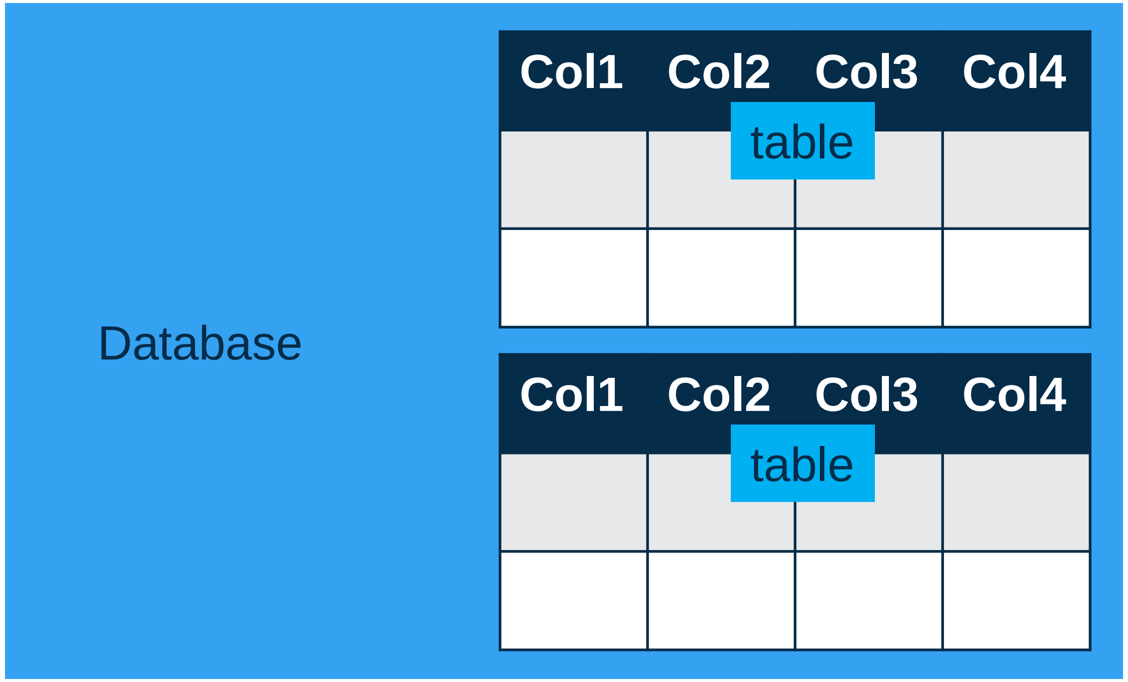
MSDS600 W8: Sentiment analysis on social media data

REGIS UNIVERSITY

Review from W7

- APIs, wrappers, SDKs
- SQL and SQLite3
- Text EDA
- Reddit PRAW API wrapper in Python



Database

table

Col1	Col2	Col3	Col4

table

Col1	Col2	Col3	Col4

Our Social Media Analysis Plan

- Collect data from the Reddit API using the PRAW Python package.
- Collect data from at least one subreddit where we want to analyze the sentiment of the text.
- Store the data in a SQLite3 database file on our computer.
- Load the data and analyze it using sentiment analysis to understand what the sentiment in the subreddit looks like.

https://www.translatemedia.com/us/blog-usa/machine-translation-multilingual-sentiment-analysis-projects/

