Web Scraping-Natural Language Processing (NLP) for Disease Outbreak Detection

Yijun Wei

Disclaimer

The findings and conclusions in this presentation are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy.

Background and motivation for hog disease outbreak detection

1. The impact of disease can be huge, but initial event can be local and/or small
2. Detection of the early stages is challenging
3. Current diagnostics lag the occurrence of the outbreak by a quarter
4. Web scraping-NLP approach is proposed to detect very early signals of the outbreak

Goals for Web scraping-NLP

1. Rapid detection
2. Geo-locating the outbreak
3. Better precision in predicting pattern and rate of spread of the disease
4. Documentation of the disease on each scale national-state-local

Web scraping-NLP Approach

Two stages: in Web scraping-NLP approach

Stage 1: Hog disease outbreak detection using web scraping in Swine Disease Global Surveillance Project (SDGSP)

Stage 2: Web scraping for related news from hog news websites, and NLP for information extraction

Output:

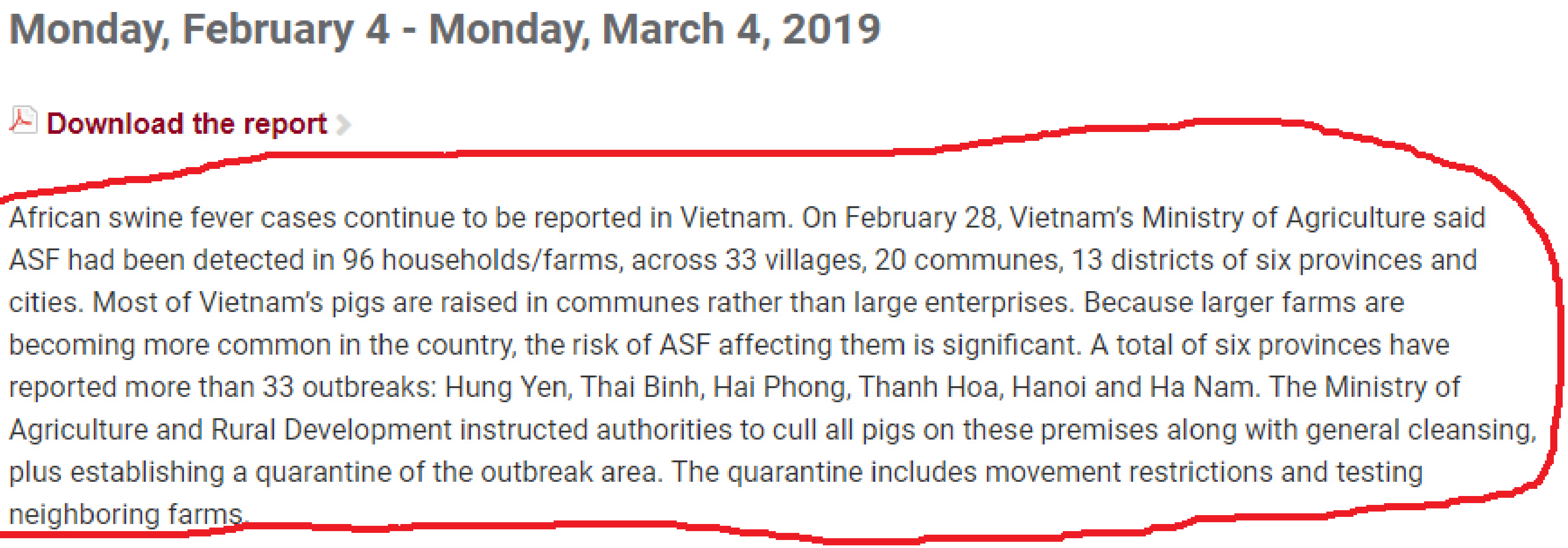
1. Input to spatial-epidemic model
2. Data prepared for experts

Stage 1: Disease Outbreak Detection

Web scraping: Sources

1. Disease Report Repositories
    1. SDGSP (Swine Disease Global Surveillance Project)
        1. University of Minnesota Swine Center
        2. Monitors hog disease outbreaks on international scale
        3. Publishes reports every two weeks
        4. **Currently used**
    2. APHIS (USDA)
2. Other media sources
    1. News feeds – national, state, local
    2. Extension service websites
    3. Producers' organizations websites
    4. Blogs

African Swine Fever Outbreak: Vietnam



Monday, February 4 - Monday, March 4, 2019

Download the report

African swine fever cases continue to be reported in Vietnam. On February 28, Vietnam's Ministry of Agriculture said ASF had been detected in 96 households/farms, across 33 villages, 20 communes, 13 districts of six provinces and cities. Most of Vietnam's pigs are raised in communes rather than large enterprises. Because larger farms are becoming more common in the country, the risk of ASF affecting them is significant. A total of six provinces have reported more than 33 outbreaks: Hung Yen, Thai Binh, Hai Phong, Thanh Hoa, Hanoi and Ha Nam. The Ministry of Agriculture and Rural Development instructed authorities to cull all pigs on these premises along with general cleansing, plus establishing a quarantine of the outbreak area. The quarantine includes movement restrictions and testing neighboring farms.

Affected: 96 households/farms in 6 provinces and cities (listed)

Response: Ministry of Agriculture and Rural Development Instructions:

Cull all

Quarantine outbreak area

Test all neighboring farms

Stage 2: NLP

As soon as a hog disease outbreak is detected:

1. Related news will be scraped from the website

2. Information will be extracted from related news, using a 4-step Information Extraction (IE)

   1. Normalize time

      Different temporal formats transformed to a single form

   2. Normalize word

      Different word formats converted to a singular form

   3. Keywords identification

      Keywords are identified

   4. Named Entity Recognition and information extraction

      Recognize the pertinent information

Illustration – China, February, 2019

Input, News item web-scraped from **The Pig Site**

'The Ministry of Agriculture and Rural Affairs said the first outbreak is on a farm in the Xushui district of Baoding city which has 5,600 hogs, some of which died because of the swine fever, though it did not provide a death toll.

\n\nThe farm has been quarantined and the herd slaughtered, it added.\n\n

Illustration – China, February, 2019

Input, News item web-scraped from **The Pig Site**

Reuters reports that the second outbreak is in the remote Greater Khingan Mountains in Inner Mongolia, where 210 of the 222 wild boar raised on the farm died, the ministry said in a separate statement. The rest have been slaughtered, it said.

China has reported more than 100 cases of African swine fever in 27 provinces and regions since last August. The disease is deadly for pigs but does not harm humans.'

NLP Steps

1. Normalize word:

    raised to raise

    slaughtered to slaughter

    quarantined to quarantine

1. Keyword identification:

    “outbreak”

2. Named Entity Recognition and information extraction:

    The Ministry of Agriculture and Rural Affairs

    Xushui district of Baoding city

    swine fever

Result

Output:

1. Noun: 'outbreak', Source: 'The Ministry of Agriculture and Rural Affairs', Location: 'a farm in the Xushui district of Baoding city', Stats: 'has 5,600 hogs'

2.Noun: 'outbreak', Source: 'Reuters', Location: 'remote Greater Khingan Mountains in Inner Mongolia', Stats: '210 of the 222 wild boar died'

Potential for information

Time and location of disease references:

1. Fine scale (state, county) incidence allowing spatial disease modeling and mapping

2. Time course of spread

3. External documentation confirming disease and response to outbreak

4. Data for pre-board, ASB, or other experts

5. Information to incorporate into the model system