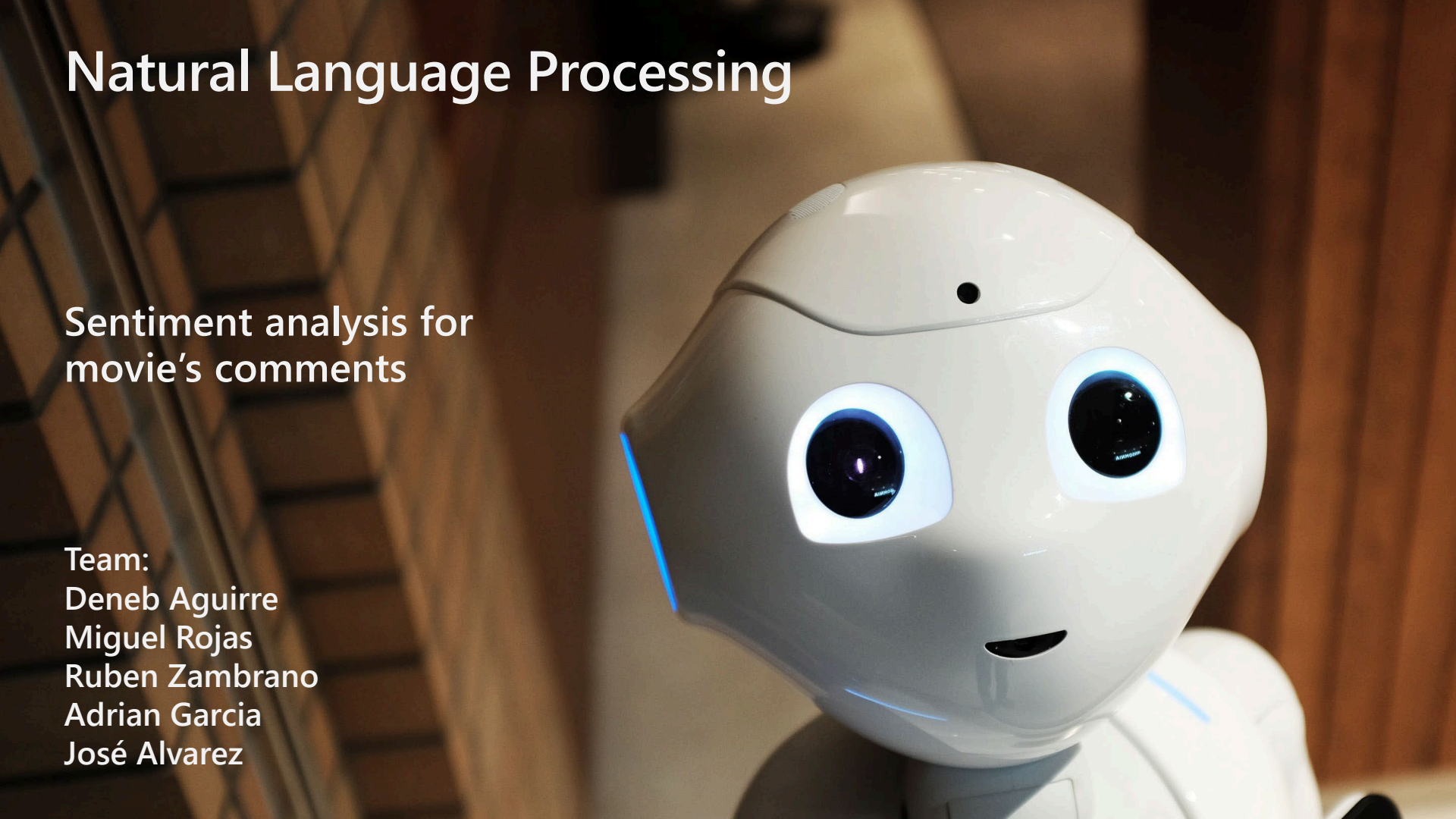
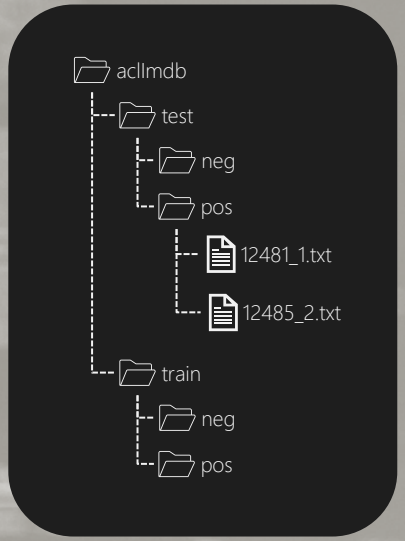
Natural Language Processing



