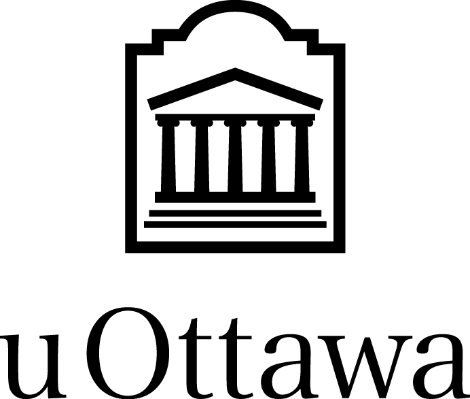
|  |
| --- |
| E-commerce Intelligent bot |
| Final Project |
|  |



**Mohamed Fouad Abdelnaby**

**Sameh Mohamed Ismail**

**Mohamed Helmy Mostafa**

**Mostafa Mohamed Mahmoud Elseidy**

12/11/2021

Contents

[Problem Formulation: 2](#_Toc90204063)

[Goal 2](#_Toc90204064)

[Project Pipeline 3](#_Toc90204065)

[Data Collection 4](#_Toc90204066)

[Dataset Features 5](#_Toc90204067)

[Exploratory Data analysis 6](#_Toc90204068)

[Text Transformation and Feature Engineering 7](#_Toc90204069)

[Classification Algorithms: 7](#_Toc90204070)

[Clustering Algorithms 9](#_Toc90204071)

[Evaluation 11](#_Toc90204072)

[Classification Evaluation 11](#_Toc90204073)

[Clustering Evaluation 11](#_Toc90204074)

[Error Analysis 11](#_Toc90204075)

[Deployment 12](#_Toc90204076)

[References 12](#_Toc90204077)

# Problem Formulation:

There is a need for a chatbot app for smartphones that allows users to have online conversation. Based on the information provided by the user, the chatbot can recommend which medications to take. The development of the application has the potential to improve consumer-pharmacy communication while also assisting the pharmacy in developing a better customer management system.

# Goal:

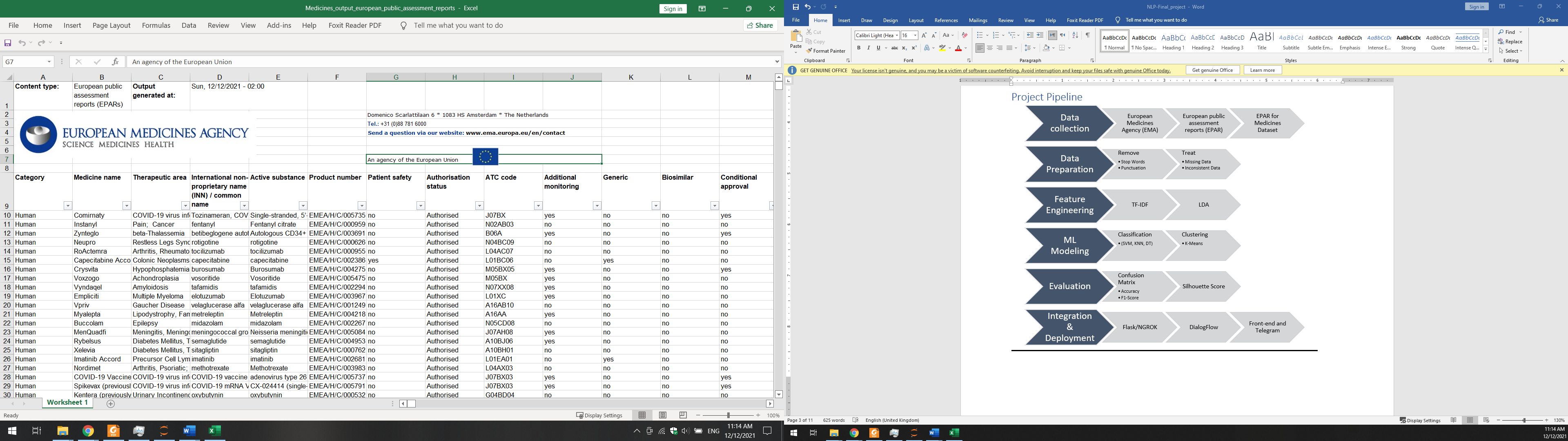
A chatbot system that not only can understand and analyze the clients’ symptoms but also diagnose the disease as well as recommend the suitable drugs based on machine learning algorithms and Natural Language Processing techniques.

