**CSC 690 CAPSTONE PROJECT**

**A Project Report**

**of**

**Clinical Expert**



**International Technological University**

**Department of**

**Computer Science and Software Engineering**

**Fall 2020**

**GROUP 21**

Submitted By:

1. Aishwarya Prakash Kembhavi(95672) - kembhaviaishw3138@students
2. Priya Phapale (94662) - phapalepriya2694@students.itu.edu
3. Shivani Sachdeva (88785) - [88785.sachdeva@students.itu.edu](mailto:88785.sachdeva@students.itu.edu)

Submitted To:

Prof. Mamoun Samaha

**ACKNOWLEDGEMENT**

## We would like to acknowledge the following for their support and help regarding this project. To our ever-loving parents who are always there to give us courage to pursue our goals and provide financial and emotional support. To our teachers and mentors who taught us principles that helped us with our study. To our friends who are always capable of giving us enough faith in doing this research proposal.

**GROUP 21 and Team members**

**Department of Computer Science and Software Engineering**

|  |  |
| --- | --- |
| **Aishwarya Prakash Kembhavi**  **Student Id:** 95672  **Masters:** Computer Science  [kembhaviaishw3138@students.itu.edu](mailto:kembhaviaishw3138@students.itu.edu)  **Years of experience-** 1.7 years |  |
| **Priya Phapale**  **Student Id**: 94662  **Masters:** Computer Science  [phapalepriya2694@students.itu.edu](mailto:phapalepriya2694@students.itu.edu)  **Years of experience-** 5 years |  |
| **Shivani Sachdeva**  **Student Id:** 88785  **Masters:** Software Engineering  [88785.sachdeva@students.itu.edu](mailto:88785.sachdeva@students.itu.edu)  **Years of experience-** 5 years |  |

## 

## 

## 

## 

## 

## 

## **Purpose**

## 

## The purpose of the project is to provide health care providers (HCPs) with automated access to articles that can support clinical decision-making and diagnosis. By providing HCPs with summarized versions of the latest publications relevant to patient diseases and acute health concerns. The source of our ER Admission Records and PubMed Articles. We chose this database because it contains all of the necessary clinical data, while still being a manageable size. When we previously attempted to use the MIT EHR database, we found that it was nearly 1TB in size, making it impractical to use for this project. Furthermore, we will use the MeSH database as our ontology form to pull relevant keywords. This may change as we further define the scope of our project, but for now it seems to contain a suitable listing of medical conditions which we can extract from our ER admission records.

## Once we have extracted the relevant words from a given admission record, we will look for relevant PubMed articles. Next, we will summarize the articles and bold sentences that contain the relevant conclusions and findings of the article. If we have enough time after accomplishing these goals, we hope to add an additional element, such as extracting the conflicting conclusions out of the larger set of conclusions that we have found within the articles. We plan to work with Dimeji Farri in order to identify an achievable and clinically relevant goal and will discuss our options with him in the 3rd week of November. We believe that this project will have real-world applications in the medical field and has the potential to provide HCPs with key information that will enable them to make more informed decisions in a clinical setting.

## 

## 

## 

## 

## 

## **Table of Contents**

## **Introduction**

## **Objective**

## **2.2 What is the problem**

**2.3 Motivation and Problem Statement**

1. **Related research**

**3.2 Unsupervised Method**

1. **Scope**
2. **Methodology**
3. **How to generate/collect input data**
4. **How to solve the problem**
5. **Algorithm design**
6. **Language and Tools used**
7. **How to generate output**
8. **How to test against hypotheses**
9. **Implementation**
10. **Mockups**
11. **Conclusion and Recommendations**

**14.1 Drawback**

**14.2 How could we improve performance**

**14.3 Future work**

1. **Bibliography**

## 

## **1.Introduction**

Study shows that more than 200 million surgeries are performed worldwide each year and overall an average of 58.1 clinical articles published per year to help doctors to give better treatment to patients. But nowadays doctors are so overloaded that they could not keep up with the latest articles about treatments, diseases, and the interactions between them. They don't get enough time to find and read such valuable artifacts before going for surgery especially when there is an emergency. As each patient represents a new combination of history, symptoms, and potential disease the process of searching for precise information becomes time taking and practically impossible. It would be a great idea if doctors get automated access to articles related to that particular case which will support them in clinical decision-making and to diagnosis accurately.

## 

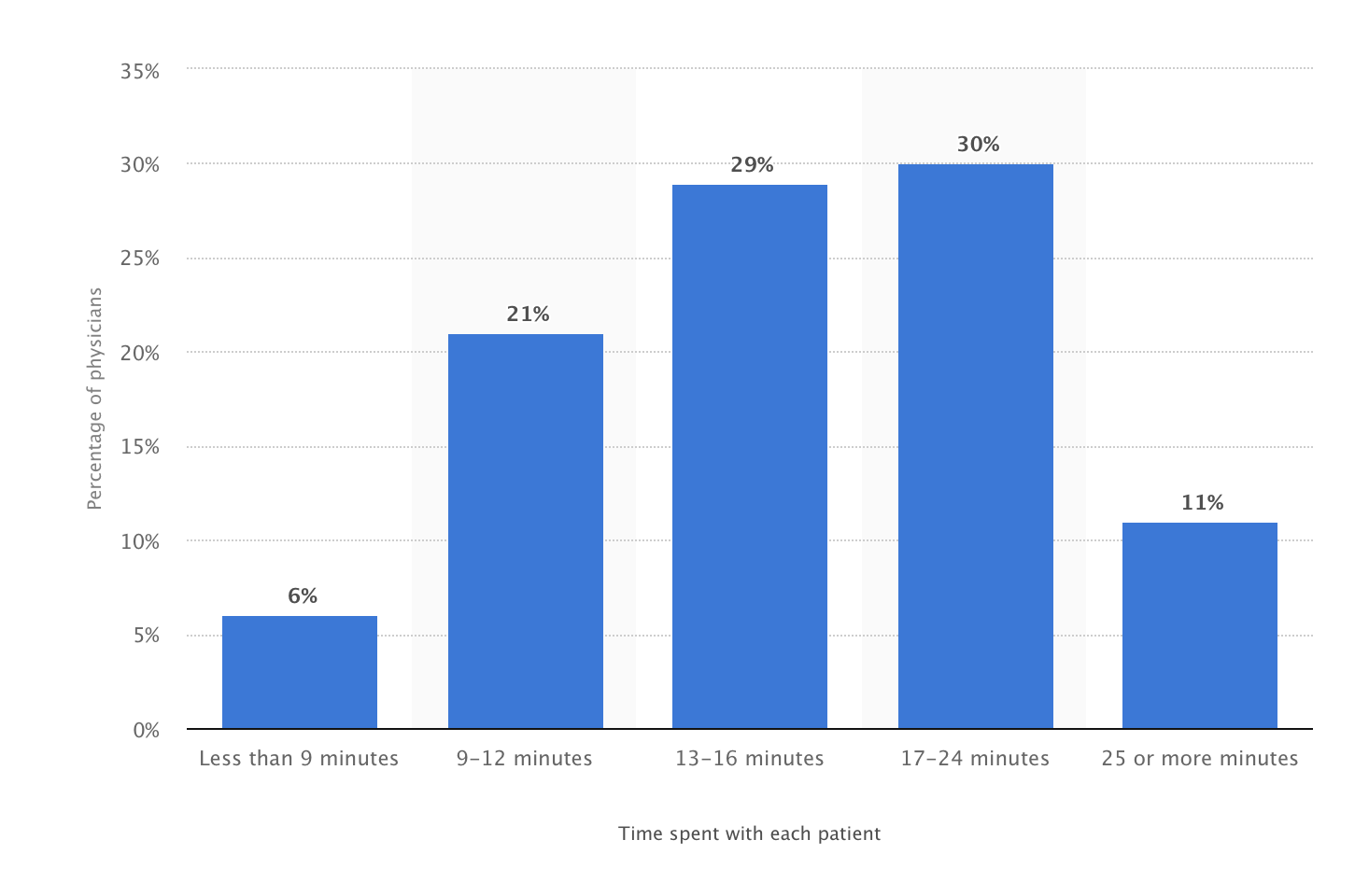
## **2. Objective**

With the help of machine learning techniques, we have developed a tool which can help doctors to make better and fast decisions in emergency situations. This tool provides rapid and accurate research papers, articles relevant to the patient by analyzing a patient's history, symptoms and diseases.

