Automated Survey Comment Coding with Naïve Bayes & Python

By Tyler Waite Advisory Data Scientist, AI, Automation & Data Platform Finance and Operations

October 2024





What is Naïve Bayes?

- Uses a csv training file of previously coded survey comments to predict how a new statement should be coded.
- Requires minimal knowledge of programming.
- Works best with longitudinal surveys where the same question is asked each time.
- Training file can be used across multiple questions that tend to elicit similar topic categories.
- Can be used for sentiment analysis to determine if comments are positive, negative or neutral.
- Accuracy of comment category prediction should increase as number of comments in training file increases.
- It helps classify more comments faster with human defined categories.

"Naive Bayes is the most straightforward and fast classification algorithm, which is suitable for a large chunk of data. Naive Bayes classifier is successfully used in various applications such as spam filtering, text classification, sentiment analysis, and recommender systems. It uses Bayes theorem of probability for prediction of unknown class." -DataCamp.com

Naïve Bayes Limitations

- Does not work with new questions that lack coded comments from prior surveys.
- It requires a minimum of 3,000 previously coded comments, and categories with at least 100-200 examples.
- While the analysis of new comments is fast, creating and validating the training file can be time consuming.
- It is still necessary to review the output for accuracy. I have found it to have an accuracy of 60-70%.
- It has a bias towards larger comment categories and works best with more conceptually discrete topics.
- It does not split up longer comments that touch on multiple topics.
- Accuracy of categorization can vary depending on the size training file, the size of the category examples, and the accuracy of the coded comments.



Categorization Method Comparison

Topics	Manual Coding (Total = 1547)	Comment Sampling (K-α =.591)	ΚΡΑ (K-α =.436)	Regex (Κ-α =.556)	Naïve Bayes (K-α = .585)
Laptops & Accessories	329	24 70% match	95 28% match	409 81% match	249 85% match
Network / CICD	346	17 45% match	140 33% match	216 36% match	168 68% match
Satisfied	209	19 73% match	210 58% match	191 17% match	118 43% match
Tools	407	40 80% match	140 56% match	480 42% match	357 82% match
IT Support	182	19 54% match	101 31% match	95 18% match	117 59% match
Documentation	18			22 54% match	1 6% match
Not categorized		1420	235	472	0

"Alpha [0.67 - 0.79]: This range is often considered the lower bound for tentative conclusions. A Krippendorff's Alpha in this range suggests moderate agreement; thus, outcomes should be interpreted with concern, questioning the roots of such diverging ratings." - www.k-alpha.org

Creating the Training File

Preparing the training file

- Open Excel
- Paste in comments from previous survey.
- Label first column Topic and second column Comment.
- You can use other column labels if you prefer (e.g., Category, Text), just remember to use those in the python script.

• •	AutoSave 🕥 🕜	S B → C ··· D AAPOR_NB_Training_File ~ Q Sea
Home	Insert Draw Pa	age Layout Formulas Data Review View Automate
Paste	→ X IBM Plex Sans → B I U →	$\begin{array}{c c c c c c c c c c c c c c c c c c c $
A1	🗘 🗶 🗸 🎵 To	pic
	A	B C D E F G
1	Торіс	Comment
2	U.S. NEWS	Woman Who Called Cops On Black Bird-Watcher Loses Lawsuit Against Ex-Employer
3	U.S. NEWS	Reporter Gets Adorable Surprise From Her Boyfriend While Live On TV
4	WORLD NEWS	Puerto Ricans Desperate For Water After Hurricane Fiona\u2019s Rampage
5	CULTURE & A	How A New Documentary Captures The Complexity Of Being A Child Of Immigrants
6	WORLD NEWS	Biden At UN To Call Russian War An Affront To Body's Charter
7	WORLD NEWS	World Cup Captains Want To Wear Rainbow Armbands In Qatar
8	WORLD NEWS	Man Sets Himself On Fire In Apparent Protest Of Funeral For Japan's Abe
9	WORLD NEWS	Fiona Threatens To Become Category 4 Storm Headed To Bermuda
10	TECH	Twitch Bans Gambling Sites After Streamer Scams Folks Out Of \$200,000
11	U.S. NEWS	Virginia Thomas Agrees To Interview With Jan. 6 Panel
12	WORLD NEWS	Russian Cosmonaut Valery Polyakov Who Broke Record With 437-Day Stay In Space Dies
13	CULTURE & A	'Reboot' Is A Clever And Not Too Navel-Gazey Look Inside TV Reboots

1/ SDODTS Maury Wills Base-Stealing Shortston For Dodgers Dies At 80

Creating a Pivot Table

• Select the two columns and click Pivot Table.