# Project Pipeline

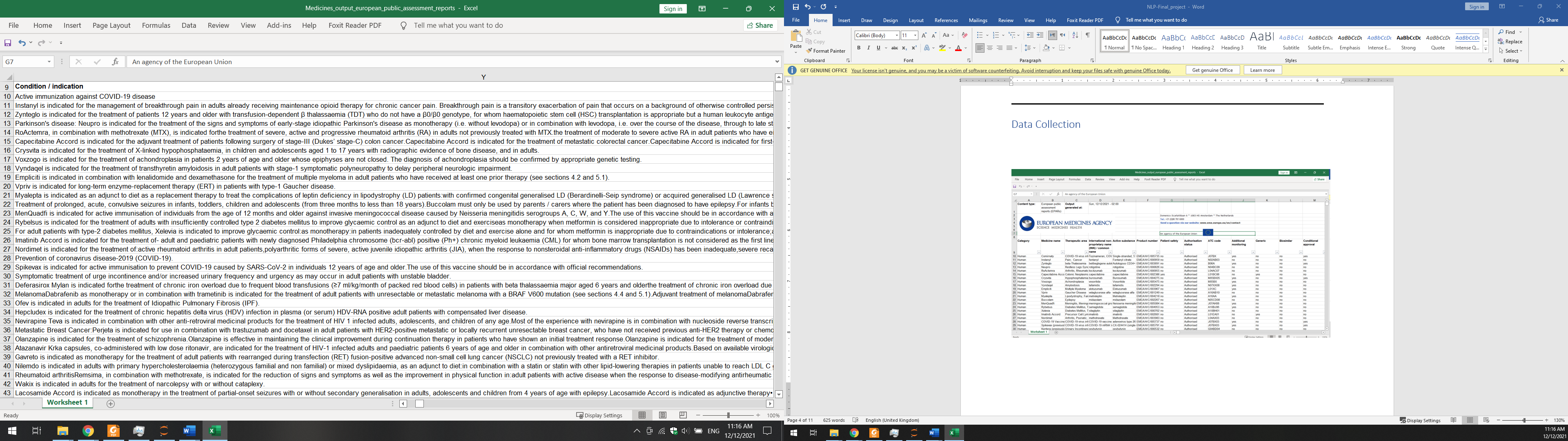
# Data Collection

We use EPAR for medicines open-source dataset and (EPAR) European public assessment medical reports from European Medicines Agency (EMA) to build our system.

Sample of the dataset



Sample of the description or indication



The “Condition / indication” feature will be used in NLP steps.

