Система аналізу і категоризації текстових медичних даних з використанням SAS технологій

Автор: Юрчук Максим Віталійович

Група КА-53, Факультет ІПСА, КПІ ім. Сікорського

Науковий керівник: к.т.н., м.н.с Терентьєв О.М.

Об’єкт дослідження

- Колекція документів: медичні текстові дані у вигляді відгуків користувачів препаратів, збережені у різних форматах (pdf, txt, docx).

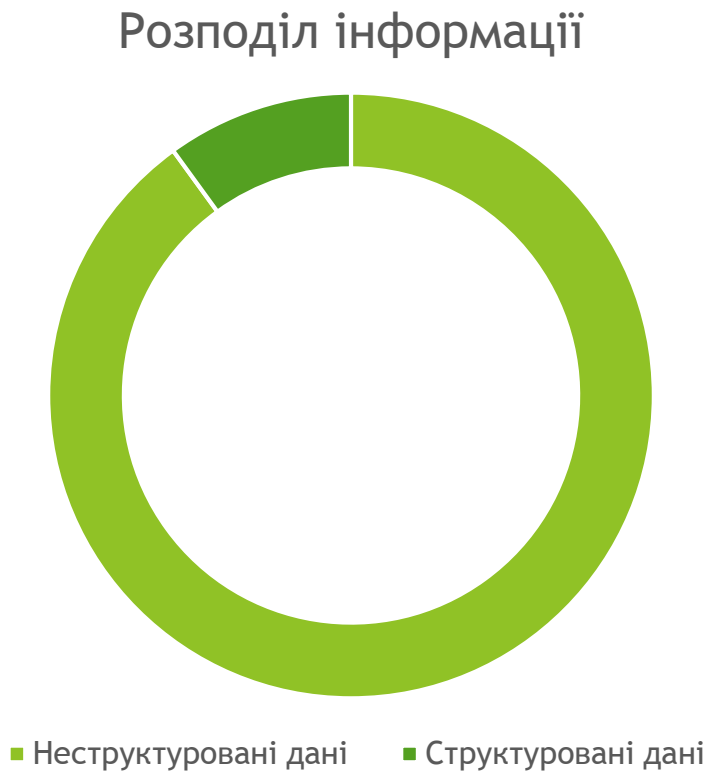
Предмет дослідження

- інформаційний пошук;
- інтелектуальний аналіз;
- метод максимальної правдоподібності;
- латентно-семантичний аналіз;
- булеві правила.

Актуальність

На сьогодні більша частина інформації (90%) знаходиться у неструктурованому вигляді, тому використання її звичними аналітичними моделями являється неможливим.

В медичній сфері, обробляючи текстові дані, можливо значно покращити якість зворотнього зв'язку з пацієнтами, а тому і якість препаратів.



Мета роботи

Дослідження існуючих методів обробки неструктурованих текстових даних та їх впровадження у системі аналізу та категоризації текстової медичної звітності.

Перелік основних методів

- Класифікація тексту на основі генерації булевих правил;
- Кластеризація тексту з застосуванням методу максимальної правдоподібності;
- Латентно-семантичний аналіз

Застосовані математичні методи

- Метод максимальної правдоподібності
- Сингулярний розклад матриці
- Метод головних компонент

Статистичні міри, що використовувались для оцінки якості моделей

- RMSSTD
- F1-score
- Загальна точність
- Відстань Кульбека-Лейблера

F1 – score та загальна точність

- $\text{precision} = \frac{TP}{TP+FP}$
- $\text{recall} = \frac{TP}{TP+FP}$
- $F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision}+\text{recall}}$