	AutoSave 🕥 🕼	A ₽ 7 • C …		(AAPOR_NB_Training_File	~		Q Search
Home	Insert Draw P	age Layout Formulas	Data Review	View Automate				
PivotTa	the Recommended Table	From Pictures	v 🕜 3D Models ∨ v 🛅 SmartArt ∨ 🔄 Screenshot ∨	Checkbox Recommended Charts	0 × 0 × 1 × 0 × 0 × 1 × 0 × 1 × 0 × 0 ×	PivotChart Sparklines	Slicer Slicer	New Comment
A1	\checkmark $f_{\rm X}$ To	opic						
	A	В	С	D	E	F	G	Н
1	Topic	Comment						
2	U.S. NEWS	Woman Who	Called Cops	On Black Bird	Watcher Loses	s Lawsuit Agai	nst Ex-Emplo	yer
3	U.S. NEWS	Reporter Gets	Adorable S	urprise From H	ler Boyfriend V	Vhile Live On 1	ΓV	
4	WORLD NEWS	Puerto Ricans	Desperate	For Water Afte	r Hurricane Fio	na\u2019s Ra	ampage	
5	CULTURE & A	How A New D	ocumentary	Captures The	Complexity Of	Being A Child	Of Immigrant	ts
6	WORLD NEWS	Biden At UN	o Call Russi	an War An Aff	ront To Body's	Charter		
7	WORLD NEWS	World Cup Ca	ptains Want	To Wear Rain	oow Armbands	In Qatar		
8	WORLD NEWS	Man Sets Him	self On Fire	In Apparent P	rotest Of Fune	ral For Japan's	s Abe	
9	WORLD NEWS	Fiona Threate	ns To Becor	ne Category 4	Storm Headed	To Bermuda		
10	TECH	Twitch Bans	Gambling Sit	tes After Stream	mer Scams Fol	ks Out Of \$20	0,000	
11	U.S. NEWS	Virginia Thom	ias Agrees T	o Interview Wi	th Jan. 6 Pane	l		
12	WORLD NEWS	Russian Cosn	10naut Valer	y Polyakov Wł	o Broke Recor	d With 437-Da	ay Stay In Spa	ce Dies /
13	CULTURE & A	'Reboot' Is A	Clever And N	Not Too Navel-	Gazey Look Ins	side TV Reboo	ts	
14	SPORTS	Maury Wills, I	Base-Stealin	g Shortstop Fo	or Dodgers, Die	s At 89		
15	WORLD NEWS	4 Russian-Co	ntrolled Ukr	ainian Regions	Schedule Vote	es This Week 1	Fo Join Russia	ı
16	WORLD NEWS	Fiona Barrels	Toward Tur	ks And Caicos	Islands As Cat	egory 3 Hurric	cane	
17	U.S. NEWS	American Airl	ines Flyer C	harged, Banne	d For Life After	Punching Flig	ght Attendant	On Video
18	COMEDY	23 Of The Fu	niest Tweet	ts About Cats A	And Dogs This A	Week (Sept. 1	7-23)	

Using the Pivot table

- Drag the Topic column into the Rows and Values area.
- Look at the number of comments in each topic.

A3	$ $ \times \checkmark f_{X} Row Labels			•
	A	В	PivotTable Fields	8
1			FIELD NAME	Q Search fields
2			Tania	
3	Row Labels 🛛 🕞	Count of Topic	Comment	
4	BLACK VOICES	49		
5	BUSINESS	51		
6	COMEDY	188	0	•
7	CRIME	125	Y Filters	III Columns
8	CULTURE & ARTS	40		
9	EDUCATION	10		
10	ENTERTAINMENT	1139		
11	ENVIRONMENT	102		
12	FOOD & DRINK	102		
13	HOME & LIVING	118	E Rows	Σ Values
14	IMPACT	22	: Topic 🕜	: Count of Topic 🍈
15	LATINO VOICES	1		
16	MEDIA	90	••	
17	MONEY	42		
18	PARENTING	113		
19	POLITICS	2376	Drag fields	between areas
P	Sheeti AAFON_NB_Irai	mig_rile T		

Sorting the category count

- Click on a cell that contains a topic count.
- Select "Sort" and "Sort Smallest to Largest."

Count of Topic Row Labels BLACK VOICES 19 Copy жC BUSINESS # 1 Format Cells... COMEDY Number Format... Refresh CRIME Sort > CULTURE & ARTS Select > EDUCATION Formulas > ENTERTAINMENT Remove "Count of Topic" ENVIRONMENT Summarize Values By > FOOD & DRINK Show Values As > HOME & LIVING Show Details IMPACT Value Field Settings... PivotTable Options... LATINO VOICES ✓ Hide Field List MEDIA AutoFill > MONEY Services > PARENTING 113 POLITICS 2376 OUEER VOICES 24

Sort Smallest to Largest

Sort Largest to Smallest

More Sort Options...

Combining Categories

- Look at the number of comments in each topic.
- Look for smaller categories that can be combined into larger categories.
- Look for large categories that you might want to split.
- Category counts should reflect typical distribution of topics, but small categories are less likely to be predicted.



Save the training file

- Once you are happy with your categories delete the sheet with the Pivot Table.
- Be sure to save the file as csv.

Cloud			Save As: AAPOR_NB_Training_File	
Contract Con	<>		AAPOR Mini Conference	Q Search
Tyler's Mac	Pictures	AAPOR Mi	ni Conference > AAPOR Mini Cn Matrix.ipynb	
Tags	Poll.atp	AAPOR Mi	ni Conference.zip	
	🚞 postgres_data	Decision T	ree > AAPOR_10_7eadlines.ipynb	
Favorites	Public	GenAl	> AAPOR_CateHeadlines.csv	
Applications	public.pem	MLCourse	> AAPOR_Confredictions.csv	
🚍 Desktop	PycharmProjects	> 📄 venv	> AAPOR_NB_Training_File.csv	
Documents	pythonProject	>	AAPOR_Workeywords.ipynb	
Downloads		Online Locations	File Format: CSV UTF-8 (Comma delimited) (.csv)	Θ
tyler_waite				_
Creative Clo			Options	
	New Folder			Cancel Save

Anaconda, NLTK, Scikit-Learn & Jupyter Notebooks

Installing Anaconda

 Anaconda is a free open-source online application that you can use to launch a Jupyter notebook.

ANACONDA

Solutions Resources Partners Company

Distribution

Products

Free Download*

Register to get everything you need to get started on your workstation including Cloud Notebooks, Navigator, Al Assistant, Learning and more.

Easily search and install thousands of data science, machine learning, and Al packages

Manage packages and environments from a desktop application or work from the command line

Opploy across hardware and software platforms

Oistribution installation on Windows, MacOS, or Linux

 Use of Anaconda's Offerings at an organization of more than 200 employees requires a Business or Enterprise license. See Pricing



https://www.anaconda.com/download

Using Anaconda online

 Anaconda now has an online version of jupyter notebooks you can use if you don't want to install any libraries on your computer.

 After creating an account with Anaconda sign in and select the "Code Online" option. ○ A == https://anaconda.cloud

ANACONDA.

