

Single Day Event

Exploration of NLP Technology and Its Usage in Clinical Trials

Luo Anqi
Peng Vanessa
Novartis

Disclaimer

- The views and opinions expressed in this [document/presentation] are those of the author and do not necessarily reflect the official policy or position of Novartis or any of its officers.



Background

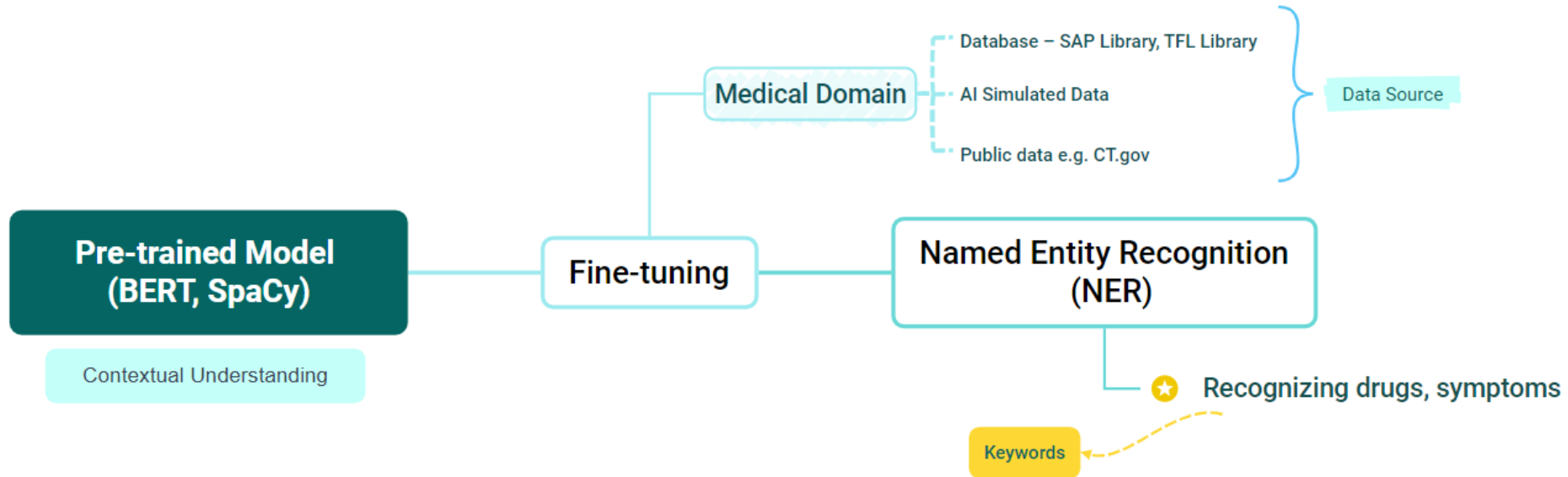
- ❑ **What is Natural Language Processing (NLP)**
Branch of artificial intelligence (AI) & Human Language
- ❑ **Why NLP Exploration**
Unstructured Text Data & Automated Extraction
- ❑ **A Recommendation System for TFL Shell TOC**
Keywords Extraction
Similarity Matching



Algorithm Exploration

Feature	Sumy/Summa/YAKE/ RAKE	NLTK	SpaCy 	BERT
Performance	Fast, task-specific	Slow	Fast	Slow, GPU*
Pre-trained Model Support	No	No	Yes	Yes
Extensibility	Moderate	Moderate	Strong	Strong
Industrial Application	No	No	Yes	Yes
Multilingual Support	Limited	Limited	Extensive	Extensive
Visualization Support	None	Moderate	Strong 'Display'	Moderate, requires tools like TensorBoard

phuse Fine-tuning on SpaCy



phuse NER in SpaCy

scispacy



SpaCy models for biomedical text processing

[View the Project on GitHub](#)
allenai/scispacy

Download
ZIP File

Download
TAR Ball

View On
GitHub

Models

Model

en_core_sci_sm

en_core_sci_md

en_core_sci_scibert

en_core_sci_lg

en_ner_craft_md

en_ner_jnlpba_md

en_ner_bc5cdr_md

en_ner_bionlp13cg_md

```
9 # Load BC5CDR model
10 nlp = spacy.load('en_ner_bc5cdr_md')
11
12 text = """
13 Aspirin is frequently used to prevent heart disease,
14 while insulin is a critical treatment for diabetes.
15 """
16
17 doc = nlp(text)
18
19 for ent in doc.ents:
20     print(f"Entity: {ent.text}, Label: {ent.label_}")
21
```

