

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 Терентьев 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.

Розподіл інформації



■ Неструктуровані дані ■ Структуровані дані

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

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

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

RMSSTD

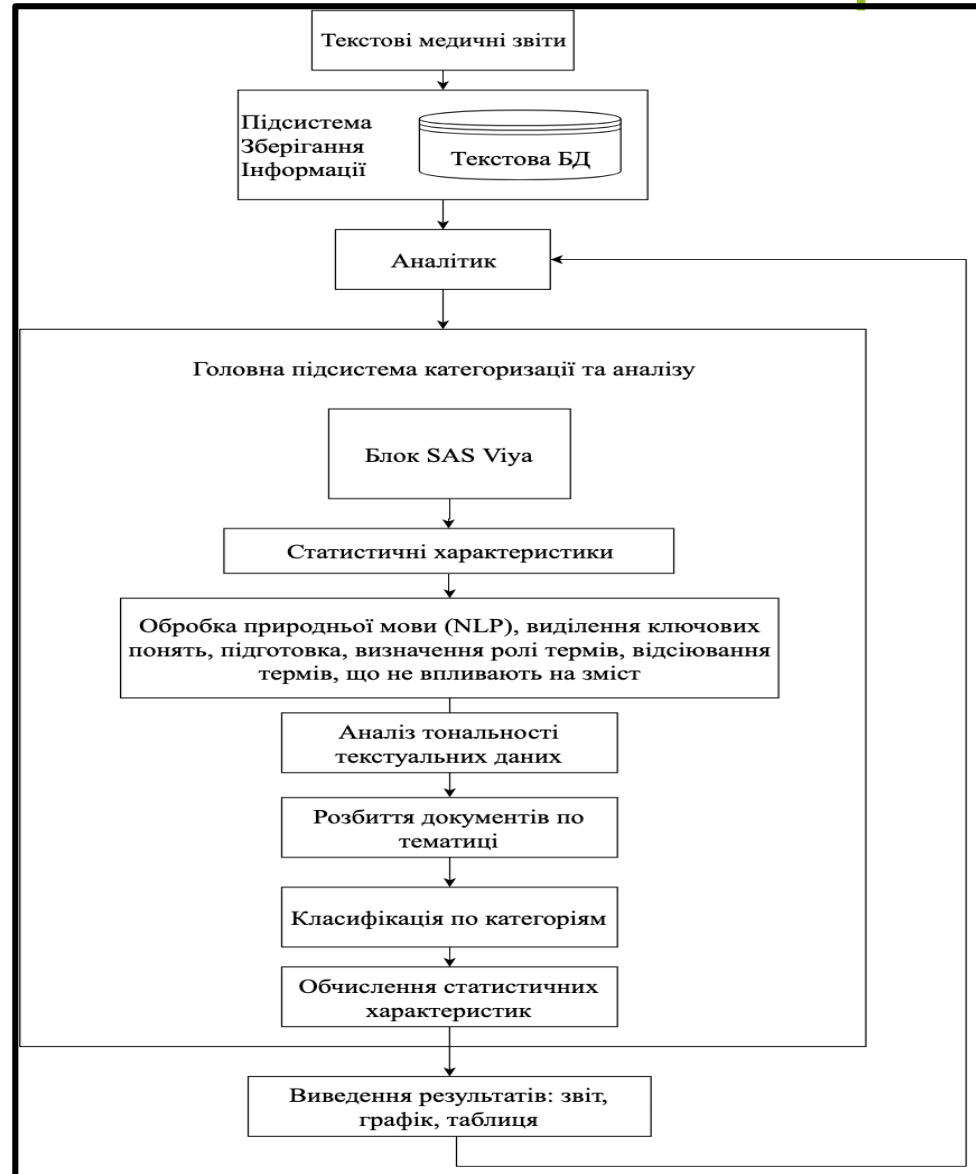
RMSSTD є мірою однорідності в межах кластерів, статистичний показник:

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

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

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

СТРУКТУРА ІНФОРМАЦІЙНОЇ СИСТЕМИ



Програмне забезпечення, яке було використано в дипломній роботі

SAS Viya

(хмарні обчислення)

SAS Enterprise Miner

(локальний інструмент)

