Identifying Egyptian Arabic websites using mahcine learning during a web crawl

Sara Elshobaky & Youssef Eldakar

BIBLIOTHECA ALEXANDRINA
مكتبة الإسكندرية

Problem

- Crawl the web for Egyptian content
- Initialize the crawl with a list of curated seeds
- Websites that were not in the initial seeds are discovered, are they Egyptian?
- Check for ccTLD = .eg?
  - Many “Egyptian” websites have .com etc. domain names
- Check for content language = Arabic?
  - 25 countries have speak Arabic
  - #6 most spoken language

العربية

Example: Ministry of Interior

How to tell the origin country of a website?

- A human curator is able to read the “spirit” of a website’s homepage and distinguish the country of origin
- Clues for such judgement include:
  - Topics discussed
  - Localized characteristics, e.g., month naming system
  - Term usage
- Not codifiableusing conventional programming methods

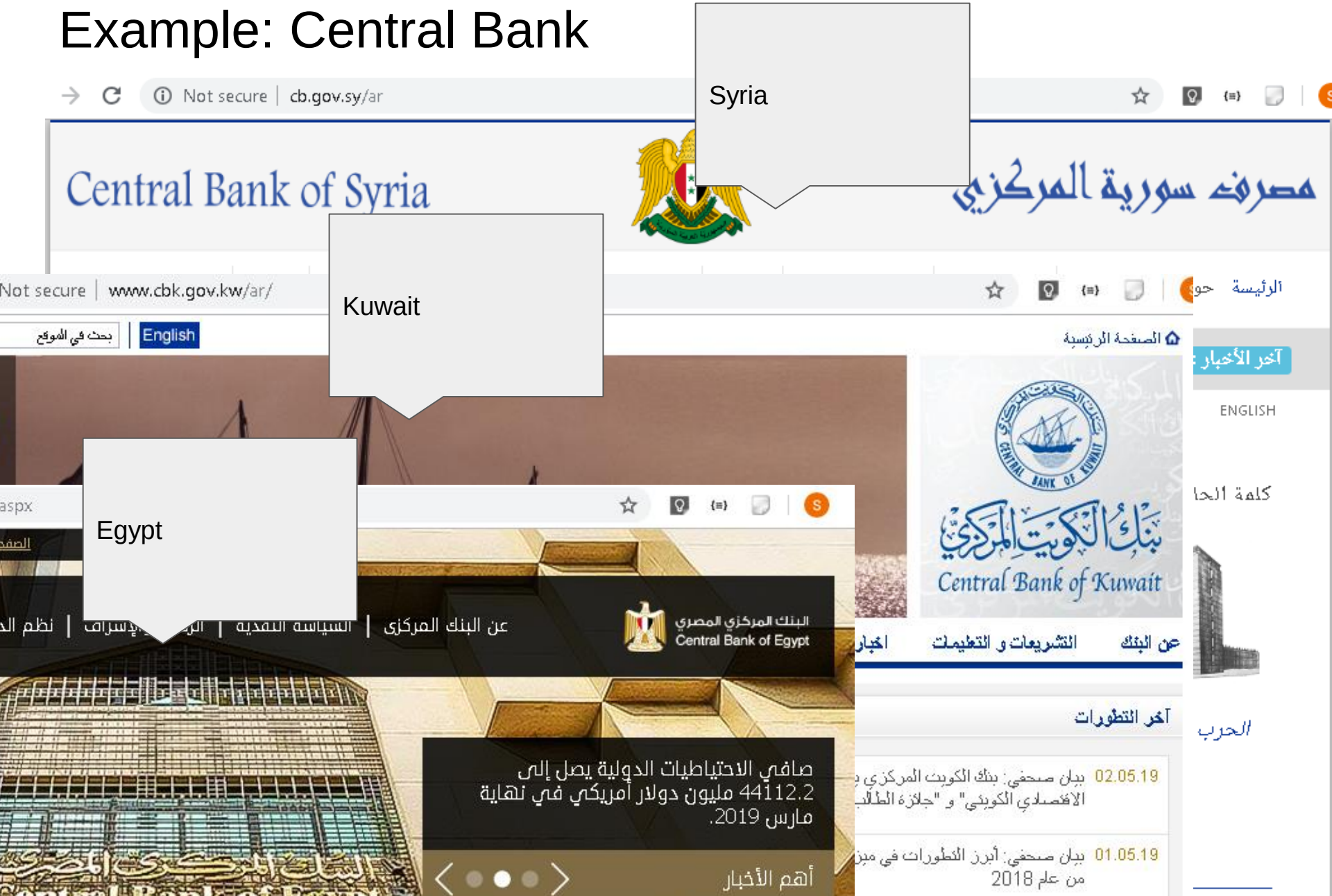
Distinguishing features

- Month naming system
  - Levant, used in Iraq, Syria, Lebanon, Palestine, and Jordan: Nisan, Iyar, Hzirin, Tammuz, etc.
  - Gregorian: April, May, June, July
  - See Wikipedia, [Arabic names of calendar months]
- Term usage
  - For example, in certain countries, is transliterated as-is (“بنك”)
  - Formal Arabic translation used elsewhere (“مصرف”)