Sentiment analysis for movie's comments

Team:
Deneb Aguirre
Miguel Rojas
Ruben Zambrano
Adrian Garcia
José Alvarez

Database Construction

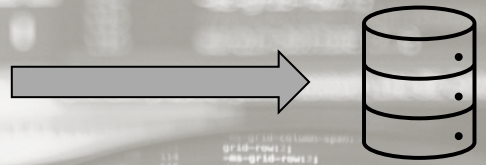


'Sorry everyone,,, I know this is supposed to be an "art" film,, but wow, they should have handed out guns at the screening so people could blow their brains out and not watch. Although the scene design and photographic direction was excellent, this story is too painful to watch. The absence of a sound track was brutal. The loooonnnnng shots were too long. How long can you watch two people just sitting there and talking? Especially when the dialogue is two people complaining. I really had a hard time just getting through this film. The performances were excellent, but how much of that dark, sombre, uninspired, stuff can you take? The only thing i liked was Maureen Stapleton and her red dress and dancing scene. Otherwise this was a ripoff of Bergman. And i\'m no fan f his either. I think anyone who says they enjoyed 1 1/2 hours of this is,, well, lying.'

'Sorry everyone I know this is supposed to be an art film but wow they should have handed out guns at the screening so people could blow their brains out and not watch Although the scene design and photographic direction was excellent this story is too painful to watch The absence of a sound track was brutal The loooonnnnng shots were too long How long can you watch two people just sitting there and talking Especially when the dialogue is two people complaining I really had a hard time just getting through this film The performances were excellent but how much of that dark sombre uninspired stuff can you take The only thing i liked was Maureen Stapleton and her red dress and dancing scene Otherwise this was a ripoff of Bergman And im no fan f his either I think anyone who says they enjoyed 1 12 hours of this is well lying'

	text	label
0	It sounds a bit awkward to call a film about w...	1
1	This had a great cast with bigname stars like ...	0
2	This movie is so bad it hurts The car doing 30...	0
3	Justine cannot find the perfect mate to make h...	0
4	For some reason my fatherinlaw gave me a copy ...	0
...	...	...
49995	Disappearance is set in the Mojave desert as J...	0
49996	Although theres Flying Guillotines as part of ...	0
49997	I thought this was a very good movie Someone s...	1
49998	Almost no information is available about this ...	1
49999	The laughs are few and far between in this dul...	0

50000 rows × 2 columns



Sentiment analysis – Naive Bayes

Naive Bayes: Is a **Probabilistic Classifie**r algorithm based on Bayes' Theorem. Well know as a ML model for **Sentiment Analysis.**

Tokenize
stopremove
Hashing
idf

Naive Bayes

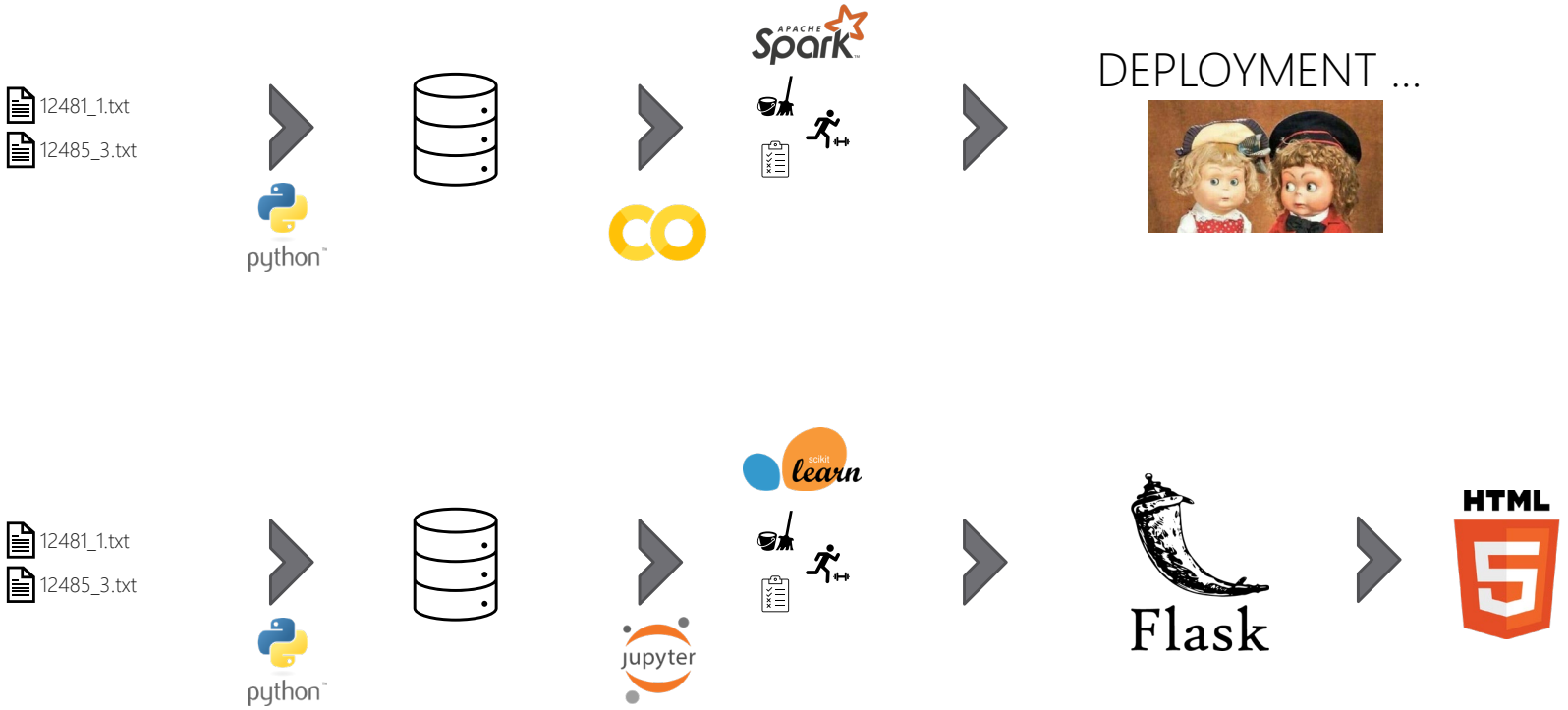
Accuracy:
0.8178
(Test data)