Develop SAS Code

Manage Data

Prepare Data

Explore and Visualize Data

Build Models

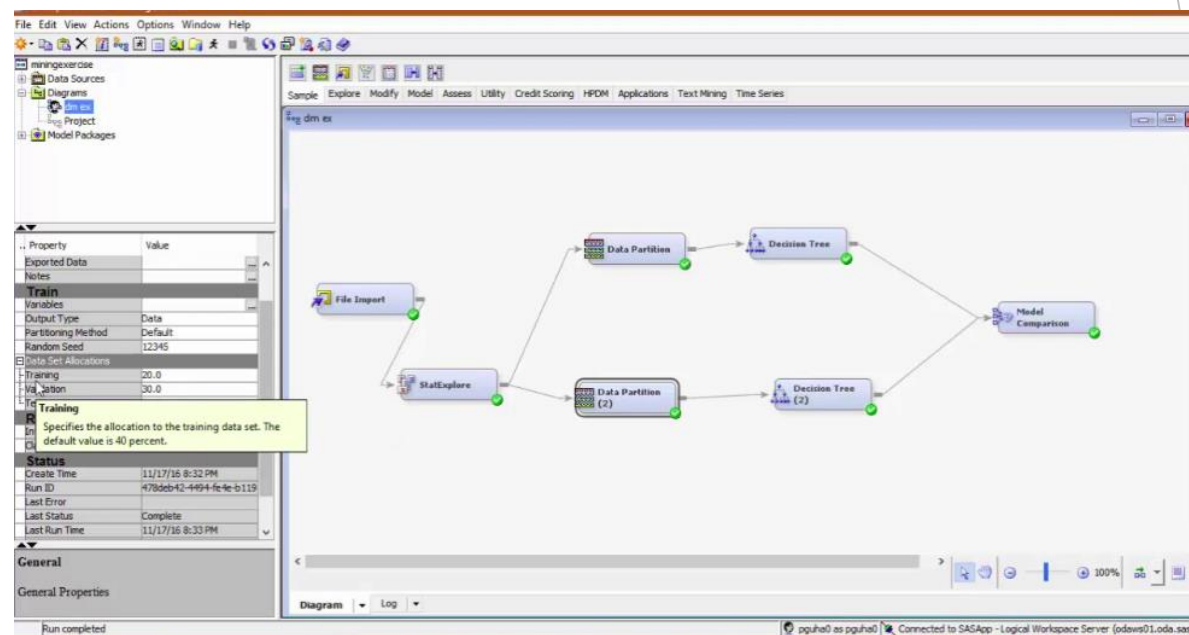
Manage Models

Manage Decisions

Explore Lineage

Build Graphs

Manage Workflows



SAS технології

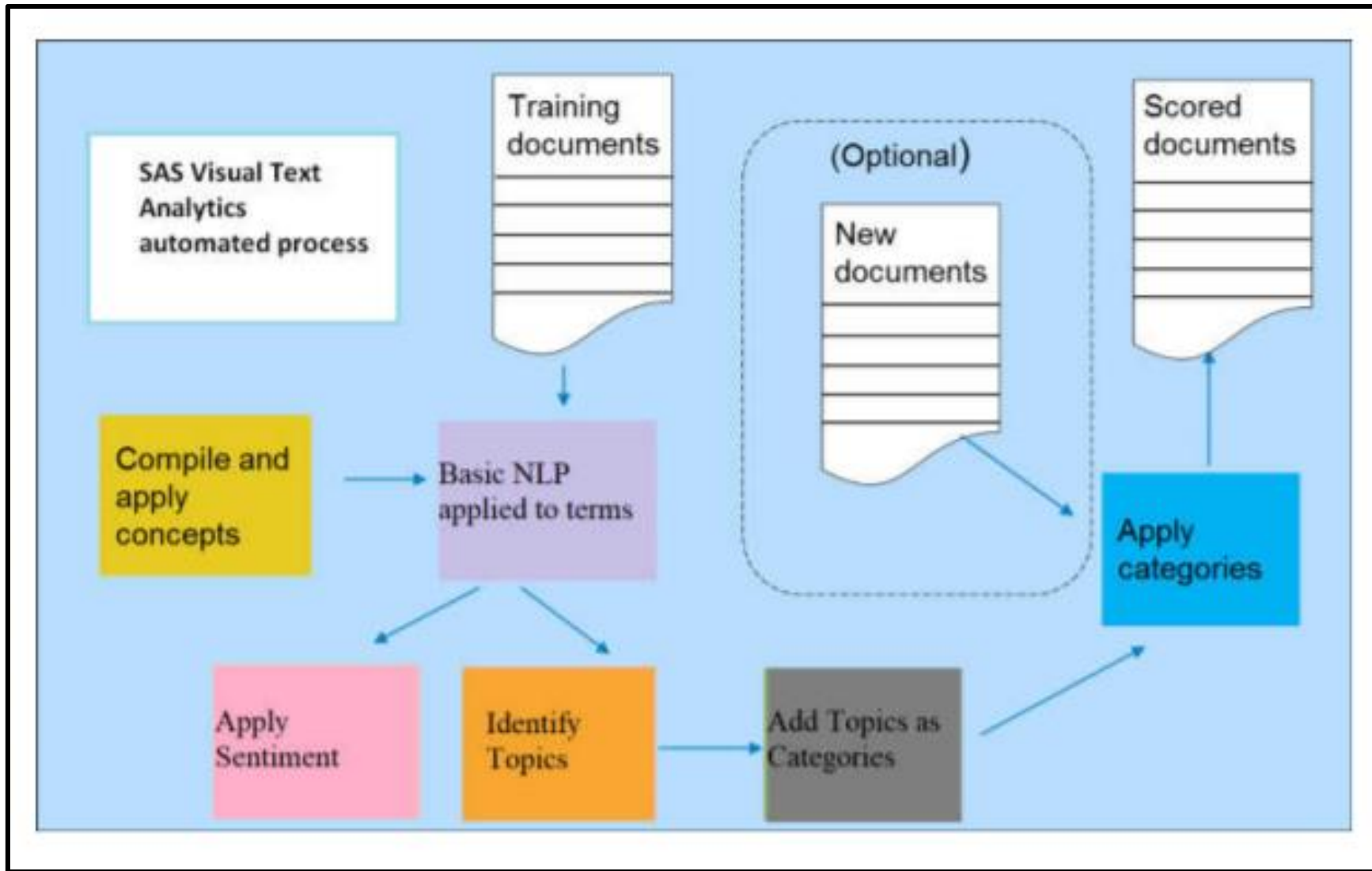
- ▶ Нормалізація, стемінг
- ▶ Виокремлення понять
- ▶ Створення тематик з поєднанням важливих понять
- ▶ Розподіл документів по створеним категоріям
- ▶ Аналіз тональності текстів