مصرف

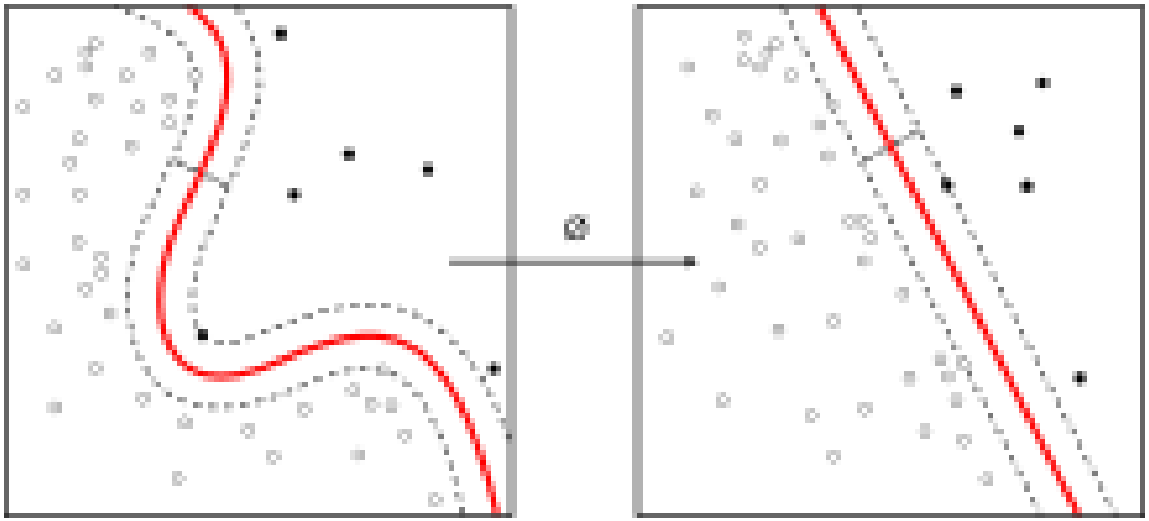
بنك

Example: Central Bank

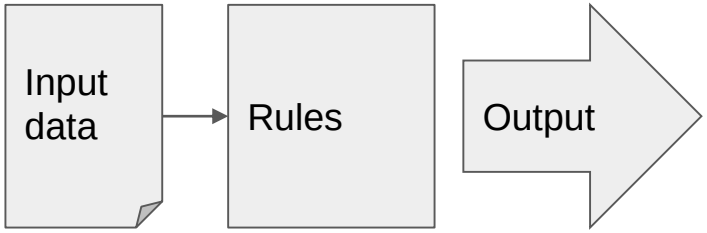


Machine learning

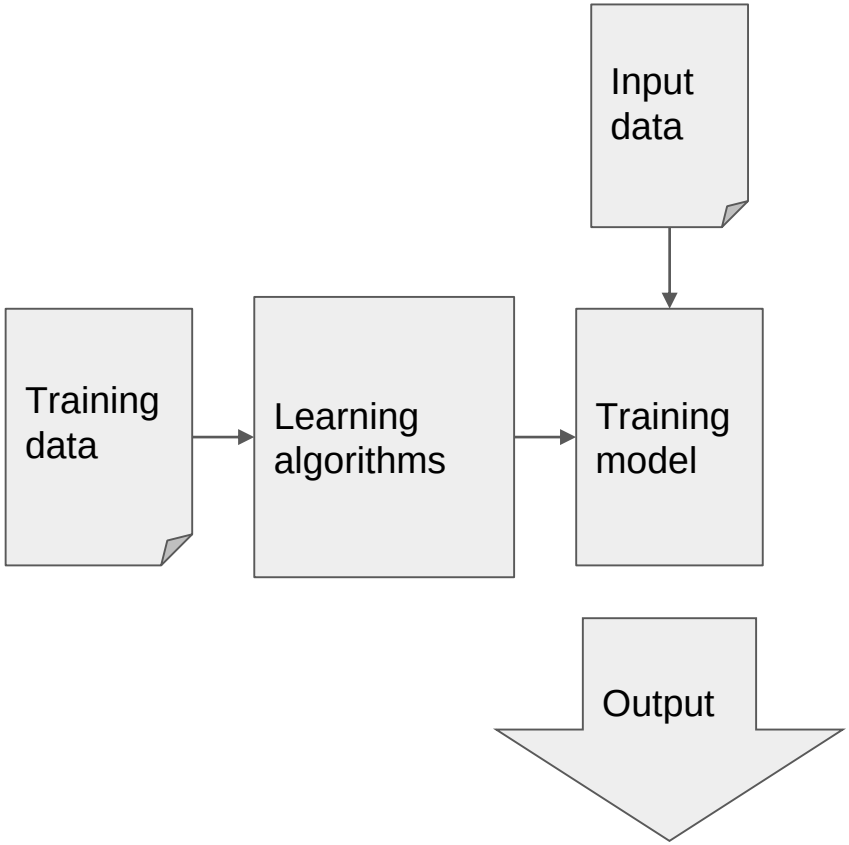
- In Recent years, machine learning, which falls under the general domain of artificial intelligence, has been making significant proress
- Enabling machines to make better sense of context and derive meaning from data
- See Wikipedia, [Machine learning]