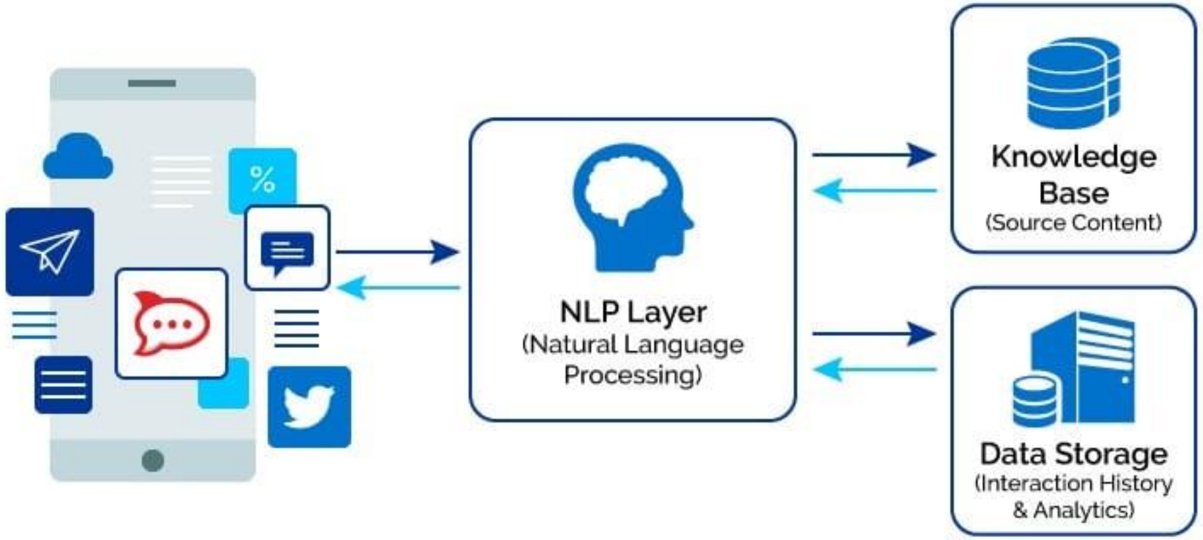
REGIS UNIVERSITY

Natural language processing (NLP)

- NLP is a field that some data scientists specialize in.
- It is using computers to understand human language.
- Contains things like chat bots, language translation, text classification, topic modeling, sentiment analysis, information retrieval, and more.
- Advanced NLP often uses neural networks.

Ethical issues:

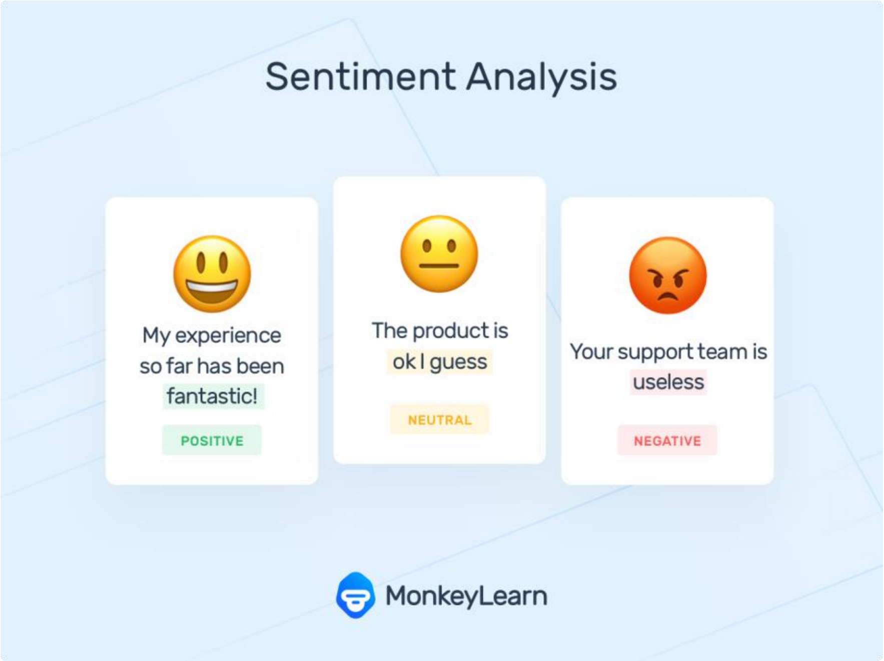
- Creating chat bots based on deceased people: https://www.washingtonpost.com/technology/2021/02/04/chat-bots-reincarnation-dead/
- Bias in chat bots


NLP Layer
(Natural Language Processing)
Knowledge Base
(Source Content)
Data Storage
(Interaction History & Analytics)
%

https://research.aimultiple.com/natural-language-platforms/

Sentiment analysis

- NOT “sentimental” analysis
- Measures the positivity and negativity of text
  - Quantified, often from a value of -1 (negative) to +1 (positive), with 0 being neutral
- Strategies:
  - Keyword analysis: match words to + or – scores from a dictionary
  - Machine learning: train a classifier on text features (e.g. word counts or other more advanced methods like TFIDF or word vectors)
  - Rule-based: combine keywords with rules (e.g. negation from the word ‘not’)


