ASCO®
AMERICAN SOCIETY OF CLINICAL ONCOLOGY
KNOWLEDGE CONQUERS CANCER

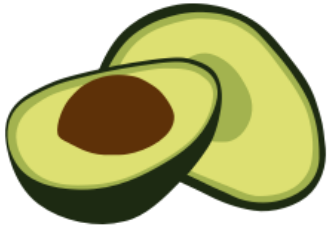
Intern Showcase Presentation 2023

Lia Arakal, Data Science Intern, IT Department

August 10, 2023

Overview of Internship

- **Learned the basics of AWS**
  - S3 – storing data in buckets, ingesting data for ML models
  - SageMaker – creating notebooks and writing code
  - DynamoDB – noSQL database
- **Worked with SQL**
  - Data cleaning with SQL Server
- **Graph Databases**
  - Worked with ArangoDB
- **Text Classification**
  - Machine Learning lifecycle
  - Data preparation in Python
  - Trained and deployed ML models



ArangoDB

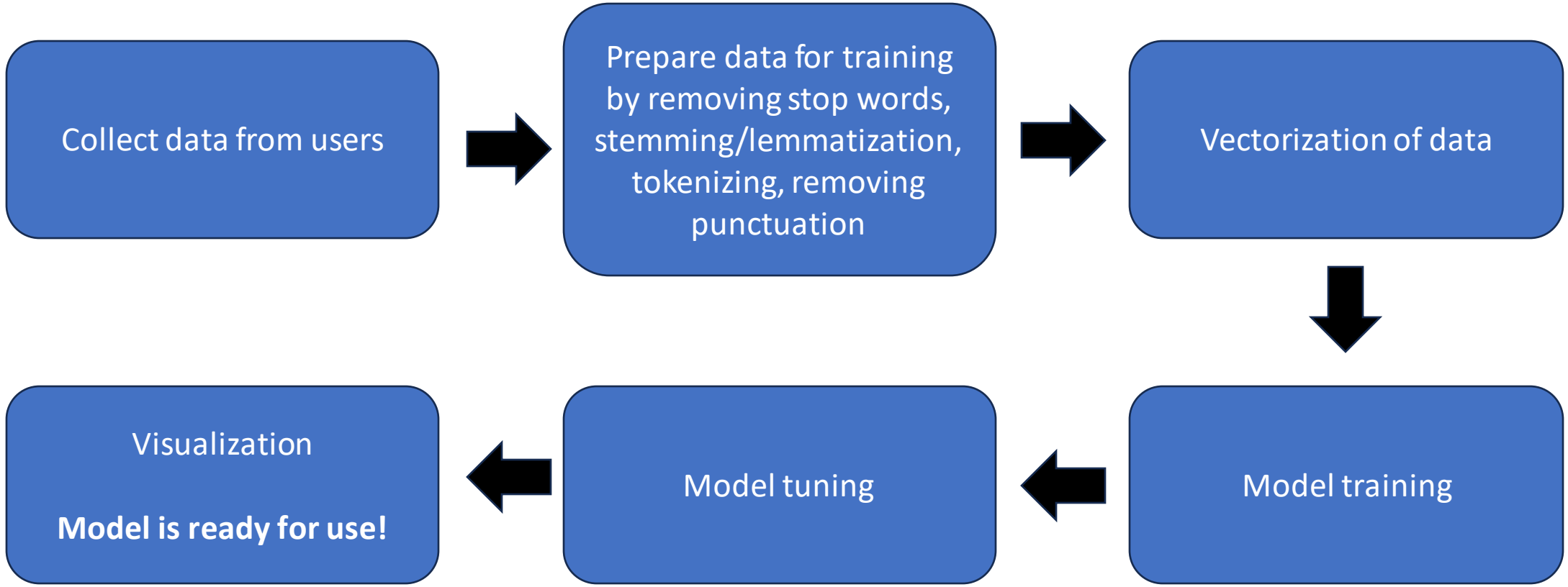
Amazon SageMaker

Main Project: Text Classification

- Given survey feedback from ASCO website
- **Problem**: Manual categorization of feedback was time-consuming
- **Goal:** Create and train a Machine Learning model that can categorize the data for us