PROBLEMS

OUTPUT

DEBUG CONSOLE

TERMINAL

PORTS

```
PS C:\> python M1_bc5cdr.py
Entity: Aspirin, Label: CHEMICAL
Entity: heart disease, Label: DISEASE
Entity: insulin, Label: CHEMICAL
Entity: diabetes, Label: DISEASE
```

SpaCy Model Pipelines

✓ custom_medical_ner

> attribute_ruler

> lemmatizer / lookups

> ner

> parser

> sender

> tagger

> tok2vec

> vocab

⚙ config.cfg

{ } meta.json

≡ tokenizer

Component	Function
attribute_ruler	Adjusts token attributes by rules
lemmatizer	Maps words to their root forms
ner	Named Entity Recognition
parser	Dependency parsing of sentences
sender	Sentence boundary detection
tagger	Part-of-Speech (POS) tagging
tok2vec	Transforms tokens into vectors
vocab	Stores vocabulary and word embeddings
config.cfg	Configuration file for the model
meta.json	Metadata for the model
tokenizer	Splits text into tokens



Similarity Matching

Method Type	Technique/Method	Description
Simple String Matching	Jaccard Similarity	Ratio of common to total words
	Cosine Similarity	Cosine of angle between vectors
Edit Distance	Levenshtein Distance	Minimum edit operations
	Damerau-Levenshtein Distance	Levenshtein with transpositions
Word Embedding	Word2Vec	Word vectors and cosine similarity
	GloVe	Word vectors from co-occurrence matrix
	FastText	Word2Vec with subword information
Deep Learning Models	BERT	Context-aware word embeddings
Regex Matching	Regular Expressions	Pattern matching with regex

SAP – ToC

Table of contents

List of abbreviations

1 Introduction

1.1 Study design

Type	Sec	Nu	Title
Heading	A		Shells and specifications for Sections 14 and 16 of the CSR
Section	14		Tables, figures and graphs referred to but not included in the text
Section	14.1		Demographic data
Section	14.2		Efficacy and other non-safety data (e.g. PK, PK/PD, Health Econ., QoL)
Section	14.3		Safety data
Section	14.3.1		Displays of adverse events
Section	14.3.2		Listings of deaths, other serious and significant adverse events
Section	16		Appendices

2.4 Treatments (study treatment, rescue medical...

2.4.1 Study treatment / compliance

2.4.2 Prior, concomitant and post therapies

2.5 Analysis supporting primary objective(s)

2.5.1 Primary endpoint(s)

2.5.2 Statistical hypothesis, model, and metho...

2.5.3 Handling of intercurrent events

2.5.4 Handling of missing values not related t...

Statistical Analysis Plan (SAP)

Release date: dd-Mmm-yyyy

Number of pages:

General Workflow

2.7 Safety analyses

- 2.7.1 Adverse events (AEs)
- 2.7.2 Deaths
- 2.7.3 Laboratory data
- 2.7.4 Other safety data
- 2.8 Pharmacokinetic endpoints
- 2.9 PD and PK/PD analyses
- 2.10 Patient-reported outcomes
- 2.11 Biomarkers
- 2.12 Other Exploratory analyses

2 Statistical methods

2.1 Data analysis general information

2.1.1 General definitions

2.2 Analysis sets

2.2.1 Subgroup of interest

- 2.3 Patient disposition, demographics and other baseline characteristics
- 2.4 Treatments (study treatment, rescue medication, concomitant therap...

2.5 Analysis supporting primary objective(s)

- 2.5.1 Primary endpoint(s)
- 2.5.2 Statistical hypothesis, model, and method of analysis
- 2.5.3 Handling of intercurrent events
- 2.5.4 Handling of missing values not related to intercurrent event
- 2.5.5 Sensitivity analyses
- 2.5.6 Supplementary analyses

2.6 Analysis supporting secondary objectives

- 2.6.1 Secondary endpoint(s)
- 2.6.2 Statistical hypothesis, model, and method of analysis
- 2.6.3 Handling of intercurrent events
- 2.6.4 Handling of missing values not related to intercurrent event
- 2.6.5 Sensitivity analyses
- 2.6.6 Supplementary analyses