C

Home Get Started Learning Notebooks C Community Packages

Welcome to Anaconda Cloud



Installing NLTK

 If you want to launch Jupyter Notebooks from the Anaconda application, you will need to install some libraries.

• The NLTK is a natural language processing library that we will use to help categorize the comments.

Mac/Unix

1. Install NLTK: run pip install --user -U nltk

2. Install Numpy (optional): run pip install --user -U numpy

3. Test installation: run python then type import nltk

For older versions of Python it might be necessary to install setuptools (see https://pypi.python.org/pypi/ setuptools) and to install pip (sudo easy_install pip).

Windows

These instructions assume that you do not already have Python installed on your machine.

32-bit binary installation

1. Install Python 3.8: https://www.python.org/downloads/ (avoid the 64-bit versions)

2. Install Numpy (optional): https://numpy.org/install/

3. Install NLTK: https://pypi.python.org/pypi/nltk

4. Test installation: Start>Python38, then type import nltk

https://www.nltk.org/install.html

Installing SciKit-Learn libraries

 Python is a powerful language because there are so many libraries that can be used to perform various functions on your data.

Installing the latest release

pip	conda
Install conda using the <u>Anaconda or miniconda installers</u> or required for any of those). Then run:	the miniforge installers (no administrator permission
\$ conda create –n sklearn–env –c conda–forge sciki \$ conda activate sklearn–env	t-learn
In order to check your installation, you can use:	
<pre>\$ conda list scikit-learn # show scikit-learn ver \$ conda list</pre>	rsion and location ackages in the environment s()"

https://scikit-learn.org/stable/install.html

Installing Numpy

 Numpy should be installed with Anaconda, but if for some reason it isn't try these commands. CONDA

If you use conda, you can install NumPy from the defaults or conda-forge channels:

Best practice, use an environment rather than install in the base env conda create -n my-env conda activate my-env # If you want to install from conda-forge conda config --env --add channels conda-forge # The actual install command conda install numpy

PIP

If you use pip, you can install NumPy with:

pip install numpy

https://numpy.org/install/

Jupyter Notebooks

Launching Jupyter Notebook

- Once Anaconda is installed you should be able to open it and launch a Jupyter notebook.
- Jupyter Notebooks are where we will be creating the Python script.



Creating a new Python notebook

- Navigate to the folder where you want to keep your Python scripts.
- Select the New dropdown menu and Python 3 (ipykernel).

💭 Jupyter	Quit Logout		
Files Running Clusters			
Select items to perform actions on them.	Upload New - 2		
0 v pythonProject / AAPOR_Mini_Conference	Name Votebook: Python 3 (ipykernel)		
۵.	Other: Create a new notabook with Putton 2 /inukarnall		
AAPOR Mini Course Confusion Matrix.ipynb	Runn Text File kB		
APOR_10_7_AI_Comment_Categorization-NB-with-Tokenizer-Headlines-Copy1.ipynb	Folder kB		
APOR_10_7_AI_Comment_Categorization-NB-with-Tokenizer-Headlines.ipynb	Runr KB		

Exploring the Jupyter Notebook Interface

- The Jupyter Notebook interface is very similar in many ways to a text editor.
- You can cut, copy and paste.
- You can search and replace text.
- You can reorder sections of the script.
- You can add sections for explanatory text (markdown) or code.
- When you want to run the code in a section you press Run.



Markdown text

• With Jupyter Notebooks you can add explanatory text outside of the code sections and format it using Markdown.

This is a level 1 heading

This is a level 2 heading

This is some plain text that forms a paragraph. Add emphasis via **bold** or __bold__, and *italic* or _itali

Paragraphs must be separated by an empty line.

* Sometimes we want to include lists.* Which can be bulleted using asterisks.

1. Lists can also be numbered.

2. If we want an ordered list.

[It is possible to include hyperlinks](https://www.dataquest.io)

Inline code uses single backticks: `foo()`, and code blocks use triple backticks:

bar()

Or can be indented by 4 spaces:

...

foo()

And finally, adding images is easy: ![Alt text](https://www.dataquest.io/wp-content/uploads/2023/02/DQ-Logo.s

https://www.dataquest.io/blog/jupyter-notebook-tutorial/

Markdown text

• Press play to view your formatted text.

This is a level 1 heading

This is a level 2 heading

This is some plain text that forms a paragraph. Add emphasis via **bold** or **bold**, and *italic* or *italic*.

Paragraphs must be separated by an empty line.

- · Sometimes we want to include lists.
- Which can be bulleted using asterisks.
- Lists can also be numbered.
 If we want an ordered list.

It is possible to include hyperlinks

Inline code uses single backticks: foo(), and code blocks use triple backticks:

bar()

Or can be indented by 4 spaces:

foo()

And finally, adding images is easy:



DATAQUEST

https://www.dataquest.io/blog/jupyter-notebook-tutorial/

Writing code

- Use a # If you want to add a comment or prevent a line of code from being run.
- All code in a section will be run at the same time.
- Breaking up code into smaller chunks can help with bug checks and validating that data was loaded properly.
- Python remembers the most recent value of the variables in each section that has been run.
- The number in brackets to the left of the code section tells you the order in which the sections have been run.

In [1]:	print('Hello World!')
	Hello World!
In [2]:	<pre>import time time.sleep(3)</pre>
In [3]:	<pre>def say_hello(recipient): return 'Hello, {}!'.format(recipient) say_hello('Tim')</pre>
Out[3]:	'Hello, Tim!'
In [4]:	<pre>import numpy as np def square(x): return x * x</pre>
In [5]:	<pre>x = np.random.randint(1, 10) y = square(x) print('%d squared is %d' % (x, y))</pre>
	7 squared is 49

https://www.dataquest.io/blog/jupyter-notebook-tutorial/

Creating the script

Loading the libraries

• Don't change these libraries.

• Don't worry about the pink section after you click run.