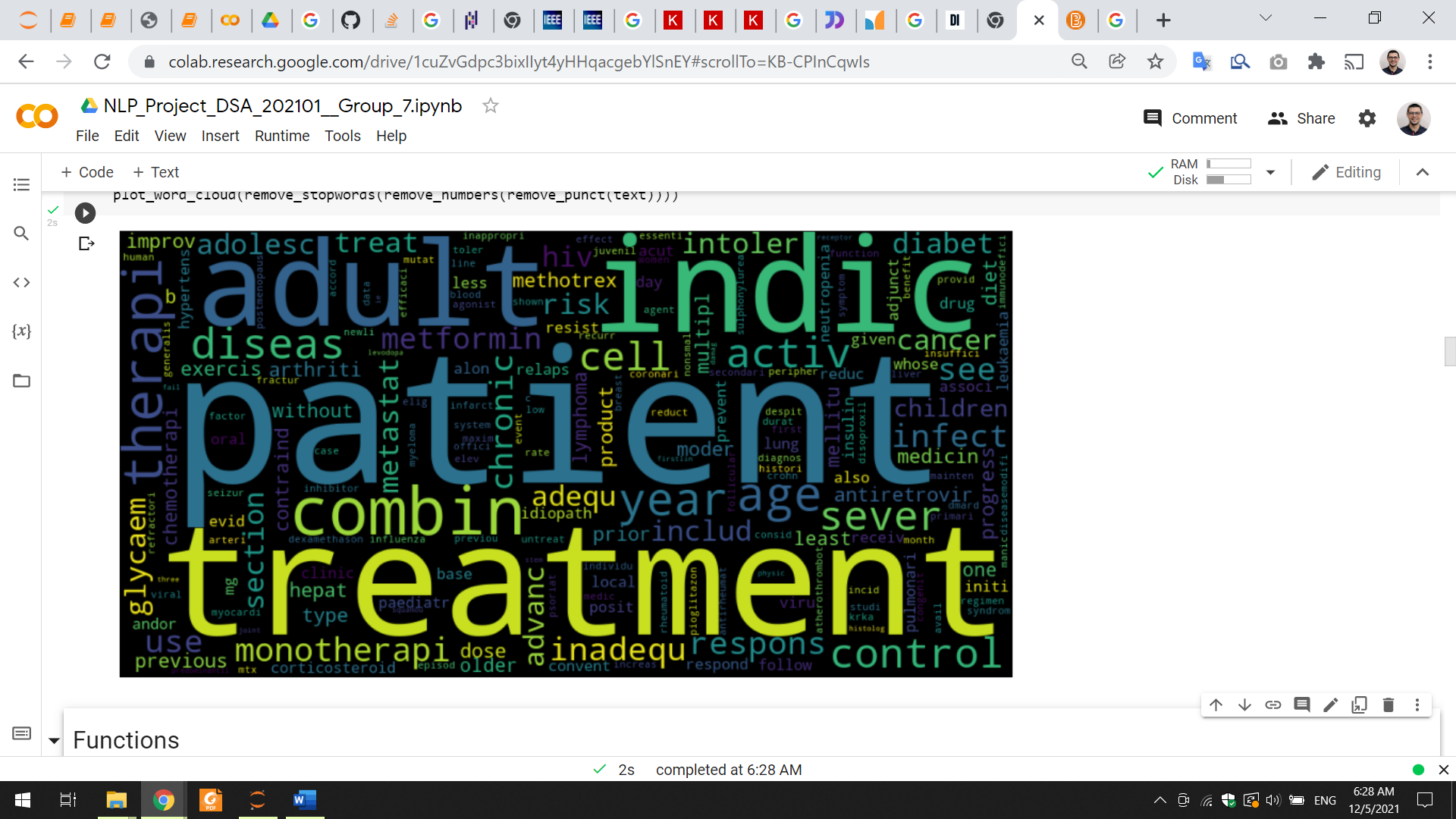
## Dataset Features

Description of Some features:

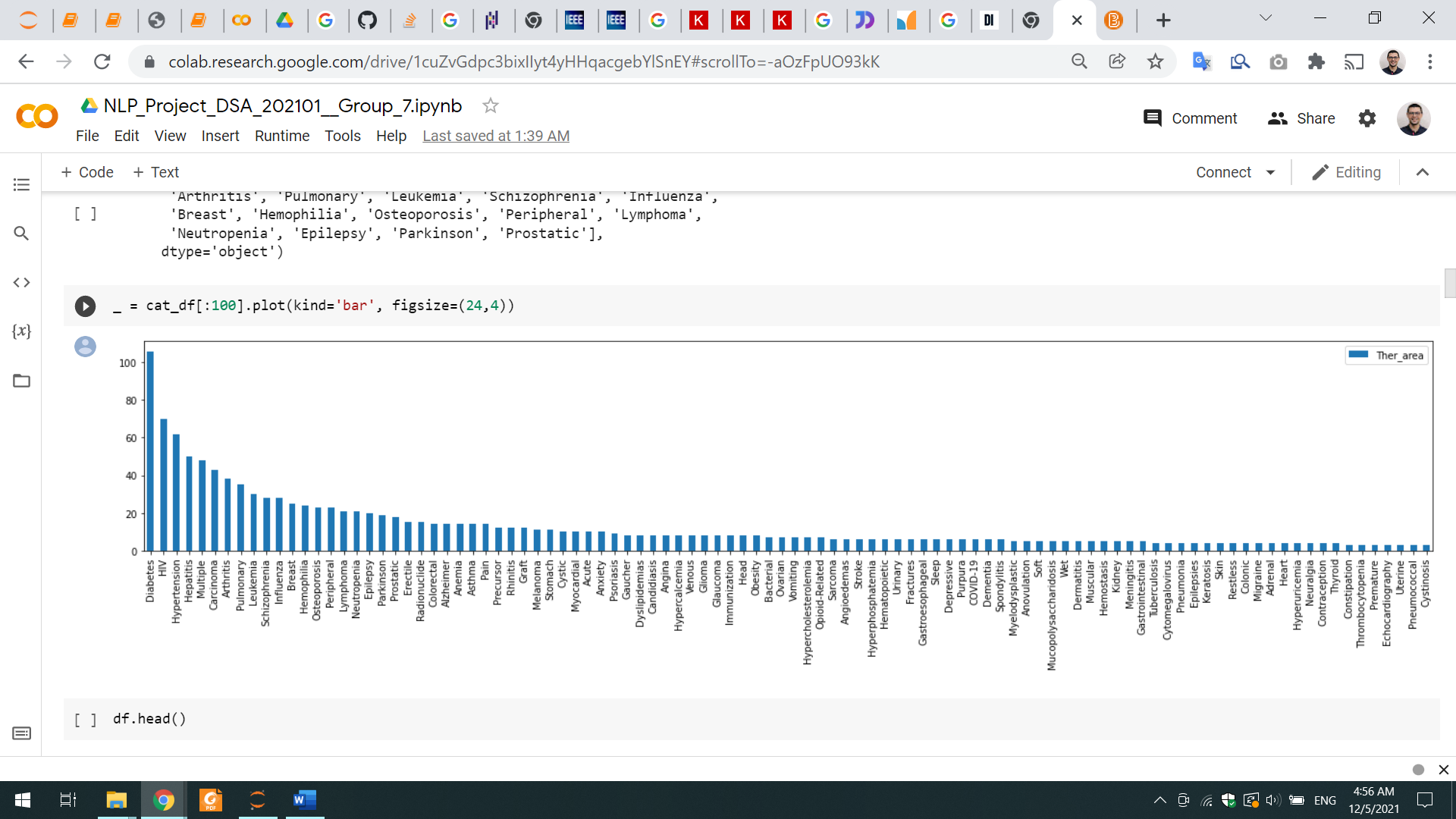
* Category: Human or Veterinary
* Medicine name: about 600 unique drugs
* Therapeutic area: disease like Anaemia, Diabetes, Bacterial Infections and more
* International non-proprietary name (INN) / common name
* Patient safety: a binary feature with yes or no
* Orphan medicine: a binary feature with yes or no
* Condition / indication: text descriptions of drugs and diseases
* URL: for each drug

