Project Proposal: Exploring Real and Fake News Articles with Data Science Techniques and Knowledge Graph Analysis

By Evan McLaughlin and Vladimir Nimchenko

Introduction:
In this project, we aim to analyze a dataset of real and fake news articles obtained from Kaggle. Our goal is to understand the characteristics of fake news articles and develop models to distinguish between real and fake news with high accuracy. Additionally, we will build a knowledge graph to gain insights into the relationships between news articles, topics, entities, and their credibility.

Data Source:
We will be using the Kaggle dataset containing labeled real and fake news articles. The dataset includes features such as the title, text, and source of each article, along with the label indicating whether it is real or fake.

Objectives:
1. Preprocessing and Cleaning:
   - Remove noise, such as HTML tags and punctuation, from the text data.
   - Tokenize and lemmatize the text.
   - Handle missing values and duplicates.

2. Word Embeddings:
   - Utilize pre-trained language models to generate word embeddings for the text dataset.
   - Explore different embedding techniques, such as Word2Vec and GloVe.

3. Dimensionality Reduction and Visualization:
   - Apply techniques like UMAP, LSA, and t-SNE to reduce the dimensionality of the word embeddings.
   - Visualize the embeddings in lower-dimensional space to observe clusters and patterns.

4. Knowledge Graph Construction:
   - Build a knowledge graph to represent relationships between news articles, topics, entities, and credibility.
   - Use the graph for analysis, including topic clustering, entity relationship analysis, and timeline analysis.

5. Machine Learning Models:
   - Train classical machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines on the dataset.
   - Evaluate the models using K-Folds Cross Validation and compare their accuracy and F1 scores.

6. Further Evaluation:
   - Perform additional evaluation methods on one of the models, including confusion matrices and ROC curves.
   - Analyze the model's performance metrics to understand its strengths and weaknesses.

Use Cases for Knowledge Graph:
- Topic Clustering:
  - Cluster news articles based on their topics or themes to understand the distribution of fake and real news across different subjects.

- Entity Relationship Analysis:
  - Explore relationships between entities mentioned in news articles (people, organizations, locations) to understand how they are connected and influence the credibility of news.

- Timeline Analysis:
  - Construct timelines of events based on news articles and identify patterns of fake and real news dissemination over time.

Plan:
1. Data Preprocessing and Cleaning (All):
   - Responsibilities: Data cleaning, tokenization, lemmatization.

2. Word Embeddings (All):
   - Responsibilities: Utilize pre-trained language models to generate word embeddings.

3. Dimensionality Reduction and Visualization (All):
   - Responsibilities: Apply UMAP, LSA, and t-SNE for visualization.

4. Knowledge Graph Construction (Team Members 1 and 2):
   - Responsibilities: Build the knowledge graph and conduct analysis.

5. Machine Learning Models (Team Members 1 and 2):
   - Responsibilities: Train classical ML models, perform cross-validation.

6. Further Evaluation (Team Member 3):
   - Responsibilities: Perform additional evaluation methods and analysis of model performance.

Concerns:
- Ensuring the quality of pre-processing and cleaning to maintain the integrity of the text data.
- Selecting appropriate parameters for dimensionality reduction techniques to achieve meaningful visualizations.
- Building an accurate knowledge graph that captures relevant relationships between news articles, topics, entities, and credibility.

Conclusion:

By exploring this dataset and employing various data science techniques, including word embeddings, dimensionality reduction, and machine learning models, along with knowledge graph analysis, we aim to gain insights into the characteristics of real and fake news articles. Our analysis will contribute to understanding the spread and impact of fake news and aid in the development of methods to combat misinformation.

Team Members:

1. Vladimir Nimchenko: Responsible for data preprocessing, cleaning, word embeddings, and machine learning model training.
2. Evan McLaughlin: Responsible for word embeddings, dimensionality reduction, knowledge graph analysis, further evaluation methods and analysis of model performance.