	# Python librari	ies used to run the scripts. Do not change.
	import os	
	import io	
	import numpy	
	import pandas as	s pd
	import re	
	<pre>import os.path</pre>	
	from pandas impo	ort DataFrame
	from sklearn.fea	ature_extraction.text import CountVectorizer
	from sklearn.nat	ive_bayes import MultinomialNB
	<pre>from sklearn.pip</pre>	beline import Pipeline
	import nltk	
	nltk.download('a	all')
	from nltk.corpus	s import stopwords
	from nltk.stem i	import WordNetLemmatizer
1	[nltk_data] Down	nloading collection 'all'
	[nltk_data]	
	[nltk_data]	Downloading package abc to
	[nltk_data]	/Users/tyler_waite/nltk_data
	[nltk_data]	Package abc is already up-to-date!
	[nltk_data]	Downloading package alpino to
	[nltk_data]	/Users/tyler_waite/nltk_data
	[nltk_data]	Package alpino is already up-to-date!
	[nltk_data]	Downloading package averaged_perceptron_tagger to
	[nltk_data]	/Users/tyler_waite/nltk_data
	[nltk_data]	Package averaged_perceptron_tagger is already up-
	[nltk_data]	to-date!
	[nltk_data]	Downloading package averaged_perceptron_tagger_eng to
	[nltk_data]	/Users/tyler_waite/nltk_data
	[nltk_data]	Package averaged_perceptron_tagger_eng is already
	[nltk_data]	up-to-date!
	[nltk_data]	Downloading package averaged_perceptron_tagger_ru to
	[nltk_data]	/Users/tyler_waite/nltk_data
	[nltk_data]	Package averaged_perceptron_tagger_ru is already
	[nltk_data]	up-to-date!

Copying file pathname

 An easy way to copy a files path name is to right click or control click on the file name at the bottom of the finder window, then click Copy.

	$\langle \rangle$ tyler_waite 😬 🗄			… ~ ☆ ⊘ … ~	Q
Shared	Name		Size	Date Modified	Kind
ocations	v pythonProject			Today at 5:38 PM	Folder
Tulor's M	AAPOR_Mini_Conference			Today at 5:28 PM	Folder
	AAPOR_Categorized_Headlines2.csv		173 KB	Today at 7:24 PM	comma
ans.	AAPOR_10_7_AI_Commentith-Tokenizer-Headlines.i	oynb	47 KB	Today at 5:28 PM	Documer
.8.	Uncategorized_headlines.csv		83 KB	Today at 5:24 PM	comma
avorites	AAPOR_NB_Training_File.csv		596 KB	Today at 5:18 PM	comma
AirDrop	AAPOR_10_7_AI_Commentenizer-Headlines-Copy1.i	pynb	195 KB	Sep 5, 2024 at 10:02 AM	Documer
Annlientie	Tools-Keywords.ipynb		162 KB	Jul 22, 2024 at 11:29 PM	Documer
Applicatio	coded_comments.csv	92 KB	Jul 22, 2024 at 11:24 PM	comma	
Desktop	AAPOR_Workshop-Tools-Keywords.ipynb		159 KB	Jul 16, 2024 at 8:06 AM	Documer
Recents	AAPOR Mini Course Confusion Matrix.ipynb		100 KB	Jul 15, 2024 at 3:08 PM	Documer
10001110	AAPOR_Confusion_Matrix_Headline_Predictions.csv		2 KB	Jul 15, 2024 at 3:07 PM	comma
Documents	AAPOR_Categorized_Headlines.csv		596 KB	Jul 15, 2024 at 9:34 AM	comma
Downloads	AAPOR Mini Conference.zip		440 KB	Jul 16, 2024 at 8:09 AM	ZIP archi
tudas unaita	> MLCourse			Jul 15, 2024 at 3:04 PM	Folder
tyler_waite	> Decision Tree			Jul 12, 2024 at 8:34 AM	Folder

Open		
Open With		>
Show in Enclosing Fo	older	
Get Info		
Copy "AAPOR_NB_T	raining File.csv" as Pathname	
Open URL in BibDes	k Web G	

Loading the training file

• Replace the file name with the path to your training file.

 If you want to see more rows, put the number of rows you want between the parenthesis: head(15). Loads file containing categorized comments from previous surveys. Make sure file is in the same folder as your python script or use full pathname for file.

comments_df = pd.read_csv(r"AAPOR_NB_Training_File.csv")

Use this to view the dataframe so you can make sure the comments were correctly loaded.

comments_df.head()

	Topic	Comment
0	U.S. NEWS	Woman Who Called Cops On Black Bird-Watcher Lo
1	U.S. NEWS	Reporter Gets Adorable Surprise From Her Boyfr
2	WORLD NEWS	Puerto Ricans Desperate For Water After Hurric
3	CULTURE & ARTS	How A New Documentary Captures The Complexity
4	WORLD NEWS	Biden At UN To Call Russian War An Affront To

Vectorizing the training file

 This is the Naïve Bayes process that creates the prediction values for the categories from the training file.

This creates the comment classifier.

vectorizer = CountVectorizer()

When using this with your own files, update 'Comment' to match the column name for your comments.

counts = vectorizer.fit_transform(comments_df['Comment'].values)
classifier = MultinomialNB()

When using this with your own files, update 'Topic' to match the column name for your categories or topics.

targets = comments_df['Topic'].values
classifier.fit(counts, targets)

MultinomialNB

MultinomialNB()

Loading the uncategorized comments

 If you have demographic data associated with the comments, it is OK to include those columns in the uncategorized file. The Naïve Bayes category column is appended to the existing columns in the uploaded file. This loads the file containing the new uncategorized comments. Make sure file is in the same folder as your python script or use full pathname for file.

new_comments_df = pd.read_csv(r"Uncategorized_headlines.csv")

View the dataframe to make sure the comments were correctly loaded.

new_comments_df.head()

Comment

0	brown labels democrats new slogan awful
1	victorians honoured with queens birthday awards
2	charges to be laid over barcaldine train crash
3	jim morrison surfaces in pre doors short film
4	scientists converge on nt for eclipse

This removes all rows without a comment

new_comments_df['Comment'].fillna(0, inplace = True)
new_comments_df = new_comments_df[new_comments_df.Comment != 0]

Lemmatization

 This takes the comments and removes words like "of" "or" "and" "the" which are very common and not useful for categorization. It also converts all capital letters to lowercase and removes endings like "ing" "ed" to convert words to their root form which will be easier to match with categorized text.