We will use “Condition / indication” feature for NLP stages and “Medicine name”, “Therapeutic area” and other features for fulfilment.

# Exploratory Data analysis



We apply a word cloud EDA on the “Condition / indication” feature and we got most frequent words as shown in the figure.



This bar chart shows the frequencies of each disease with repect to number of drugs. We will use the most 20 diseases and their drugs in future steps.

After applying some filter and explore most feature manually, we found that this data set has many blank or missing values and some in consistent values. We will deal with these issues in the data preparation step.

# Text Transformation and Feature Engineering

We implement 2 approaches of text transformation:

* TF-IDF:

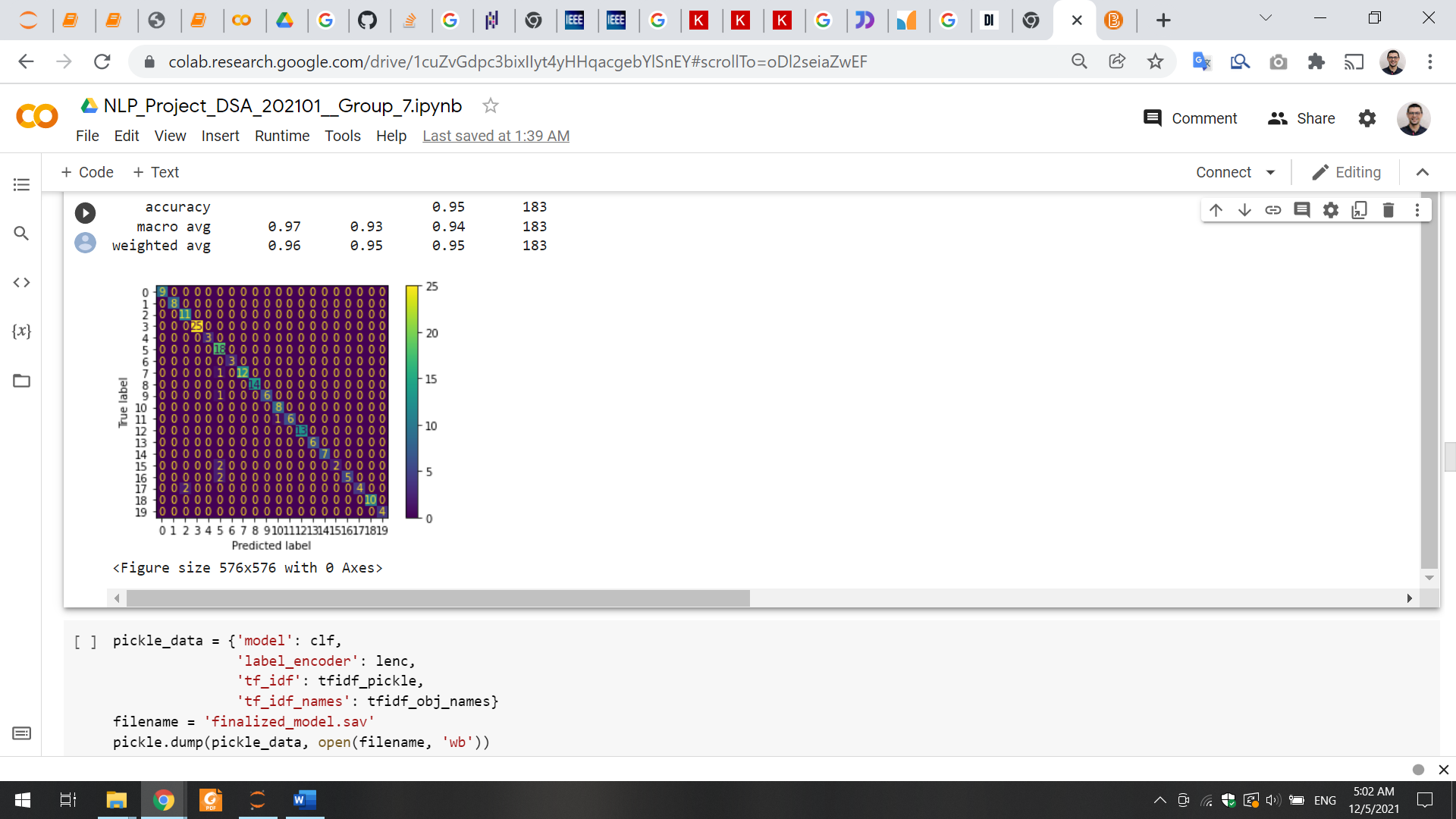
Bag of word doesn't capture the importance of the word it gives you the frequency of the word. TF-IDF resolves this matter through computation of two values. TF is count of occurrences of the word in a document. IDF of the word across a set of documents. It tells us how common or rare a word is in the entire document set. The closer it is to 0, the more common is the word. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm. We then multiply these two values TF and IDF. We used sklearn tf-idf-transformer which transforms a count matrix to a normalized tf or tf-idf representation.