Traditional programming vs. machine learning

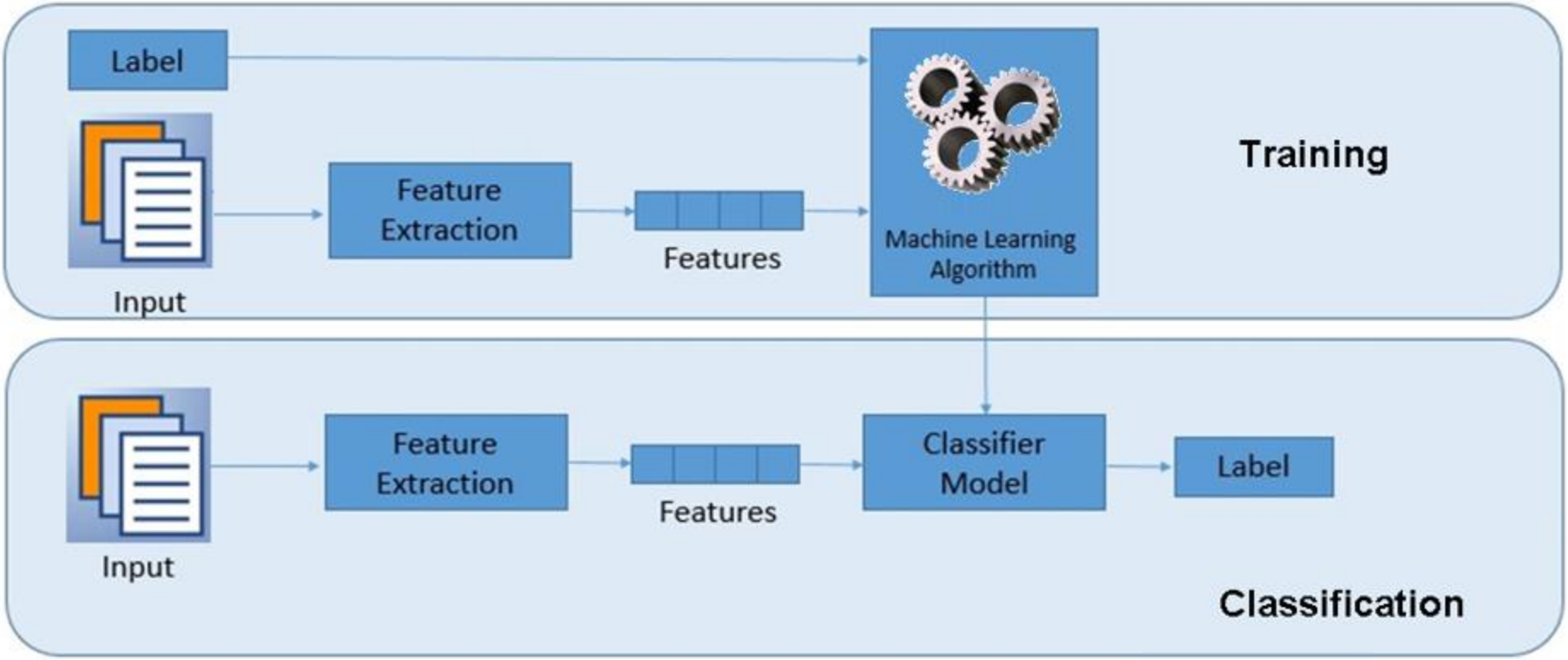


Traditional programming



Machine learning

Supervised classification



Proposed solution

Enhance the accuracy of Egyptian web crawls using machine learning

Steps:

1. Training data preparation
2. Feature extraction
3. Training the model
4. Evaluating the model

1. Training data preparation

- Select a few seed URLs from different Arabic websites with Arabic content
  - 300 URLs from ‘.eg’ domain
  - 300 URLs from other Arab country domains, e.g., ‘.sa’, ‘.ly’, ‘.iq’
- Harvest content from the selected domains (homepages only)
- Parse HTML to extract plain text
- Apply preprocessing and normalization rules
  - Remove punctuation
  - Normalize certain Arabic characters
  - etc.

2. Feature extraction

- Why feature extraction?
  - Text initially comes with a large number of features (n-grams)
  - Need to eliminate noise and avoid overfitting
  - Decreasing the number of features reduces training time
  - Features were extracted from the text based on their TF-IDF
- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Well-known method for evaluating how important a word is in the document
  - Helps adjust for the fact that some words tend to generally appear more frequently

3. Training the model

- The extracted features and their labels were used to train a binary linear classifier
  - Label = 1 for Egyptian
  - Label = 0 for not Egyptian
- The output of this process is a model that can be used to classify newly discovered websites as Egyptian or otherwise

4. Evaluating the model

- This initial experiment used the content of 300 Egyptian websites and 300 non-Egyptian Arabic websites
- Of that dataset, 90% of URLs were used to train the model, 10% to evaluate it
- The resulting average F1-score was approximately 84%

Future work

- In the future, we would like to increase the size of the training and evaluation dataset and experiment with alternative machine learning algorithms and parameters in attempt to improve classification accuracy and make more thorough evaluation
- In addition, we hope to extend our experiment to identifying Egyptian websites in languages other than Arabic, i.e., English and French
- Integration into a live web crawl vs. post-crawl

We hope that was useful...

- Questions?
- See also the Python code used to run the experiment:
  - https://github.com/arcalex/content-scoper
- Contacts:

  - sara.elshobaky@bibalex.org
  - youssef.eldakar@bibalex.org