1 Introduction

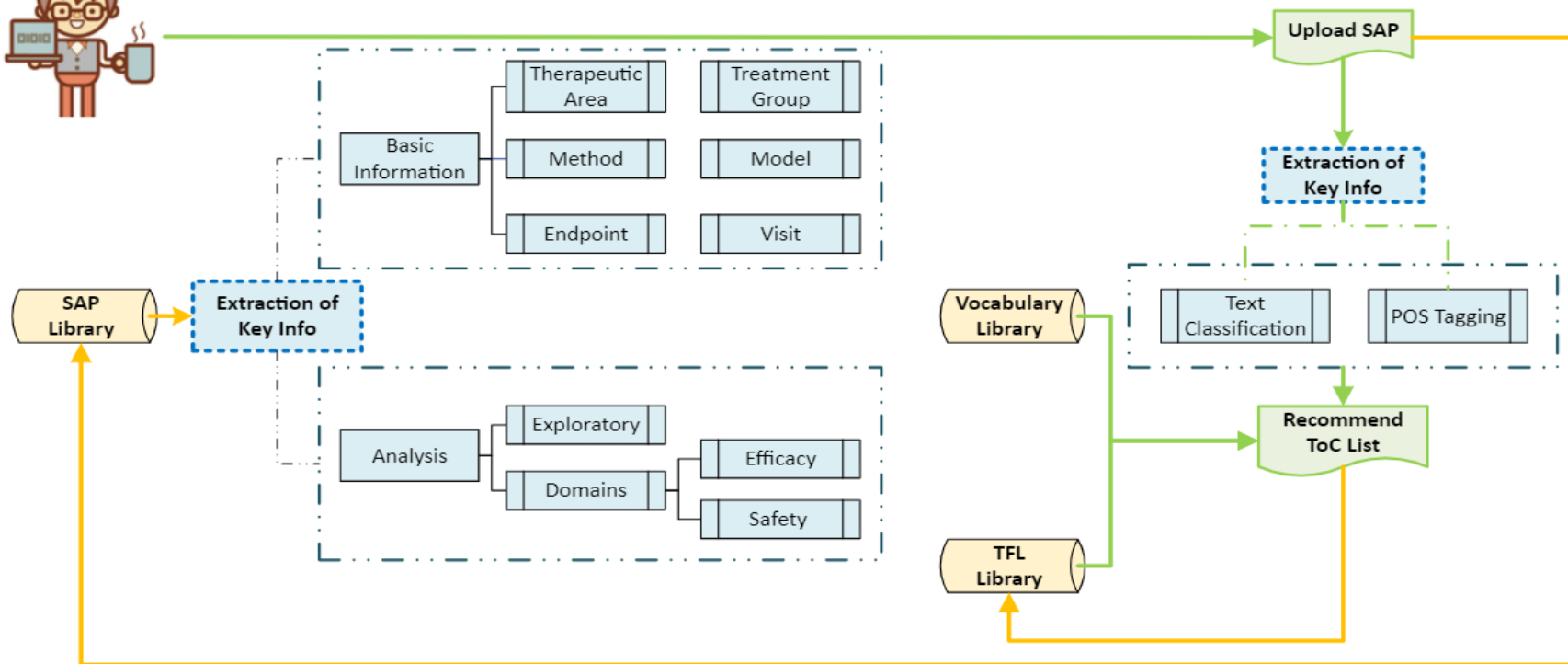
1.1 Study design

1.2 Study objectives, endpoints and estimands

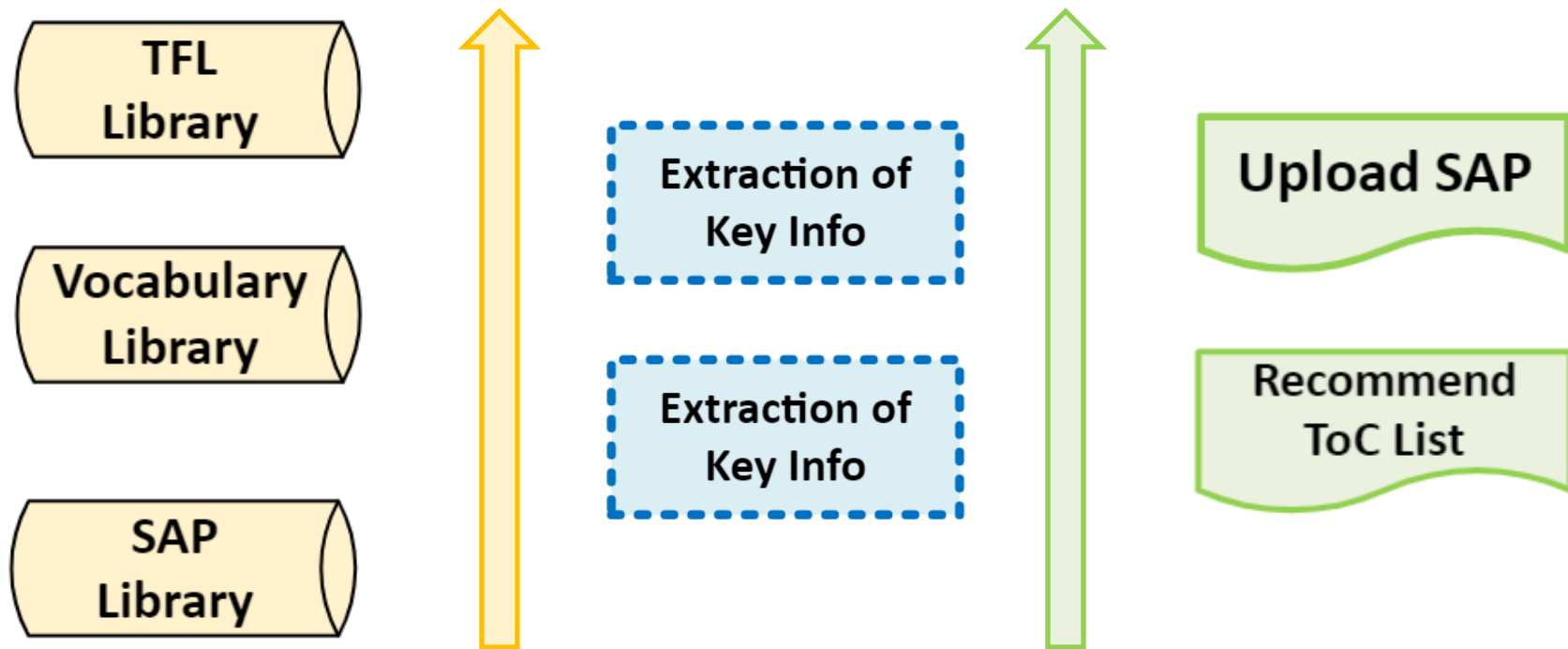