- $\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$

RMSSTD

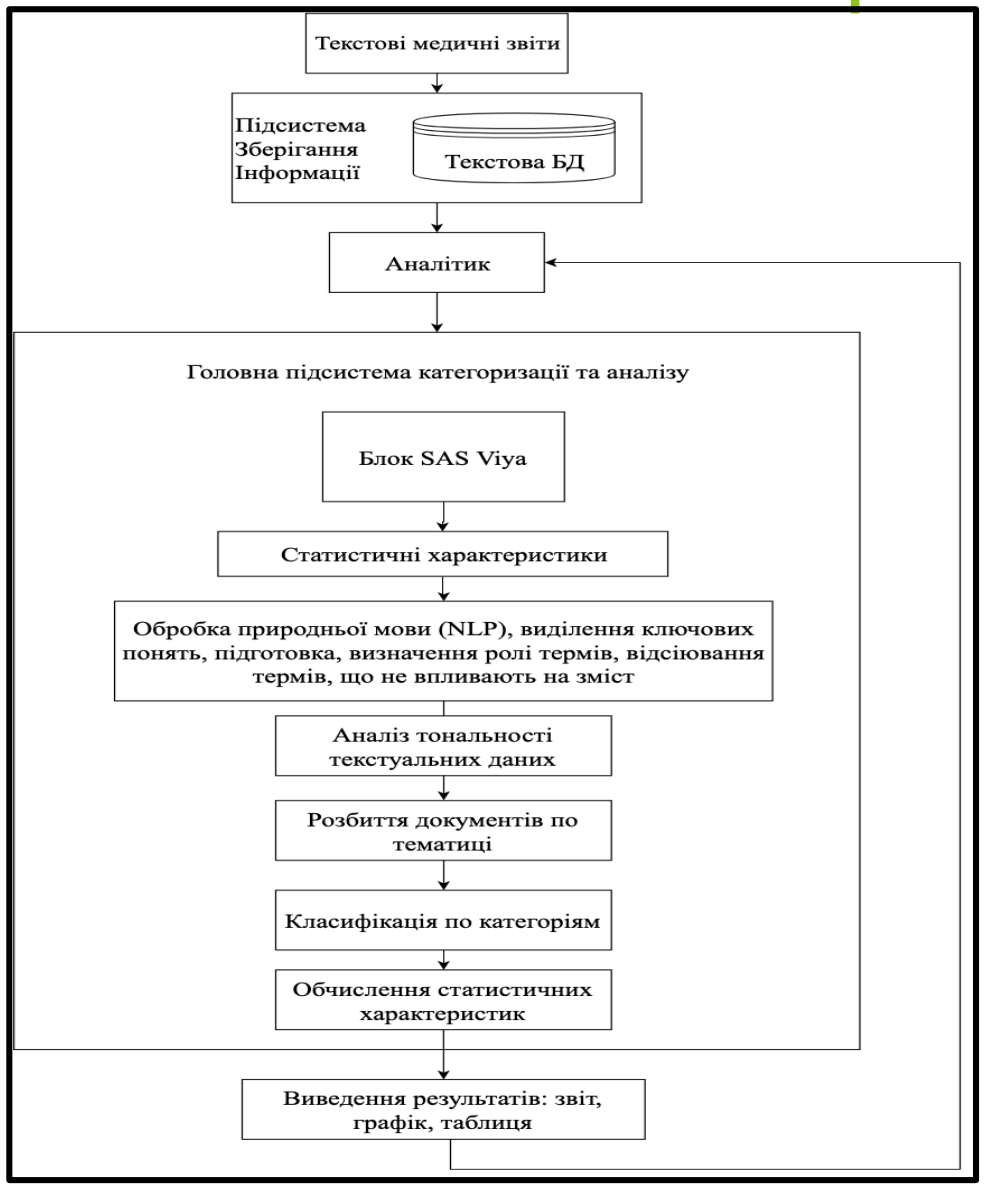
RMSSTD є мірою однорідності в межах кластерів, статистичний показник:

$$\sqrt{\frac{SS_1 + \cdots + SS_p}{df_1 + \cdots + df_p}}$$

тобто об'єднане стандартне відхилення всіх змінних. Термін SS_j - сума квадратів j-ої змінної, обчислюється за формулою:

$$SS_j = \sum_{i=1}^{n} (x_{ij} - \overline{x_{ij}})^2$$

СТРУКТУРА ІНФОРМАЦІЙНОЇ СИСТЕМИ



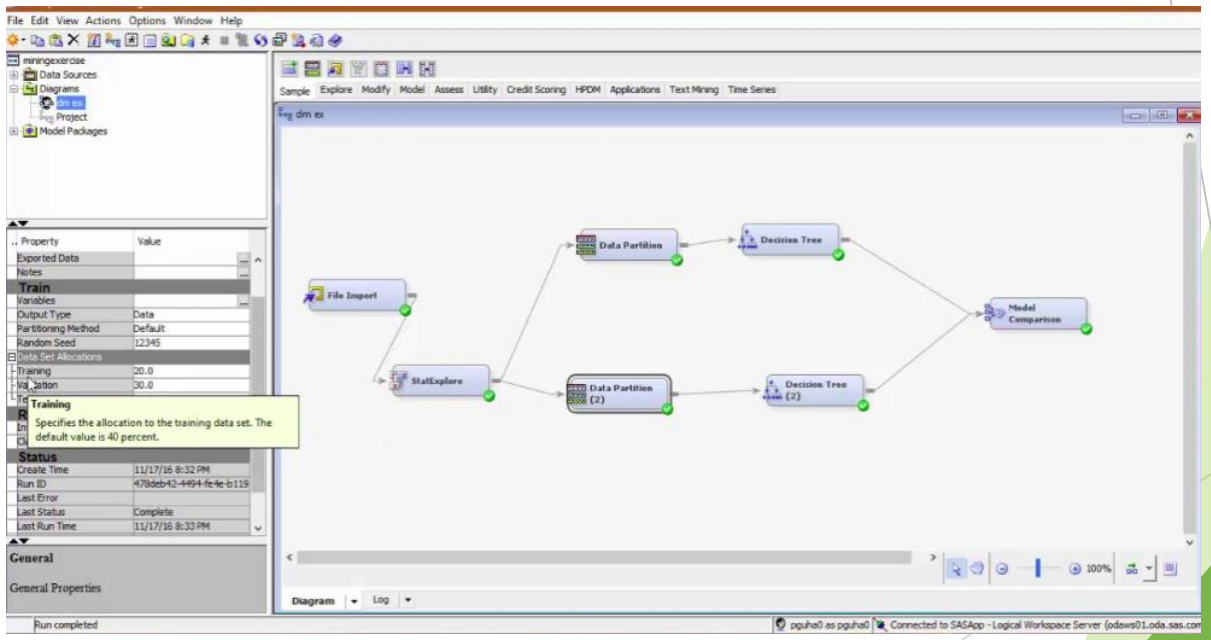
Програмне забезпечення, яке було використано в дипломній роботі

SAS Viya

(хмарні обчислення)

- Develop SAS Code
- Manage Data
- Prepare Data
- Explore and Visualize Data
- Build Models
- Manage Models
- Manage Decisions
- Explore Lineage
- Build Graphs
- Manage Workflows

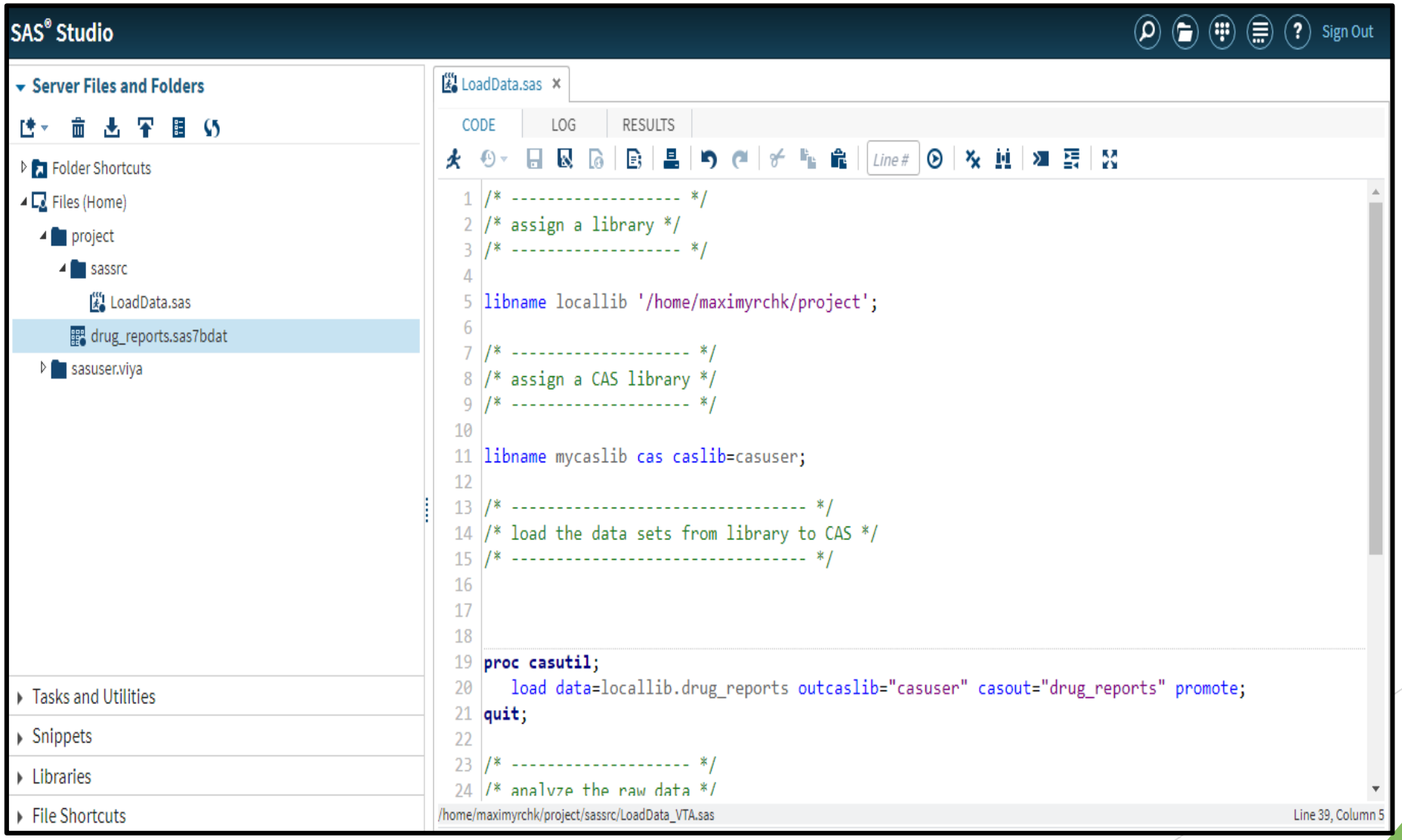
SAS Enterprise Miner
(локальний інструмент)