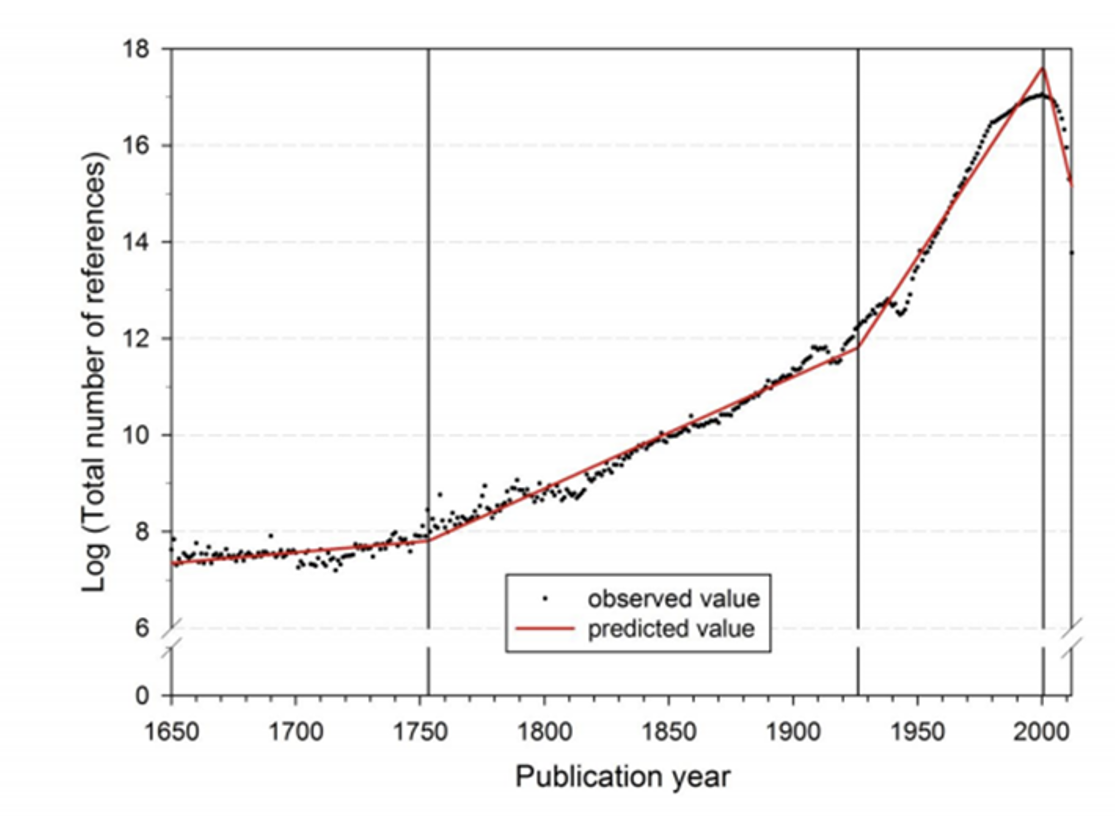
## **2.1 What is the problem**

There is currently no readily-available software that will link electronic health records to relevant scientific articles. While there are existing tools which summarize articles and extract keywords, it does not have the capability to combine these with medical terminology to identify clinically relevant information. Our approach will identify actionable keywords in patient’s summaries and use these tokens to identify articles that can provide support in clinical decision making by delivering articles that provide information relevant to a patient’s condition.

Below statistic shows that physicians have to spend most of the time with patients to understand their health issues, symptoms, to know their present and past medications, effects and side effects to find out the root cause of disease.



Sources: statista.com (above), nature.com (below)



Segmented growth of the annual number of cited references from 1650 to 2012 (citing publication from 1980 to 2012)

**2.3 Motivation and Problem Statement**

To provide health care providers (HCPs) with automated access to articles that can support clinical decision-making and diagnosis.

Each patient represents a new combination of history, symptoms, and potential disease.

Health care providers are already overloaded and often have trouble keeping up with the latest literature about treatments, diseases, and the interactions between them.

Our Project plan is to develop a tool which can be used to provide HCPs with rapid and accurate reports of research relevant to each patient who enters the hospital.

**3. Related research**

**Unsupervised Method**

Chosen Method: Unsupervised Learning.

Looking at the different models we trained, on the different data sets, we got the best performance from the unsupervised learning.

Unsupervised Learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning). A task involving machine learning may not be linear, but it has a number of well known steps:

* Problem definition.
* Preparation of Data.
* Learn an underlying model.
* Improve the underlying model by quantitative and qualitative evaluations.
* Present the model.

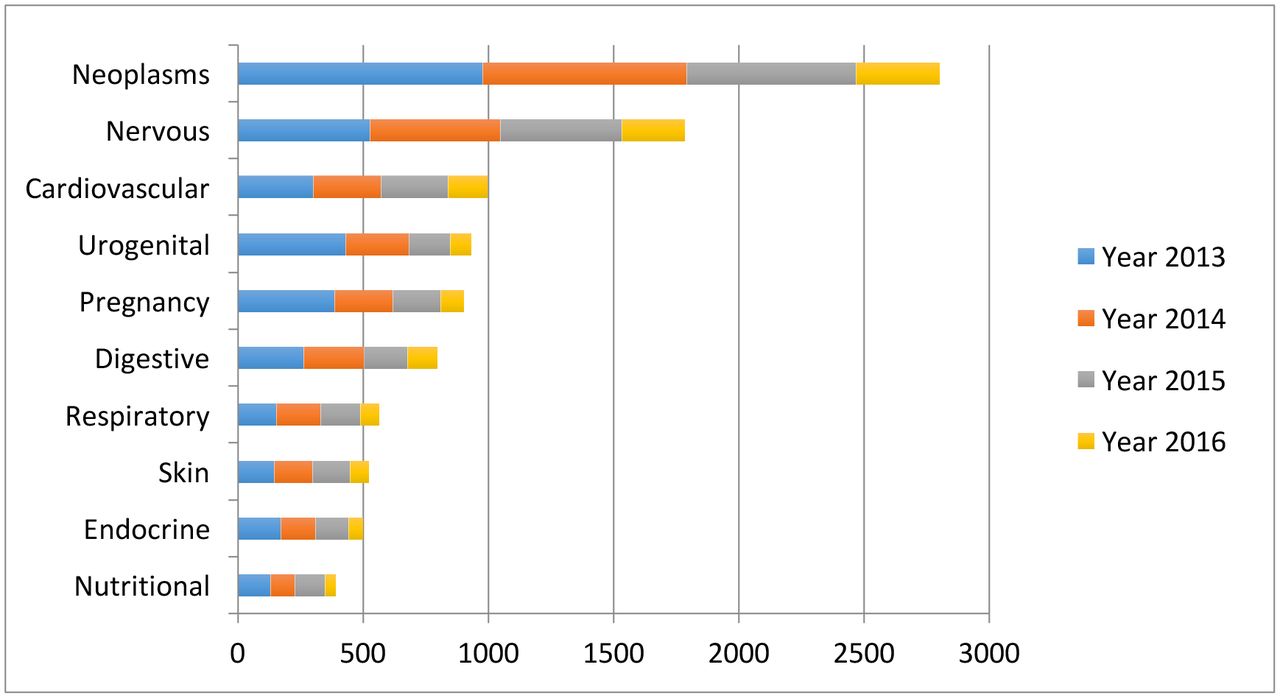
Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.

We use machine learning in day to day life almost everywhere. We use Virtual Personal Assistant, predictions while commuting, using social media, and online shopping, etc.

These days, machine learning plays a key role in many health-related fields, including the development of new medical procedures, the handling of patient data and records and the treatment of chronic diseases.

Machine learning can definitely assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare. That's the reason nowadays researchers use ML techniques in studying different diseases.

Chosen Method:

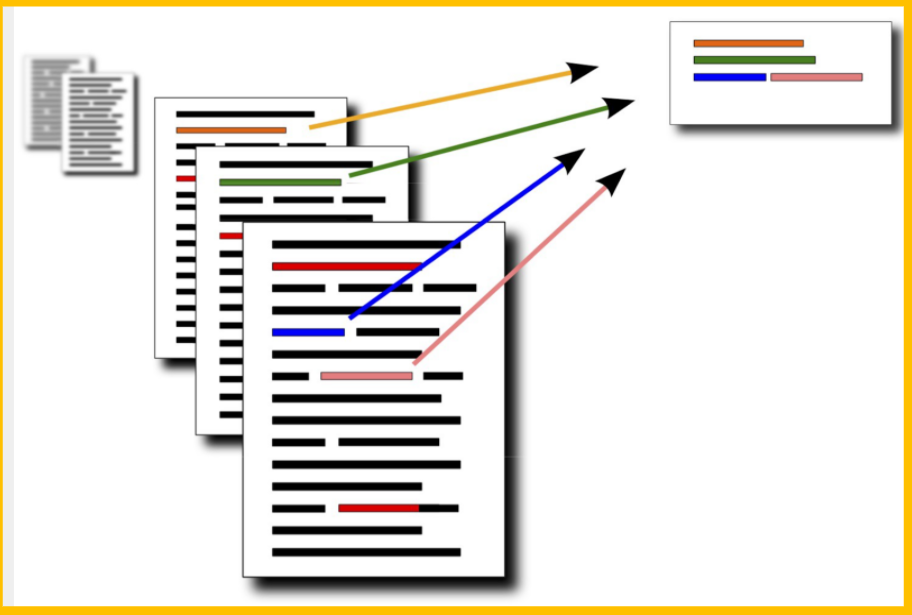


**Research about Text Summarization**

* Automatic Text Summarization gained attention as early as the 1950’s.
* A research paper, published by Hans Peter Luhn in the late 1950s, titled “The automatic creation of literature abstracts”, used features such as word frequency and phrase frequency to extract important sentences from the text for summarization purposes.
* Another important research, done by Harold P Edmundson in the late 1960’s, used methods like the presence of cue words, words used in the title appearing in the text, and the location of sentences, to extract significant sentences for text summarization.
* Since then, many important and exciting studies have been published to address the challenge of automatic text summarization.
* The objective is to save a prospective reader time and effort in finding useful information in a given article or report.
* Whereas research paper published by Harsh Desai:
* The product of this procedure still contains the most important points of the original text and is generally referred to as an abstract or a summary

**Text Summarization Approaches**

* Text Summarization is one of those applications of Natural Language Processing (NLP) which is bound to have a huge impact on our lives.
* With growing digital media and ever growing publishing – who has the time to go through entire articles / documents / books to decide whether they are useful or not?
* What is Text Summarization? Automatic summarization is the process of shortening a text document with software, to create a summary with the major points of the original document. –Wikipedia •
* Why is it Important? It can quickly extract accurate content and help readers understand large volumes of information. (reviews, news headline, biography)

****

* Identify most useful and necessary
  + information in the text.
* NLP application.
* Huge impact.