Sentiment Analysis
My experience so far has been fantastic!
POSITIVE
The product is ok I guess
NEUTRAL
Your support team is useless
NEGATIVE
MonkeyLearn

https://monkeylearn.com/sentiment-analysis/

Sentiment analysis

- We can use it to evaluate customer satisfaction, check reactions to products, policies, services, or other things through social media or other text data.
- We can also use it with a chat bot to adapt the bot to the person’s mood.
- It could be used with customer support systems.

Sentiment Analysis

My experience so far has been fantastic!

The product is ok I guess

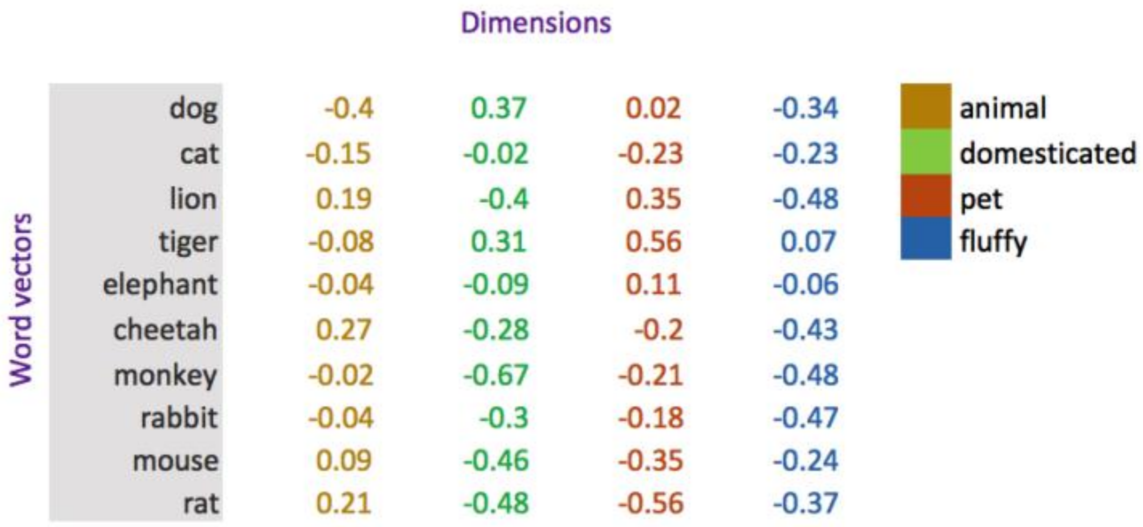
Your support team is useless

POSITIVE
NEUTRAL
NEGATIVE
MonkeyLearn

https://monkeylearn.com/sentiment-analysis/

Other NLP methods

- Word vectors
  - Represents words as a list of numbers; can be used a features in ML and NLP methods
- Chat bots
  - Can include question-answer applications or general chat bots
- Language generation
  - E.g. GPT3 model or other generative models
- Language translation – Google translate
- Text classification – e.g. bank transaction classification for budgeting
- Topic modeling – e.g. grouping news stories into topics
- Summarization – e.g. generating an abstract from a paper