1.2.1 Primary estimand(s)

1.2.2 Secondary estimand(s) Optional

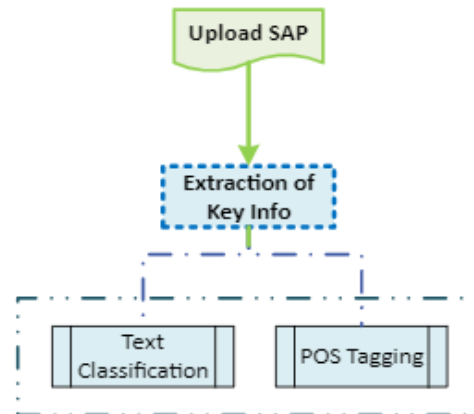
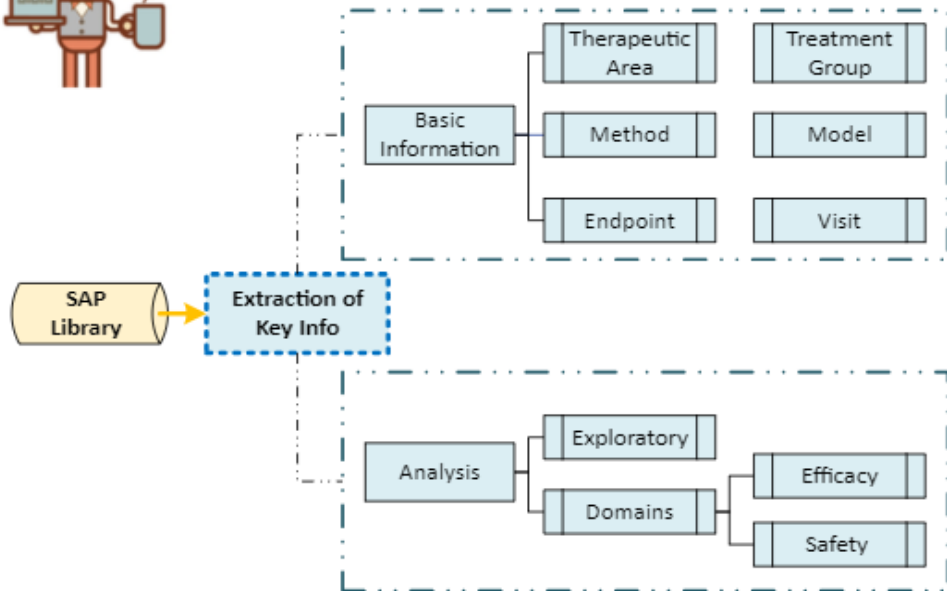
NLP Workflow in ToC