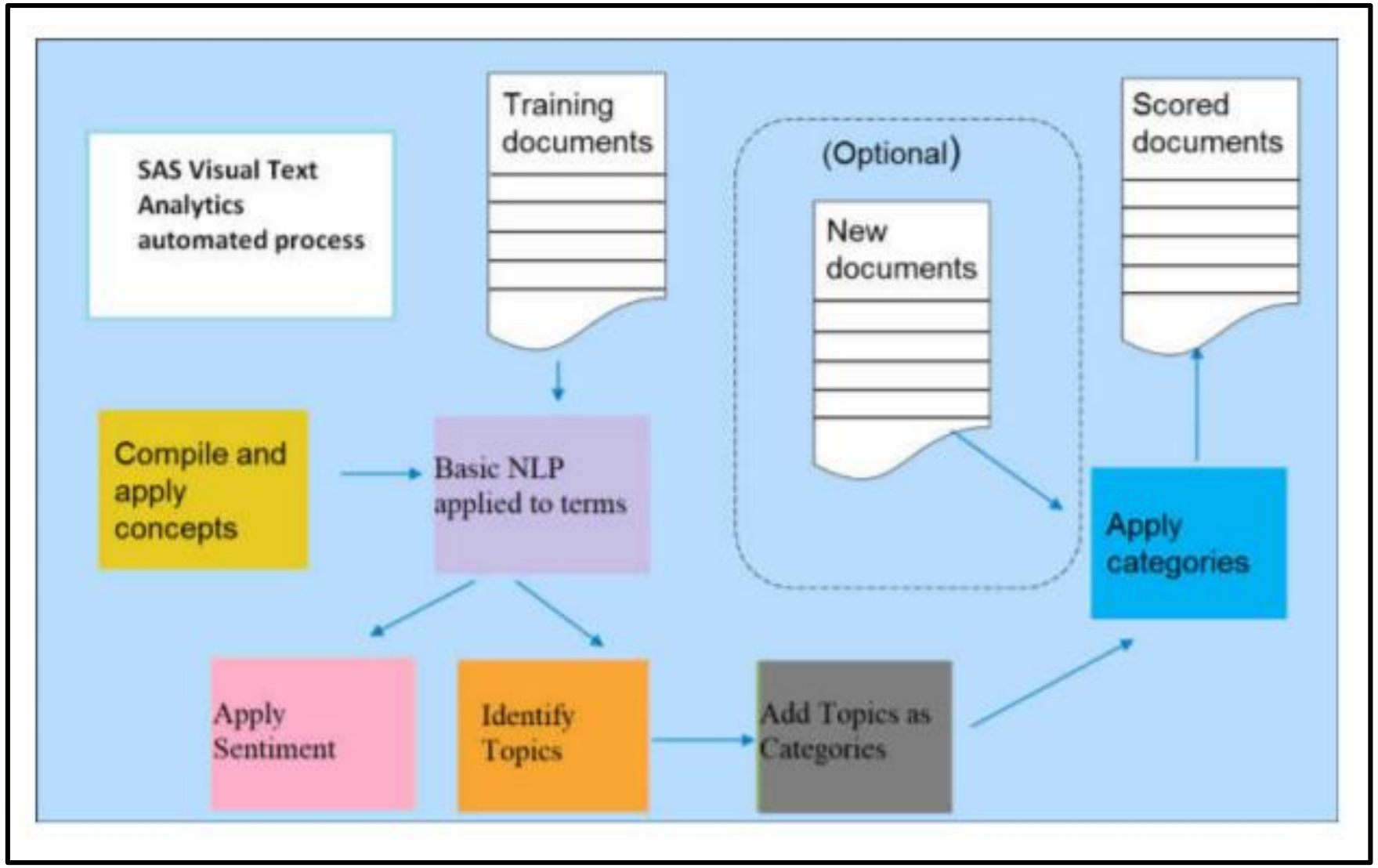


SAS технології

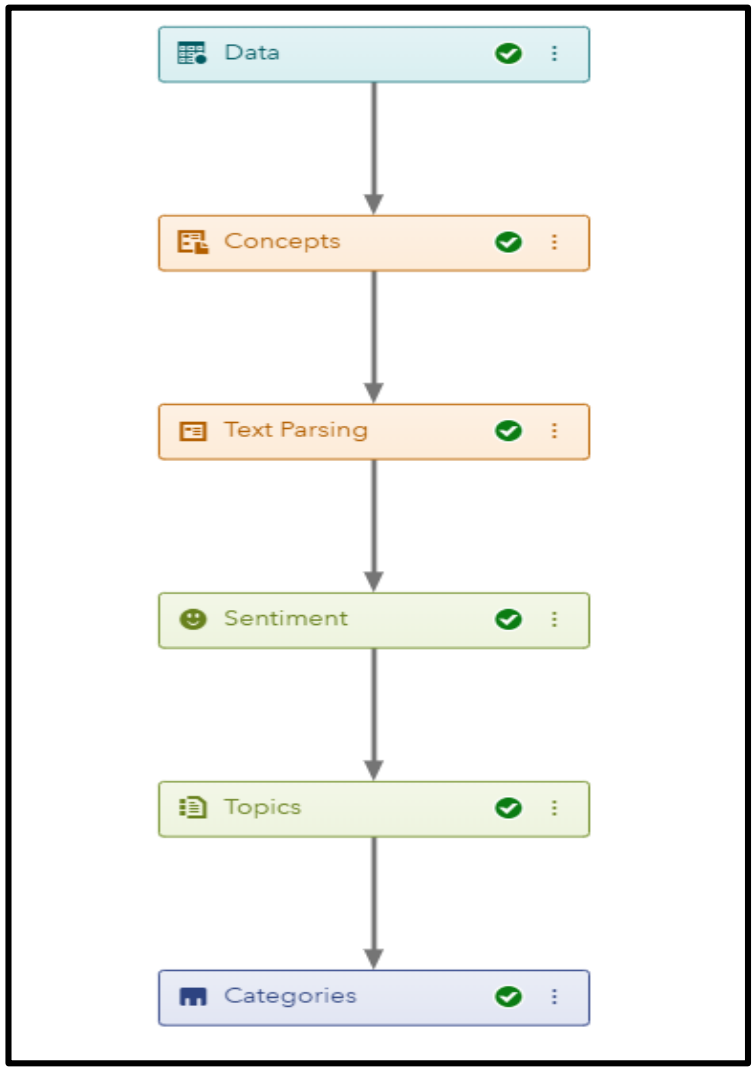
- Нормалізація, стемінг
- Виокремлення понять
- Створення тематик з поєднанням важливих понять
- Розподіл документів по створеним категоріям
- Аналіз тональності текстів