* LDA:

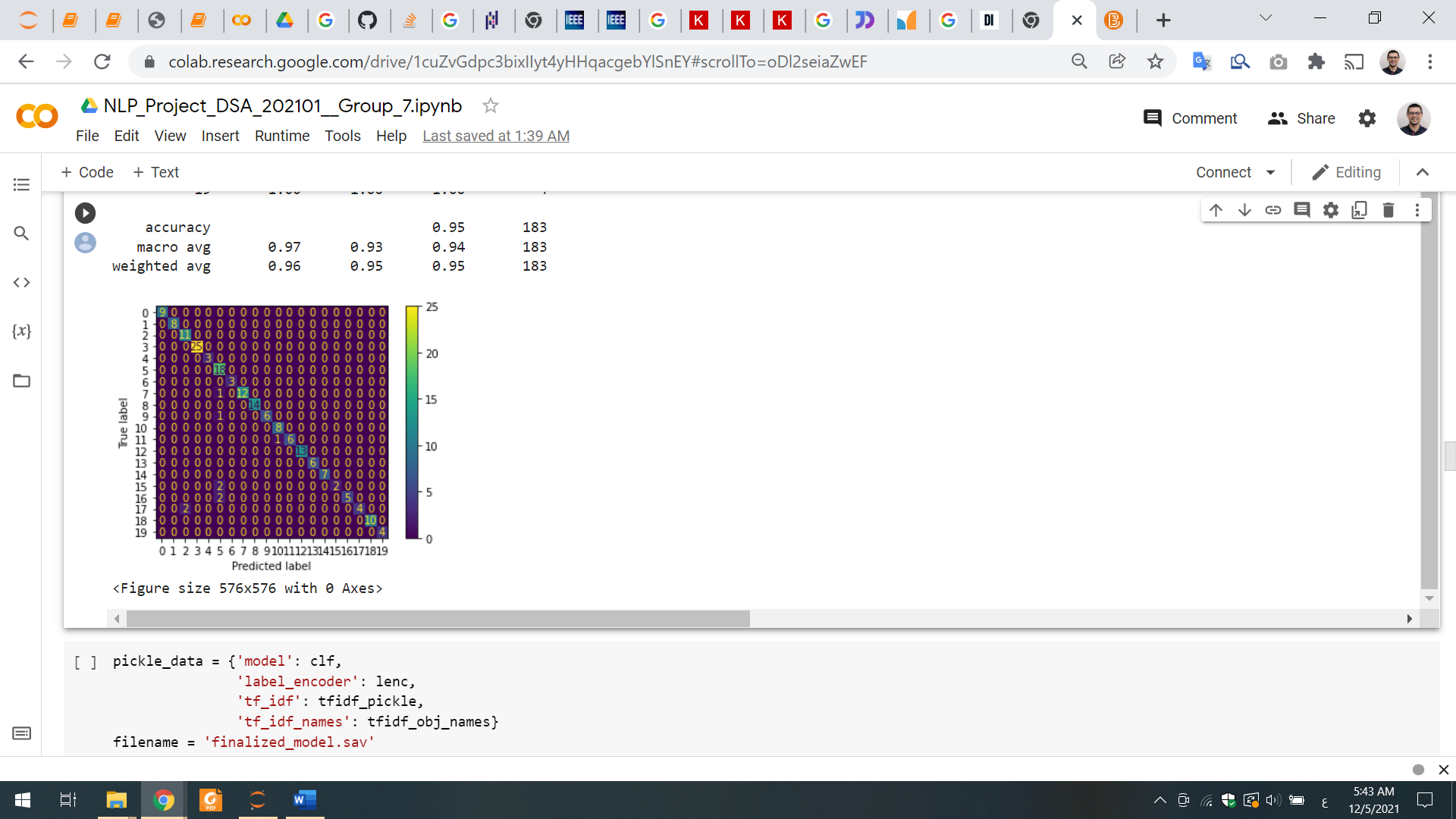
Latent Dirichlet Allocation (LDA) algorithm is an unsupervised learning algorithm that attempts to describe a set of observations as a mixture of distinct categories. LDA is most commonly used to discover a user-specified number of topics shared by documents within a text corpus. Here each observation is a document, the features are the presence (or occurrence count) of each word, and the categories are the topics. Since the method is unsupervised, the topics are not specified up front, and are not guaranteed to align with how a human may naturally categorize documents. The topics are learned as a probability distribution over the words that occur in each document. Each document, in turn, is described as a mixture of topics.