**Text summarization can broadly be divided into two categories**

* **Extractive Summarization**: These methods rely on extracting several parts, such as phrases and sentences, from a piece of text and stack them together to create a summary. Therefore, identifying the right sentences for summarization is of utmost importance in an extractive method.
* **Abstractive Summarization**: These methods use advanced NLP techniques to generate an entirely new summary. Some parts of this summary may not even appear in the original text.
* We are using Extractive Summarization in our project.



## **4. Scope**

Stage1 : Develop working medical ontology

Stage 2: Extract clinical summaries

Stage 3: Extract keywords from summaries

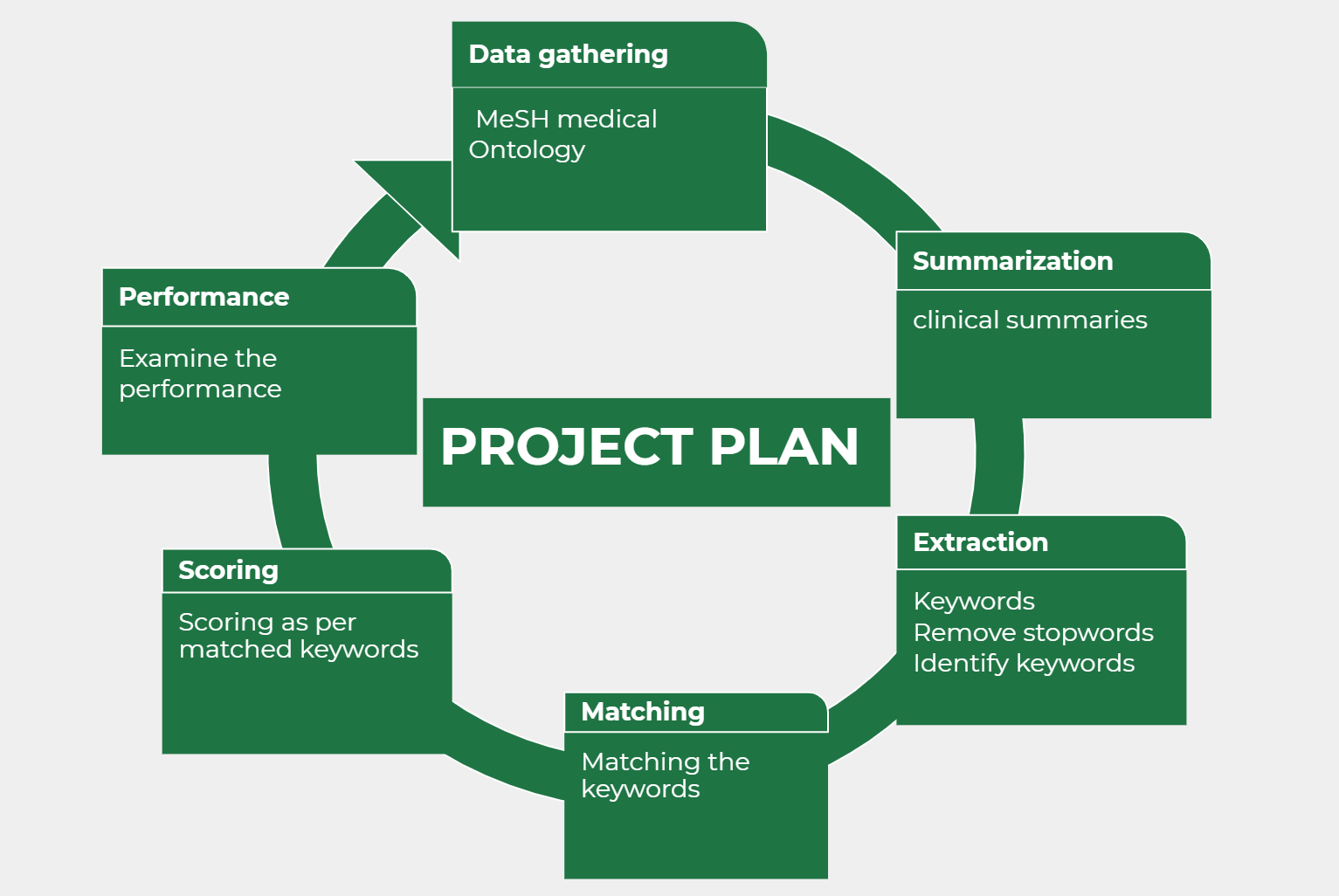
Stage 4: Lookup articles

Stage 5: Summarize articles

Stage 6: Examine performance

## **5. Methodology**

Clinical expert follows step by step process



## 

## 

## 

## 

## 

## 

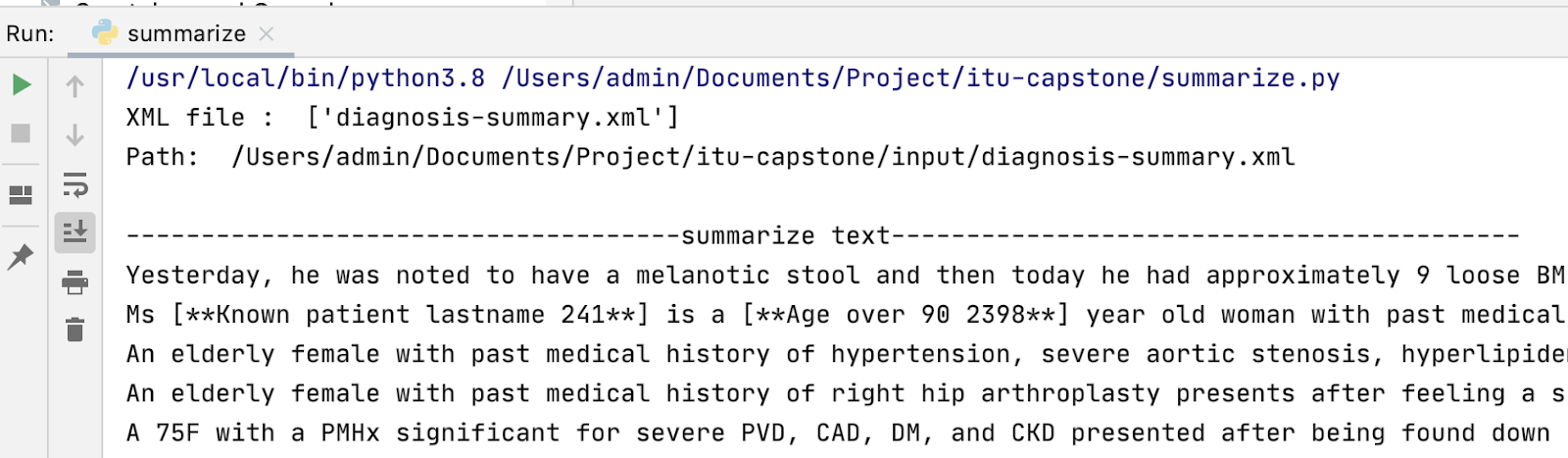
## 

## 

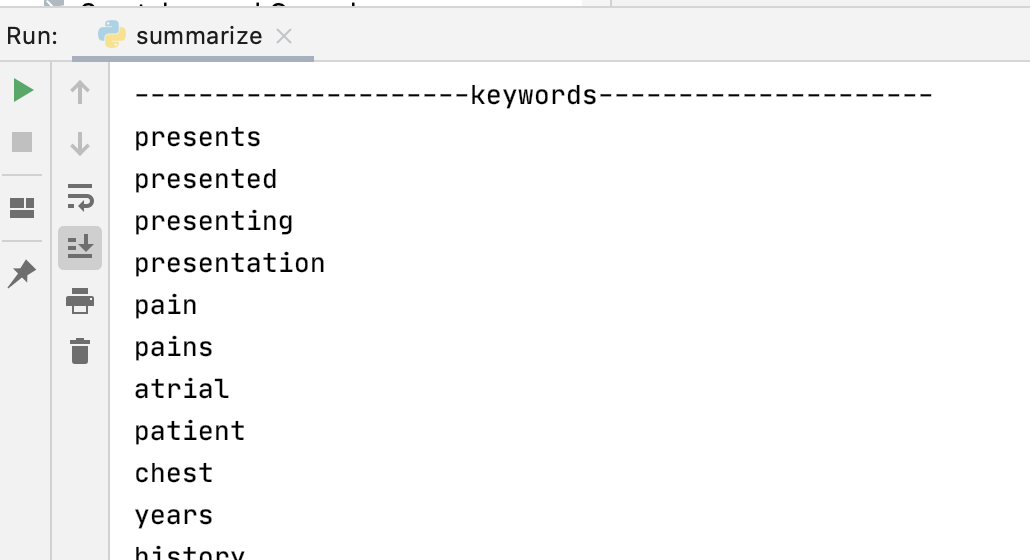
## 

## **6. How to generate/collect input data**

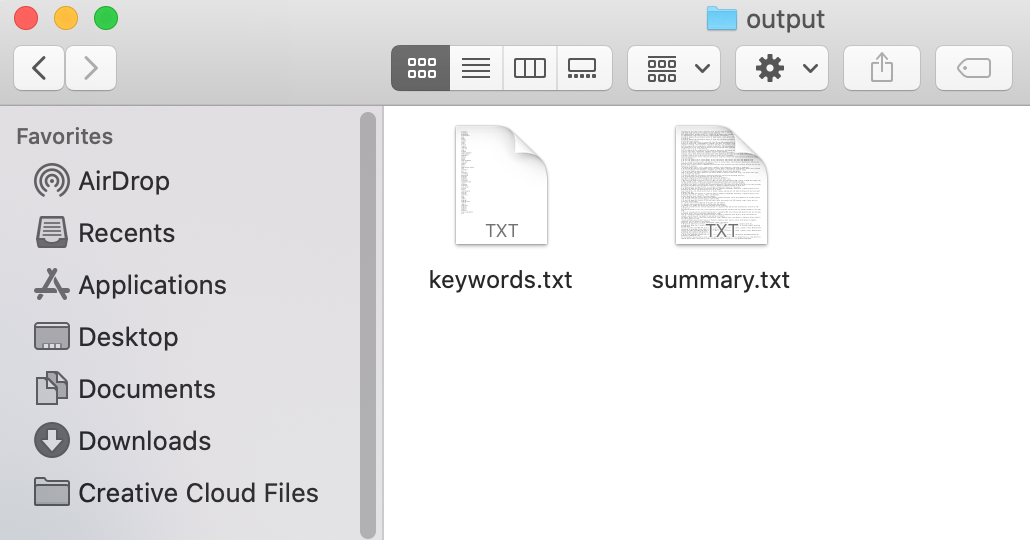
* Output of summerize.py

****

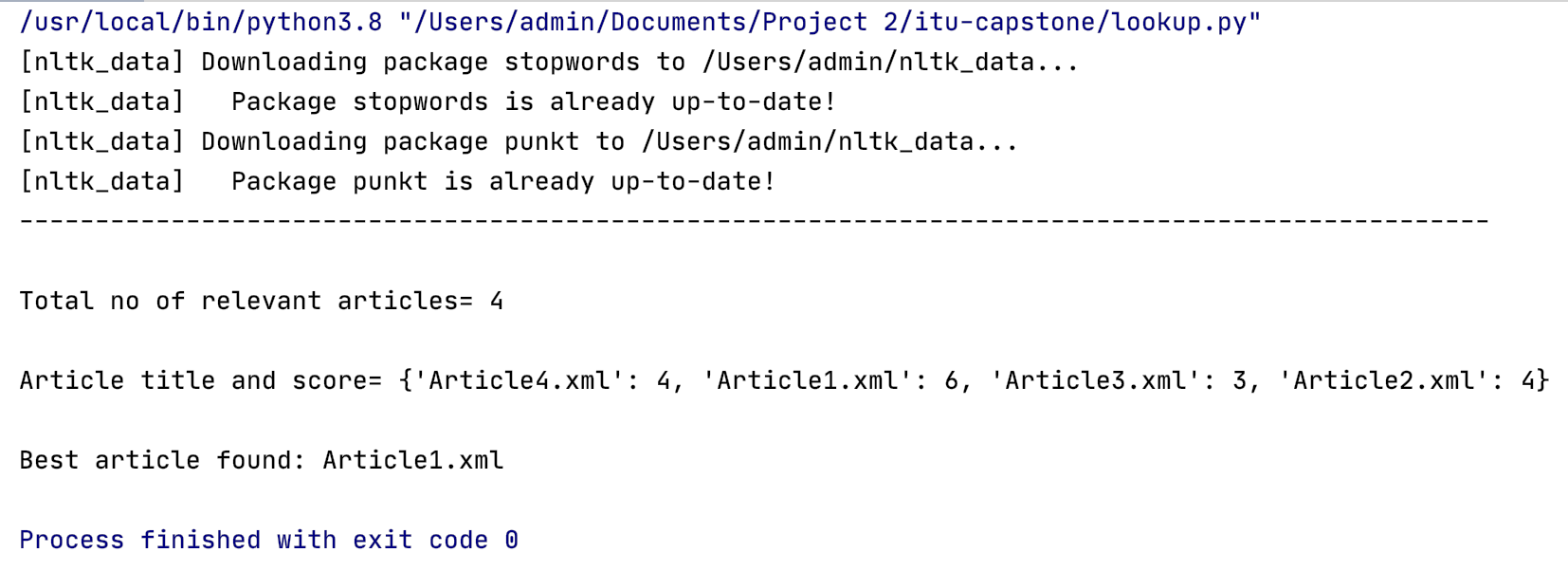