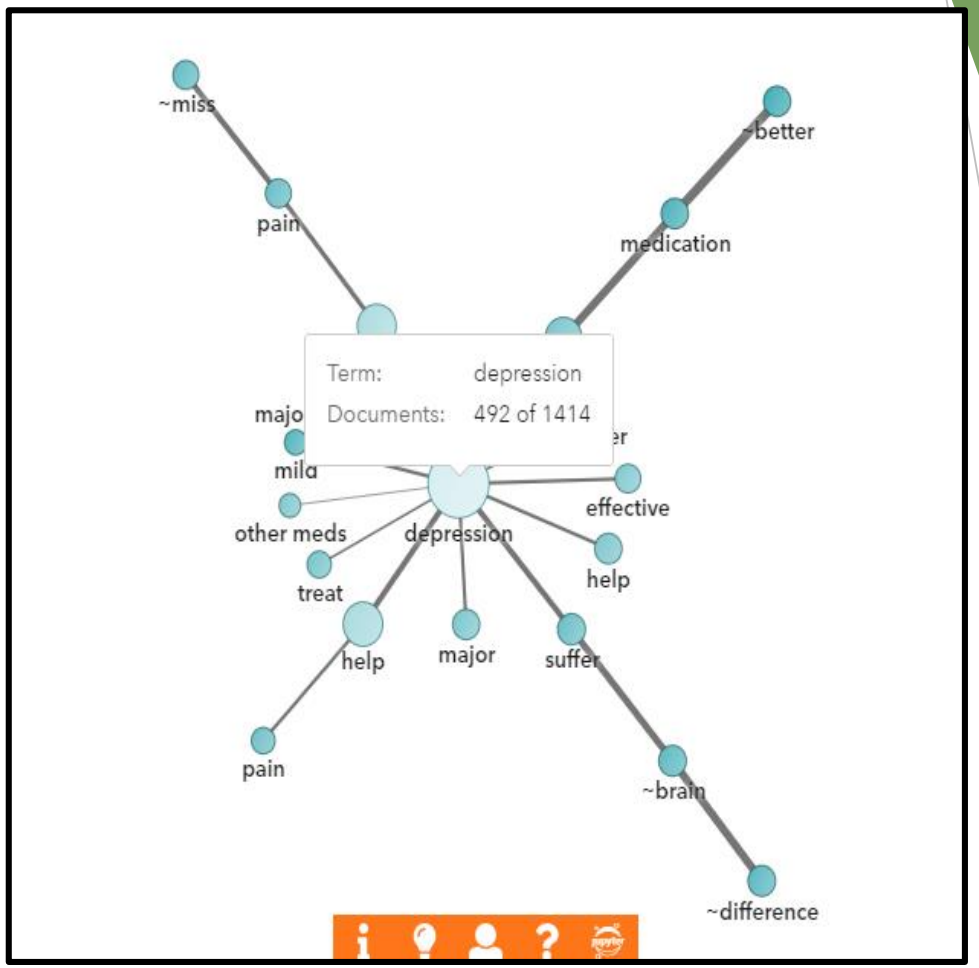
Огляд хмарного інструменту SAS Viya




SAS Visual Text Analytics automated process
Training documents
(Optional)
New documents
Scored documents
Compile and apply concepts
Basic NLP applied to terms
Apply categories
Apply Sentiment
Identify Topics
Add Topics as Categories