 This helps improve matching new comments to similar previously categorized comments.

```
text = list(new_comments_df['Comment'])
lemmatizer = WordNetLemmatizer()
corpus = []
for i in range(len(text)):
    r = re.sub('[^a-zA-Z]', ' ', text[i])
    r = r.lower()
    r = r.split()
    r = [word for word in r if word not in stopwords.words('english')]
    r = [lemmatizer.lemmatize(word) for word in r]
    r = ' '.join(r)
    corpus.append(r)
```

This puts the lemmatized comment into a new column so that you still have the original, easier to read, version.

new_comments_df['lemm_text'] = corpus
new_comments_df.head()

lemm_text	Comment
brown label democrat new slogan awful	brown labels democrats new slogan awful
victorian honoured queen birthday award	victorians honoured with queens birthday awards
charge laid barcaldine train crash	charges to be laid over barcaldine train crash
jim morrison surface pre door short film	jim morrison surfaces in pre doors short film
scientist converge nt eclipse	scientists converge on nt for eclipse

Outputting the categorized comments

 This converts the lemmatized new comments into an array and a predicted category classification is assigned.

 Once the categorization of the new comments is complete it writes the results to a new column that is added to the data frame with the comments.

• The last step outputs the data frame to a csv file. This writes the dataframe to an array

<pre>new_comments = new_comments_df['Comment'].to_numpy() new_comments</pre>
array(['brown labels democrats new slogan awful', 'victorians honoured with queens birthday awards', 'charges to be laid over barcaldine train crash',, 'wheatbelt towns propose amalgamation', 'shoppers boost property trusts bottom line', 'elton john becomes a father'], dtype=object)
Assigns predicted categories to comments
new_comments_counts = vectorizer.transform(new_comments) new_categories = classifier.predict(new_comments_counts) new_categories
array(['POLITICS', 'ENTERTAINMENT', 'U.S. NEWS',, 'POLITICS', 'STYLE & BEAUTY', 'ENTERTAINMENT'], dtype=' <u14')< td=""></u14')<>
This adds a column with the new predicted topic category.
categories_df = pd.DataFrame(new_categories, columns=[' <mark>Topics</mark> ']) categorized_comments = categories_df.join(new_comments_df)
This outputs comment classifications to csv

categorized_comments.to_csv(r"AAPOR_Categorized_Headlines.csv", index=False)

Viewing the results

Reviewing the categorization

Notice how Lemmatization changed the text by removing words like "with", "to", "be", "over", "in" and "on."

• You can delete this column if you like.

Topics	Comment	lemm_text
POLITICS	brown labels democrats new slogan awful	brown label democrat new slogan awful
ENTERTAINMENT	victorians honoured with queens birthday awards	victorian honoured queen birthday award
U.S. NEWS	charges to be laid over barcaldine train crash	charge laid barcaldine train crash
ENTERTAINMENT	jim morrison surfaces in pre doors short film	jim morrison surface pre door short film
POLITICS	scientists converge on nt for eclipse	scientist converge nt eclipse

Reviewing the categorization

- Notice how the NB was biased towards the larger categories.
- Notice how the top four categories map to the top four predicted categories.
- Notice how only categories with over 100 examples had any predictions.
- This is why it is important to roll up smaller categories in your training data into larger groups and why you may want to split up large categories into smaller sets.

\mathbf{O}	rio	inal	Cateo	iories
$\mathbf{\nabla}$	IЦ	mai	Cuice	

Row Labels	💶 Count of	Торіс		
POLITICS		2376		П
U.S. NEWS		1142	$\langle $	ĸ
ENTERTAINME	NT	1139	$\langle \rangle$	
WORLD NEWS		839	$\langle \langle \rangle$	Ρ
COMEDY		188		
SPORTS		173	$\sqrt{\sqrt{2}}$	U
STYLE & BEAU	ΓY	144		
CRIME	-	125	\ \ \¥	W
HOME & LIVIN	G	118	$\langle \rangle \langle \rangle$	
WELLNESS		113	$\langle \langle \langle \rangle \rangle$	F
PARENTING		113	$\langle \rangle$	-
-OOD & DRINK	_	102		S
ENVIRONMENT		102	11R	5
WEIRD NEWS		95		D
		90		Г
		6Z		C -
		10		Э
		49	///	0
CUI TURE & AR	TS	40		C
SCIENCE	10	24		-
OUEER VOICES		24	*	\E
		22		\
TECH		19		Η
RELIGION		14		
TRAVEL		13		(t
EDUCATION		10		V
LATINO VOICE	S	1		G
WEDDINGS		1		
(blank)				
Grand Total		7231		

Predicted Categories

	Row Labels	Count of Topics
	POLITICS	969
	U.S. NEWS	517
	WORLD NEWS	340
	ENTERTAINMENT	160
	SPORTS	5
	PARENTING	2
	STYLE & BEAUTY	2
	COMEDY	2
Ž	ENVIRONMENT	1
	HOME & LIVING	1
	(blank)	
	Grand Total	1999

Improving the accuracy of the training file

Confusion Matrix

- The Confusion Matrix takes a sample of the training file and then tries to predict what the category should be.
- It then creates a table that show you how many predictions were correct and how many predictions were assigned to other categories.
- This can help you identify categories that might need closer review to make sure the comments are correctly categorized.