Generated keywords

****

* Generated keyword.txt and summary.txt



* Output of lookup.py

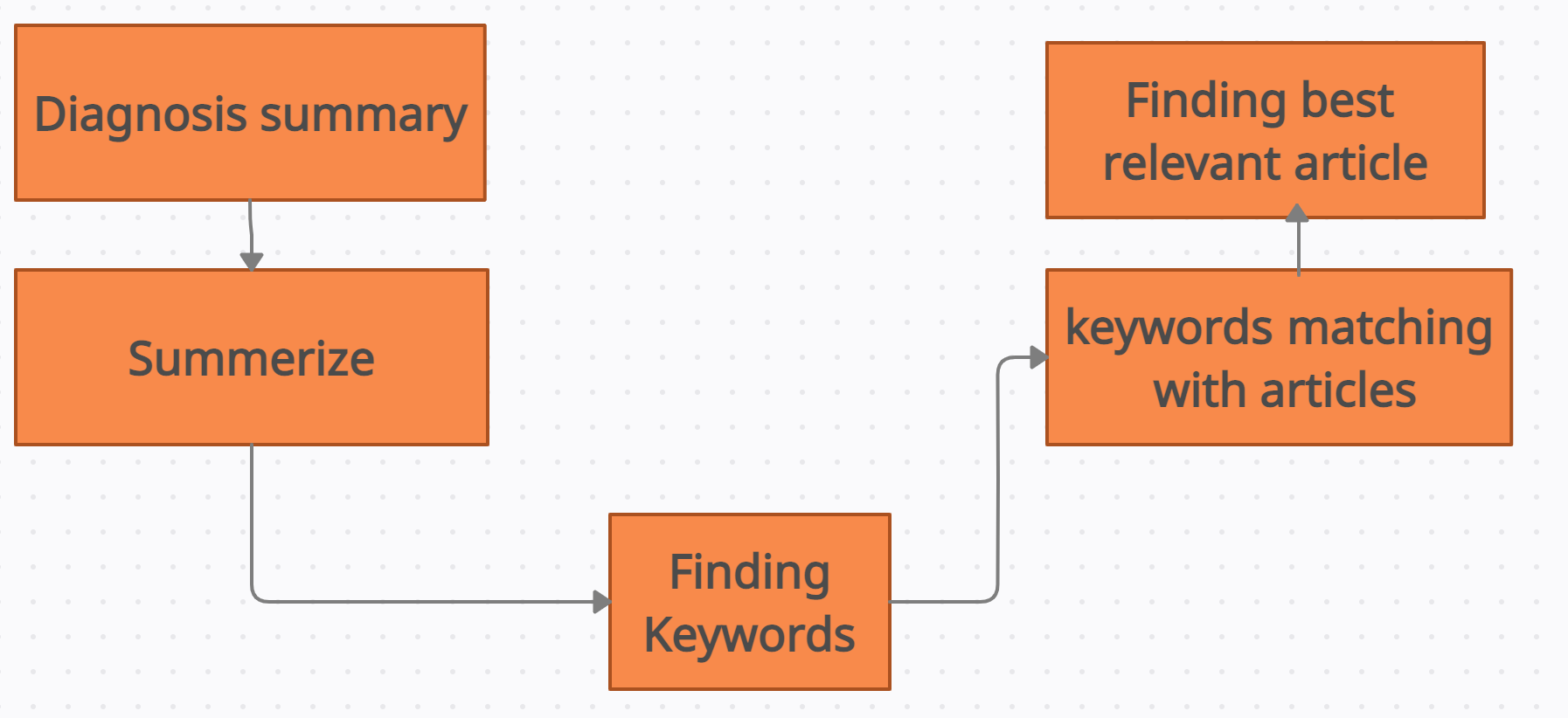


Data gathering

* We started to gather the data from MeSH(Medical subject headings) (.csv version) developed by the National Library of Medicine <https://bioportal.bioontology.org/ontologies/MESH/?p=summary>.
* The full tree version of the Mesh was very complex and does not use standard ontology tree semantics so it would be hard for us to use in our project.
* So, we used the .csv version provided by the national center for biomedical ontology.
* The files was too large about 250,000terms many were chemical compounds that had no relevance to our project.
* To increase the efficiency and accuracy of the project we filtered out the ontology down to a size of about 50,000 terms.
* The extraction of the aforementioned terms was performed in R due to its ease of use in manipulating tabular data.
* Used simple filtering commands and regular expressions in R. Sequentially filtered out unwanted entries until the ontology was the desired size.
* We focused on conditions, medications, symptoms and the like— filtered out chemical compounds, molecules, etc