NLP Workflow in ToC



NLP Workflow in ToC



Adverse events (AEs)

Text	Label Group	Vocabulary	Start Position	End Position
The number and percentage of subjects with AEs will be summarized for ocular and		lar	71	76
The number and percentage of subjects with AEs will be summarized for ocular and		and percentage	82	91
The number and percentage of subjects with AEs will be summarized for ocular and		vely	5	25
The number and percentage of subjects with AEs will be summarized for ocular and		ment	93	104
The number and percentage of subjects with AEs will be summarized for ocular and		SOC	107	118
primary SOC and PT term			1	11
primary SOC and PT term		SOC	17	23
primary SOC, PT and maximum severity			1	11
primary SOC, PT and maximum severity			14	15
primary SOC, PT and maximum severity		m severity	21	36


```

file_path = r'C:\...'
sheet_name = 'Sheet1'

df = pd.read_excel(file_path, sheet_name=sheet_name)
df['Start Position'] = None
df['End Position'] = None

for idx, row in df.iloc[2:].iterrows():

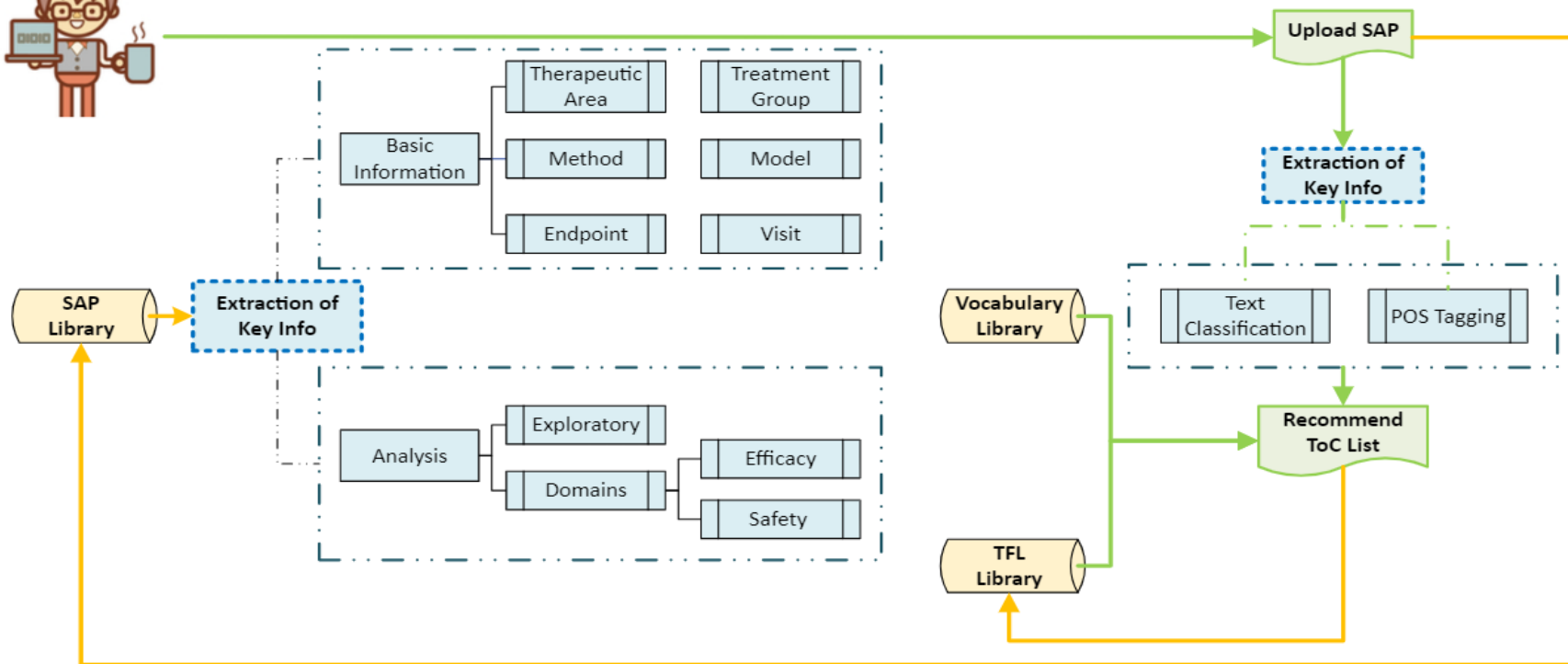
df = df.fillna({'Text': '', 'Vocabulary': '', 'Start Position': 0, 'End Position': 0, 'Label Group': ''})
grouped = df.groupby('Text', group_keys=False).apply(lambda x: x[['Start Position', 'End Position', 'Label Group']].values.tolist()).to_dict()

TRAIN_DATA = [
    (text, {'entities': entities})
    for text, entities in grouped.items()
]

df.at[idx, 'Start Position'] = start_pos+1
df.at[idx, 'End Position'] = end_pos