#This imports the python libraries that will be used in this script. Do not change.

import matplotlib.pyplot as plt import numpy as np import pandas as pd from sklearn import metrics from sklearn import datasets, svm from sklearn.metrics import ConfusionMatrixDisplay from sklearn.model_selection import train_test_split

#Loads the training file containing categorized comments. Replace the pathname with the pathname to your file. data = pd.read_csv('AAPOR_Categorized_Headlines.csv', encoding='latin-1')

#Column names from your training file. Change 'Headline' and 'Category' to match your data file if needed.
data.columns = ['Category', 'Headline']

#This lets you view your imported data. Change number to view more or fewer rows. data.head(10)

Category	Headline
U.S. NEWS Woman V	Vho Called Cops On Black Bird-Watcher Lo
U.S. NEWS Repo	rter Gets Adorable Surprise From Her Boyfr
ORLD NEWS Pue	to Ricans Desperate For Water After Hurric
TURE & ARTS How A N	ew Documentary Captures The Complexity
ORLD NEWS Bide	en At UN To Call Russian War An Affront To
ORLD NEWS World Cu	p Captains Want To Wear Rainbow Armban
ORLD NEWS Man	Sets Himself On Fire In Apparent Protest O
ORLD NEWS Fiona T	hreatens To Become Category 4 Storm Hea
TECH Twitch	Bans Gambling Sites After Streamer Scam
U.S. NEWS Virg	inia Thomas Agrees To Interview With Jan

Frequency Chart

 Creating a bar chart of the comment category frequencies shows that there are a few large categories. # This checks the frequency of the categories in the training file. # Because NB uses probability, predictions will be skewed towards larger categories. # Change 'Category' to match your column name.

data['Category'].value_counts(normalize = True).plot.bar()



Predicting the training categories

- This code randomly selects 33% of the training file to put into the test data.
- It then vectorizes the test file to predict the categories using the remaining training data.
- The predicted category is then compared with the original category.

<pre>X = data['Headline'] y = data['Category']</pre>	
<pre># train test split (66% train - 33% test) from sklearn.model_selection import train_test_split X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=123)</pre>	
<pre>print('Training Data :', X_train.shape) print('Testing Data : ', X_test.shape)</pre>	
Training Data : (4844,) Testing Data : (2387,)	
# Train Bag of Words model. Do not change.	
<pre>from sklearn.feature_extraction.text import CountVectorizer cv = CountVectorizer()</pre>	
X_train_cv = cv.fit_transform(X_train) X_train_cv.shape	
(4844, 10272)	
# Training Logistic Regression model. Do not change this block.	
from sklearn.linear_model import LogisticRegression	
lr = LogisticRegression() lr.fit(X_train_cv, y_train)	
<pre># transform X_test using CV X_test_cv = cv.transform(X_test)</pre>	
<pre># generate predictions predictions = lr.predict(X_test_cv)</pre>	
<pre>#predictions predictions_df = pd.DataFrame(data = predictions) predictions_df.head(20)</pre>	

This creates the feature and label sets. Only change 'Heading'and 'Category' to match your file.

Labeling the rows and columns

- When outputting the matrix you need to name all the columns.
- The column names should be in alphabetical order even if that is not the order of the columns in the original file.
- The index (rows) names should match the order of the column names.

# i	nis creates the confusion matrix table. ndex and columns text should match the categories in your training file.
# T # T	ne same names should be used in both the index and columns section.
# I # T	te category names should be in alphabetical order.
n I	ie tast category shouth not have a ,
df	<pre>= pd.DataFrame(metrics.confusion_matrix(y_test,predictions),index=['BLACK VOICES','BUSINESS','COMEDY','CRIME','CULTURE & ARTS','EDUCATION','ENTERTAINMENT','ENVIRONMENT', 'FOOD & DRINK','HOME & LIVING','IMPACT','MEDIA','MONEY','PARENTING','POLITICS','QUEER VOICES','RELIGION', 'SCIENCE','SPORTS','STYLE & BEAUTY','TECH','TRAVEL','U.S. NEWS','WEIRD NEWS','WELLNESS','WOMEN', 'WORLD NEWS'</pre>
],	columns=[
1,	columns=['BLACK VOICES','BUSINESS','COMEDY','CRIME','CULTURE & ARTS','EDUCATION','ENTERTAINMENT','ENVIRONMENT', 'FOOD & DRINK','HOME & LIVING','IMPACT','MEDIA','MONEY','PARENTING','POLITICS','QUEER VOICES','RELIGION', 'SCIENCE','SPORTS','STYLE & BEAUTY','TECH','TRAVEL','U.S. NEWS','WEIRD NEWS','WELLNESS','WOMEN', 'WORLD NEWS'

Interpreting the Confusion Matrix

- The rows are the original categories. The columns are the predicted categories.
- The higher the number at the intersection of the category column and row pairs the more accurate the prediction.
- If the count is low look for high counts in other categories (e.g., Comedy).

	BLACK VOICES	BUSINESS	COMEDY	CRIME	CULTURE & ARTS	EDUCATION	ENTERTAINMENT	ENVIRONMENT	FOOD & DRINK	HOME & LIVING	 SCIENCE	SPORTS
BLACK VOICES	0	0	0	0	0	0	7	0	0	0	 0	1
BUSINESS	0	1	0	0	0	0	1	0	0	0	 0	0
COMEDY	0	0	29	0	0	0	20	1	0	0	 0	0
CRIME	0	0	0	5	0	0	0	0	0	0	 0	0
CULTURE & ARTS	0	0	0	0	0	0	9	o	0	0	 0	0
EDUCATION	0	0	0	0	0	0	0	0	0	0	 0	0
ENTERTAINMENT	1	0	9	0	0	0	248	0	1	1	 0	4
ENVIRONMENT	0	0	0	0	0	0	3	0	0	0	 0	0
FOOD & DRINK	0	0	0	0	0	0	5	O	10	0	 0	0
HOME & LIVING	0	0	0	0	0	0	5	O	1	27	 0	0
IMPACT	0	0	0	0	0	0	0	o	1	0	 0	0
MEDIA	0	1	1	0	0	0	3	0	0	0	 0	0
MONEY	0	0	0	0	0	0	2	0	2	0	 0	0
PARENTING	0	0	0	0	0	0	8	0	1	1	 0	0

Outputting the Confusion Matrix

- This script will output the matrix to a csv.
- When we look at the four largest categories in Excel and highlight the cells with the largest counts, we see that the majority of headlines in those categories were correctly predicted.
- If we then take the count at the intersection of the categories and divide it by the total for the column we can see the overall accuracy.
- The "U.S. News" category in the training file should probably be reviewed to see if the categorization was accurate.

