
Week 7: Deliverables

Group Name: NLP_Task_Force – Document Classification

Name: Konstantinos Soufleros

Email: soufleros.kostas@gmail.com

Country: Serbia

Company: Data Glacier

Specialization: NLP

Github: https://github.com/kostas696/DG_Intern/tree/main/week7

Internship Batch: LISUM30

Date: 19/03/2024

PROBLEM DESCRIPTION

The problem revolves around analyzing a collection of newsgroup posts to gain insights into the topics, themes, and sentiments expressed within them. Newsgroups are online discussion forums where users share information, ask questions, and engage in conversations related to specific topics.

The key challenges in this project include:

- **Data Understanding and Cleaning:** The dataset comprises a vast amount of unstructured text data. Cleaning and preprocessing this data to remove noise, irrelevant information, and standardize its format are essential steps.
- **Topic Modeling:** Identifying the main topics or themes discussed across different newsgroups is crucial for understanding the content.
- **Sentiment Analysis:** Understanding the sentiment expressed in the posts can provide valuable insights into the users' opinions and attitudes towards various topics.
- **Visualization:** Visualizing the data through word clouds, topic distributions, and sentiment heatmaps can facilitate better interpretation and communication of the findings.

BUSINESS UNDERSTANDING

Leveraging the rich source of information available in newsgroup posts can provide businesses with valuable insights that can drive strategic decision-making, enhance customer satisfaction, and foster business growth.

- **Market Research:** Understanding the topics and sentiments prevalent in newsgroups related to specific industries or products can provide valuable market insights.
- **Customer Feedback Analysis:** Newsgroup posts often contain valuable feedback and opinions from customers about products or services.
- **Competitive Analysis:** Monitoring discussions in newsgroups related to competitors' products or industry trends can provide valuable competitive intelligence.
- **Community Engagement:** Engaging with users in newsgroup communities can help businesses build brand awareness and foster relationships with potential customers.

PROJECT LIFECYCLE

Weeks	Date	Deliverables
Week 7	19 March 2024	Problem Understanding
Week 8	26 March 2024	EDA of Dataset
Week 9	2 April 2024	Data Preprocessing
Week 10	9 April 2024	Model Building & Training
Week 11	16 April 2024	Performance Evaluation & Reporting
Week 12	23 April 2024	Model Deployment
Week 13	29 April 2024	Model Inference