Результати



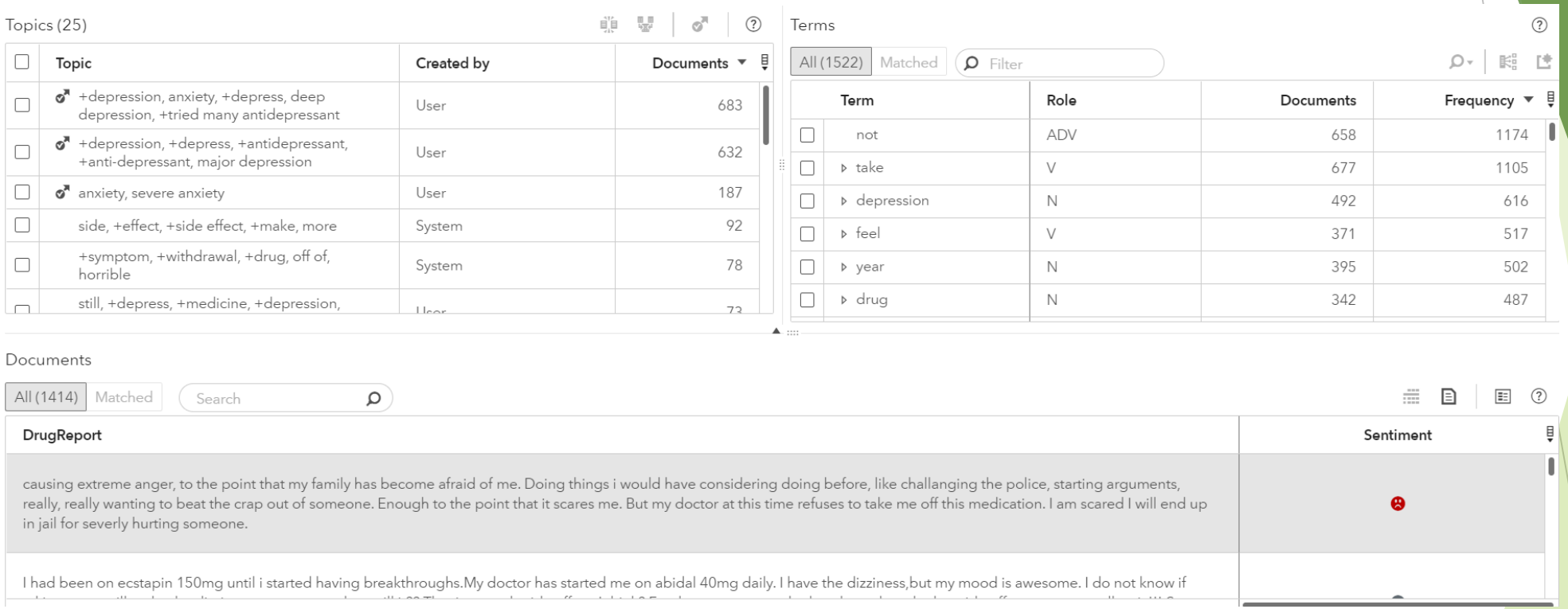


Взаємозв’язок понять

Результати

- Проведено аналіз сучасних методів інформаційного пошуку;
- Досліджені можливості використання наявних інструментів текстової аналітики;
- Налаштування їх під конкретну галузь;
- Розроблена система, яка дозволяє виконувати задачу класифікації, аналізу тональності, розподілу текстової інформації по категоріям отримуючи на вхід велику кількість медичних звітів.

Тематики



Topics (25)

Topic	Created by	Documents