This saves the confusion matrix to a csv file.

Change the pathname to where you want to save the file and the file name.

Divide the number where the row and column names match by the total count for the column to determine accuracy.

matrix = df
matrix.to_csv(r"AAPOR_Confusion_Matrix_Headline_Predictions.csv", index=True)

BLACK VOICES 7 3 5 0 BUSINESS 1 2 5 0 COMEDY 20 9 0 0 CRIME 0 7 25 4 CULTURE & ARTS 9 3 5 0 EDUCATION 0 2 4 0 ENTERTAINMENT 248 45 37 10 ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 QUEER VOICES 2 1 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 US. NEWS 31 91<		ENTERTAINMENT	POLITICS	U.S. NEWS	WORLD NEWS
BUSINESS 1 2 5 0 COMEDY 20 9 0 0 CRIME 0 7 25 4 CULTURE & ARTS 9 3 5 0 EDUCATION 0 2 4 0 ENTERTAINMENT 248 45 37 10 ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 STYLE & BEAUTY 12 7 5 0 US. NEWS 31 91	BLACK VOICES	7	3	5	0
COMEDY 20 9 0 0 CRIME 0 7 25 4 CULTURE & ARTS 9 3 5 0 EDUCATION 0 2 4 0 ENTERTAINMENT 248 45 37 10 ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 U.S. NEWS 31 91	BUSINESS	1	2	5	0
CRIME 0 7 25 4 CULTURE & ARTS 9 3 5 0 EDUCATION 0 2 4 0 ENTERTAINMENT 248 45 37 10 ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 US. NEWS 31 91	COMEDY	20	9	0	0
CULTURE & ARTS 9 3 5 0 EDUCATION 0 2 4 0 ENTERTAINMENT 248 45 37 10 ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 QUEER VOICES 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 US. NEWS 31 <td< td=""><td>CRIME</td><td>0</td><td>7</td><td>25</td><td>4</td></td<>	CRIME	0	7	25	4
EDUCATION 0 2 4 0 ENTERTAINMENT 248 45 37 10 ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 QUEER VOICES 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 <t< td=""><td>CULTURE & ARTS</td><td>9</td><td>3</td><td>5</td><td>0</td></t<>	CULTURE & ARTS	9	3	5	0
ENTERTAINMENT 248 45 37 10 ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 US. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOREN D NEWS 3 <td< td=""><td>EDUCATION</td><td>0</td><td>2</td><td>4</td><td>0</td></td<>	EDUCATION	0	2	4	0
ENVIRONMENT 3 8 23 6 FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 US. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOREN D NEWS 3 4	ENTERTAINMENT	248	45	37	10
FOOD & DRINK 5 8 3 0 HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 US. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOREN N 3 4 4 0	ENVIRONMENT	3	8	23	6
HOME & LIVING 5 2 2 1 IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOMEN 3 4 4 0	FOOD & DRINK	5	8	3	0
IMPACT 0 2 0 0 MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	HOME & LIVING	5	2	2	1
MEDIA 3 8 3 0 MONEY 2 4 3 1 PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 US. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOMEN 3 4 4 0	IMPACT	0	2	0	0
MONEY 2 4 3 1 PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELINESS 1 4 1 2 WOMEN 3 4 4 0	MEDIA	3	8	3	0
PARENTING 8 5 1 2 POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOMEN 3 4 4 0	MONEY	2	4	3	1
POLITICS 21 670 69 22 QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOMEN 3 4 4 0	PARENTING	8	5	1	2
QUEER VOICES 2 1 2 0 RELIGION 0 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOMEN 3 4 4 0	POLITICS	21	670	69	22
RELIGION 0 1 1 SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WOMEN 3 4 4 0	QUEER VOICES	2	1	2	0
SCIENCE 3 0 2 0 SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	RELIGION	0	0	1	1
SPORTS 16 6 12 8 STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	SCIENCE	3	0	2	0
STYLE & BEAUTY 12 7 5 0 TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	SPORTS	16	6	12	8
TECH 0 3 4 1 TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	STYLE & BEAUTY	12	7	5	0
TRAVEL 1 0 0 0 U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	TECH	0	3	4	1
U.S. NEWS 31 91 221 27 WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	TRAVEL	1	0	0	0
WEIRD NEWS 8 5 15 0 WELLNESS 1 4 1 2 WOMEN 3 4 4 0	U.S. NEWS	31	91	221	27
WELLNESS 1 4 1 2 WOMEN 3 4 4 0 WORLD NEWS 16 40 55 178	WEIRD NEWS	8	5	15	0
WOMEN 3 4 4 0 WORLD NEWS 16 40 55 179	WELLNESS	1	4	1	2
WORLD NEWS 16 40 55 179	WOMEN	3	4	4	0
WORLD NEWS 10 40 55 176	WORLD NEWS	16	40	55	178
Total 425 939 507 263	Total	425	939	507	263
Accuracy 58% 71% 44% 68%	Accuracy	58%	71%	44%	68%

Contact

Tyler Waite Advisory Data Scientist CIO, AI, Automation & Data Platform Tyler.Waite@ibm.com

More reading on Python, Naïve Bayes and comment coding

2022, November). Understanding Text Classification in Python [Review of Understanding Text Classification in Python]. Data Camp. <u>https://www.datacamp.com/tutorial/text-classification-python</u>