```

NLP Workflow in ToC





STUDYID	POTYPE	IREF	Title	Keywords Extraction
	Table	14.3.1-15	Overview of adverse events of special interest	Overview,adverse events,special interest
	Table	14.3.1-15s	Overview of adverse events of special interest	Overview,adverse events,special interest
	Listing	14.3.2-6	Adverse events of special interest	Adverse events,special interest
	Table	14.3.1-13	Overview of adverse events of special interest	Overview,adverse events,special interest
	Table	14.3.1-11.1	Overview of adverse events of special interest	Overview,adverse events,special interest
	Table	14.3.1-11.2	Overview of adverse events of special interest	Overview,adverse events,special interest
	Listing	14.3.2-5.1	Adverse events of special interest	Adverse events,special interest
	Listing	14.3.2-5.2	Adverse events of special interest	Adverse events,special interest
	Table	14.3.1-3.1.3	Summary of Adverse events of special interest related to	Summary,Adverse events,special interest,treatment,system organ class, preferred term
	Table	14.3.1-5.1.4	Adverse events of special interest by severity and treatm	Adverse events,special interest,severity,treatment
	Table	14.3.1-1.5	Exposure adjusted incidence rates for adverse events of s	Exposure adjusted incidence rates,adverse events,special interest
	Table	14.3.1-11.1	Absolute and relative frequencies for adverse events of s	Absolute and relative frequencies,adverse events,special interest,Treatment Period
	Table	14.3.1-11.1	Absolute and relative frequencies for adverse events of s	Absolute and relative frequencies,adverse events,special interest,Entire Study Period
	Table	14.3.1-9	Adverse events of special interest by grouping and prefe	Adverse events,special interest,grouping,preferred term
	Listing	14.3.2-6	Adverse events of special interest	Adverse events,special interest
	Table	14.3.1-8	Treatment emergent adverse events of special interest, k	Treatment emergent adverse events,special interest,grouping,preferred term
	Listing	14.3.2-5	Adverse events of special interest	Adverse events,special interest
	Table	14.3.1-5.7	Adverse events of special interest by primary system org	Adverse events,special interest,primary system organ class and preferred term—treatment period
	Table	14.3.1-7.1.5	Absolute and relative frequencies for treatment emerge	Absolute and relative frequencies,treatment emergent adverse events,special interest,safety topic,preferred term, treatment period
	Table	14.3.1-7.1	Absolute and relative frequencies for treatment emerge	Absolute and relative frequencies,treatment emergent adverse events,special interest,safety topic,preferred term, treatment period
	Table	14.3.1-8.1	Exposure-adjusted incidence rate and 95% confidence int	Exposure-adjusted incidence rate,95% confidence intervals,treatment,special interest,safety topic,preferred term, treatment period