## **7. How to solve the problem?**

## We have follow the below process

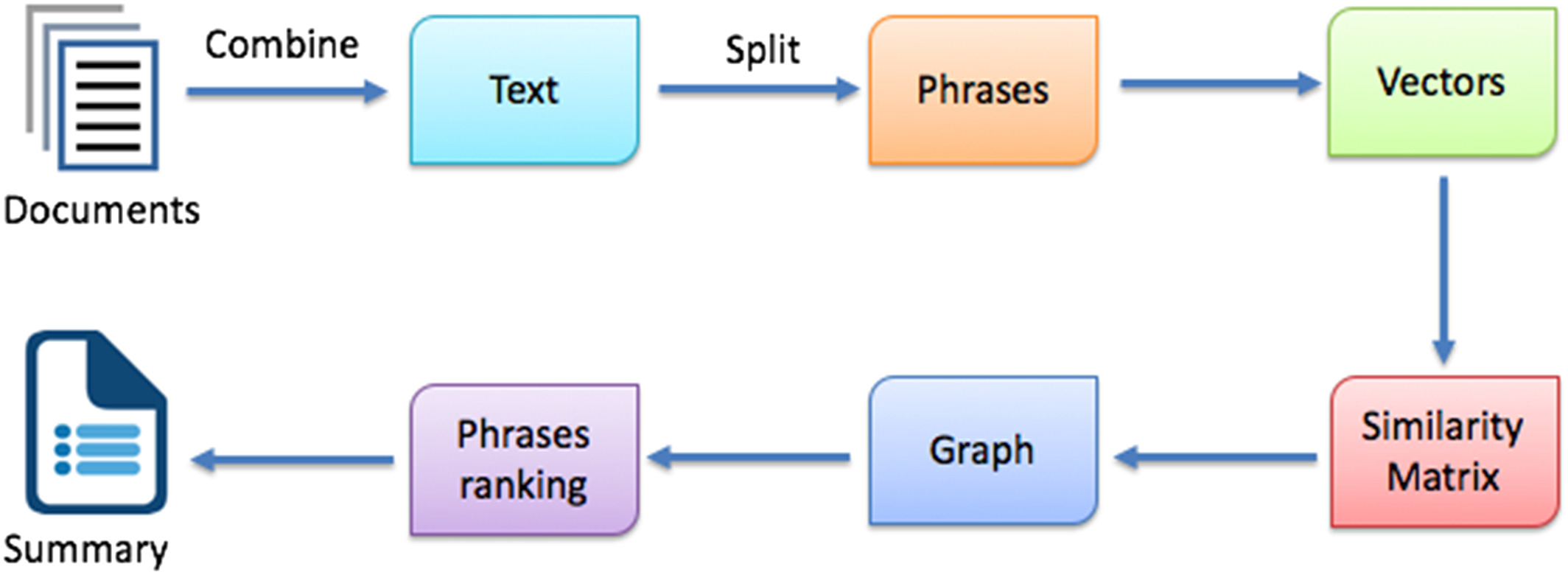


## **8. Algorithm**

**Textrank is an extractive and unsupervised text summarization technique**

* Textrank is an extractive and unsupervised text summarization technique.
* The TextRank algorithm is based on a graph-based ranking algorithm. Generally used in web searches at Google, but have many other applications.
* In the NLP case we need to define what we want to use as vertices and edges. In our case will we be using sentences as the vertices and words as the connection edges. So sentences with words that appear in many other sentences are seen as more important.
* **PageRank Algorithm -**It is the foundation of TextRank. PageRank used by Google search. Used to compute the rank of web pages.
* **Fundamentals -**  The PageRank value of a page is the probability of a user visiting that page. In TextRank, the only difference is that we consider sentences instead of pages.
* TextRank leverages NLP and graph algorithms at scale, improving means to integrate Deep Learning plus Knowledge Graphs – for enhanced machine intelligence in text handling

Let’s take a look at the flow of the Textrank algorithm that we will be following:We are using a text rank algorithm. It is a very popular graph-based ranking model used for text processing which can be used in order to find the most relevant sentences in text and also to find keywords.



* The first step would be to concatenate all the text contained in the articles
* Then split the text into individual sentences
* In the next step, we will find vector representation (word embeddings) for each and every sentence
* Similarities between sentence vectors are then calculated and stored in a matrix
* The similarity matrix is then converted into a graph, with tion
* Finally, a certain number of top-ranked sentences form the final summary.

## **9. Language and Tools used**

* Programming language used- Python
* Libraries- Gensim and nltk

**NLP**

* **Integral part of Deep learning**

We all know that It’s not an easy task teaching machines to understand how we communicate. Natural Language Processing is the branch of artificial intelligence that deals with the interaction between computers and humans using the natural language.

* **Focuses on extracting meaningful information from text**

The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. Most NLP techniques rely on machine learning to derive meaning from human languages.

For example : We have a **Capstone Presentation** today.

could I build a model that could extract these relevant pieces from any given text? Turns out the answer is yes – thanks to a concept called Information Extraction.

* **Train data models**

This rapid increase in NLP adoption has happened largely thanks to the concept of [transfer learning enabled through pretrained models](https://www.analyticsvidhya.com/blog/2017/06/transfer-learning-the-art-of-fine-tuning-a-pre-trained-model/?utm_source=blog&utm_medium=top-pretrained-models-nlp-article). Transfer learning, in the context of NLP, is essentially the ability to train a model on one dataset and then adapt that model to perform different NLP functions on a different dataset. This breakthrough has made things incredibly easy and simple for everyone, especially folks who don’t have the time or resources to build NLP models from scratch.

* **T**[**ext mining**](https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/)

Text mining (also referred to as text analytics) is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

As a technology, natural language processing has come of age over the past ten years, with products such as Siri, Alexa and Google's voice search employing NLP to understand and respond to user requests.

* **Text classification & text analysis**

Text classification is the process of assigning tags or categories to text according to its content.It’s one of the fundamental tasks in natural language processing with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection.

NLP is a field within artificial intelligence (AI) and machine learning that combines linguistics and computer science to break down language, so it can be analyzed by machines.

* **Sentiment analysis, word sequencing**

*Sentiment analysis* in NLP is about deciphering such sentiment from text. Is it *positive*, *negative, both, or neither*? If there is sentiment, *which* objects in the text the sentiment is referring to and the actual sentiment phrase such as *poor*, *blurry*, *inexpensive*, … (Not just *positive* or *negative*.) This is also called *aspect-based analysis*

* **Speech recognition & generation**

NLP works closely with speech/voice recognition and text recognition engines. While text/character recognition and speech/voice recognition allows computers to input the information, NLP allows making sense of this information.

Alibaba Cloud system has uses NLP and AliGenie voice assistant to receive customers' requests in Mandarin Chinese.

* **Machine translation and dialog systems**

Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

We know that An ontology is a formal representation of knowledge that includes the concepts in a domain and some relations between them. In NLP, ontologies can be used as a source of knowledge for machine translation systems. With access to a large knowledge base, systems can be enabled to resolve many (especially lexical) ambiguities on their own.For NLP, various libraries and frameworks are available like Gensim, Spacy, Core NLP and etc. that are available.



Depending on our use case, We needed a library that was best suitable for our purpose i.e. extraction, summarization and matching. We chose Gensim and nltk.

**Gensim**

* **Admiral & Specialized**

Gensim is not for all challenges, but what it does do, it does them well. We don’t send your admiral to a land battle, and we don’t use gensim for general NLP. Gensim is a well-optimized library for topic modeling and document similarity analysis. Among the various NLP libraries , it’s the most specialized one.

* **Provides Vectorization**

From a mathematical perspective, a vector is a geometric object which has magnitude and direction. We don't need to pay as much attention to this, and rather think of vectors as a way of projecting words onto a mathematical space while preserving the information provided by these words.

* **Platform Independent**

Gensim runs on Linux, Windows and OS X, as well as any other platform that supports Python and NumPy.

* **Open source**

Gensim is licensed under the OSI-approved GNU LGPL license which allows it to be used for both personal as well as commercial use for free. Any modifications made in Gensim are in turn open-sourced and have abundance of community support too.

* **Scalable**

Gensim can easily process large and web-scale corpora by using its incremental online training algorithms. It is scalable in nature, as there is no need for the whole input corpus to reside fully in Random Access Memory (RAM) at any one time. In other words, all its algorithms are memory-independent with respect to the corpus size.

* **Extracts automatically**

We use the summarization.summarizer from gensim.This summarising is based on ranks of text sentences using a variation of the TextRank algorithm.

* **Robust & Efficient**

Gensim is robust in nature and has been in use in various systems by various people as well as organisations for over 4 years. We can easily plug in our own input corpus or data stream. It is also very easy to extend with other Vector Space Algorithms.

**Nltk**

The Natural Language Toolkit (NLTK) is a platform used for building programs for text analysis.It is free, opensource, easy to use, large community, and well documented. It is super fast and is extensible.

* **Full NLP library**

NLTK consists of the most common algorithms such as tokenizing, part-of-speech tagging, stemming, sentiment analysis, topic segmentation, and named entity recognition. NLTK helps the computer to analyze, preprocess, and understand the written text.