Friedman, R., Dankin, L., Hou, Y., Aharonov, R., Katz, Y., & Slonim, N. (2021, November 1). Overview of the 2021 Key Point Analysis Shared Task (K. Al-Khatib, Y. Hou, & M. Stede, Eds.). ACLWeb; Association for Computational Linguistics. <u>https://doi.org/10.18653/v1/2021.argmining-1.16</u>

How to interpret a confusion matrix for a machine learning model. (n.d.). Www.evidentlyai.com. <u>https://www.evidentlyai.com/classification-metrics/confusion-matrix</u>

Kane, F., & Kane, F. (2020). Machine Learning, Data Science and Deep Learning with Python. Udemy; Udemy. <u>https://www.udemy.com/course/data-science-and-machine-learning-with-python-hands-on/</u>

Krippendorff's Alpha Calculator (K-Alpha) - Official Website. (n.d.). Www.k-Alpha.org. Retrieved May 8, 2024, from <u>https://www.k-alpha.org/krippendorffs-alpha-calculator-k-alpha-official-website</u>

MeasuringU: How to Code & Analyze Verbatim Comments. (2017). Measuringu.com. <u>https://measuringu.com/code-verbatim/</u>

pandas.Series.str.contains — pandas 1.4.2 documentation. (n.d.). Pandas.pydata.org. <u>https://pandas.pydata.org/docs/reference/api/pandas.Series.str.contains.html</u>

Reaching saturation point in qualitative research. (2016, July 21). Quirkos Blog. <u>https://www.quirkos.com/blog/post/saturation-qualitative-research-guide/</u>

S, Y. (2020, May 8). An Introduction to Naïve Bayes Classifier. Medium. https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf

Pryke, B. (2019, August 22). *Jupyter Notebook for Beginners Tutorial*. Dataquest. <u>https://www.dataquest.io/blog/jupyter-notebook-tutorial/</u>

Keyword Search

 This script searches the comments for terms that are frequently found in survey comment responses.

- If a comment contains one of the keywords the comment is assigned a category label and written to a data frame for that category.
- This script assigns both primary and secondary categories to comments.
- If a comment does not match any of the defined categories it is placed in the "Other" category so that the researcher can review those comments to identify new topics or new string patterns to add to the existing categories.
- This script is useful for responses that contain specific product names.

Import Python libraries import pandas as pd import numpy as np import re

Reading csv file into a data frame df = pd.read_csv(r"comment_file.csv")

#How to add column to data frame if you want to df['Survey Period'] = '1Q24'

#Removes all rows without a comment df['Comment_Column'].fillna(0, inplace = True)

 $df = df[df.Comment_Column != 0]$

#Remove unwanted columns from data frame and only keeps the ones specified between the brackets df = df.reindex(columns=['Survey Period', 'response_id', 'Comment_Column'])

#Finds comments that mention these products and copies them to a new data frame with their product label. anaconda_df = all_surveys_df[all_surveys_df['Tools_Needed'].str.contains('anaconda', na=False, case=False)] anaconda_df['Topic'] = 'Anaconda'

angular_df = all_surveys_df[all_surveys_df['Tools_Needed'].str.contains('angular', na=False, case=False)] angular_df['Topic'] = 'Angular'

#Merging all the topic data frames together

frames = [anaconda_df, angular_df]

#Writes labeled comments to csv

coded_tools_df = pd.concat(frames)
coded_tools_df.to_csv(r"/coded_comments.csv",index=False)

#Identifies comments that were not coded using existing keyword scripts so script can be updated to find new keywords

cond = all_surveys_df['Tools_Needed'].isin(coded_tools_df['Tools_Needed'])
all_surveys_df.drop(all_surveys_df[cond].index, inplace = True)
uncat_df = all_surveys_df
uncat_df['Topic'] = 'Other'
uncat_df.to_csv(r"/uncategorized_comments.csv",index=False)

Writes all comments to single csv file

frames = [coded_tools_df, uncat_df]
all_comments_df = pd.concat(frames)
all_comments_df = all_comments_df.reindex(columns=['Survey Period', 'response_id', 'Comment_Column', 'Topic'])
all_comments_df.to_csv(r"/Tools_Needed_all.csv", index=False)

Exploring Jupyter Notebooks

• Go to: <u>https://jupyter.org/try</u> <u>-jupyter/lab/</u>

Intro.ipynb						•	+				
Э	+	\approx	Ō	۵	۲		С	**	Markdown ~	Ŷ	

Notebook 🖸 Python (Pyodide) 🖸 🗏

Introduction to the JupyterLab and Jupyter Notebooks

This is a short introduction to two of the flagship tools created by the Jupyter Community.

JupyterLab 🧪

JupyterLab is a next-generation web-based user interface for Project Jupyter. It enables you to work with documents and activities such as Jupyter notebooks, text editors, terminals, and custom components in a flexible, integrated, and extensible manner. It is the interface that you're looking at right now.

For an overview of the JupyterLab interface, see the JupyterLab Welcome Tour on this page, by going to Help -> Welcome Tour and following the prompts.

See Also: For a more in-depth tour of JupyterLab with a full environment that runs in the cloud, see the JupyterLab introduction on Binder.

Jupyter Notebooks 📓

Jupyter Notebooks are a community standard for communicating and performing interactive computing. They are a document that blends computations, outputs, explanatory text, mathematics, images, and rich media representations of objects.

JupyterLab is one interface used to create and interact with Jupyter Notebooks.

For an overview of Jupyter Notebooks, see the JupyterLab Welcome Tour on this page, by going to Help -> Notebook Tour and following the prompts.

See Also: For a more in-depth tour of Jupyter Notebooks and the Classic Jupyter Notebook interface, see the Jupyter Notebook iPython tutorial on Binder.

An example: visualizing data in the notebook 🚼

Below is an example of a code cell. We'll visualize some simple data using two popular packages in Python. We'll use NumPy to create some random data, and Matplotlib to visualize it.

Note how the code and the results of running the code are bundled together.