ ΕΦΑΡΜΟΓΗΣ

Κατά την έναρξη της εφαρμογής, τα .txt αρχεία που βρίσκονται στον φάκελο covid19\_articles επεξεργάζονται ώστε να μετατραπούν σε αντικείμενα τύπου Article, τη θεμελιώδη κλάση που αναπαριστά ένα άρθρο με τα σχετικά πεδία του. Έπειτα η λίστα με αυτά τα αντικείμενα 'μεταφέρεται' στη Lucene, προκειμένου να δημιουργηθούν indexes για αυτά. Για το παραπάνω (indexing) απαιτείται κάποιος χρόνος, για αυτό και καθυστερεί η εφαρμογή κατά την εκκίνηση της.

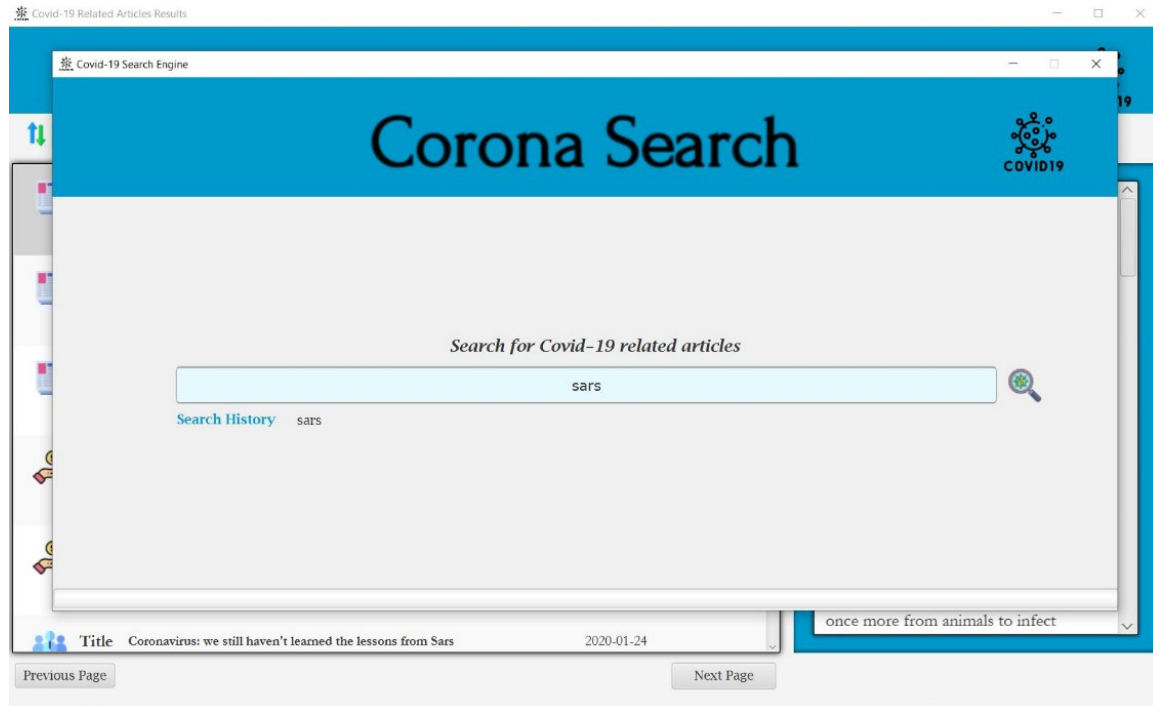
Μόλις φορτωθεί το γραφικό περιβάλλον και το βασικό παράθυρο της εφαρμογής, ο χρήστης μπορεί να πληκρολογήσει τον όρο που θα αναζητηθεί. Μέσω της mainengine που συνδέει back με front-end καλείται η μέθοδος searchForDocument με όρισμα τον παραπάνω όρο, αναζητώντας στα πεδία «TEXT» και «TITLE» για τον όρο αυτό. Αν βρεθεί, επιστρέφεται μια λίστα με τα αντικείμενα τύπου Article που τον περιέχουν είτε στον τίτλο είτε στο κείμενο.

Για την παραπάνω λειτουργία, γίνεται η χρήση του StandardAnalyzer ο οποίος μετατρέπει σε lowercase τα token και αφαιρεί κοινές και σημεία στίξης, αν αυτά υπάρχουν.

