Darah Isabel E. Manayaga | DatSci – 4A

**Topic Modeling Analysis on X (formerly Twitter) Data**

**Introduction**

In this report, I analyzed X (formerly known as Twitter) data using Latent Dirichlet Allocation (LDA), a popular technique for topic modeling. The objective is to uncover latent topics within the dataset, which could provide insights into the prevalent themes or subjects of discussion on Twitter.

**Methodology**

I utilized the Python programming language along with the pandas, gensim, and scikit-learn libraries for data manipulation, natural language processing, and topic modeling. The dataset was loaded from a CSV file named 'twitter\_training.csv'.

**Data Preprocessing**

Text data was preprocessed using the simple\_preprocess function from the gensim.utils module. Tokenization was performed to convert text documents into a list of tokens. A dictionary and a corpus were created using the Dictionary and doc2bow functions from the gensim.corpora module, respectively.

**Topic Modeling**

I applied the LDA model with the number of topics set to 3. This parameter can be adjusted based on specific requirements or domain expertise. The LdaModel class from gensim.models was employed, with 15 passes over the corpus for model convergence.

**Results and Analysis**

The LDA model identified three distinct topics within the Twitter data. Each topic is represented by a list of key terms, which are indicative of the underlying theme or subject matter.

* Topic 0:

Words: *user, people, time, today, love, work, life, day, world, great*

This topic seems to revolve around general discussions about life, work, and daily experiences.

* Topic 1:

Words: *trump, biden, president, election, vote, news, america, country, american, political*

This topic appears to be focused on political discussions, particularly related to the 2024 U.S. presidential election.

* Topic 2:

Words: *covid, vaccine, pandemic, health, cases, mask, people, virus, omicron, booster*

This topic likely pertains to discussions surrounding the COVID-19 pandemic, including vaccination efforts, public health measures, and the emergence of new variants.

**Conclusion**

The application of LDA topic modeling on Twitter data revealed three prominent topics: general life experiences, political discussions related to the U.S. presidential election, and conversations regarding the COVID-19 pandemic. These findings could be valuable for understanding the prevalent discourse on Twitter and extracting actionable insights from social media data. Adjusting parameters such as the number of topics or the number of passes during model training can further refine the analysis and uncover additional insights.