+depression, anxiety, +depress, deep depression, +tried many antidepressant	User	683
+depression, +depress, +antidepressant, +anti-depressant, major depression	User	632
anxiety, severe anxiety	User	187
side, +effect, +side effect, +make, more	System	92
+symptom, +withdrawal, +drug, off of, horrible	System	78
still, +depress, +medicine, +depression,	User	73

Terms

All (1522) Matched Filter

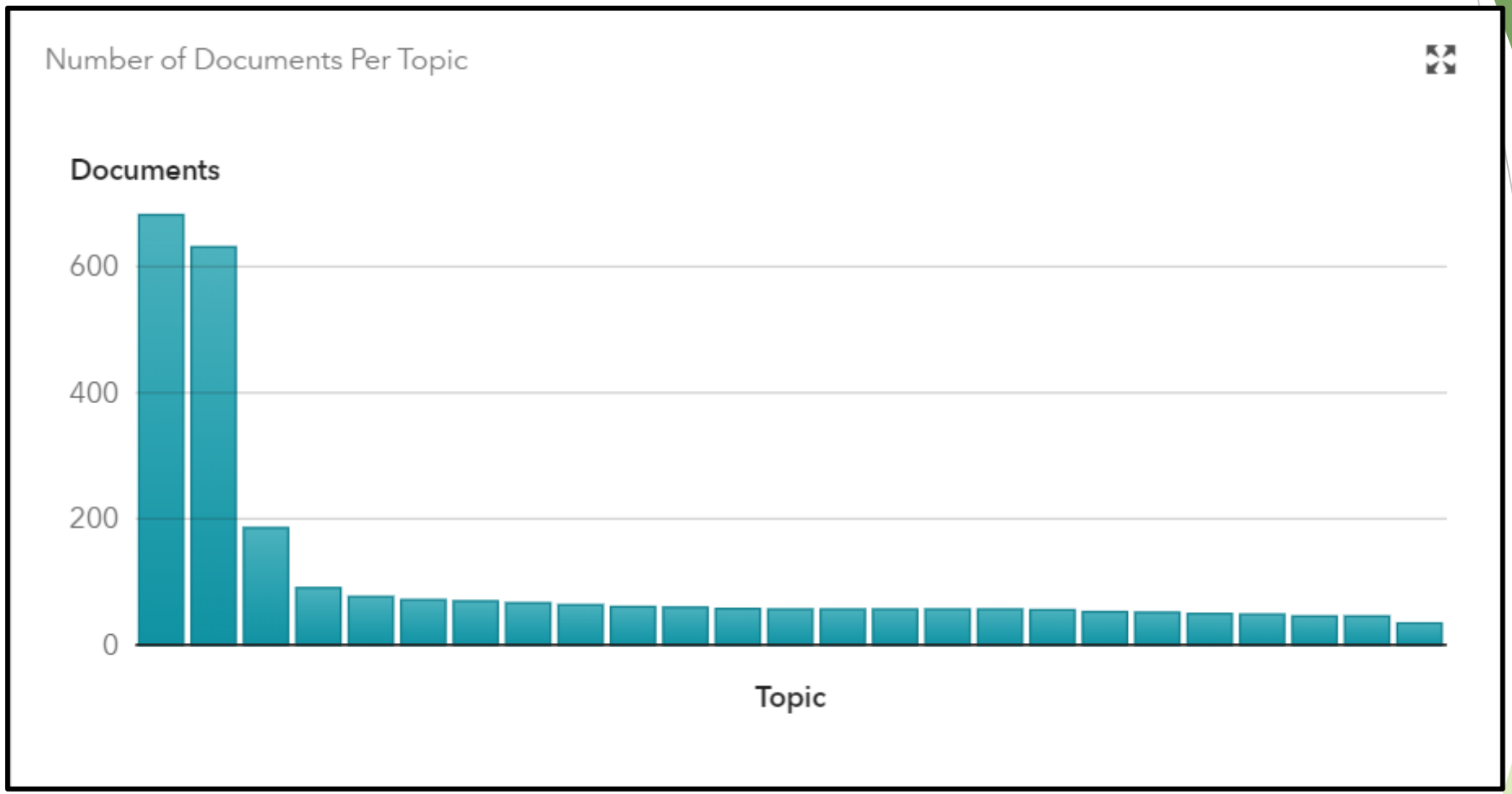
Term	Role	Documents	Frequency
not	ADV	658	1174
take	V	677	1105
depression	N	492	616
feel	V	371	517
year	N	395	502
drug	N	342	487

