Project Handout – Fake News Detector

Welcome to classification for fake news.

What you will build

In this program you will learn how to build an end to end Machine learning pipeline that will:

- Ingest raw text data
- Process the raw text data into paragraph vectors
- Apply trained supervised learning classifiers to the paragraph vectors to label the original text as fake or not fake

What you will learn

Students successfully completing the program will learn how to:

- Compare different methods for word embedding applications used today
- Use neural embedding implementations like Gensim on both for
    - word vectorization and for
    - paragraph vectorization
- Hyper-tune neural embedding algorithms as part of an end-to-end pipeline
- Use standard industry classifiers and integrate them with the end-to-end pipeline
- Troubleshoot multi stage Machine Learning pipelines

How the course is structured

The course is broken into 3 content lessons:

(Lesson 1) Classification for fake news:

This section will cover

- Classifier applications to fake news text.
- Embedding code is prepared in advance for students so they can focus on applying classifier fundamentals.
- Attention will be given to metrics (precision, recall, F1), and model selection.

(Lesson 2) Text Embedding techniques:

This section will cover

- What Word2Vec is and what Paragraph2vec is
- Reviews historical strategies and why word2vec works better

    - TF IDF (brief for history)
    - Keyword presence VSM (brief for history)
    - Neural embeddings (mainline)
- Lab sessions students focus on implementing Gensim

(Lesson 3) Putting it all together:

This section will focus on putting together the complete pipeline

- The lesson covers the strategies for hypertuning
    - Grid search vs automated search (not too deep)
    - How to priorities your time with searching
    - which parameters are important and what their impact is in typical classifiers
- Troubleshooting
    - Managing and preparing imbalanced data sets
    - Information leakage and hold out for Test as well as Validation
- Lab sessions hands on with troubleshooting and developing search technique loops"

Student responsibilities

Participation

Students are expected to attend and participate in every session of the program. **Students will be expected to answer questions posed by the instructor during lesson and lab times, and they will be expected to interrupt and ask the instructors during the lessons if something does not make sense.** Students will be expected to participate from a quite environment where they can keep their microphones open for questions and **where they can use their computer video cameras to see each other during the lessons.**

Final projects

In order to successfully complete this course your student group will need to

1. Develop an end to end pipeline that accomplishes the 3 tasks in the "What you will build" section.
2. Capture your results in a jupyter notebook, including
3. Data exploration, feature manipulation other EDA
4. Execution of the pipeline
5. An articulation of tactics used in achieving final performance metrics
6. Final performance metric results

- Students will be broken into groups of 3 or 4 students per group.
- Each group will present their final presentation on the last day of the program and will have 10 min to present.