# Classification Algorithms:

**SVM**



**KNN**

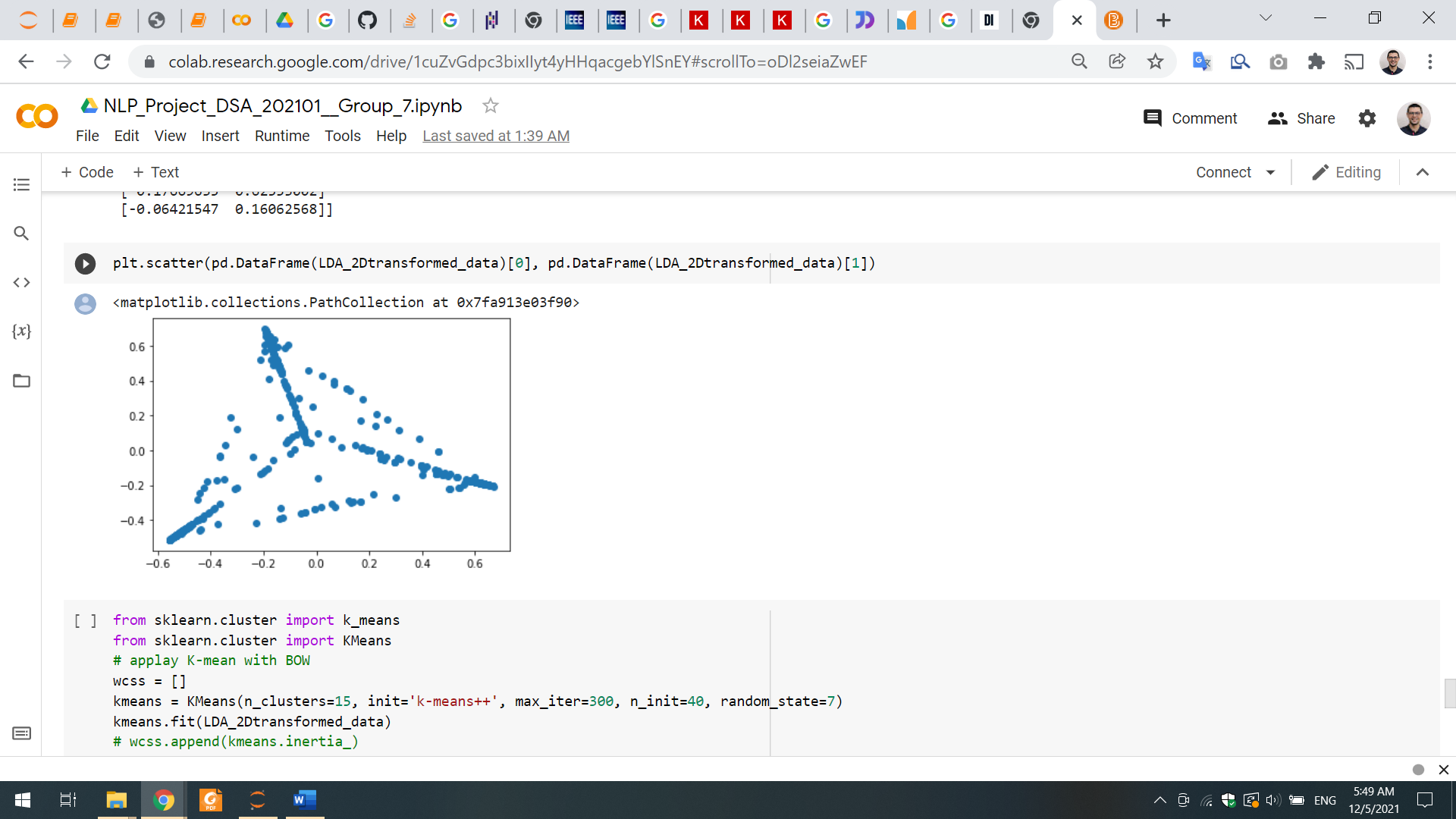


**DT**

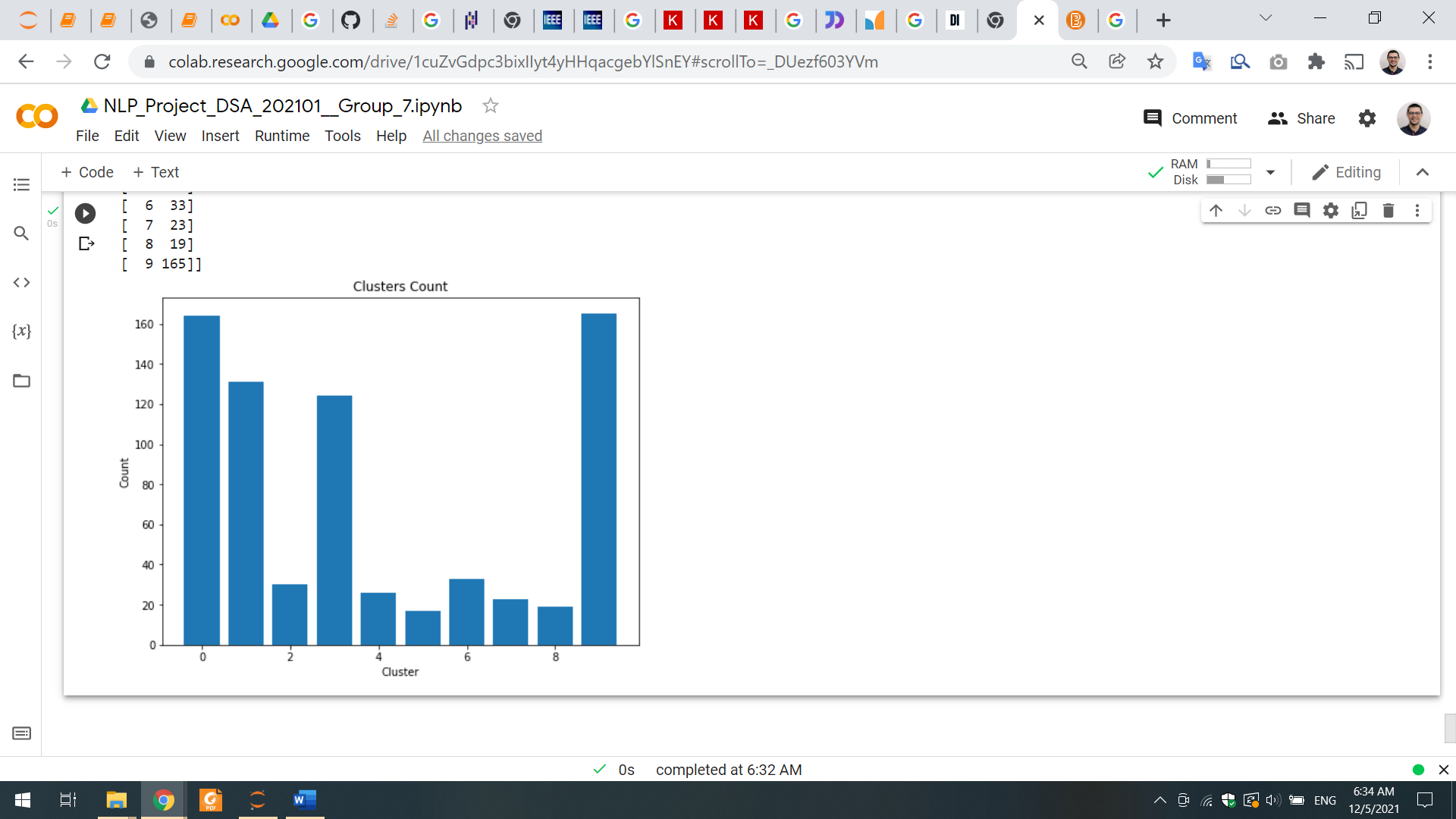


# Clustering Algorithms

**K-Means + LDA**







We implement 3 algorithms:

* K-Means:

The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

* + - The centroids of the K clusters, which can be used to label new data
    - Labels for the training data (each data point is assigned to a single cluster)

# Evaluation

## Classification Evaluation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| SVM | 0.95 | **0.97** | 0.93 | 0.94 |
| KNN | **0.96** | 0.96 | **0.96** | **0.96** |
| DT | 0.95 | 0.95 | 0.94 | 0.94 |

## Clustering Evaluation

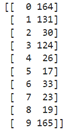
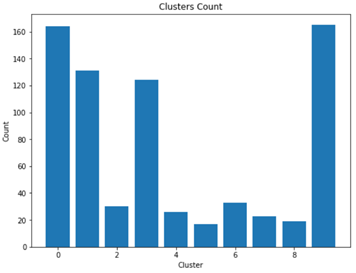
|  |  |  |
| --- | --- | --- |
| Model | # Clusters | Silhouette Score |
| K-Means | **10** | 0.6956 |

|  |  |  |