* **Stopwords**

Stopwords considered as noise in the text. Text may contain stop words such as is, am, are, this, a, an, the, etc.

* **Fast sentence tokenization**

Nltk provides fast Tokenization. The process of breaking down a text paragraph into smaller chunks such as words or sentences is really fast.

* **Supports the largest no. of languages**

Languages supported by NLTK depends on the task being implemented. For stemming, we have RSLPStemmer (Portuguese), ISRIStemmer (Arabic), and SnowballStemmer (Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish). For sentence tokenization, PunktSentenceTokenizer is capable of multilingual processing. Stopwords are also available in multiple languages.

**Packages**

* **re package**

Checks if a particular string matches with regular expression. We have used re to remove tags, apostrophe, and words that have special characters in them.

* **os package**

To provide a portable way of using operating system dependent functionality.

* **Glob package**

To find all path names matching to a specified pattern.

* **Sys package**

The python sys module provides functions and variables which are used to manipulate different parts of the Python Runtime Environment

* **Csv package**

The so-called CSV (Comma Separated Values) format is the most common import and export format for spreadsheets and databases.

**Functions**

* **summarize()**

Returns a summarized version of the given text using a variation of the TextRank algorithm

* **keywords()**

Returns the keywords

* **Word\_tokenize()**

Return a tokenized copy of *text*

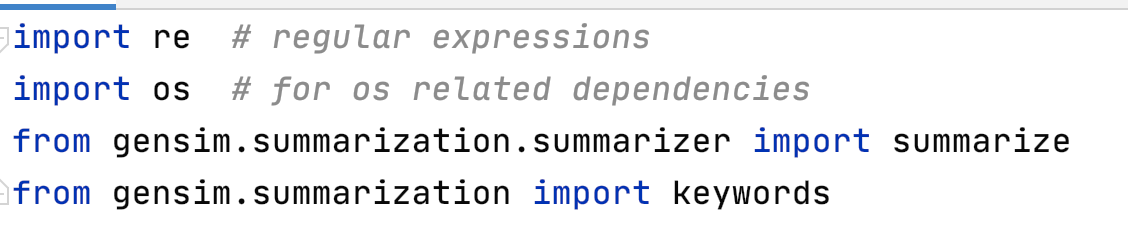
* **Punkt**

This tokenizer divides a text into a list of sentences, by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences.

## **10. How to generate output**

## Step by step implementation is as follows

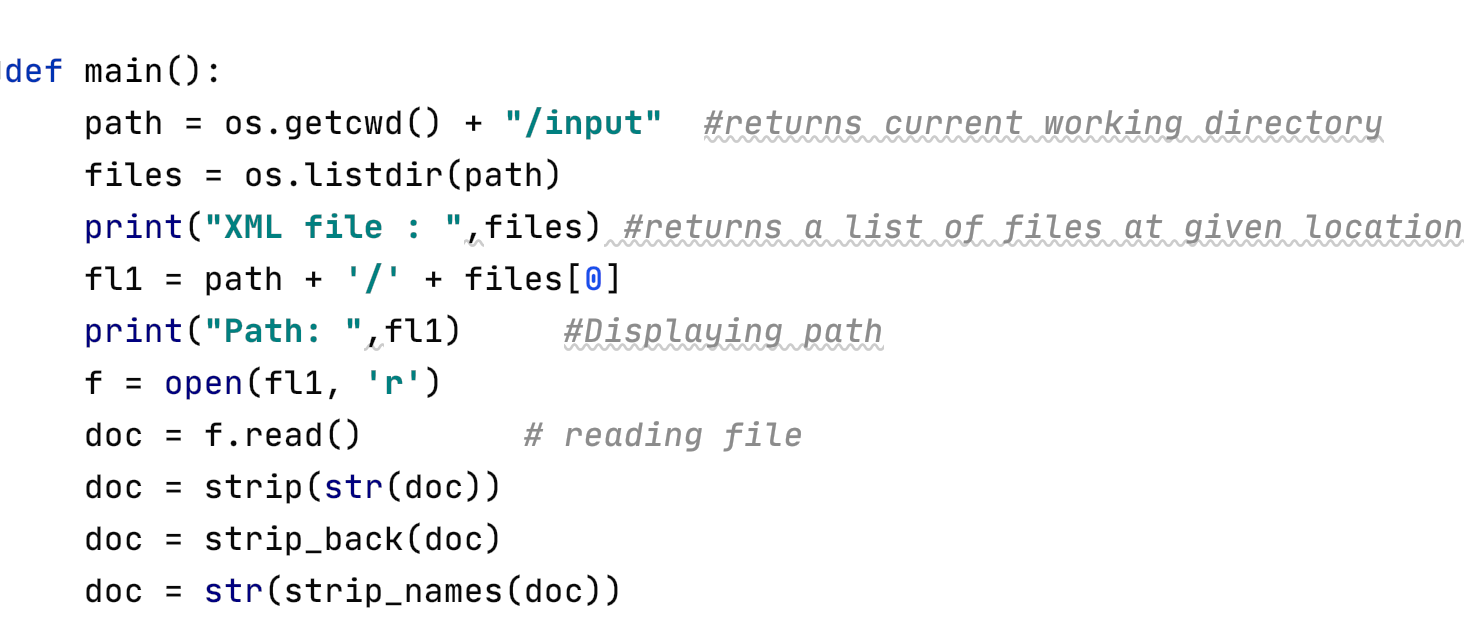
1. Importing packages



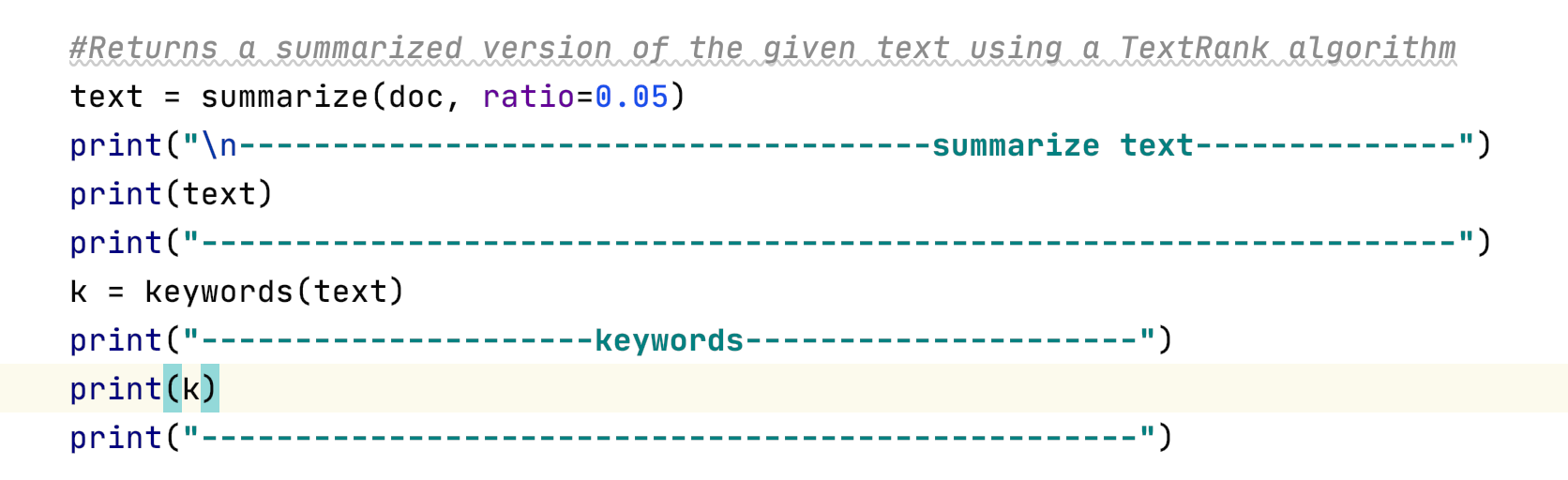
1. Using regular expressions to find the keywords



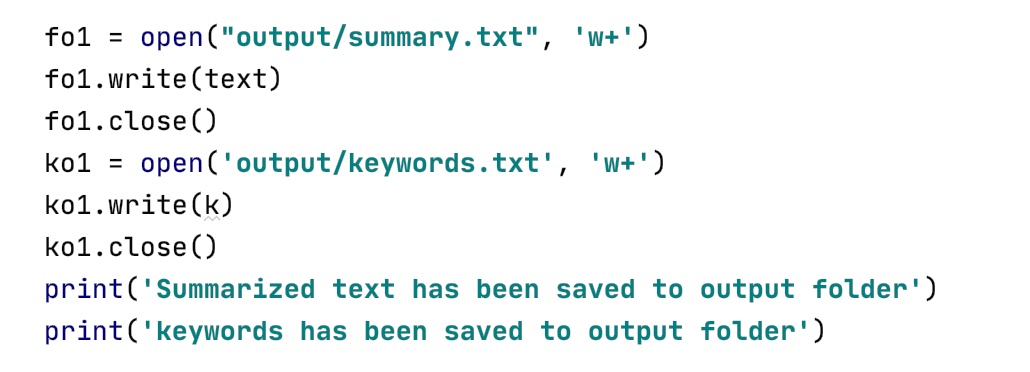
1. Working with Diagnosis summary of patient (XML file)



