Assignment

- Fake news Text Classification

WATCH: Mass Shooting Occurs During #TrumpRiot Media Ignores (Video)

“The five victims range in age from their 20s to 50s, and they have gunshot wounds to their legs, chest and neck.”– The Seattle Times , November 10, 2016 Seattle police are assuring citizens that a virtually unreported mass shooting has nothing to do with an anti-Trump rally incited by Seattle socialists, featuring Kshama Sawant, the Marxist Seattle city council

Black Hawk crashes off Florida human remains found

(CNN) Thick fog forced authorities to suspend the air search Wednesday for seven Marines and four Army aircrew, feared dead after their Black Hawk helicopter crashed into waters off the Florida Panhandle.

The helicopter was first reported missing at about 8:30 p.m. (9:30 p.m. ET) Tuesday. Hours later, searchers found debris around Okaloosa Island near Eglin Air Force Base, base spokesman Andy Bourland said.

How to solve it

- **Classification problem**
  - News Report (***document***) → ***Class***: [FAKE, REAL]
- **Try text-related classifiers**
  - Naive Bayes
  - MaxEnt
  - SVM
- **NLTK+SKLearn provides you anything you need**
  - NLP Pre-processing
  - Classifiers
  - N-grams

Dataset

- **fake_or_real_news_training:**
  - **ID**: ID of the news
  - **Title**: Title of the news report
  - **Text**: Textual content of the news report
  - **Label**: Target Variable [FAKE, REAL]
  - **X1, X2:** additional fields
- **fake_or_real_news_test:**
  - **ID, title and text**

Advices

- **Take a look to the data**
  - Check your data loading process
  - News have 2 levels of text (title and text)
- Try the **pre-processing methodologies** we have seen **in class**
- **TF-IDF** seems to be better (but try it!)
- **N-grams** could pay the effort
- Less than 90-92%? **Try again**

Advices/Warnings

- Avoid ML mistakes
- Mind parsing data issues (commas)
- Explain anything you do
- Try different approaches and compare results
  - Classifiers
  - NLP Pipelines
- Analyze your results
- **Individually or in pairs**

Submission

- **Due: 2nd June**
- **Submission** (Send **everything** please):
  - **CSV with your predictions**
    - **News_id (ID), prediction[FAKE, REAL]**
  - **Notebook**
- Send me something that **actually works**
- **Grading:** 50% results – 50% notebook

Resources

- **NLTK Book Chapter**
  - http://www.nltk.org/book/ch06.html
- **Examples of NLTK + SkLearn for Text Classification**
  - https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a
  - http://billchambers.me/tutorials/2015/01/14/python-nlp-cheatsheet-nltk-scikit-learn.html
  - http://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/
  - https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/
  - https://github.com/bonzanini/nlp-tutorial
  - https://www.analyticsvidhya.com/blog/2015/10/6-practices-enhance-performance-text-classification-model/
  - https://link.springer.com/article/10.1023/A:1012491419635
- **Resources in the class slides**