Documents

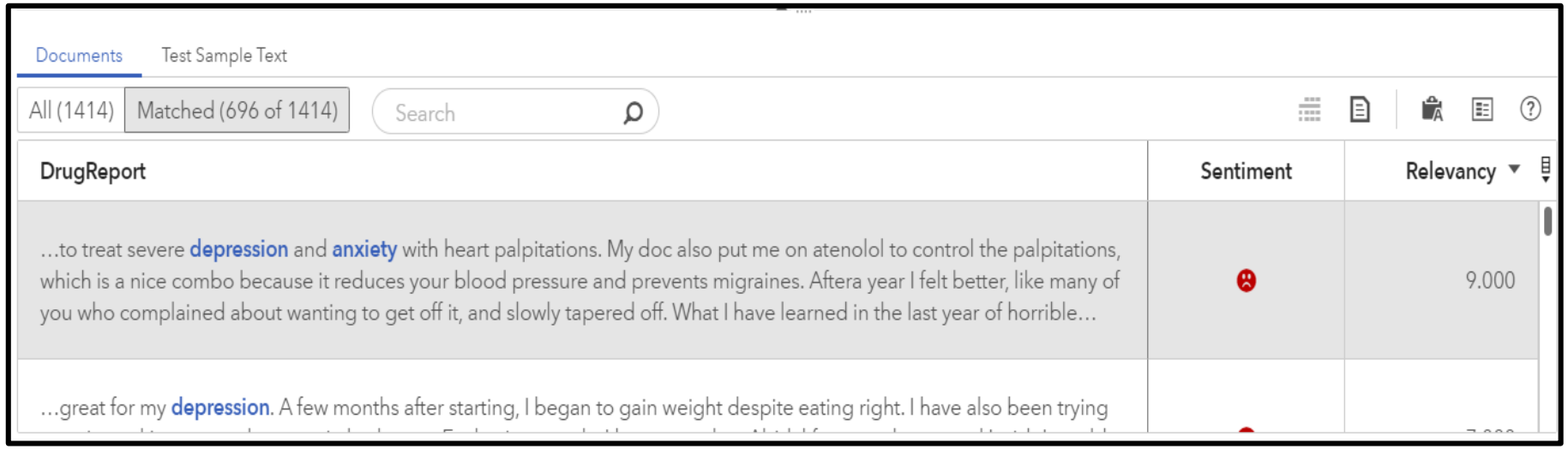
All (1414) Matched Search

DrugReport	Sentiment
causing extreme anger, to the point that my family has become afraid of me. Doing things i would have considering doing before, like challanging the police, starting arguments, really, really wanting to beat the crap out of someone. Enough to the point that it scares me. But my doctor at this time refuses to take me off this medication. I am scared I will end up in jail for severly hurting someone.	
I had been on ecstapin 150mg until i started having breakthroughs.My doctor has started me on abidal 40mg daily. I have the dizziness,but my mood is awesome. I do not know if	

Об’єднання понять та підвищення їх до категорій


Number of Documents Per Topic
Documents
600
400
200
0
Topic

Категоризація



Створення категорій на основі булевих правил

Порівняльна таблиця характеристик моделей класифікації з використанням WOE та без використання WOE

	Навчальна вибірка (Train)		Тестова вибірка (Test)	
	Загальна точність	F1	Загальна точність	F1
Модель без використання WOE	0.774	0.673	0.728	0.631
Модель з використанням WOE	0.803	0.701	0.785	0.682

Дякую за увагу