Word vectors	Dimensions			
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

🟫 animal
🟩 domesticated
🟥 pet
🟦 fluffy

https://dzone.com/articles/introduction-to-word-vectors

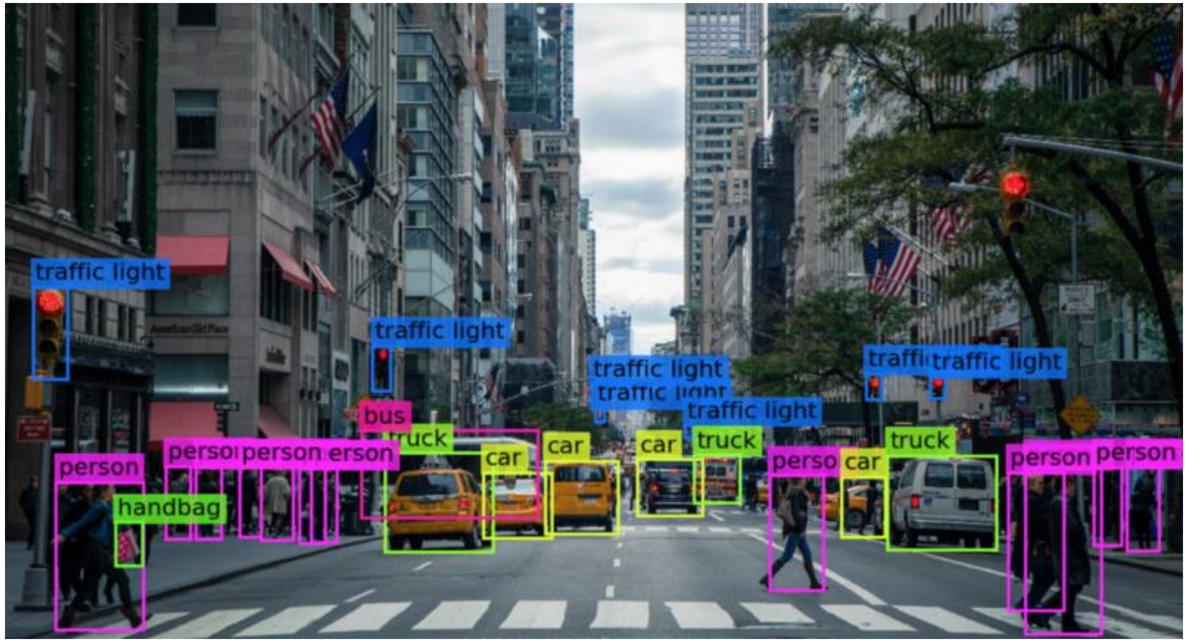
Recap of the course

- Week 1: Getting started with Jupyter, the data science process (CRISP-DM), EDA, and data analysis
- Week 2: More advanced use of Jupyter (hotkeys), terminals and the command line, preparing data (missing values, cleaning data, outliers, feature engineering)
- Week 3: ML for binary classification with logistic regression (overfitting, underfitting)
  - ML types: supervised learning, unsupervised, reinforcement, semi-supervised
- Week 4: Decision trees and random forests, feature importances, feature selection, curse of dimensionality
- Week 5: DS automation with Python and autoML, using code editors and GitHub
- Week 6: Recommender systems and distance metrics; big data; graph analysis
- Week 7: Collecting social media data with APIs and some text EDA; SQL and SQLite
- Week 8: Analyzing social media data with NLP and sentiment analysis

Other data science methods we didn’t discuss

- Deep learning and neural networks – complex machine learning models which were inspired by the human brain. They work well for NLP, image recognition, and more.
- Computer vision – this is a specialty field which can use lots of computer science techniques. It overlaps with neural networks and AI.
- Artificial intelligence – includes topics such as machine learning, path planning, chat bots, and more.
- Forecasting and regression – we can use ML models or other classic regression and forecasting models to predict future values of things like the weather and sales. fbprophet is an interesting Python package to check out for this
- Statistics – t-tests, statistical tests, etc
- Data engineering, web scraping, and more…

- We have many electives that cover these topics, and we will cover some of these topics in future core courses as well



https://towardsdatascience.com/everything-you-ever-wanted-to-know-about-computer-vision-heres-a-look-why-it-s-so-awesome-e8a58dfb641e

The future of data science

- More automation is coming (more advanced autoML, data cleaning, etc)
- Data engineering is breaking off into its own field (we have a specialization and MSDE electives and an MSDE practicum)
- SQL will still be around, but NoSQL will continue to be used as well
- More cloud tools for automation and scaling up ML models and data science workflows
- Neural networks may become more widespread due to decreases in GPU and TPU costs
- Specializing helps – neural networks, NLP, computer vision, statistical testing and experimental design, big data, machine learning

- Verbal and written communication very important
- Keep on eye on the Julia programming language, but also stay sharp on Python and/or R

https://towardsdatascience.com/the-future-of-data-science-14653afb52f5