feedback	category 1
Everything, super complex, unfriendly & convoluted	General
Better way of navigating to create an itinerary for the meeting	General
Time Zone setting for Chicago instead!	Timezone
show all names on abstracts, increase search functionality, release abstracts sooner	Search
Facilitate the registration to the virtual congress	Meeting Reg
Make it easier to search abstracts	Search
clinical practice guidelines not available	
terrible navigation	General
pages don't load	Technical issue
Leave out the profile page if a longstanding member logins.	Request
Improve search engine	Search
findingwho is attending and when more easily for setting up meets	Meeting Networking
Faster downloads	Technical issue
ability to delete abstract from dashboard and to view entire abstract	Request
Registration for live verses remote attendance	
random glitches	
lagged so badly :(	Technical issue
Can change the time zone from local to Chicago, otherwise it is extremely difficult to plan.	Timezone
Higher percentage to use	
the CommsPolicies email to contact for ASCO Meeting is bouncing back. Permissions email has not responded to attempts to get in touch.	
Very hard to find times for specific presentations on this website	Abstracts/presentations/sessions/program
searchability - for all authors to appear on content they are involved in. And for rich search, so we can search certain words in different fields - institution, author, abstract title, etc.	Search
speed, usability	Technical issue
user interface, searching for things and finding things easily and less	Search
A way to export list of posters or presentations	Request
To be able to search by sponsor on the abstract/publications listings	Search
resore functionality prior to beta version	
I cannnot enter the registration page	Technical issue

Text Classification Lifecycle



Cleansing and Lemmatization

```
documents = []

nlp = spacy.load("en_core_web_sm")
english_stopwords = stopwords.words('english')
stemmer = WordNetLemmatizer()
for sen in range(0, len(feedback)):
    # Remove all the special characters
    document = re.sub(r'\W', ' ', str(feedback[sen]))
    # remove all single characters
    document = re.sub(r'\s+[a-zA-Z]\s+', ' ', document)
    # Remove single characters from the start
    document = re.sub(r'\^[a-zA-Z]\s+', ' ', document)
    # Substituting multiple spaces with single space
    document = re.sub(r'\s+', ' ', document, flags=re.I)
    # Removing prefixed 'b' (not sure if necessary)
    document = re.sub(r'^b\s+', '', document)
    # Converting to Lowercase
    document = document.lower()
     # Lemmatization + remove stopwords
    document = document.split()
    document = [t for t in document if not t in english_stopwords]
    document = [stemmer.lemmatize(word) for word in document]
    document = ' '.join(document)
    d = nlp(document)
    document = " ".join([word.lemma_ for word in d])
    document = document.lower()
    #print(document)
    documents.append(document)
```

making it easier to search through posters, presentations, etc. Separating the different ASCOs (GI, GU, etc.).
Have different tabs for each day of ASCO and within those having separate tabs for poster, abstracts, etc.
Have a search tool that just list out all of the presenters instead of having to search for them
Change it back to the old style! This website doesn't really seem to work properly, and searching for things doe
the search function doesn't clear my original search, I wasn't able to save posters to a collection
searchability of abstracts - its not user friendly at all, in fact its awful
Functionality, filters for things like abstract lists don't work
search function for abstracts/presentations
The searchability of abstracts without having to enter each session individually
More searchable

After cleaning and lemmatization

```
['make easy search poster presentation etc separate different
ascos gi gu etc different tab day asco within separate tab pos
ter abstract etc search tool list presenter instead search',
'change back old style website really seem work properly searc
h thing always find thing definitely', 'search function clear
original search able save poster collection', 'searchability a
bstract user friendly fact awful', 'functionality filter thing
like abstract list work', 'search function abstract presentati
on', 'searchability abstract without enter session individuall
y', 'searchable', 'give downloadable csv file whole abstract l
ist website user friendly filter find need', 'ability search a
sco abstract title', 'make search easy make like google', 'cou
ld put title pdf search navigate difficult', 'get list abstrac
t title rather clunky system require open session separately',
'make easy search', 'locate abstract title search lead submiss
ion guideline', 'often fail page go blank renew membership dep
end phone call helpdesk please simplify automate', 'find recei
pt', 'site load', 'none', 'able select rct', 'make easy access
```

Vectorization and Model Training

Vectorization

In [15]:
```
# initialize vectorizer
vectrz = CountVectorizer()

# run the vectorizer
feedback_matrix = vectrz.fit_transform(documents)

# extract column names
column_names = vectrz.get_feature_names_out()

# total number of columns
len(column_names)

# convert to array
feedback_array = feedback_matrix.toarray()
```