Key Libraries/Database

SAP Library

UDYID	filename	heading	section	keyword
		Adverse events (AEs)	Adverse events (AEs)	
		Adverse events (AEs)	Summary tables for AEs will sum	AEs, Summary tables, TEAEs, emergent
		Adverse events (AEs)	Pre-treatment AEs are adverse e	AEs, pre-treatment AEs, treatment
		Adverse events (AEs)	TEAEs are adverse events started	ICF signature, Pre-treatment AEs, adverse events, first administration, study treatment
		Adverse events (AEs)	primary SOC and PT term;	TEAEs, adverse events, events, first administration, preferred term, severity, study treatment, treatment
		Adverse events (AEs)	primary SOC, PT and maximum se	PT term, primary SOC
		Adverse events (AEs)	Standardized MedDRA Query (SM	PT, maximum severity, primary SOC
		Adverse events (AEs)	Separate summaries for TEAEs w	SMQ, Standardized MedDRA Query, preferred term
		Adverse events (AEs)	A subject with multiple adverse	SAE, Separate summaries, TEAEs, death, discontinuation, other significant AEs, serious adverse
		Adverse events (AEs)	A subject with multiple adverse	events, study medication related AEs
		Adverse events (AEs)	All adverse events reporting will	A subject, primary SOC, adverse events, total
		Adverse events (AEs)		MedDRA version
		Adverse events (AEs)		adverse events, by subject, listings

Vocabulary Library

Label	Label Group	Vocabulary
Statistics	Descriptive Statistics	
	Descriptive Statistics	mean
	Descriptive Statistics	median
	Descriptive Statistics	standard deviation
	Descriptive Statistics	variance
	Descriptive Statistics	minimum
	Descriptive Statistics	maximum
	Descriptive Statistics	range
	Descriptive Statistics	interquartile range (IQR)
	Descriptive Statistics	quartiles
	Descriptive Statistics	25th and 75th percentiles
	Descriptive Statistics	confidence interval (CI)

Label	Label Group	Vocabulary
Set Group	Set Group	
	Set Group	Enrolled Set
	Set Group	Evaluable Set
	Set Group	Safety Set
	Set Group	Full Analysis Set
	Set Group	PD Set
	Set Group	Adult Set
	Set Group	Pediatric Set

Model	Model	
	Model	t-test
	Model	chi-square test
	Model	Fisher's exact test
	Model	Wilcoxon signed-rank test
	Model	Mann-Whitney U test
	Model	Kolmogorov-Smirnov test
	Model	ANOVA (analysis of variance)
	Model	ANCOVA
	Model	Kruskal-Wallis test
	Model	Z-test
	Model	linear regression
	Model	logistic regression
	Model	Poisson regression
	Model	Cox proportional hazards model
	Model	mixed-effects model



Recommendation Results

Adverse events (AEs)

Summary tables for AEs will summarize for pre-treatment AEs and treatment emergent AEs (TEAEs) separately:

1. Pre-treatment AEs are adverse events started between ICF signature and prior to the first administration of study treatment.
2. TEAEs are adverse events started after the first administration of study treatment or events present prior to start of the treatment but increased in severity based on preferred term.

The number and percentage of subjects with AEs will be summarized for ocular and non-ocular respectively, by treatment, in the following ways:

- primary SOC and PT term.
- primary SOC, PT and maximum severity.
- Standardized MedDRA Query (SMQ) and preferred term.



Recommendation Results

```
def get_sections(document, filename, specific_headings=None):  
    section_data = {"filename": [], "heading": [], "section": [], "sentence": [], "keyword": []}  
    current_heading = specific_headings  
  
    for para in document.paragraphs:  
        if para.style.name.startswith('Heading'):  
            current_heading = para.text
```

filename	heading	section	sentence	keyword
SAP.docx	Adverse events (AEs)	The number and percentage of subjects with AEs will be summarized for ocular and non-ocular respectively, by treatment, in the following ways:	The number and percentage, ocular, non-ocular	
SAP.docx	Adverse events (AEs)	primary SOC and PT term.	primary SOC, PT term	
SAP.docx	Adverse events (AEs)	primary SOC, PT and maximum severity.	primary SOC, PT, maximum severity	
SAP.docx	Adverse events (AEs)	Standardized MedDRA Query (SMQ) and preferred term.	Standardized MedDRA Query, SMQ, preferred term	

```
        section_data["filename"].append(filename)  
        section_data["heading"].append(current_heading)  
        section_data["section"].append(para.text)  
        section_data["sentence"].append(sentence)  
        keyword = get_SpaCy_Label(sentence)  
        section_data["keyword"].append(keyword)  
  
    return section_data
```