Огляд хмарного інструменту SAS Viya

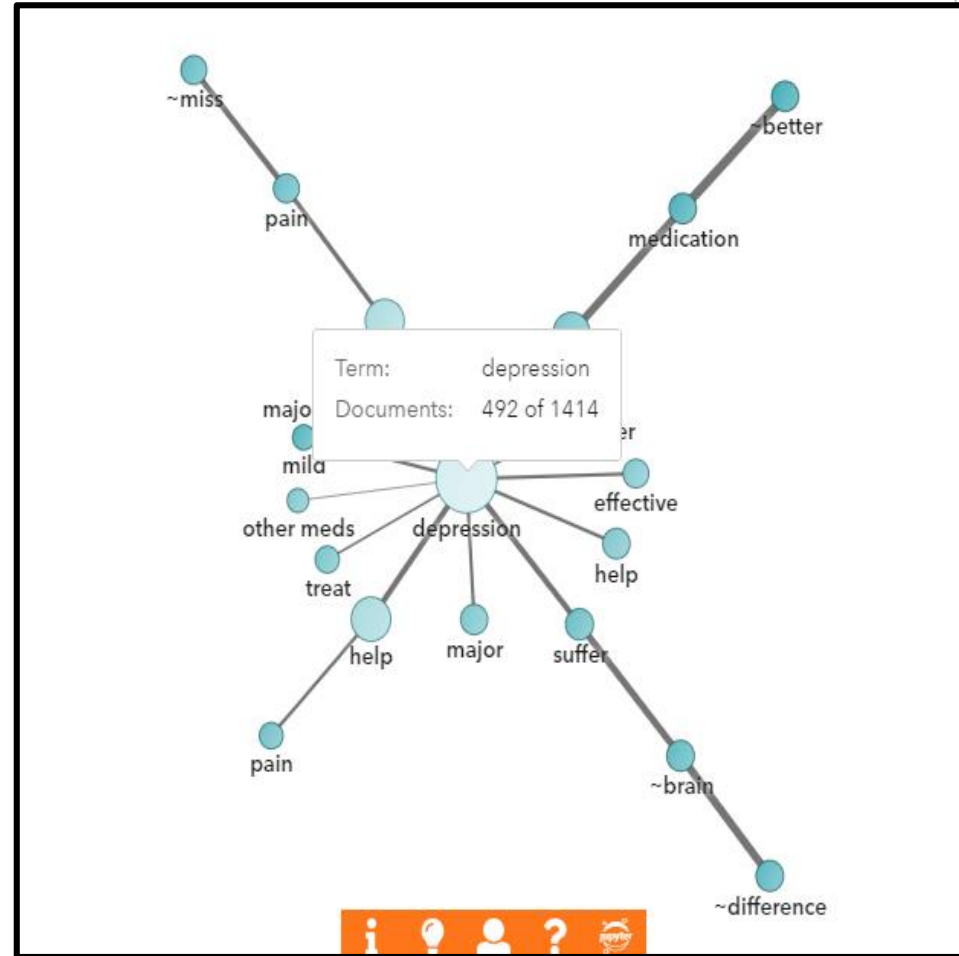
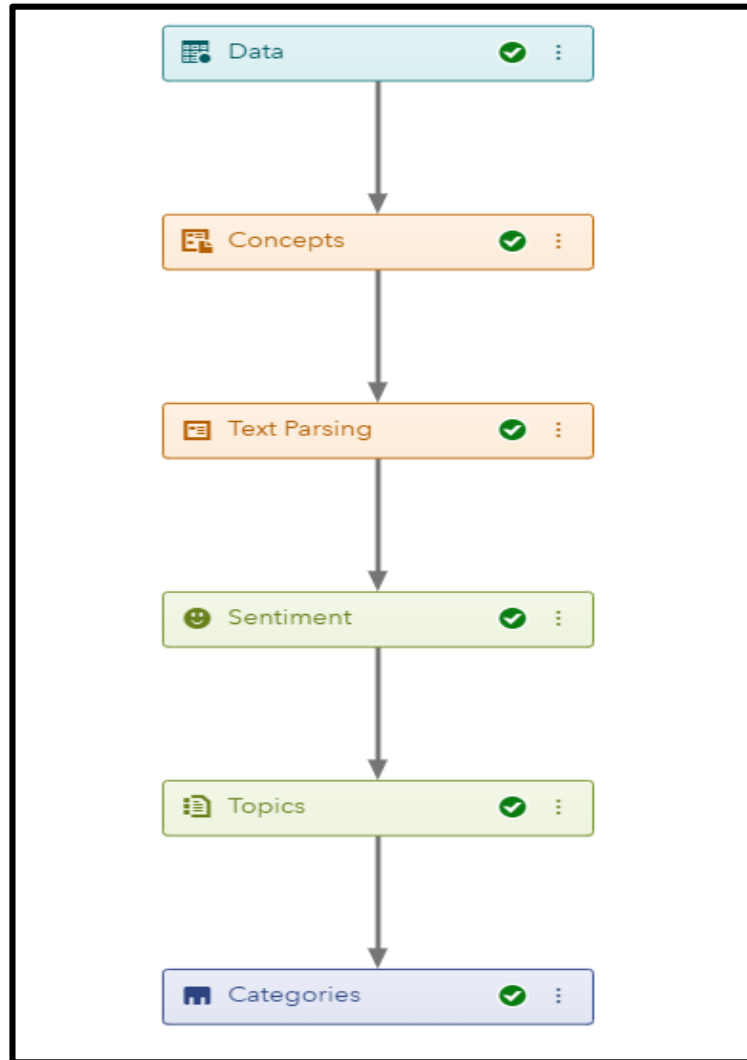
The screenshot displays the SAS Studio web interface. On the left, the 'Server Files and Folders' pane shows a tree view with 'Files (Home)' expanded to 'project', then 'sassrc', where 'LoadData.sas' is selected. The main editor area shows the code for 'LoadData.sas' with line numbers 1 through 24. The code includes comments and SAS commands for library assignment and data loading. The status bar at the bottom indicates the current file path and cursor position.

```
1 /* ----- */
2 /* assign a library */
3 /* ----- */
4
5 libname locallib '/home/maximyrchk/project';
6
7 /* ----- */
8 /* assign a CAS library */
9 /* ----- */
10
11 libname mycaslib cas caslib=casuser;
12
13 /* ----- */
14 /* load the data sets from library to CAS */
15 /* ----- */
16
17
18
19 proc casutil;
20   load data=locallib.drug_reports outcaslib="casuser" casout="drug_reports" promote;
21 quit;
22
23 /* ----- */
24 /* analyze the raw data */
```

Line 39, Column 5



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)

<input type="checkbox"/>	Topic	Created by	Documents
<input type="checkbox"/>	+depression, anxiety, +depress, deep depression, +tried many antidepressant	User	683
<input type="checkbox"/>	+depression, +depress, +antidepressant, +anti-depressant, major depression	User	632
<input type="checkbox"/>	+anxiety, severe anxiety	User	187
<input type="checkbox"/>	side, +effect, +side effect, +make, more	System	92
<input type="checkbox"/>	+symptom, +withdrawal, +drug, off of, horrible	System	78
<input type="checkbox"/>	still, +depress, +medicine, +depression,	User	72