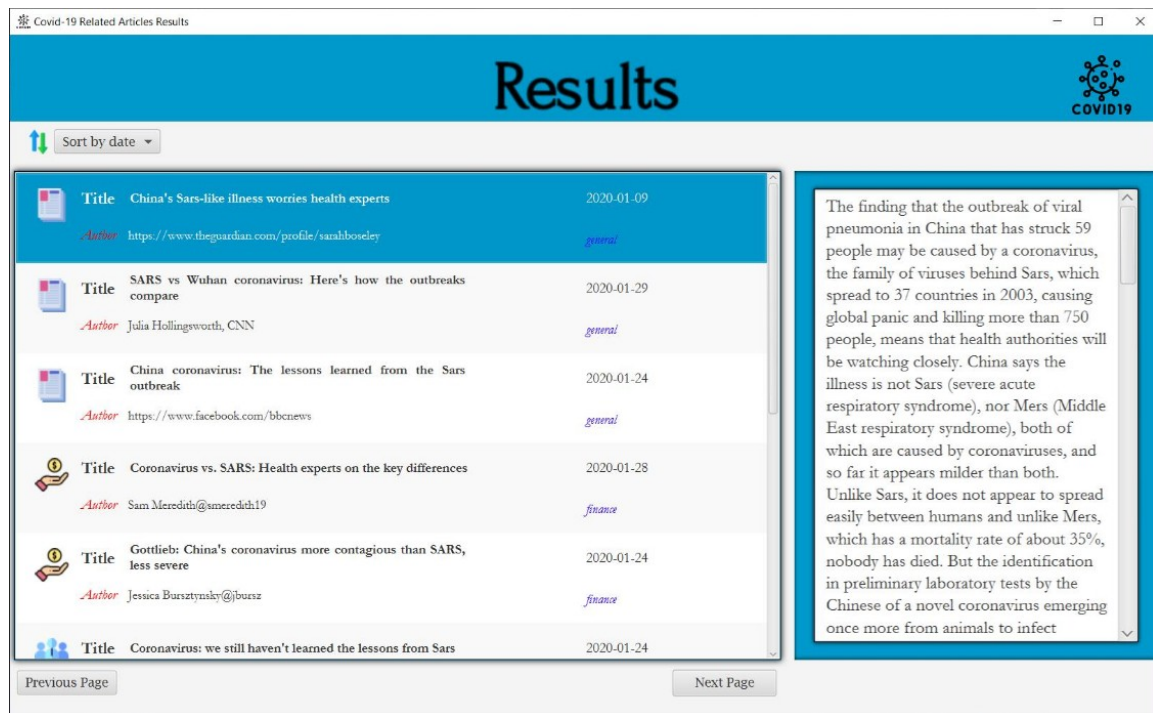
Τα αποτελέσματα εμφανίζονται στο 2<sup>ο</sup> παράθυρο της εφαρμογής ανα 10, με τον χρήστη να έχει τη δυνατότητα να δει τα 10 επόμενα ή/και 10 προηγούμενα με τη χρήση των πλήκτρων *Next page* και *Previous Page* αντίστοιχα. Το κάθε αποτέλεσμα εμφανίζει πληροφορίες για τον τίτλο, συγγραφέα, ημερομηνία και κατηγορία άρθρου ενώ επιλέγοντας το, εμφανίζεται και το κείμενο του. Τέλος ο χρήστης μπορεί να διατάξει τα άρθρα που βρέθηκαν με βάση την ημερομηνία συγγραφής τους, είτε κατά αύξουσα είτε φθίνουσα σειρά.

## ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ

Έστω ότι ο χρήστης ψάχνει την λέξη sars



Το αποτέλεσμα είναι ανα 10





Όπου στα αριστερά βρίσκεται ο συγγραφέας, ο τίτλος, η ημερομηνία και ο τομέας (οικονομικός, εργασιακός. ...) και στα δεξιά το κείμενο του άρθρου.

Υπάρχει η επιλογή για ταξινόμηση με βάση την ημερομηνία

**Covid-19 Related Articles Results**

## Results

Sort by date ▾

Ascending  
Descending

	<b>Title</b> China's Sars-like illness worries health experts	2020-01-09
	<b>Author</b> <a href="https://www.theguardian.com/profile/sarahboseley">https://www.theguardian.com/profile/sarahboseley</a>	<a href="#">general</a>
	<b>Title</b> SARS vs Wuhan coronavirus: Here's how the outbreaks compare	2020-01-29
	<b>Author</b> Julia Hollingsworth, CNN	<a href="#">general</a>
	<b>Title</b> China coronavirus: The lessons learned from the Sars outbreak	2020-01-24
	<b>Author</b> <a href="https://www.facebook.com/bbcnews">https://www.facebook.com/bbcnews</a>	<a href="#">general</a>
	<b>Title</b> Coronavirus vs. SARS: Health experts on the key differences	2020-01-28
	<b>Author</b> Sam Meredith@smereidit19	<a href="#">finance</a>
	<b>Title</b> Gottlieb: China's coronavirus more contagious than SARS, less severe	2020-01-24
	<b>Author</b> Jessica Burzutyusky@jburz	<a href="#">finance</a>
	<b>Title</b> Coronavirus: we still haven't learned the lessons from Sars	2020-01-24

Previous Page Next Page

The finding that the outbreak of viral pneumonia in China that has struck 59 people may be caused by a coronavirus, the family of viruses behind Sars, which spread to 37 countries in 2003, causing global panic and killing more than 750 people, means that health authorities will be watching closely. China says the illness is not Sars (severe acute respiratory syndrome), nor Mers (Middle East respiratory syndrome), both of which are caused by coronaviruses, and so far it appears milder than both. Unlike Sars, it does not appear to spread easily between humans and unlike Mers, which has a mortality rate of about 35%, nobody has died. But the identification in preliminary laboratory tests by the Chinese of a novel coronavirus emerging once more from animals to infect

Τα αποτελέσματα φαίνονται ανά 10, κάνοντας κλικ στο κουμπί Next Page, προχωράμε στα επόμενα 10 (πχ 11-20), αντίστοιχα με το κουμπί Previous Page.

Επίσης αποθηκεύεται το ιστορικό των αναζητήσεων του χρήστη.

**Covid-19 Search Engine**

## Corona Search

Search for Covid-19 related articles

covid

Search History sars covid

UK presses China to let dual nationals join coronavirus

coronavirus can spread between humans