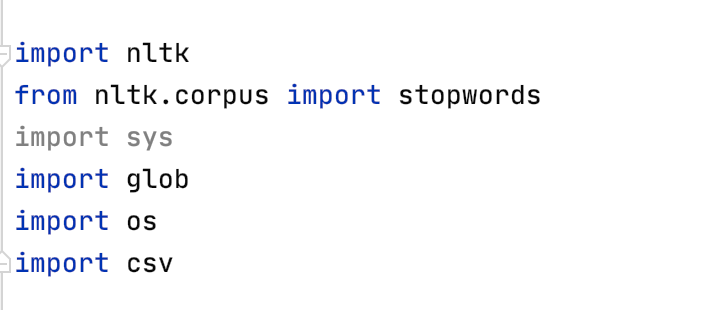
1. Text summarization



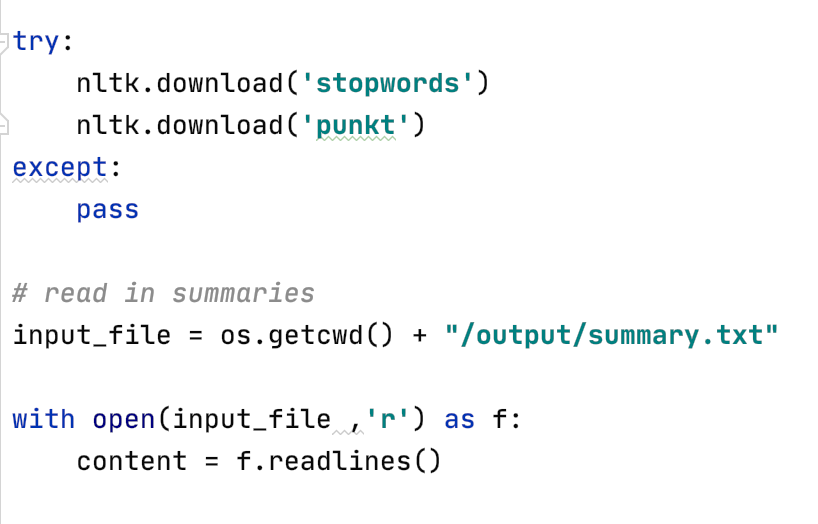
1. Writing into the files-



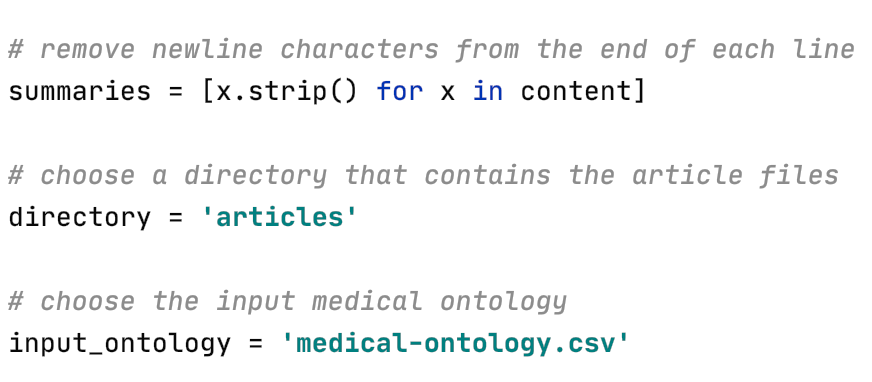
1. Lookup.py



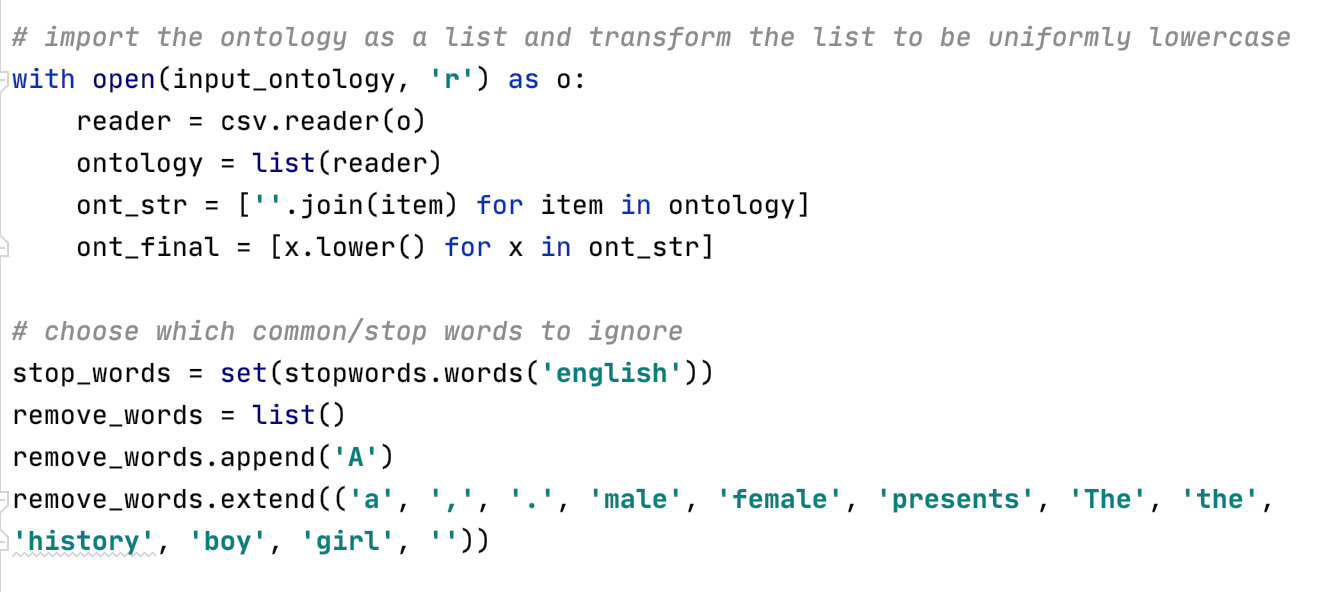
1. Reading summary



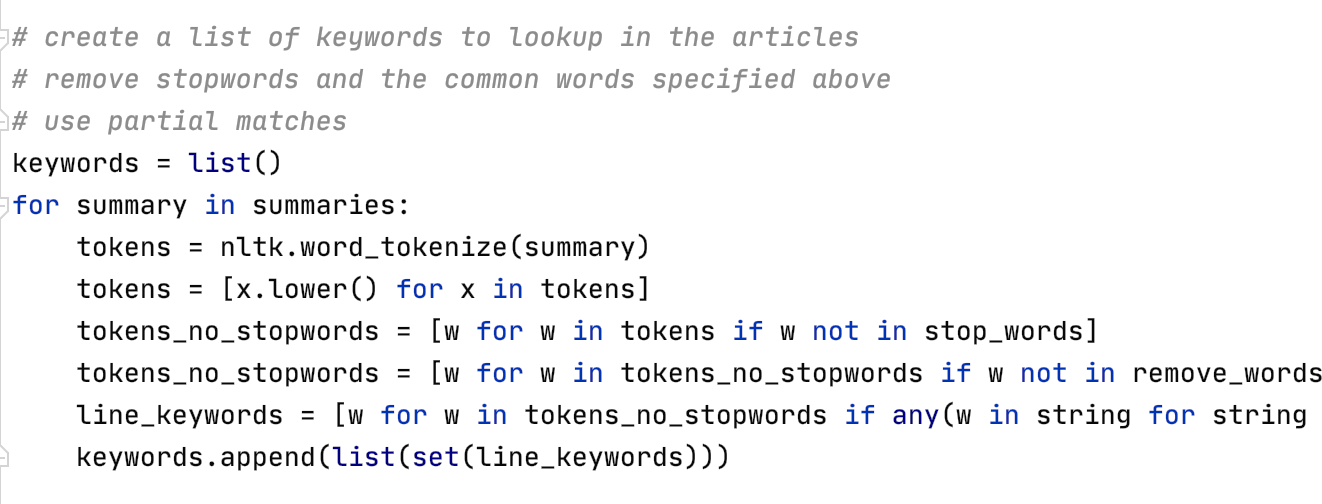
1. Taking input data



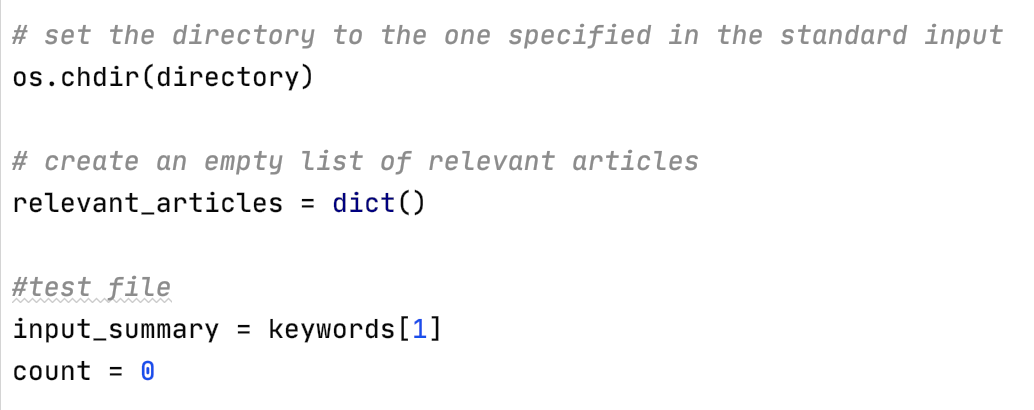
1. Working on input data



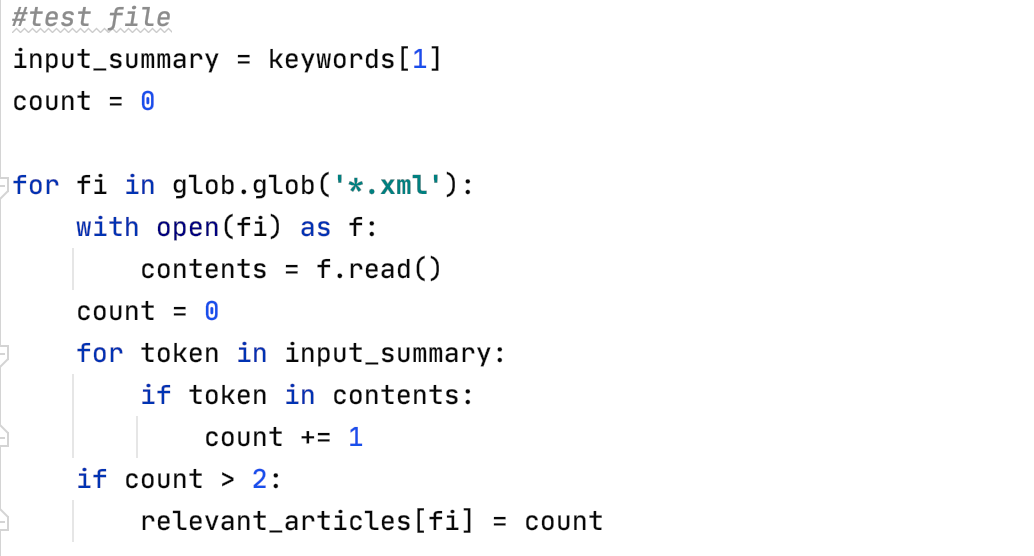
1. Remove stopwords and common words

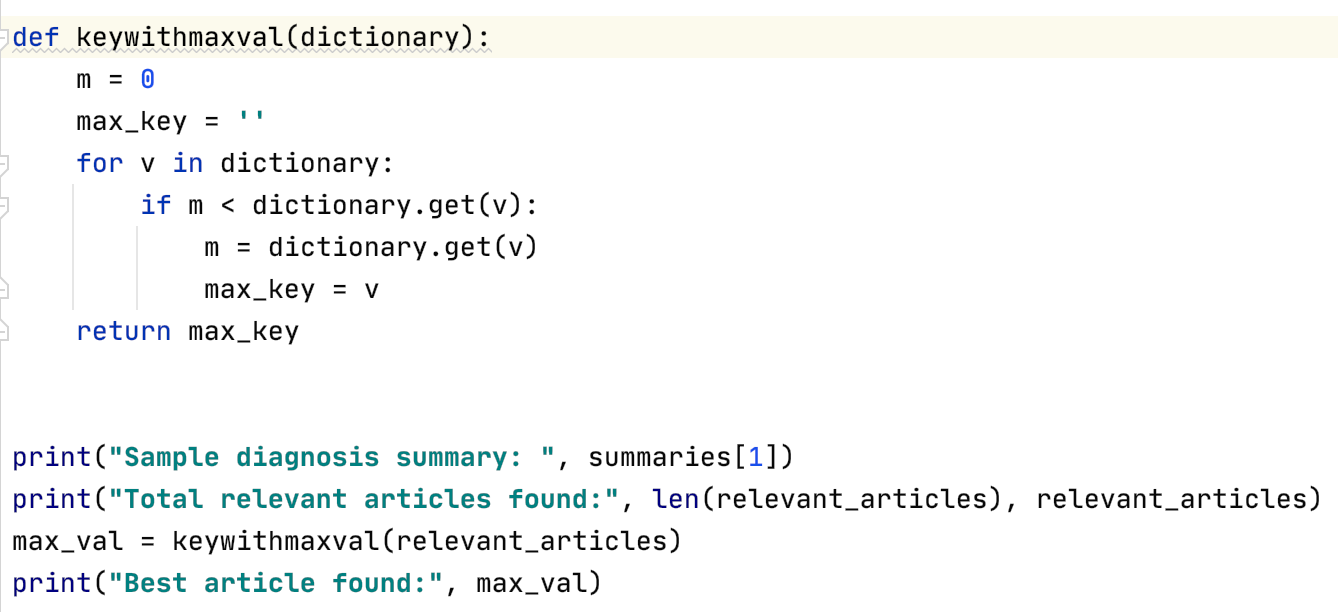


1. Creating directory for relevant articles



**11. How to test against the hypothesis**

1. 
2. Finding most relevant article and display



**12. Implementation**

Source code is at the bottom of the report

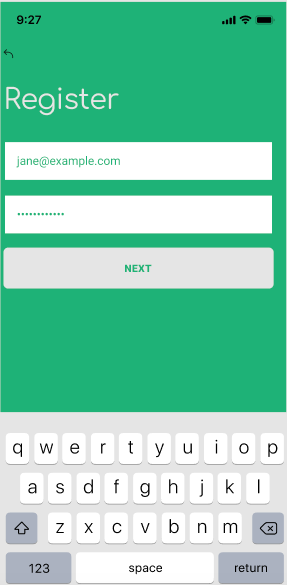
**13. Mockup screens**

In future we are going to create an App for doctors which will help doctors to find the best relevant article among a bunch of articles available over the internet. Below are mockup screens

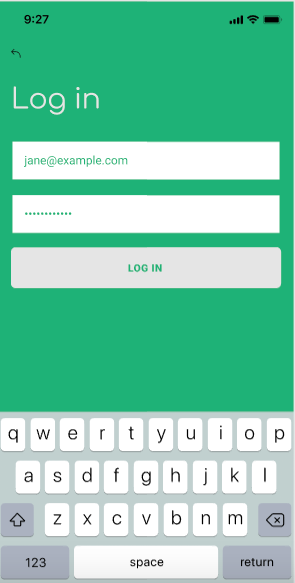
* Home page



* Register page



* Login Page



* Result Page



## **14.Conclusion and recommendation**

* The project tool will provide an efficient way of decision-making on the emergency status of the patient to explain the factors influencing the emergency specialists’ decision-making regarding the emergency status of patients’ conditions. The findings of this study led to the emergence of a major research objective and two main classes. The objective of the study was representative of the main focus on how emergency decision-makers decided on the emergency conditions of the patients. In this study, factors such as clinical conditions, clinical experiences, diagnostic procedures, pre arrival conditions to the emergency department, work considerations, and patients and their families’ approaches were shaped around the axial factor, and it was reported what parameters are taken into consideration by an emergency specialist at the time of the decision-making about the emergency conditions of patients. . Not only surgeons but also any emergency specialists, medical professionals, members of the medical team for quicker and better results. Doctors may consider different approaches to make decisions about the patients’ emergency conditions based on their level of experience and responsibility; however, our main focus in the project is to understand the acute health threats of patients which would help to make critical decisions for patients in an unstable and stressful situation. Therefore, their decisions may be regarded as pessimistic from the perspective of other members of the medical team.
