# Key Agriculture:

* Mediterranean fruit fly (aka Medfly): Apples, avocados, bell peppers, melons, tomatoes.
  + <https://newsforkids.net/articles/2023/11/08/fighting-fruit-flies-withfruit-flies/> This article discusses a recent fruit fly outbreak in California and the steps they are taking to mitigate the outbreak (and how they have in the past). The content of this article is more cautionary than informative. While it poses the outbreak as a threat, it strikes down concern at the same time. Words of importance: threat, solution, method, problem.
  + <https://www.elsegundo.org/Home/Components/News/News/3260/268327> Same as above. It poses medfly as a threat but outlines how citizens can help prevent the spread of the pests. Key words: protect,
* Melon fly: Vegetable crops, especially cucurbits like melon, pumpkin, squash, zucchini, and cucumber.
  + <https://assets.ippc.int/static/media/files/pestreport/2023/08/15/report__zeugodacus_DNSAB_03_augst_2023_English.pdf> More informative…and it states so in the paper’s objective. **This is perhaps a good thing to look at when it comes to classifying a paper.**
* Oriental fruit fly: Notable ones include apricots, cherries, figs, tomatoes.
* Caribbean fruit fly: Mango, Annona, papaya, avocado, west Indian cherry, guava, coffee, citrus, tomato, sapote
* South American Cucurbit fruit fly: citrus, malus, prunus
* Spotted Lanternfly: Tree of heaven, black walnut maples, willow river birch, black cherry, tulip poplar, + more.
* Emerald ash borer: bark of ash trees.
* Orange tortrix: leaves, buds, and surface of fruits.
* Oriental fruit moth: peach, plum, apricot, apples, pears, cherry

# Weekly Outline

**Week 1 (06/10 – 06/14):**

* Project setup. Create GitHub Repository for the project, finalize the intended goals of the project, and obtain all resources needed.
* Web scraper. Choose the web scraping packages to be used for the project and at least identify all documentation needed for the project.
* Identify internet sources to use for article scraping and/or what sources actually allow for scraping.
* Talk to Morgan and Sagrario to find out what they have already done with their portion of the project. Maybe center goals around what they have done.

**Week 2 (06/17 – 06/21):**

* Article research. Using the sources decided on in week one, formulate a set of articles that the application can be used to classify and interpret. There should be articles in each of the following categories. There will be overlap between each category.
  + Relevant to the topic, not relevant to the topic.
* **The following categories only apply if the article is relevant to the topic.** 
  + Informative or suggestive. Does the article merely inform the reader of a particular species, or does it inform and also give a suggestion to the reader?
  + Positive or negative. Is the content within the article a concern or Mexico? Is it a good thing for Mexico?
* Web scraper. For a subset of the articles chosen (or all of them!), build the web scraper to scrape the data from the article and parse it in a way such that it maintains its structure.
* NLP Libraries. Look at possible language processing libraries that can be used to interpret the articles as a human would. There are many potential pre-made models out there, but they will have to be fine-tuned using my list of articles and my categories.

**Week 3 (06/24 – 06/28):**

**Week 4 (07/01 – 07/05):**

**Week 5 (07/08 – 07/12):**

**Week 6 (07/15 – 07/19):**

**Week 7 (07/22 – 07/26):**

**Week 8 (07/29 – 08/02):**

# Week One

Project goals: Automate the process of interpreting an article. While we are easily able to do so manually, there would be a lot of benefit from automating this process for researchers.

* + Natural language processing.
  + APIs.
* The end goal of the project is to have a working classifier that is able to get the information from an article about a certain pest and create a table of all of the useful information as well as a summary of all of the information gathered.
* **I. Identify** possible risks and threats from monitoring **unstructured** information and **structured** (national and international information monitoring).
* **II. Evaluation** of the degree of **certainty of the monitored information** (level of reliability), verification and validation of information sources.

Top 15 Countries **Mexico Imports** From: USA, China, South Korea, Germany, Japan, Malaysia, Canada, Brazil, Vietnam, Thailand, Italy, India, Spain, France, Philippines. **(~87% of all imports as of 2022** [**https://tradingeconomics.com/mexico/imports-by-country**](https://tradingeconomics.com/mexico/imports-by-country)**).**

* Given this, we will mostly focus on imports coming from those countries.

Can break up articles into categories.

* Informative and cautionary. Informative has more scientific undertones, while cautionary may be scientific as well as suggestive. **(second)**
* Positive and negative. I anticipate most articles to be negative. **(third)**
* Classifying articles as relevant or not relevant **(first)**

Rough outline of the project:

* Client can start by putting an article into a file dropper in the form of a website link (??). The software then scrapes the website, and gives a summary and several classifications. (Create as a website using HTML, CSS, JavaScript/TypeScript)
* Will need to figure out how to efficiently display the information.

All the articles are considered unstructured data as the data is not organized in a tabular format. Cannot use something like a SQL query to gather data. Will have to gather data by scraping, parsing, and analyzing using ML.

* Some risks of using unstructured data it can be difficult to analyze even when using ML techniques. Storage and extraction are also other challenges.

# Week Two

Resources to pull articles from:

* Google Scholar
* Academia.edu
* ResearchGate
* SSRN
* Science Direct
* Journal of Plant Pathology
* New Disease Reports

Highlighted resources are behind login screens…requests library should be able to handle that kind of block.

Post-Meeting with Sagrario and Morgan:

* Try to focus on pulling info from social media (Twitter is the most important).
* Store info in a database (MongoDb is recommended, can use MySQL).
* Hugging Face for NLP Models

### X API v2 rate limits - Free

The following table lists the rate limits for the X API v2 Free access. These rate limits are also documented on each endpoint's API Reference page and also displayed in the  [developer portal](https://developer.x.com/en/docs/developer-portal)'s products section.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Endpoint** | **#Requests** | **Window of time** | **Per** | **Part of the Post pull cap?** | **Effective 30-day limit** |
| POST\_2\_tweets | 50 | 24 hours | per user | no | 1,500 |
| 50 | 24 hours | per app | no | 1,500 |
| DELETE\_2\_tweets\_id | 50 | 24 hours | per user | no | 1,500 |
| 50 | 24 hours | per app | no | 1,500 |
| GET\_2\_users\_me | 25 | 24 hours | per user | no | 750 |

Social Media Sites to Scrape Data From:

* Twitter (needs API, rate limits are lowkey crazy)
* Meta (Need to use some plugins)
* Instagram (200/hr) rate?

Meeting With Dr.Guzman

Once you find a relevant article, need to create a summary of the article or post. Include things like:

* URL
* Country
* Date
* Name of the plague
* How many acres/hectacres
* What/how many animals have been affected

Also need to classify the article using information like:

* Is it harmful?
* Is it a concern?
* Use natural language processing methods
* Negation will be a hard problem.

There are primary and secondary words to pay attention to when classifying articles. For example -> Primary word = screw worm, and secondary word = animals or people or cattle.

In order to interpret the articles that I encounter…

* Tag different parts of speech (like nouns, verbs, pronouns, adjective, etc). This is called position of speech.
* Then you must lemmatize (There are packages available)
* Take note of synonyms in both English and Spanish. This is synonymizing (need to build our own).
* ^ Use the methods above to help concentrate information
* Best resource to use for synonyms is a thesaurus (Rogets) (look online)
* For hiperonimos, use wordnet 3