Introduction

Social media analytics leverages the ability to gather and find meaning in data gathered from social channels to support business decisions — and measure the performance of actions based on those decisions through social media.

Objectives

- Discover breaking news from various sources as the events they are describing are unfolding.
- Understand breaking news such as natural disasters, political unrest and big happenings in real-time.
- Make system more efficient disaster relief.
- Generate insights leveraging on NLP techniques in social media, newsfeed, etc.

Advantages

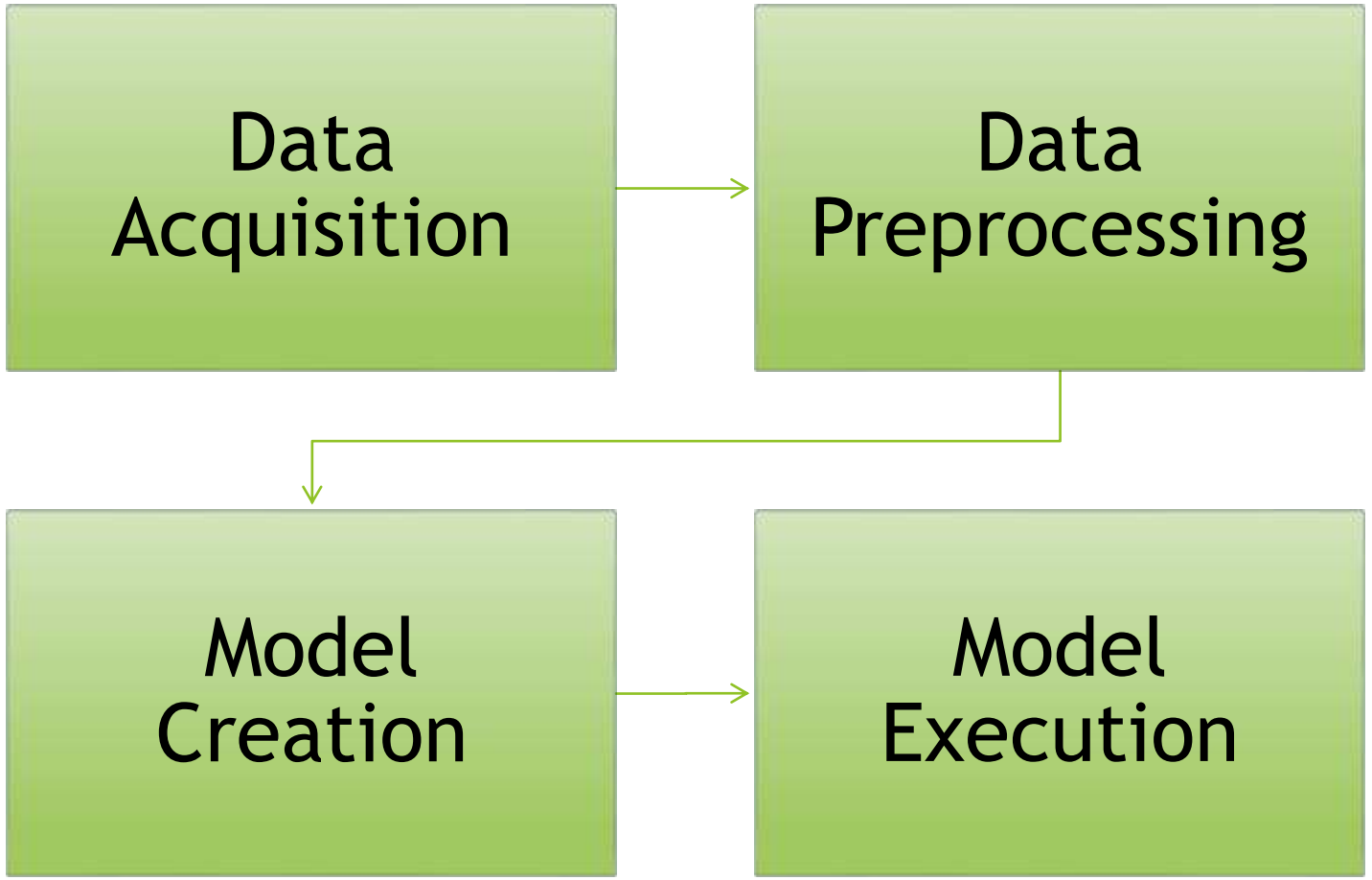
Social media analytics helps government organizations and companies to:

- Spot trends related to topics and articles
- Understand conversations — what is being said and how it is being received
- Derive peoples' sentiment towards news and articles
- Gauge response to social media and other communications
- Identify highly-important news regarding a certain topic
- Uncover world politics and its effectiveness
- Map how third-party partners and channels may affect credibility of the news

Sub-problems?

- Find those sources that are news relevant in near real-time.
- Given a set of news concerning the same event, decide which of them is the most representative of the set.
- Geo-tagging news to see their reach and local/global effect.
- Sentimental Analysis on the crowd response to the events that news describe.

Project Flowchart



Acquiring Data

- Twython is actively maintained, pure Python wrapper for the Twitter API. It supports both normal and streaming Twitter APIs which are used to fetch live tweets data from twitter.
- Pandas - to store the data into dataframe.

Data Pre-processing

- Cleaning - removing irrelevant characters, URLs, emoji’s, etc.
- Exploratory data analysis on number of mentions, hashtags, name titles, words, characters, average word length, stop word, part of speech tags, named-entity-recognition.

Text Preprocessing Activity for NLP Modeling

During text preprocessing the following stages are performed:

- **Tokenization** is a process when we split the text into sentences and the sentences into words. We lowercase the words and remove punctuation.
- **Lemmatization** looks beyond word reduction and considers a language's full vocabulary to apply a morphological analysis to words. For example, words in third person are changed to first person and verbs in past and future tenses are changed into present.

Text Representation Models in NLP

- The bag-of-words model is simple to understand and implement and has seen great success in problems such as language modeling and document classification.
- In practice, the Bag-of-words model is mainly used as a tool of feature generation. After transforming the text into a "bag of words", we can calculate various measures to characterize the text.

Topic Modeling

- LDA (**Latent Dirichlet Allocation**) is an example of a topic model and belongs to the machine learning toolbox and in wider sense to the artificial intelligence toolbox. A frequently used methodology in topic modeling, generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

Text Classification

- Named entity recognition (NER)is probably the first step towards information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. NER is used in many fields in Natural Language Processing (NLP).