* In future we will try additional summarization approaches & compare performance​

## **Drawback**

* Does not support multithreading.
* Can be slow if in the worst case the size of the graph is quadratic.
* Currently only works for English datasets.

**How could we improve performance?**

* Stemming/lemmatization (problem: we don’t have a dictionary)
* Implement a machine learning algorithms (problem: most require gold/labeled data, can be time-consuming)

**Future Work**

* Improve matching: perhaps try both matching based on unique words and total word matches and compare lists
* Use of synonyms in the ontology
* Match part of keyword.
* Try additional summarization approaches & compare performance.
* Provides an efficient way of decision-making on emergency status.
* Not only surgeons but also any emergency specialists, medical professionals, medical team members
* Quicker and better results.
* In future we will try additional summarization approaches & compare performance​

## 

## 

## **15. Bibliography**

* B. Sun, W. Ma, and H. Zhao, “An approach to emergency decision making based on decision-theoretic rough set over two universes,” Soft Computing, vol. 19, no. 5, pp. 1–12, 2015.
* A. Ebrahimian, N. Khalesi, G. Mohamadi, M. Tordeh, and M. Naghipour, “Transportation management in pre-hospital emergency with physiological early warning scores,” Journal of Health Administration, vol. 15, no. 49, pp. 7–13, 2012.
* J. N. Fullerton, C. L. Price, N. E. Silvey, S. J. Brace, and G. D. Perkins, “Is the Modified Early Warning Score (MEWS) superior to clinician judgement in detecting critical illness in the pre-hospital environment?” Resuscitation, vol. 83, no. 5, pp. 557–562, 2012.
* S. A. Mulholland, B. J. Gabbe, and P. Cameron, “Is paramedic judgement useful in prehospital trauma triage?” Injury, vol. 36, no. 11, pp. 1298–1305, 2005.
* <https://www.programiz.com/python-programming/regex>
* <https://www.debuggex.com/cheatsheet/regex/python>
* <https://svn.bmj.com/content/2/4/230>