Convert to a more readable form using dataframes

In [16]:
```
# set Pandas to show all columns
pd.set_option('display.max_columns', None)

# convert to the dataframe
df_feedback_postVect = pd.DataFrame(data=feedback_array,columns = column_names)

# see how many rows/columns you get total
df_feedback_postVect.shape
```

Out[16]: (533, 862)

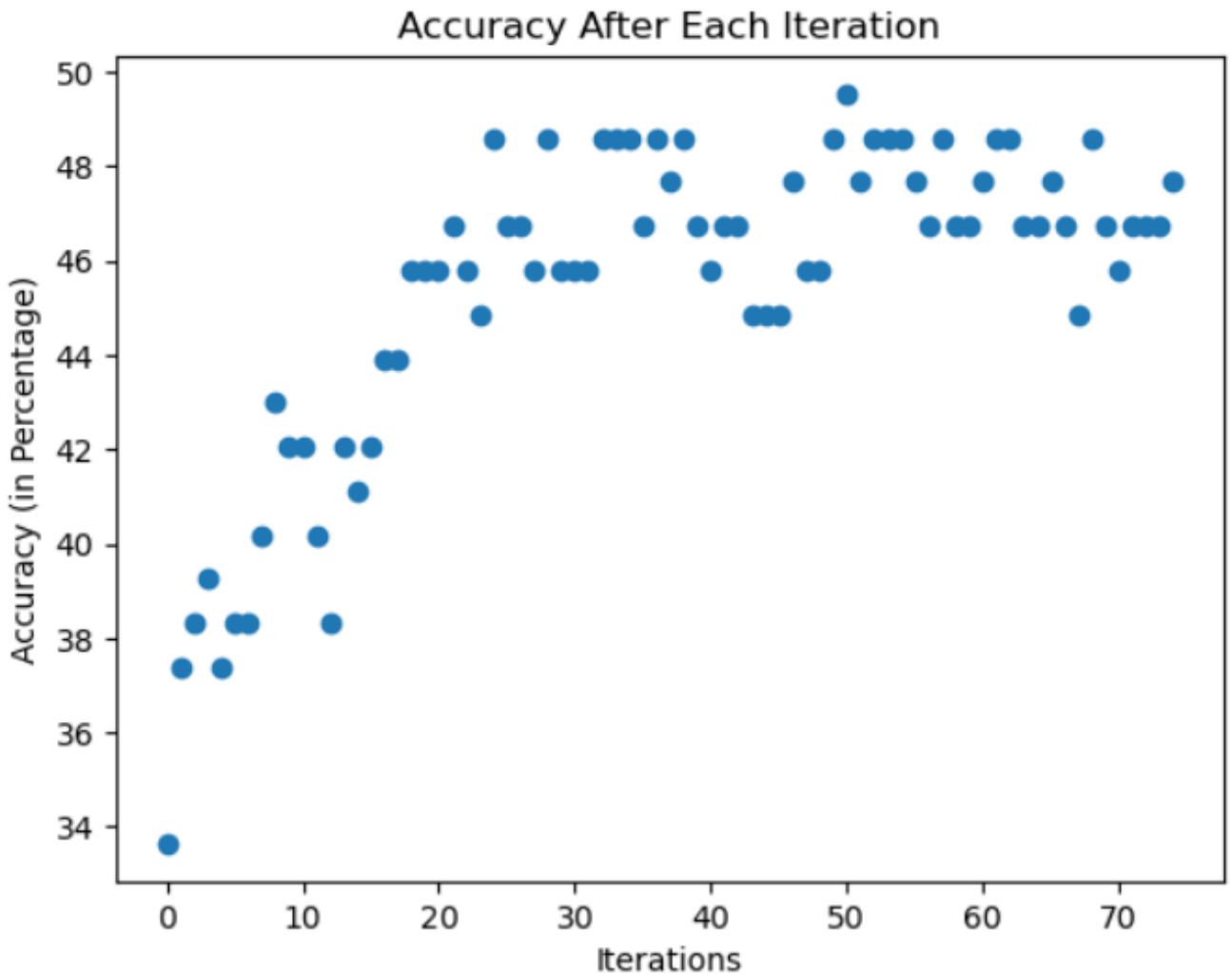
In [17]:
```
df_feedback_postVect.head()
```

Out[17]:

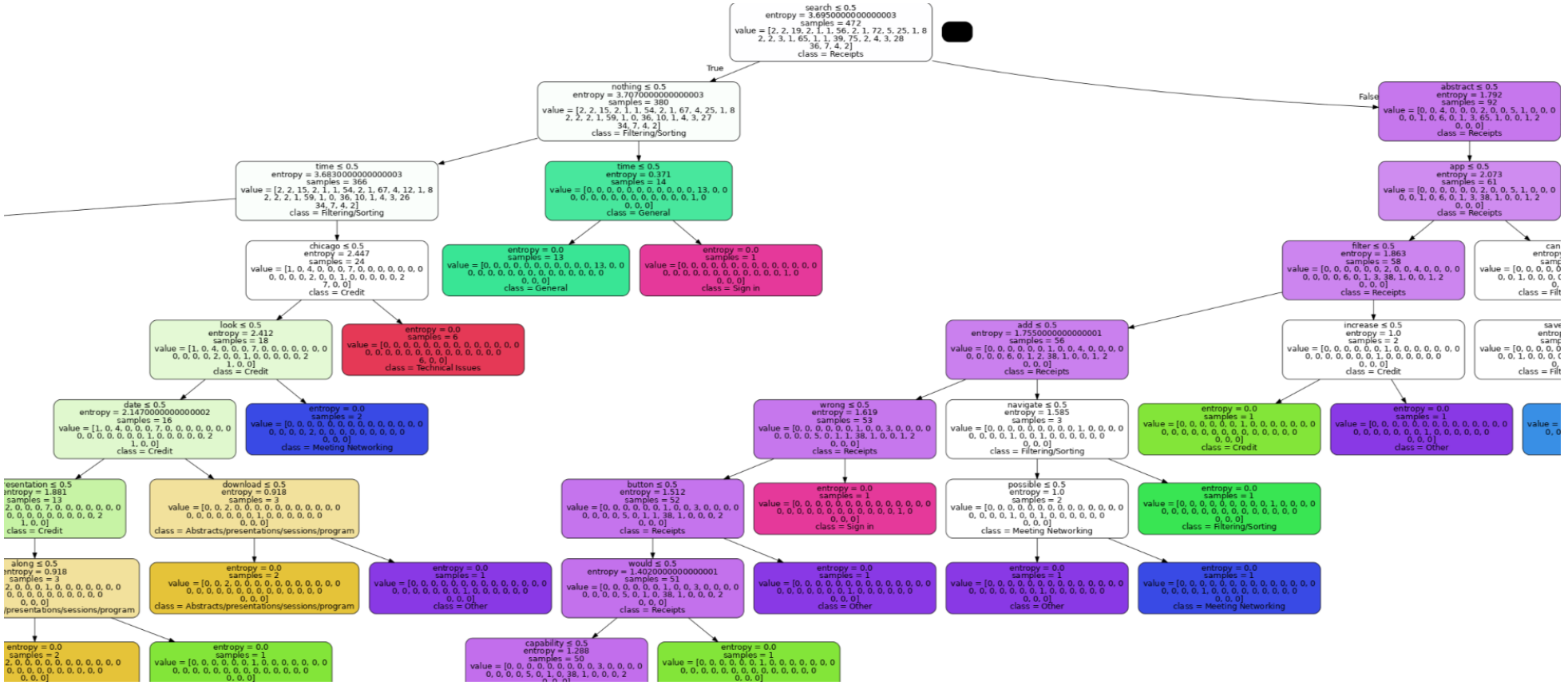
right	run	safari	save	say	schedule	scheduler	scheduleâ	scheduling	science	scientific	scrap	screen	scroll	seach	search	searchability	searchable	s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Model Training

:
```
xArr = []
yArr = []
for i in range(75):
    xArr.append(i)
    i = i + 1
    ### Fitting the model ###
    # Entropy - impurity in a group of examples. Information gain is the decrease in entropy
    # GINI criterion is much faster because it is less computationally expensive.
    # ENTROPY criterion provides slightly better results because it's more computationally intensive

    feedback_cls = DecisionTreeClassifier(max_depth = i, criterion = "entropy", random_state=42)
    feedback_cls.fit(X_train, Y_train)

    #Predict the response for test dataset
    y_pred = feedback_cls.predict(X_test)

    # Calculate accuracy, to see how often the classifier is correct
    from sklearn import metrics
    accuracy = metrics.accuracy_score(Y_test, y_pred)*100
    print("Accuracy for " + str(i) + " iteration: %.2f%%" % accuracy)
    yArr.append(accuracy)
```



Decision Tree



ASCO®
AMERICAN SOCIETY OF CLINICAL ONCOLOGY
KNOWLEDGE CONQUERS CANCER

Questions?

Thank you for a great experience!