Terms

All (1522) Matched Filter

<input type="checkbox"/>	Term	Role	Documents	Frequency
<input type="checkbox"/>	not	ADV	658	1174
<input type="checkbox"/>	take	V	677	1105
<input type="checkbox"/>	depression	N	492	616
<input type="checkbox"/>	feel	V	371	517
<input type="checkbox"/>	year	N	395	502
<input type="checkbox"/>	drug	N	342	487

Documents

All (1414) Matched Search

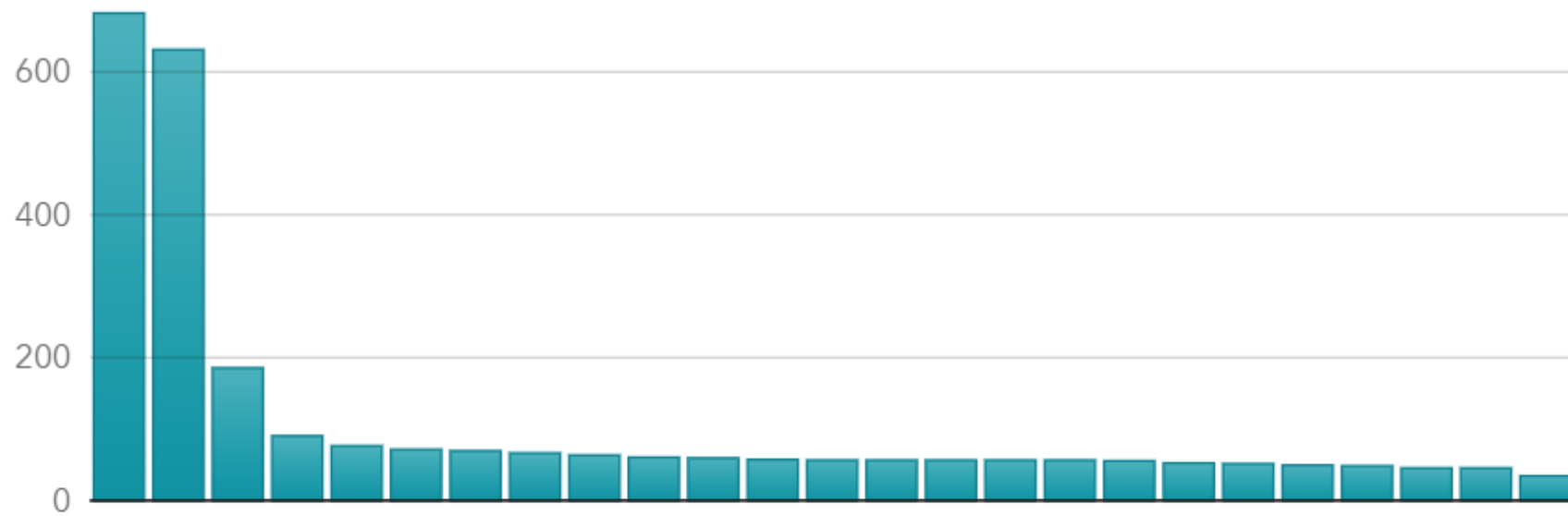
DrugReport	Sentiment
causing extreme anger, to the point that my family has become afraid of me. Doing things i would have considering doing before, like challanging the police, starting arguments, really, really wanting to beat the crap out of someone. Enough to the point that it scares me. But my doctor 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



Topic

Категоризація

Documents Test Sample Text

All (1414) Matched (696 of 1414) Search

DrugReport	Sentiment	Relevancy
...to treat severe depression and anxiety with heart palpitations. My doc also put me on atenolol to control the palpitations, which is a nice combo because it reduces your blood pressure and prevents migraines. After a year I felt better, like many of you who complained about wanting to get off it, and slowly tapered off. What I have learned in the last year of horrible...	☹️	9.000
...great for my depression . A few months after starting, I began to gain weight despite eating right. I have also been trying	☹️	7.000

Categories

- All Categories (4)
 - Anxiety
 - Depression
 - Anxiety&Depression**

```
OR,  
  (AND,(OR,"depression","depressions")),  
  (AND,"anxiety"),  
  (AND,(OR,"depressing","depressed","depress")),  
  (AND,(OR,"antidepressants","antidepressant")),  
  (AND,(OR,"anti-depressant","anti-depressants")),  
  (AND,"depressed"),  
  (AND,"depressive"),  
  (AND,(OR,"depressants","depressant")),  
  (AND,"major"))
```

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