PHiSH-ing for News

Helping you find the Politics, Health, intense Sports and Hollywood news you care about with PHiSH

Amber Dsilva (173199200), Kieara Miranda (170802500),
Alexander Ojo (170750510), Hitanshi Shroff (170492430)

Problem Statement

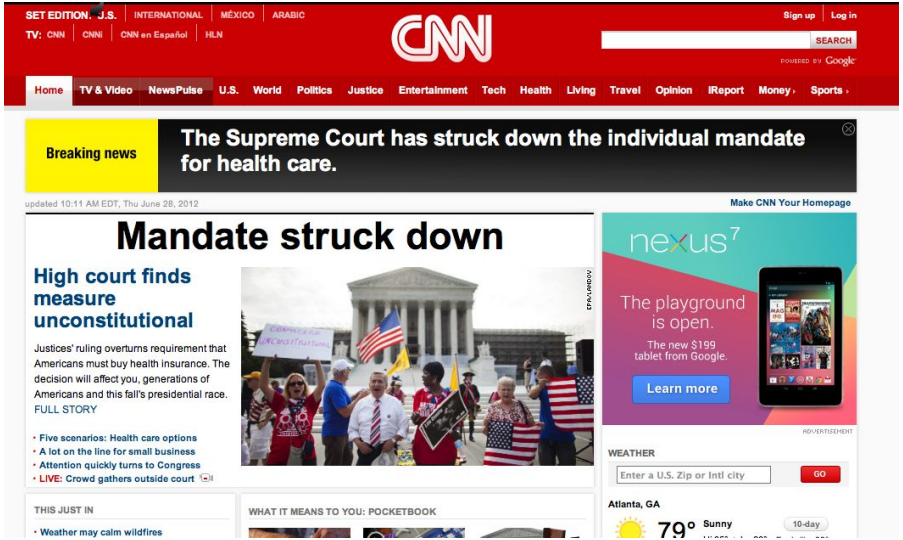
- Problem: you want to know what topics are trending nowadays
- Solution: Automatically tags and classify CNN’s recent news based on recent news that we find of interest with PHiSH
- Program requirements:
  - Python
  - CNN

Overview of PHiSH

- It uses a web crawler to examine all news articles on CNN’s home page and adds relevant articles to csv database
- Features vectors are created with frequency vectors and TF-IDF vectors
- Utilize KNN to generate a classification report which determines the whether an article is about Politics, Health, Sports or Hollywood

Input Generation

- Web crawler
  - examines all news articles on CNN
  - filters them into categories based off the articles' text
- The article's title, text and category name is saved in a csv database
- 61 articles were searched



Machine Learning Application

- Feature Vectors
  - Word frequency
  - TF IDF = term frequency * inverse document frequency
- Training set
  - Randomly separated article - 60% training set and 40% testing set
- K-Nearest Neighbour (KNN)
  - Supervised classification algorithm
  - Classfies articles by the distance between each feature vectors’ data points

KNN

- How K nearest applies to our project

Category	Value code
Politics	0
Health	1
Sports	2
Hollywood	3

Output

- Classification report

	precision	recall	f1-score	support
0	0.95	0.90	0.92	19
1	0.88	0.84	0.86	17
2	0.90	0.92	0.91	11
3	0.93	0.87	0.90	14
avg / total	0.92	0.88	0.90	61

Results

					TOTAL
Predicted Politics	18	1	0	0	19
Predicted Health	2	15	0	0	17
Predicted Sports	0	0	10	1	11
Predicted Hollywood	1	0	0	13	14
	Actual Politics	Actual Health	Actual Sports	Actual Hollywood	61

What to do with data

- Ie. with this data, politics is a hot topic rn as approx. ⅓ of the articles talk about it

Project Management

- Project planning:
  - Met on a regular basis to discuss progress
  - Took an Agile approach
  - Documented code
- Primary communication via facebook messenger and email

Challenges:

- Topic changes
- Varying schedules

Conclusion

PHISH: Helping you keep up with the news that you care about and determine PHISH market trends

Thanks for watching!

Logical Partitioning

- Visual Studio Live Share was used so everyone could collaborate and work on the code together.
- We used google docs to work on the report:
  - Introduction, Conclusion - Alexander Ojo
  - Project Description - Amber Dsilva
  - Experimental Analysis - Kieara Miranda,Hitanshi Shroff
- PowerPoint Slides:
  - Everyone

Appendix

Sadangi, Siddhant. “Introduction to Text Classification in Python.” Medium, Analytics Vidhya, 29 June 2020, https://medium.com/analytics-vidhya/introduction-to-text-classification-in-python-659eccf6b2e.

Kulshrestha, Utkarsh. “NLP‑Feature Selection Using TF-IDF.” Medium, Analytics Vidhya, 2 June 2020, https://medium.com/analytics-vidhya/nlp-feature-selection-using-tf-idf-db2f9eb484fb#:~:text=TF-IDF%20acronym%20for%20Term%20Frequency%20%26%20Inverse%20Document,Classification%2C%20Information%20Retrieval%20Systems%2C%20Text%20Data%20Mining%20etc.

“K-Nearest Neighbors (KNN) For Iris Classification Using Python.” Indowhiz, 25 June 2020, www.indowhiz.com/articles/en/implementation-of-k-nearest-neighbors-knn-for-iris-classification-using-python-3/.

“Sklearn.metrics.confusion_matrix¶.” Scikit, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html