| --- | --- | --- |
|  | Bot Operation | Accuracy |
| 1 | Disease Diagnosis | 95% |
| 2 | Drug Recommendation | 70% |
| 3 | Disease Diagnosis and Drug Recommendation | 66.5% |

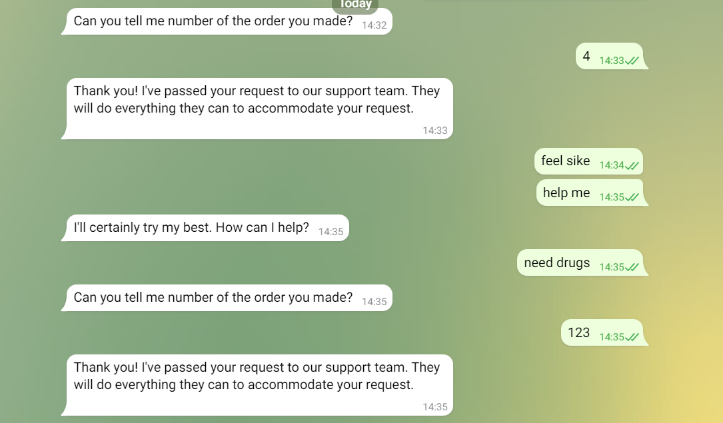
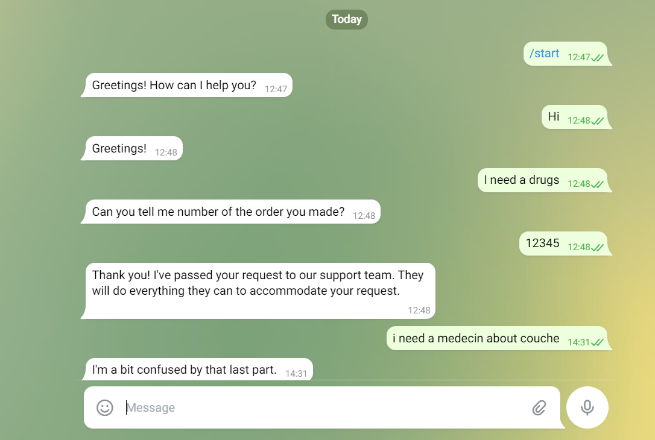
Accuracy of the system is **66.5%**

# Error Analysis

* Remove the very beginning and the very last paragraphs of each book
* Remove the most frequent and the least weighted words from TF-IDF Transformation
* Try different collections of books



# Deployment



# References

* <https://data.europa.eu/euodp/en/data/dataset/epar-human-medicines/resource/34ccc778-81e3-4ee3-86f4-a59604e116be>
* https://www.cs.cmu.edu/~schneide/tut5/node42.html
* https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering?ex=9
* https://docs.aws.amazon.com/sagemaker/latest/dg/lda.html
* https://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html