Public Resource — CT.gov

Focus Your Search
(all filters optional)

Study Results ⓘ
☐ With results (67,227)
☐ Without results (446,596)

Study Documents ⓘ
☐ Study protocols (35,570)
☐ Statistical analysis plans (SAPs) (33,419)
☐ Informed consent forms (ICFs) (10,251)

Funder Type ⓘ
☐

Hide

None Selected

	Study Title	NCT Number	Status	Conditions	Interventions
☐ 1	Auricular Neuromodulation for FESS New	NCT06662422	Recruiting	- Functional Endoscopic Sinus Surgery - Vagal Nerve Stimulation - Platelet Activation	Device: Spark Biomedical Transcutaneous Auricular Nerve Stimulation (tAN) System
☐ 2	Support for Bereaved Cancer Caregivers New	NCT06662409	Not yet recruiting	- Bereavement - Cancer - Caregiver 1 more	Other: Nature-based healing (NBHM) intervention
☐ 3	A Study of Diltiazem With a Single Dose of MK-6552 in Healthy Adult Participants (MK-6552-007) New	NCT06662396	Completed	Healthy	- Drug: MK-6552 - Drug: Diltiazem

RSS Manage



Integrated Database

重置所有筛选 研究状态

Help Stanley

Status: ☒ Active ☒ Retired

+ 添加 批量添加 查询 排序 条件颜色 清除 Remove Duplicates 刷新 显示列 快速查询

	#	编辑操作	研究编号	研究主题	研究状态	申办者	适应症	研究分期	招募人数	开始日期	完成日期	
☐	1		NCT05013879	Kinesiotape for Edema After Bilateral Total Knee Arthroplasty	COMPLETED	Montefiore Medical Center	Arthroplasty Complications Arthroplasty, Replacement, Knee	NA	65	2021-10-18	2023-11-24	
☐	2		NCT01402479	An Open-labeled Trial of Ramipril in Patients With Migraine	COMPLETED	Seoul National University Hospital	Migraine With Hypertension	PHASE2	50	2004-10	2005-07	
☐	3		NCT05600179	OCTA in Epivascular Glia After Dex Implant	COMPLETED	Federico II University	Diabetic Retinopathy		38	2021-01-01	2022-09-30	
☐	4		NCT03878979	Preoperative Immune Checkpoint Inhibitor for Patients With Primary Untreated or Recurrent/Metastatic SCCHN	COMPLETED	Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins	Head and Neck Squamous Cell Carcinoma Head and Neck Cancer Head and Neck Cancer Metastatic	PHASE2	26	2019-07-08	2023-10-17	
☐	5		NCT04175379	The Effect of Permissive Hypercapnia on Oxygenation and Post-operative Pulmonary Complication During One-lung Ventilation	UNKNOWN	Yonsei University	Thoracic Surgery	NA	279	2019-11-25	2021-10	
☐	6		NCT03058679	Trial of Specific Carbohydrate and Mediterranean Diets to Induce Remission of Crohn's Disease	COMPLETED	University of Pennsylvania	Crohn Disease	NA	197	2017-09-29	2020-03-01	
☐	7		NCT03877679	The Effect of Topical Curcumin Versus Topical Corticosteroid on Management of Oral Lichen Planus Patients	UNKNOWN	Cairo University	Oral Lichen Planus	PHASE1	40	2019-05-01	2020-06-01	

Total 509398 Items < 1 2 3 4 5 ... 16980 > 30 / page Go to



NLP Technology

- **Recommendation system**
Score element
- **Expand to other areas in statistical programming activities?**
Optimize workflows with text
- **Challenges for an effective recommendation system with NLP?**
Data security
Semantic understanding – the general label

Take Aways



Human Involvement

- **Ideation**

Conscious training of tools.

Importance of innovative ideas and technical design.

- **Manual Text Analysis**

Tedious and time-consuming.

Necessary for thorough and accurate understanding.

Acknowledgements

Yin, Xiaoran

Zhang, Cong

Jiang, Yun

Chen, Feng

Wang, Chengqian



**The Global Healthcare
Data Science Community**

Contact Channels

UK +44 1843 609600
office@phuse.global
phuse.global

Social Media

@phusetwit
/phusebook
/phusetube
/company/phuse

Thank You