Inputs:
50,000 labeled records as (+/-).
70 % of records were used for model training, 30 % for testing.
Data Cleanse process included: Drop Nulls, Stop Words, Steaming.
Pyspark Lib was used to create NaiveBayes() model.
GColab, Pyspark, aws s3 storage tools were used.


Now what?

Thought process

Pivoting

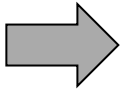

12481_1.txt
12485_3.txt
python
APACHE
Spark
DEPLOYMENT ...
12481_1.txt
12485_3.txt
python
jupyter
scikit
learn
Flask
HTML

Preprocessing in SKlearn

- **CountVectorizer()** to convert a collection of text documents to a matrix of token counts

Out[5]:

	text
0	Hello everybody
1	my name is Jose Alvarez everybody
2	my age is 32 years and have been living in CDM...

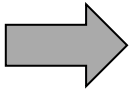


Out[28]:

	32	age	alvarez	cdmx	everybody	hello	jose	living	years
0	0	0	0	0	1	1	0	0	0
1	0	0	1	0	1	0	1	0	0
2	2	1	0	1	1	0	0	1	2

Out[1]:

	text	label
0	This movie has some things that are pretty ama...	1
1	Duchess and her three kittens are enjoying the...	1
2	The Class is a comedy series that portrays a b...	0
3	Latter days is the best gay movie of the homos...	1
4	There is part of one sequence where some water...	0
...	...	...
49995	This movie succeeds at being one of the most u...	0
49996	There is a reason Chairman of the Board got a ...	0
49997	My Favorite part was when the credits started ...	0
49998	Jack Frost no kids its not the warm hearted fa...	0
49999	Has there ever been a movie more charming than...	1



	Token 1	Token 2	Token 3	...	...	Token 108083
0	X	X	X	...	...	X
1	X	X	X	...	...	X
2	X	X	X	...	...	X
....	...	...	...	...	...	...
49997	X	X	X	...	...	X
49998	X	X	X	...	...	X
49999	X	X	X	...	...	X

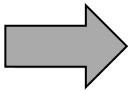
Preprocessing in SKlearn

- **TfidfTransformer()** Transform a count matrix to a normalized tf or tf-idf representation Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency. This is a common term weighting scheme in information retrieval, that has also found good use in document classification.

$$tf - idf_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

Out[28]:

	32	age	alvarez	cdmx	everybody	hello	jose	living	years
0	0	0	0	0	1	1	0	0	0
1	0	0	1	0	1	0	1	0	0
2	2	1	0	1	1	0	0	1	2



Out[30]:

	32	age	alvarez	cdmx	everybody	hello	jose	living	years
0	0.000000	0.000000	0.000000	0.000000	0.508542	0.861037	0.000000	0.000000	0.000000
1	0.000000	0.000000	0.652491	0.000000	0.385372	0.000000	0.652491	0.000000	0.000000
2	0.593683	0.296841	0.000000	0.296841	0.175319	0.000000	0.000000	0.296841	0.593683

Model Training

- *SGDClassifier()* it is a linear classifiers which can use SVM, logistic regression, among others. For our model we went with the SVM as a default of the *SGDClassifier()*

- We also divided our data set with a train_test_split() and train our model

- Test score: 0.89384

Results and Deployment

- Flask was used to interact between the app and the Front interface
- The app collects the comment from the front Interface then runs the comment through out model
- Flask finally delivers the result of the app to the Front interface
- Alternatives

What’s next?

Fine tune the model by:

- Add stop words accordingly to experience
- Use more data
- SDGClasiffier parameter tunning. Using different:
  - function i.e. SVG, Logistic regresión, etc.
  - Learning rate i.e. constant, optimal, adaptive…
  - Etc.

Resources

Web: Flask, html, css

Data Wrangling: Python Pandas

ML: Python SkLearn

Explore: Google Colab and pySpark